BMJ Open

# Measuring ability to assess claims about treatment effects: the development of the 'Claim Evaluation Tools'

Astrid Austvoll-Dahlgren,[1] Daniel Semakula,[2] Allen Nsangi,[2] Andrew David Oxman,[1] Iain Chalmers,[3] Sarah Rosenbaum,[1] Øystein Guttersrud,[4] The IHC Group

**To cite:** Austvoll-Dahlgren A, Semakula D, Nsangi A, *et al.* Measuring ability to assess claims about treatment effects: the development of the 'Claim Evaluation Tools'. *BMJ Open* 2016;**7**: e013184. doi:10.1136/bmjopen-2016-013184

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-013184).

Received 6 July 2016
Revised 12 October 2016
Accepted 26 October 2016

For numbered affiliations see end of article.

**Correspondence to**
Dr Astrid Austvoll-Dahlgren; astrid.austvoll-dahlgren@fhi.no

## ABSTRACT

**Objectives:** To describe the development of the Claim Evaluation Tools, a set of flexible items to measure people's ability to assess claims about treatment effects.

**Setting:** Methodologists and members of the community (including children) in Uganda, Rwanda, Kenya, Norway, the UK and Australia.

**Participants:** In the iterative development of the items, we used purposeful sampling of people with training in research methodology, such as teachers of evidence-based medicine, as well as patients and members of the public from low-income and high-income countries. Development consisted of 4 processes: (1) determining the scope of the Claim Evaluation Tools and development of items; (2) expert item review and feedback (n=63); (3) cognitive interviews with children and adult end-users (n=109); and (4) piloting and administrative tests (n=956).

**Results:** The Claim Evaluation Tools database currently includes a battery of multiple-choice items. Each item begins with a scenario which is intended to be relevant across contexts, and which can be used for children (from age 10  and above), adult members of the public and health professionals. People with expertise in research methods judged the items to have face validity, and end-users judged them relevant and acceptable in their settings. In response to feedback from methodologists and end-users, we simplified some text, explained terms where needed, and redesigned formats and instructions.

**Conclusions:** The Claim Evaluation Tools database is a flexible resource from which researchers, teachers and others can design measurement instruments to meet their own requirements. These evaluation tools are being managed and made freely available for non-commercial use (on request) through Testing Treatments *interactive* (testingtreatments.org).

**Trial registration numbers:** PACTR201606001679337 and PACTR201606001676150; Pre-results.

### Strengths and limitations of this study

- As far as we are aware, this is the first attempt to develop a set of evaluation tools that objectively measure people's ability to assess treatment claims.
- This development resulted from collaboration among researchers in high-income and low-income countries, and included feedback from people with methodological expertise as well as members of the public.
- Based on qualitative and quantitative feedback, the Claim Evaluation Tools were found to have face validity and relevance in the contexts studied.
- There are many ways of developing evaluation instruments. We chose to use a pragmatic and iterative approach, but the reliability of the items remains to be tested.

## BACKGROUND

There are endless claims about the effects of treatments in the mass media, advertisements and everyday personal communication.[1–4] Such claims may include strategies to prevent illness, such as changes in health behaviour or screening; therapeutic interventions; or public health and system interventions. Many claims are unsubstantiated, and patients and professionals alike may neither know whether the claims are true or false, nor have the necessary skills or tools to assess their reliability.[5–11] As a result, people who believe and act on unvalidated claims may suffer by doing things that can be harmful, and by not doing things that can help. Either way, personal and societal resources for healthcare will be wasted.[12]

The Informed Health Choices (IHC) project aims to support the use of research evidence by patients and the public, policy-makers, journalists and health professionals. The multidisciplinary group responsible for the project includes researchers in six countries—Norway, Uganda, Kenya, Rwanda, the UK and Australia. The project is funded by the Research Council of Norway. It has been responsible for developing educational resources for schoolchildren and their parents in Uganda, with the objective of improving their ability to assess claims about treatment effects (A Nsangi, D Semakula, M Oxman, *et al.* Evaluation of resources to teach children in low income countries to assess claims about treatment effects. Protocol for a randomized trial. Accepted manuscript. 2016; D Semakula, A Nsangi, M Oxman, *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). Evaluation of the effects of these educational resources is taking place in two randomised trials (the IHC trials) in 2016 and 2017.

As our starting point for developing these educational interventions, the IHC group began by developing a list of key concepts that people need to understand to assess claims about treatment effects.[13] The generation of this list was performed by using the second edition of the book 'Testing Treatments'; by doing a literature review to identify key concepts and by reviewing critical appraisal tools for the public, journalists and health professionals.[11 13] The list of concepts (box 1) that emerged from this process was revised iteratively, based on feedback from members of the project team and the IHC advisory group. The latter includes researchers, journalists, teachers and others with expertise in health literacy, and in teaching or communicating evidence-based healthcare.[13] The resulting set list of concepts serves as a syllabus or curriculum from which researchers, teachers and others may develop interventions. It is an evolving document hosted by testingtreatments.org. The list will be subject to annual review to allow for revisions of existing concepts or identification and inclusion of additional concepts. For the remainder of this paper, we will refer to these as Key Concepts.

In our search for appropriate outcome measures for the IHC randomised trials, we conducted a systematic mapping review of interventions and outcome measures used for evaluating understanding of one or more of the Key Concepts (A Austvoll-Dahlgren, A Nsangi, D Semakula. Key concepts people need to understand to assess claims about treatment effects: a systematic mapping review of interventions and evaluation tools. Accepted paper. 2016). On the basis of the findings of this review, we concluded that the procedures and instruments available covered only a handful of the Key Concepts we had identified, and were not suitable for our purposes (A Austvoll-Dahlgren, A Nsangi, D Semakula. Accepted paper. 2016). Accordingly, we set out to develop the Claim Evaluation Tools to serve as the primary outcome measure of the IHC randomised trials evaluating the effects of the educational resources.

Although our primary target groups were children and adults in Uganda, we wanted to create a set of tools—a database—which would be relevant in other settings. Four important elements underpinned the development of the Claim Evaluation Tools. These tools should (1) measure objectively people's ability to apply the Key Concepts (ie, not rely on self-assessment of own abilities); (2) be flexible and easily adaptable to particular populations or purposes; (3) be rigorously evaluated; and (4) be freely available for non-commercial use by others interested in mapping or evaluating people's ability to apply some or all of the Key Concepts.

### Objective

To describe the development of the Claim Evaluation Tools, a set of flexible tools to measure people's ability to assess claims about treatment effects.

## METHODS

The development of the Claim Evaluation Tools included four processes, using qualitative and quantitative methods, over 3 years (2013–2016). These phases were: (1) determining the scope of the Claim Evaluation Tools and development of items; (2) an expert item review and feedback (face validity); (3) cognitive interviews with end-users—including children, parents, teachers and patient representatives—to assess relevance, understanding and acceptability; and (4) piloting and practical administrative tests of the items in different contexts. For clarity, we have described the methods and findings of each of these processes separately. However, development was iterative, with the different processes overlapping and feeding into each other. Researchers affiliated with the IHC project in six countries (Uganda, Norway, Rwanda, Kenya, the UK and Australia) contributed to the development of the Claim Evaluation Tools. An overview of the development process is presented in figure 1. The roles and purposes of the different research teams are described below.

### Development of items

The Claim Evaluation Tools working group, with members of the IHC group from Norway, the UK and Uganda (AA-D, AO, IC, DS, AN), had principal responsibility for agreeing on content, including the instructions and wording of individual items. The team in Norway (AA-D and AO) coordinated the development and evaluations. The scope of the Claim Evaluation Tools was based on the list of Key Concepts[13] (see box 1).

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, of which some may be more or less relevant to certain populations or 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 to test her students by picking items from the database that specifically addressed those Key Concepts.

Multiple-choice items are well suited for assessing application of knowledge, interpretation and judgements. In addition, they help problem-based learning and practical decision-making.[14] Each of the items we

**Box 1** Short list of 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 (ie, ‘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; CIs 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-performed 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)*

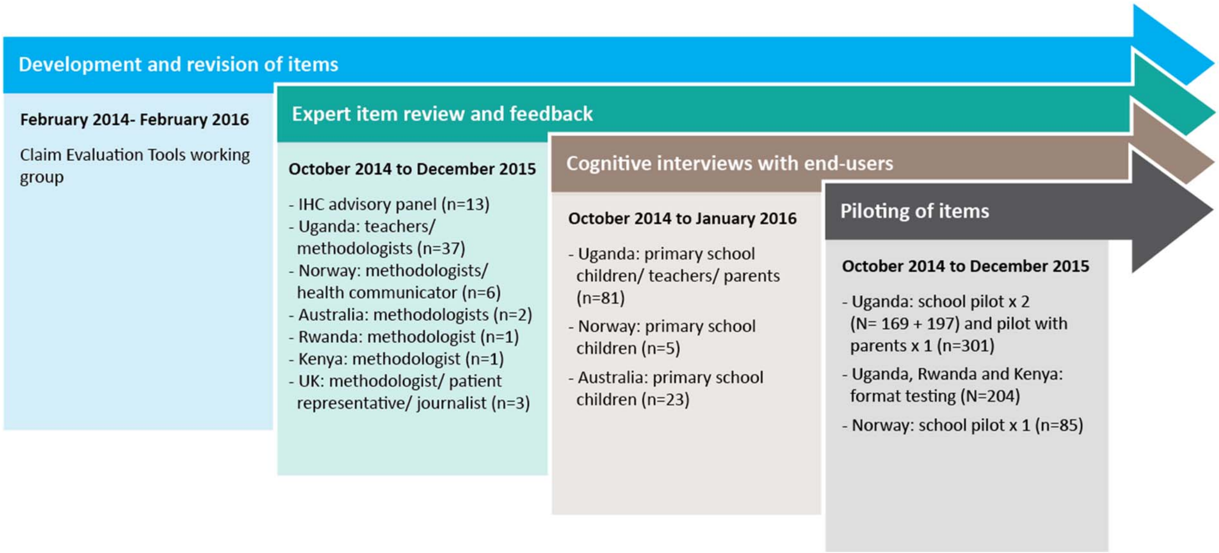

Development and revision of items
February 2014- February 2016
Claim Evaluation Tools working group
Expert item review and feedback
October 2014 to December 2015
- IHC advisory panel (n=13)
- Uganda: teachers/ methodologists (n=37)
- Norway: methodologists/ health communicator (n=6)
- Australia: methodologists (n=2)
- Rwanda: methodologist (n=1)
- Kenya: methodologist (n=1)
- UK: methodologist/ patient representative/ journalist (n=3)
Cognitive interviews with end-users
October 2014 to January 2016
- Uganda: primary school children/ teachers/ parents (n=81)
- Norway: primary school children (n=5)
- Australia: primary school children (n=23)
Piloting of items
October 2014 to December 2015
- Uganda: school pilot x 2 (N= 169 + 197) and pilot with parents x 1 (n=301)
- Uganda, Rwanda and Kenya: format testing (N=204)
- Norway: school pilot x 1 (n=85)

**Figure 1** Overview and timeline of the development process.

created opened with a scenario leading to a treatment claim and a question, followed by a choice of answers. We developed the items using two multiple-choice formats—single multiple-choice items (addressing one concept), and multiple true-false items (addressing several concepts in the same item). We developed all items with 'one-best answer' response options,[14] the options being placed on a continuum, with one answer being unambiguously the 'best' and the remaining options as 'worse'. We developed all items in English.

The initial target groups for the Claim Evaluation Tools were fifth grade children (aged 10–12 years in the next to last year of primary school) and adults (parents of primary school children) in Uganda. However, throughout the development process, our goal was to create a set of tools that we hoped would be relevant in other settings. Accordingly, we used conditions and treatments that we judged likely to be relevant across different country contexts. Where necessary, we explained the conditions and treatments used in the opening scenarios. We also decided to avoid conditions and treatments that might lead the respondents to focus on the specific treatments (about which they might have an opinion or prior knowledge), rather than on the concepts.

## Exploring relevance, understanding and acceptability of items

In order to get feedback on the relevance, understanding and acceptability of items, we used purposeful sampling of people with expertise in the Key Concepts, as well as patients and members of the public from low-income and high-income countries.[15–18]

### Item review and feedback by methodologists (face validity)

First, we circulated the complete set of multiple-choice items to members of the IHC advisory group and asked them to comment on their face validity and applicability as judged against the list of Key Concepts. Each advisory group member was assigned a set of three concepts, with associated items. A feedback form asked them to indicate to what extent they felt each item addressed the relevant Key Concept using the response options 'Yes', 'No' or 'Uncertain', together with any open-ended comments. Any items that were tagged as 'No' or 'Uncertain' by one or more of those consulted were considered for revision.

On two occasions, we also invited four methodologists associated with the Norwegian research group and with expertise in the concepts to respond to the full set of items. These experts were not involved in the project or the development of the Claim Evaluation Tools. In this element of the evaluation, the response options were randomised and the methodologists were blinded to the correct answers. They were asked to choose what they judged to be the best answer to each item's question, and were encouraged to provide open-ended comments and flag any problems they identified. Any item in which one or more of the methodologists failed to identify the 'best answer' was considered for potential revision.

We also invited people with expertise in the Key Concepts from all project partner countries to provide feedback on several occasions throughout the development of the tools. In addition to providing general feedback, an important purpose of reviewing the items in these different contexts was to identify any terminology and examples (conditions and treatments) that might be culturally inappropriate.

For all of this feedback, suggested revisions and areas of improvement were summarised in an Excel worksheet in two categories: (1) comments of a general nature relating to all items, such as choice of terminology or format; and (2) comments associated with specific items.

### Cognitive interviews with end-users on relevance of examples, understanding and acceptability

After the Claim Evaluation Tools working group and the IHC project group agreed on the instrument content, we undertook cognitive interviews with individuals from our potential target groups in Uganda, Australia, the UK and Norway.[19–21] Country representatives of the IHC project group recruited participants in their own contexts, based on purposeful sampling, in consultation with the Norwegian coordinator (AA-D). Since Uganda has been the principal focus of our interest, this was always our starting and ending point. In total, four rounds of interviews took place in Uganda. We organised interviews in Norway, the UK and Australia to assess relevance within those settings. We used these interviews to obtain feedback from potential end-users on the relevance of the scenarios (such as the conditions and treatments used in the examples), and the intelligibility and acceptability of the scenarios, formats and instructions. This was particularly important because we intended to use the items for testing children as well as adults. Throughout this process, we also piloted and user-tested several versions of the items (designs and instructions). Failure to address these issues when developing the items might increase the likelihood of missing responses, 'guessing' or other measurement errors. For example, we wanted to minimise the influence of people's cultural background on how they responded to the multiple-choice items. The effects of such confounders have been addressed in the final phase of development using psychometric testing and the Rasch analysis of the questionnaire (A Austvoll-Dahlgren, Ø Guttersrud, A Nsangi, *et al*. group. TI. Measuring ability to assess claims about treatment effects: a latent trait analysis of the 'Claim Evaluation Tools' using Rasch modelling. Submitted paper. 2016). The interviews were intended to help prevent problems resulting from confounders relatively early in the evaluation process.

Our interviews were performed iteratively between October 2014 and January 2016, allowing for changes to the items between interviews. All our interviews used a semistructured interview guide (see online supplementary appendix 1) inspired by previous research.[19–21] As part of the interviews, participants were given a sample set of the multiple-choice items and asked to respond to these. The interviews addressed questions raised during development of the items about the format of questions or the terminology used in the questions. In response, we revised the interview guide and changed the multiple-choice items when relevant. When conducting the interviews, we used the methods of 'think aloud' and 'verbal probing', two approaches to cognitive interviewing.[20] With 'think aloud', the respondent is asked to explain how they arrived at their response to each item. Such interviews are less prone to bias because of the more limited role of the interviewer. However, some respondents have difficulty in verbalising their thought processes, and in these circumstances, we followed up with 'verbal probing', which uses questions that the interviewer asks after the respondent has completed each of the items. Following each item, the interviewer began with the 'think aloud' method by asking respondents how they arrived at their response before asking more specific questions, as necessary. We audio recorded interviews when possible, and we aimed to have two people doing the interviews (with one person taking notes and the other person being the lead-interviewer). For practical reasons, this was not always possible. Each country representative summarised the key points from the interviews. Suggested revisions and areas of improvement were fed back to the Norwegian coordinator who entered these into the same Excel spreadsheet, as also the feedback from the methodologists.

### Piloting of sample sets of Claim Evaluation Tools

We conducted five small pilots in which we administered sample sets of the Claim Evaluation Tools to our target groups. As previously stated, the Key Concept list serves as a syllabus or curriculum from which researchers, teachers and others may develop interventions. Likewise, we developed the Claim Evaluation Tools, so that researchers and others can pick items that are relevant for their purposes. In other words, they can design their own instrument. The IHC interventions were initially developed to target 22 Key Concepts that were prioritised as most relevant for our target populations in Uganda. We have developed 2 instruments addressing the 22 Key Concepts targeted by the IHC interventions by selecting relevant items from the Claim Evaluation Tools database. For the pilots reported in this paper, we included items that were relevant for the IHC trials, to test how sample sets of Claim Evaluation Tools would work in a practical setting as well as to obtain an indication of the sample sizes required for the randomised trials.

The first pilot (March–April 2015) was an administrative test in a primary school in Uganda. This involved a group of children who had taken part in a pilot of the IHC primary school resources as part of the IHC project, and a comparison group of children who had not received training in the Key Concepts (in total 169 children). We included all items addressing the 22 Key Concepts. Owing to the large number of items to be tested, we divided them into four sample set questionnaires. We designed these questionnaires to be similar to the questionnaires to be used in the IHC trials. This would provide us with some feedback on how administrating a set of the Claim Evaluation Tools would work in practice, in a classroom setting. We also wanted to explore potential problems with incorrectly completed responses (through visual inspection of the responses).

The second pilot (September to December 2015) focused solely on format testing. Three different sets of formats were tested, but with the same items addressing 22 of the 32 Key Concepts kept constant across the 3 formats. We designed the formats based on lessons learnt from the feedback from methodologists, interviews with end-users and through visual inspection of the data collected in the first pilot. We recruited people in Uganda, Rwanda and Kenya to do this (N=204), using purposeful sampling of children and adults. The outcome of this test was the number of missing or incorrectly completed responses per item.

The third, fourth and fifth pilots had two objectives. The first was to compare the ability of people who had and had not received training to apply the Key Concepts. This provided an indication of the sample sizes that would be needed for the IHC randomised trials. The second objective was to estimate the frequency of missing responses as an indication of problems with understanding the item's instructions. For these purposes, we used one sample set of the Claim Evaluation Tools (addressing the 22 basic concepts). In these pilots, we also observed the time required to complete a sample set of the questionnaire. To fit an evaluation using the Claim Evaluation Tool on a busy school day as part of the IHC intervention, we hoped that it would be possible to complete a sample set questionnaire within an hour.

The third pilot (October to November 2015) and fourth pilot (November to December 2015) were conducted with Ugandan primary schoolchildren (in two schools) and their parents. The fifth pilot (December 2015) took place in Norway and included primary schoolchildren in one school. We recruited children and adults who had taken part in piloting IHC primary school materials and podcast, respectively, and children and parents who had received no such intervention. In total, 197 children took part in the Ugandan school pilot, 301 adults took part in the podcast pilot and 85 children took part in the Norwegian school pilot. The results of these pilots were summarised by calculating mean correct responses to all items addressing the same concept. We also calculated missing responses per item.

## RESULTS

We present the results thematically, beginning with the development of items, and the subsequent issues that were explored as part of the development process; judgement of relevance of the items to the Key Concepts (face validity); understanding and perceived difficulty of content; preference and understanding of instructions (formats); timing; and correct responses. An overview of the sources of feedback we used to explore these themes, our main findings and our revisions are shown in table 1.

### Development of items

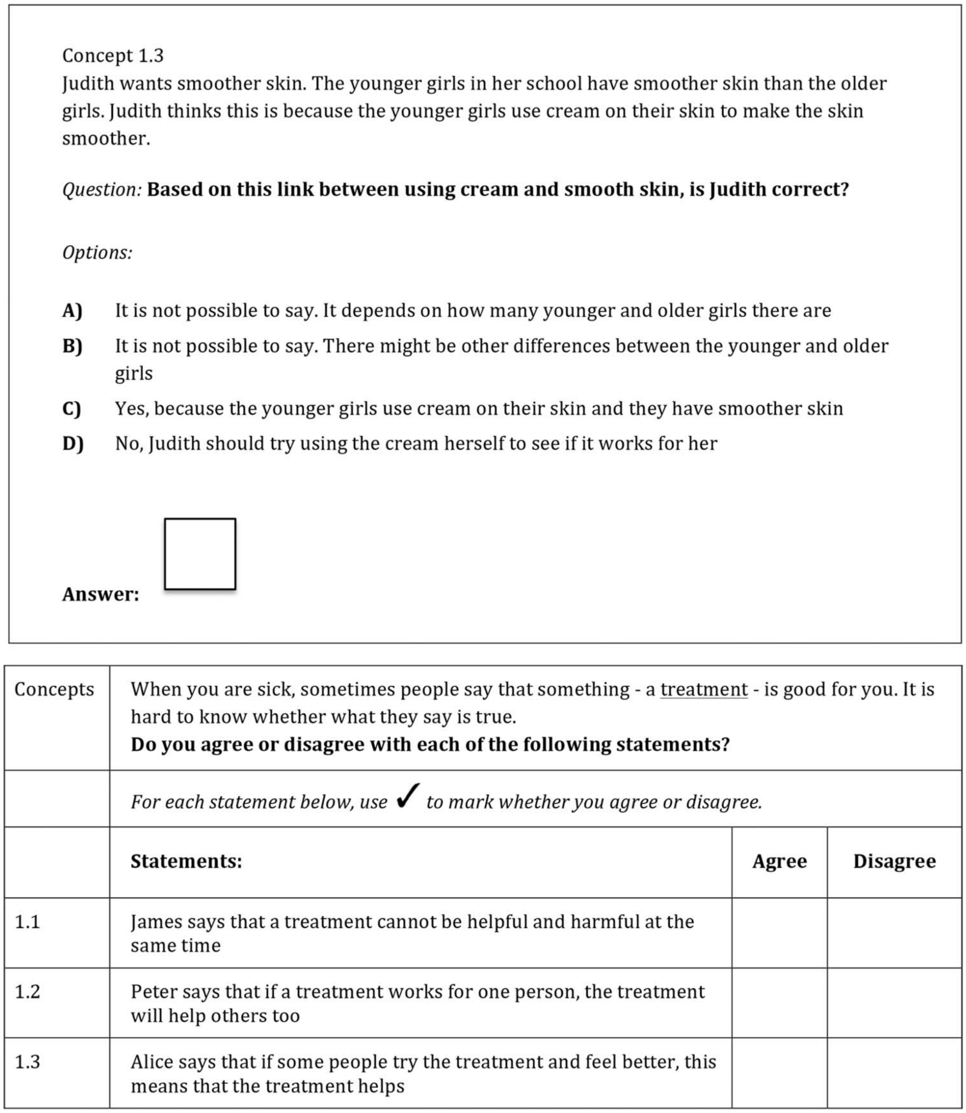
We developed items using two formats. The single multiple-choice items address only one Key Concept within each item; the multiple true–false items include questions that relate to three or more Key Concepts. The two different formats are shown in figure 2. We created an initial batch of 4–6 items addressing each Key Concept. Since we did not know which formats would be preferred by end-users, or which items would have the best psychometric properties, this allowed us to remove items based on feedback from experts, end-users and through the final psychometric testing and Rasch analysis (A Austvoll-Dahlgren, Ø Guttersrud, A Nsangi, *et al*. Submitted paper. 2016).

### Exploring relevance (face validity)

Judgements about the relevance of items to the Key Concepts were made by methodologists and people with expertise in the Key Concepts. The first phase included feedback from our advisory group: 13 members provided feedback on 135 items. Only one of these items was judged to have addressed the concept inadequately; a further 20 items were deemed to be only partly relevant. The relevance of the items was confirmed in the test-run with the Norwegian research group using the four invited methodologists, as well as by people with expertise in the Key Concepts from the project partner countries.

**Table 1** Overview of main findings and decisions about revisions, by theme

| Theme | Type of feedback | Findings | Revisions |
|---|---|---|---|
| Relevance of the items to the Key Concepts (face validity) | ▸ Methodologists and people with expertise in the Key Concepts | ▸ Most items were judged as relevant | ▸ Minor revisions, items that were found to be partly relevant (20) or not relevant (1), were considered by the working group |
| Understanding and perceived difficulty of content | ▸ Methodologists and people with expertise in the Key Concepts<br>▸ Cognitive interviews with end-users | ▸ The 'distance' between the 'best' option and the 'worse' options was considered too small<br>▸ Low literacy skills in the target audience raised as a concern<br>▸ Certain terminology identified as problematic | ▸ The worse options made more 'wrong'<br>▸ Reduction in text<br>▸ Adding explanations of terminology and rewriting of scenarios |
| Preference and understanding of instructions (formats) | ▸ Cognitive interviews with end-users<br>▸ Piloting of sample sets of the Claim Evaluation Tool (pilots 1 to 5) | ▸ A mix of the simple-multiple choice and multiple true–false formats preferred<br>▸ Formats acceptable and recognisable<br>▸ Misunderstandings of instructions; open-answers provided and checking of multiple checkboxes | Redesign of formats and instructions to remove unnecessary open spaces, avoiding use of multiple check-boxes, and the use of grids in multiple true–false options |
| Timing and correct responses | Piloting of sample sets of the Claim Evaluation Tool (pilots 3 to 5) | ▸ 30–60 min to complete a questionnaire that included demographic questions and a sample of 29 items<br>▸ Participants who had taken part in piloting of the IHC resources did slightly better than others for most of the Key Concepts | No revisions |



Concept 1.3

Judith wants smoother skin. The younger girls in her school have smoother skin than the older girls. Judith thinks this is because the younger girls use cream on their skin to make the skin smoother.

*Question:* **Based on this link between using cream and smooth skin, is Judith correct?**

*Options:*

**A)** It is not possible to say. It depends on how many younger and older girls there are

**B)** It is not possible to say. There might be other differences between the younger and older girls

**C)** Yes, because the younger girls use cream on their skin and they have smoother skin

**D)** No, Judith should try using the cream herself to see if it works for her

**Answer:** □

| | | | |
|---|---|---|---|
| Concepts | When you are sick, sometimes people say that something - a treatment - is good for you. It is hard to know whether what they say is true.<br>**Do you agree or disagree with each of the following statements?** | | |
| | *For each statement below, use ✓ to mark whether you agree or disagree.* | | |
| | **Statements:** | **Agree** | **Disagree** |
| 1.1 | James says that a treatment cannot be helpful and harmful at the same time | | |
| 1.2 | Peter says that if a treatment works for one person, the treatment will help others too | | |
| 1.3 | Alice says that if some people try the treatment and feel better, this means that the treatment helps | | |

**Figure 2** Example of formats.

### Understanding and perceived difficulty of content

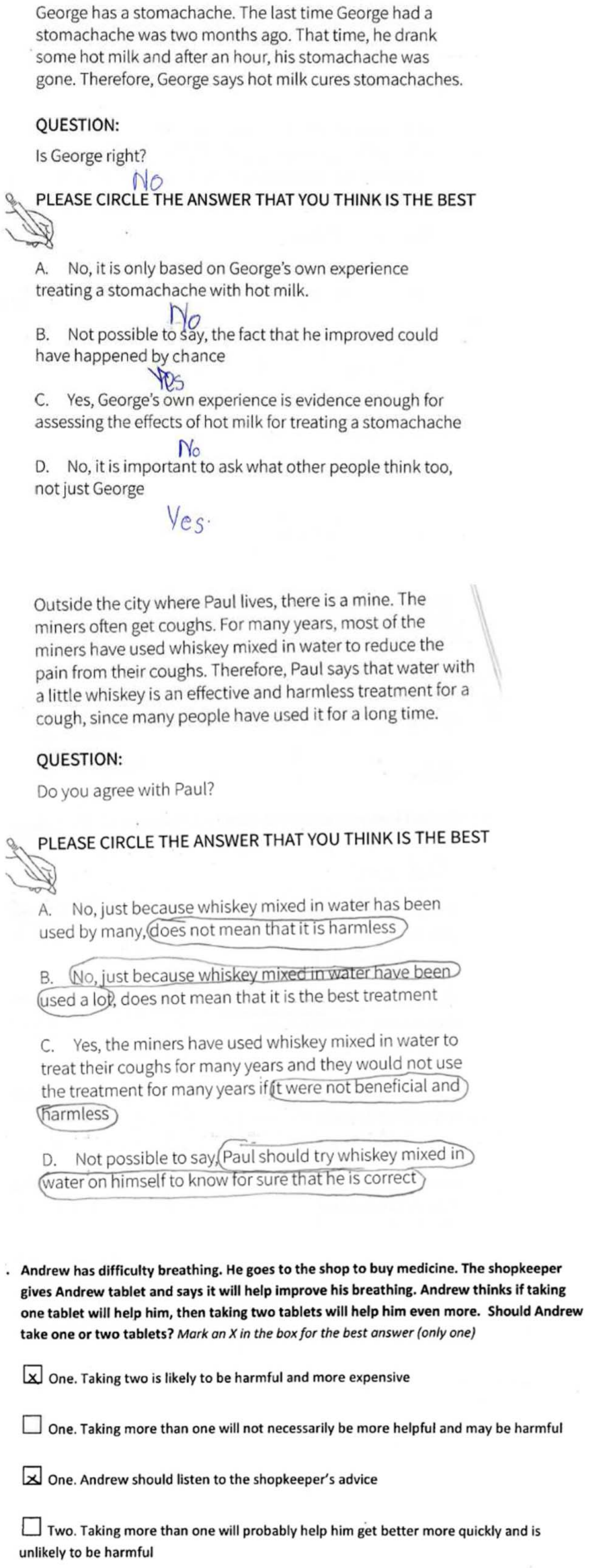
Understanding of formats and acceptability was explored by consulting methodologists and other people with expertise in the Key Concepts, as well as through cognitive interviews with end-users. Although the items were judged to be relevant, an important element of the feedback from the test-run was that the 'distance' between the 'best' option and the 'worse' options was considered too small, with the result that the judgements required were too difficult. On the basis of this feedback, we revised the 'worse' options to make them more 'wrong'.

The cognitive interviews with members of our target group also suggested that the items were too 'text heavy', and needed to be simplified. Experts and end-users in Uganda also felt that low literacy might also be a barrier. Consequently, we tried hard to make the scenarios as simple as possible without losing key content.

The end-users and the methodologists consulted in each country (Uganda, Kenya, Rwanda, the UK and Australia) also provided comments on terminology, as well as those scenarios that they felt might not be appropriate or would need to be explained. The Claim Evaluation Tools working group considered these comments and revised the items. When we were unable to avoid using certain terms (eg, 'research study'), we added explanations. Our rationale was that some terms would present a barrier to understanding the items, but were not considered to be part of the learning objectives associated with the Key Concepts. For some other terms, we used alternatives deemed acceptable by researchers, other experts and members of the target groups in each country (Uganda, Kenya, Rwanda, the UK, Australia and Norway). This process involved feeding back all changes to experts and end-users in an iterative process with continuous revisions.

### Preference and understanding of instructions (formats)

An iterative process of cognitive interviews and piloting the items using sample questionnaires informed the



George has a stomachache. The last time George had a stomachache was two months ago. That time, he drank some hot milk and after an hour, his stomachache was gone. Therefore, George says hot milk cures stomachaches.

QUESTION:

Is George right?

No

PLEASE CIRCLE THE ANSWER THAT YOU THINK IS THE BEST

A. No, it is only based on George's own experience treating a stomachache with hot milk.

No

B. Not possible to say, the fact that he improved could have happened by chance

Yes

C. Yes, George's own experience is evidence enough for assessing the effects of hot milk for treating a stomachache

No

D. No, it is important to ask what other people think too, not just George

Yes.

Outside the city where Paul lives, there is a mine. The miners often get coughs. For many years, most of the miners have used whiskey mixed in water to reduce the pain from their coughs. Therefore, Paul says that water with a little whiskey is an effective and harmless treatment for a cough, since many people have used it for a long time.

QUESTION:

Do you agree with Paul?

PLEASE CIRCLE THE ANSWER THAT YOU THINK IS THE BEST

A. No, just because whiskey mixed in water has been used by many, does not mean that it is harmless

B. No, just because whiskey mixed in water have been used a lot, does not mean that it is the best treatment

C. Yes, the miners have used whiskey mixed in water to treat their coughs for many years and they would not use the treatment for many years if it were not beneficial and harmless

D. Not possible to say, Paul should try whiskey mixed in water on himself to know for sure that he is correct

. **Andrew has difficulty breathing. He goes to the shop to buy medicine. The shopkeeper gives Andrew tablet and says it will help improve his breathing. Andrew thinks if taking one tablet will help him, then taking two tablets will help him even more. Should Andrew take one or two tablets?** *Mark an X in the box for the best answer (only one)*

☒ **One. Taking two is likely to be harmful and more expensive**

□ **One. Taking more than one will not necessarily be more helpful and may be harmful**

☒ **One. Andrew should listen to the shopkeeper's advice**

□ **Two. Taking more than one will probably help him get better more quickly and is unlikely to be harmful**

**Figure 3** Examples of incorrectly completed multiple-choice questions.

design and formats of the instructions. Our interviews with end-users were to obtain their preferences on format, to follow the steps of their reasoning when responding to the items and to assess their understanding of the items' instructions. The main message was that people preferred a mix of the simple-multiple choice and multiple true–false formats to make the questionnaire more interesting. The items were otherwise well received. The general feedback from all the different country settings was that the formats were acceptable, recognisable and similar to the multiple-choice formats they had encountered in other settings.

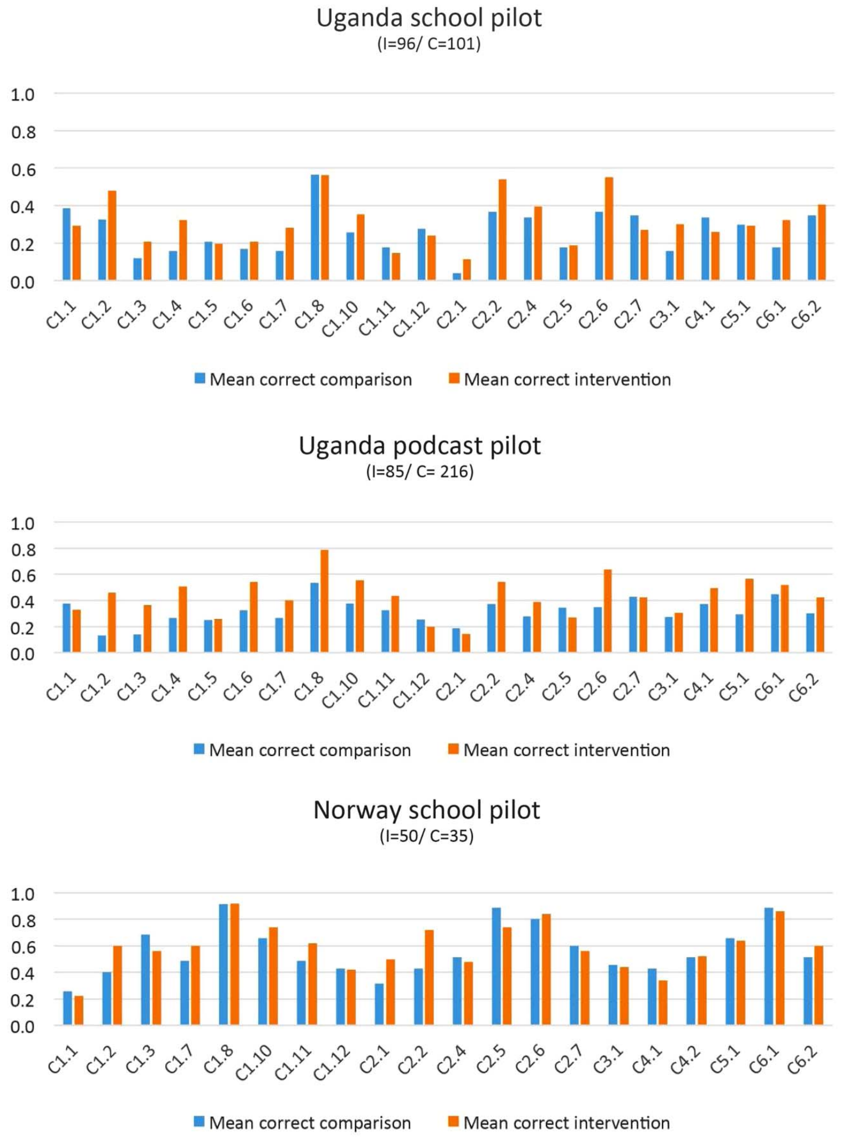
On the basis of verbal feedback in the interviews with the end-users, as well as visual inspection of how people responded to the items in the five pilots, we identified two potential problems. The first was that respondents tended to provide open-ended responses to the questions; the second was that people tended to tick more than one checkbox. Owing to these problems, the mean missing/incorrectly completed responses in the first school pilot in Uganda (March–April 2015) was 20–40%. Examples of such incorrectly completed multiple-choice items from this first pilot are shown in figure 3.

We tested revised designs (figure 2) in the second pilot in Uganda, Rwanda and Kenya (September to December 2015). This greatly improved people's responses to the questionnaire, reducing missing or incorrectly completed responses to <4% of the items. On the basis of this pilot, we made final revisions and decided on the formats to be used in the subsequent pilots.

Figure 2 shows the design changes we used to avoid these problems. These included removing blank spaces, which could be misinterpreted as inviting open (free text) responses; and avoiding use of multiple checkboxes for 'one-best answer' formats. For the multiple true–false formats, response options using an open grid design, with instructions at the top, resulted in fewer problems.

The third, fourth and fifth pilots, conducted in Uganda and Norway (October to December 2015), confirmed the appropriateness of the formats, and missing or incorrectly completed responses were <2%. These pilots also confirmed that respondents took between 30 and 60 min to complete a questionnaire that included demographic questions and a sample of 29 items. The participants' correct responses per Key Concept are shown in figure 4. This figure, in which correct answers are plotted for each Key Concept per group, shows that participants who had taken part in piloting the IHC resources were slightly more likely than others to give correct answers for most of the Key Concepts.

**Figure 4** Distribution of correct answers in pilots.



## DISCUSSION

Developing a new evaluation instrument is not straightforward, and requires rigorous testing using qualitative and quantitative methods.[22] There are many ways of doing this. We chose to use a pragmatic and iterative approach, involving feedback from experts and end-users and continuous revisions. This development work was possible because we are a multidisciplinary, international collaboration including people from high-income and low-income countries. Despite differences between countries, enabling people to assess treatment claims in their daily lives is a challenge in all countries.

We developed a battery of multiple-choice items using two formats, with several items addressing each Key Concept. An international group of people with relevant expertise considered that the items we developed addressed the Key Concepts we had identified appropriately, and end-users considered the items to be acceptable in their settings. Methodologists and end-users suggested that some items were too difficult, so we revised the answer options, reduced the amount of text used and explained terminology if necessary. Based on feedback from the interviews with end-users, the revised formats were well received, but the piloting also identified issues with understanding of instructions. We addressed these problems by further testing and redesign of instructions and formats. This resulted in a reduction in missing or incorrectly completed responses in subsequent pilots. Piloting of sample sets of Claim Evaluation Tools also confirmed that it was possible to complete a questionnaire with 29 items within an hour, and that people who had received training in the Key Concepts did slightly better than those who had not received such training.

The relevance of the items outside the contexts studied as part of this project is unclear. Feedback from end-users in other settings may be different. Researchers or teachers who would like to use the Claim Evaluation Tools in their contexts should consider the relevance of terminology and the examples used, involving end-users if possible. It should also be noted that the first phases in the development described in this paper did not include any evaluations of the reliability of the items. This requires rigorous psychometric testing including Rasch analysis, and is described in a separate paper (A Austvoll-Dahlgren, Ø Guttersrud, A Nsangi, *et al*. Submitted paper. 2016).

This paper describes the development and initial steps of validation of items addressing all 32 of the Key Concepts in 4 phases. However, in the last phase, we also did some pilot testing of items referring specifically to 22 of the 32 Key Concepts. There were several objectives of these pilots, but for development purposes, we wanted to do practical administrative tests to explore the understanding of formats and timing of Claim Evaluation Tools 'sample tests'. A limitation of these pilots is that people may respond differently to the items addressing the 10 Key Concepts not included, in terms of number of missing responses, incorrectly filled in questions or in time to completion. We judge this to be of little importance as the items addressing these two 'groups' of Key Concepts use the same formats and are similar in length and language.

As our first step in the choice of outcome measurement for the IHC trials, we conducted a systematic mapping review of interventions and outcome measures used for evaluating one or more of the Key Concepts (A Austvoll-Dahlgren, A Nsangi, D Semakula. Accepted paper. 2016). Our findings suggested that research on the Key Concepts is of interdisciplinary interest, and that a variety of assessment tools exists. However, none of the identified tools addressed more than 15 Key Concepts. The most relevant of these were instruments designed to assess competency in evidence-based medicine, the Fresno test by Ramos *et al*[23] and an instrument developed by Godwin and Seguin.[24] Assessment tools used in studies targeting patients or consumers included only seven or fewer Key Concepts. The large majority of these generally only touched on one concept—5.1 'Weigh benefits and harms of treatment'.[25] The Claim Evaluation Tools were developed to be used as the primary outcome measurement in the IHC project's randomised trials, as well as to provide a flexible measurement tool for others interested in mapping or evaluation of people's ability to apply Key Concepts when assessing claims about treatment effects. Instead of a 'set' instrument, the Claim Evaluation Tools offers the potential to tailor an instrument for specific purposes and target groups. As a consequence, educators and researchers have the opportunity to adapt the Claim Evaluation Tool by selecting a sub-sample of Key Concepts that best fit their learning goals or research aims. We envision that educators, researchers and others will use them to create their own 'tests', fitting their specific needs and contexts. The Claim Evaluation Tools also appear to be unique in that the items have been developed to be used to assess ability in children and adults, including members of the public as well as health professionals. This offers the opportunity to compare knowledge and application of the Key Concepts across populations.

The Claim Evaluation Tools were developed as objective multiple-choice items to measure understanding of the Key Concepts. A limitation of many of the instruments that have been developed to assess people's critical-appraisal skills is that they rely on self-report by respondents (subjective measurements). Typical examples are the many health literacy instruments, such as the European Health Literacy Survey (HLS-EU)[26] and instruments used to assess competence in evidence-based medicine.[27] Self-assessed abilities can be difficult to interpret, and have been found to have a weak association with objective measures of knowledge and skills.[28–30] Such instruments may be more likely to measure the confidence of respondents in their own ability rather than their knowledge or actual ability. Although improved confidence in one's own ability may be a relevant and important effect of an intervention, it may be a poor indicator of actual knowledge and ability.

## CONCLUSION

We developed the Claim Evaluation Tools to evaluate people's ability to assess claims about the effects of treatments. As far as we are aware, this is currently the only evaluation instrument designed to address most of the Key Concepts we believe people need to know to assess claims about treatment effects. This work is the result of a multidisciplinary, international collaboration including high-income and low-income countries. We have used a pragmatic and iterative approach, involving feedback from experts and end-users, and continuous revisions. Although the Claim Evaluation Tools have been developed primarily to be used as part of the IHC project in Uganda, we believe they should be useful for others interested in evaluating people's ability to apply Key Concepts when assessing treatment claims. Feedback from experts and end-users in Uganda, Kenya, Rwanda, Norway, the UK and Australia supports our hope that they will be found relevant in other contexts.

The Claim Evaluation Tools include a battery of items from which researchers can select those relevant for specific populations or purposes, and currently include ∼190 multiple-choice items. However, we anticipate that the Claim Evaluation Tools will continue to evolve. The Claim Evaluation Tools is hosted on the Testing Treatments interactive website (http://www.testingtreatments.org) and managed by the Claim Evaluation Tools working group. On request, all items will be made freely available for non-commercial use.

**Author affiliations**
[1]Norwegian Institute of Public Health, Oslo, Norway
[2]Makerere University College of Health Sciences, Kampala, Uganda
[3]James Lind Initiative, Oxford, UK
[4]Norwegian Centre for Science Education, University of Oslo, Oslo, Norway

**Acknowledgements** The authors are deeply grateful to all of the enthusiastic children, parents and teachers who contributed to this project. The authors would also like to thank the IHC advisory panel, and other experts who provided their advice. In particular, the authors would like to thank Sophie Robinson, Ruth Davis, Andrew Garratt, Chris Del Mar, Susan Munabi Babigumira, Jenny Moberg, Signe Agnes Flottorp, Simon Goudie, Esme Lynch and Gunn Vist.

**Collaborators** The IHC Group: Leila Cusack, Claire Glenton, Tammy Hoffmann, Margaret Kaseje, Simon Lewin, Leah Atieno Marende, Angela Morrelli, Michael Mugisha, Laetitia Nyirazinyoye, Kjetil Olsen, Matthew Oxman, Nelson K Sewamkambo, Anne Marie Uwitonze.

**Contributors** AA-D, ØG and AO wrote the protocol and the IHC group provided comments on the protocol. AA-D coordinated all of the development and evaluation process with support from AO. AA-D, DS, AN, IC, AO and MO drafted the items with input from the IHC group. AA-D, IC and AO were responsible for revising the items iteratively, based on the feedback received through the different processes. SR and AA-D were responsible for the format designs. AA-D and KO approached the IHC advisory group and other methodologists for feedback. AN, DS, LC, AA-D, MO, SR, TH and MK conducted the interviews with end-users. AN, DS, KO, LAM, MM, MK, NS, AMU and AA-D were involved in pilots and administrative tests in Uganda, Kenya, Rwanda and Norway, respectively. AA-D analysed the data from the pilots. AA-D authored this manuscript with significant input from the rest of the IHC group.

**Funding** The IHC project is funded in part by the Research Council of Norway-GLOBVAC project 220603.

**Competing interests** None declared.

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

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** All data are published as part of this study. All Claim Evaluation Tools are available on 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 Sociol Rev* 2010;19:57–72.
2. Glenton C, Paulsen E, Oxman A. Portals to Wonderland? Health portals lead confusing information about the effects of health care. *BMC Med Inform Decis Mak* 2005;5:7.
3. Moynihan R, Bero L, Ross-Degnan D, *et al*. Coverage by the news media of the benefits and risks of medications. *N Engl J Med* 2000;342:1645–50.
4. Wolfe RM, Sharp LK, Lipsky MS. Content and design attributes of antivaccination web sites. *JAMA* 2002;287:3245–8.
5. Woloshin S, Schwartz L, Byram S, *et al*. Women's understanding of the mammography screening debate. *Arch Intern Med* 2000;160:1434–40.
6. Fox S, Duggan M. Health Online 2013. 9 April 2013. http://www.pewinternet.org/Reports/2013/Health-online.aspx
7. Robinson E, Kerr C, Stevens A, *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, *et al*. How do patients evaluate and make use of online health information? *Soc Sci Med* 2007;64:1853–62.
9. Horsley T, Hyde C, Santesso N, *et al*. Teaching critical appraisal skills in healthcare settings. *Cochrane Database Syst Rev* 2011;(11):CD001270.
10. Stacey D, Bennett CL, Barry MJ, *et al*. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev* 2011;(10):CD001431.
11. Evans I, Thornton H, Chalmers I, *et al*. *Testing Treatments: better research for better healthcare*. 2nd edn. London: Pinter & Martin Ltd, 2011. http://www.testingtreatments.org/new-edition/
12. Chalmers I, Glasziou P, Badenoch D, *et al*. Evidence Live 2016: Promoting informed healthcare choices by helping people assess treatment claims. *BMJ* 26 June 2016. http://blogs.bmj.com/bmj/2016/05/26/evidence-live-2016-promoting-informed-healthcare-choices-by-helping-people-assess-treatment-claims/
13. Austvoll-Dahlgren A, Oxman AD, Chalmers I, *et al*. Key concepts that people need to understand to assess claims about treatment effects. *J Evid Based Med* 2015;8:112–25.
14. Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences*. 3rd edn. National Board of Medical Examiners: Philadelphia, USA, 2002.
15. Williamson PR, Altman DG, Blazeby JM, *et al*. Developing core outcome sets for clinical trials: issues to consider. *Trials* 2012;13:132.
16. Cooney RM, Warren BF, Altman DG, *et al*. Outcome measurement in clinical trials for ulcerative colitis: towards standardisation. *Trials* 2007;8:17.
17. Tugwell P, Boers M, Brooks P, *et al*. OMERACT: an international initiative to improve outcome measurement in rheumatology. *Trials* 2007;8:38.
18. Basch E, Aronson N, Berg A, *et al*. Methodological standards and patient-centeredness in comparative effectiveness research the PCORI perspective. *JAMA* 2012;307:1636–40.
19. Watt T, Rasmussen AK, Groenvold M, *et al*. Improving a newly developed patient-reported outcome for thyroid patients, using cognitive interviewing. *Qual Life Res* 2008;17:1009–17.
20. McColl E, Meadows K, Barofsky I. Cognitive aspects of survey methodology and quality of life assessment. *Qual Life Res* 2003;12:217–18.
21. Bloem EF, van Zuuren FJ, Koeneman MA, *et al*. Clarifying quality of life assessment: do theoretical models capture the underlying cognitive processes? *Qual Life Res* 2008;17:1093–102.
22. Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010;19:539–49.
23. Ramos KD, Schafer S, Tracz SM. Validation of the Fresno test of competence in evidence based medicine. *BMJ* 2003;326:319–21.
24. Godwin M, Seguin R. Critical appraisal skills of family physicians in Ontario, Canada. *BMC Med Educ* 2003;3:10.
25. O'Connor AM. Validation of a decisional conflict scale. *Med Decis Making* 1995;15:25–30.
26. Sorensen K, Pelikan JM, Rothlin F, *et al*. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *Eur J Public Health* 2015;25:1053–8.
27. Shaneyfelt T, Baum KD, Bell D, *et al*. Instruments for evaluating education in evidence-based practice: a systematic review. *JAMA* 2006;296:1116–27.
28. Dahm P, Poolman RW, Bhandari M, *et al*. Perceptions and competence in evidence-based medicine: a survey of the American Urological Association Membership. *J Urol* 2009;181:767–77.
29. Khan KS, Awonuga AO, Dwarakanath LS, *et al*. Assessments in evidence-based medicine workshops: loose connection between perception of knowledge and its objective assessment. *Med Teach* 2001;23:92–4.
30. Joffe S, Cook EF, Cleary PD, *et al*. Quality of informed consent in cancer clinical trials: a cross-sectional survey. *Lancet* 2001;358:1772–7.