**Abstract**

This manual describes the rationale and use of the Claim Evaluation Tools database and will guide you in preparing a test or questionnaire for your purposes, whether you are a teacher or a researcher. The items in the database were developed as part of the Informed Health Choices project, and can be used for creating tests to be used in school and other learning settings, as outcome measures in evaluations of educational interventions, or in surveys to map abilities in a population.

**Authored by the Claim Evaluation Tools database working group:**

**Astrid Austvoll-Dahlgren, Andy Oxman and Iain Chalmers**

**Manual for preparing
a test or questionnaire based on
 the Claim Evaluation Tools database**

**Version: 22. 11. 2016**

**Contents**

[About this document 2](#_Toc467839236)

[About the multiple-choice questions 5](#_Toc467839237)

[Who are the items suitable for? 7](#_Toc467839238)

[Have the items been validated? 7](#_Toc467839239)

[Who can have access and use the items? 7](#_Toc467839240)

[How do I get started creating my own test? 8](#_Toc467839241)

[1. *Introductory page* 8](#_Toc467839242)

[2. *Demographic items* 8](#_Toc467839243)

[3. *Selecting MCQs* 8](#_Toc467839244)

[4. *Consider including reading ability items* 9](#_Toc467839245)

[5. *Consider including intended behaviour and attitude items* 9](#_Toc467839246)

[6. *Formatting all items* 9](#_Toc467839247)

[7. *Administration* 9](#_Toc467839248)

[8. *Ordering of items:* 9](#_Toc467839249)

[How many items do I need? 9](#_Toc467839250)

[a) *For teaching* 9](#_Toc467839251)

[b) *For measuring outcomes in randomized trials, surveys or validation studies* 9](#_Toc467839252)

[How should the tests be scored? 10](#_Toc467839253)

[*How should the data be entered if I want to do a validation study?* 10](#_Toc467839254)

[What sample size is needed? 10](#_Toc467839255)

[a) *For teaching* 10](#_Toc467839256)

[b) *For randomized trials or surveys* 11](#_Toc467839257)

[c) *For validation studies* 11](#_Toc467839258)

[How is the Claim Evaluation Tools database maintained? 11](#_Toc467839259)

[References 12](#_Toc467839260)

# **About this document**

This manual describes the rationale and use of the [Claim Evaluation Tools database](#Claim), and will guide you in preparing a test or questionnaire that you can use to assess people’s ability to assess [treatment](#Treatments) claims, whether you are a teacher or a researcher.

The [items](#Item) in the database were developed as part of the [Informed Health Choices project](#IHC), and can be used for creating [tests](#Test) to be used in school and other learning settings, as outcome measures in evaluations of educational interventions, or in surveys to map abilities in a population (1, 2). In this manual we will use “test” as a generic term for tests, outcome measures, and questionnaires using items from the database. Please also note that explanations of relevant terminology are given in table 1.

The Claim Evaluation Tools database consists of four types of items:

1. Multiple-choice questions measuring people’s ability to assess treatment claims,

2. Demographic items,

3. Reading ability items, and

4. Intended behaviour and attitude items

**Table 1. Terminology**

|  |  |
| --- | --- |
| Claim Evaluation Tools database | The Claim Evaluation Tools database consists of four types of items; 1. Multiple-choice questions (MCQs) measuring people’s ability to apply the Key Concepts they need to know to be able to assess treatment claims, 2. Demographic items, 3. Reading ability items, and 4. Intended behaviour and attitude items.The database includes the following information about the MCQs including:* The Key Concept that the MCQ addresses
* The correct answer
* Translations of the MCQ to languages other than English
* Contexts in which the MCQ has been tested and the following findings from each test:
	+ Validity and reliability based on fit to the Rasch model
	+ Difficulty

The database also includes reports of evaluations of sets of MCQs and protocols for evaluations, including feedback from methodologists, interviews with end-users, and Rasch analyses.The Claim Evaluation Tools database also includes demographic questions, reading ability items and self-reported behaviour and attitudes questions.  |
| Item | “Item” refers to a question that is intended to measure a specific ability, demographic characteristic, intended behaviour, or attitude. |
| MCQ | “MCQ” is an abbreviation for multiple-choice item and is synonymous with multiple-choice question.  |
| Scenario | Each MCQ for measuring people’s ability to apply a Key Concept begins with a scenario. This scenario includes a claim about a treatment effect. When scenarios are used in MCQs they are sometimes referred to as the “stem”. |
| Question | Following the scenario, there is a question about the scenario in each MCQ. |
| Options | Following the question, there are between two and five answers to the question in each MCQ. For all the MCQs currently in the Claim Evaluation Tools database, respondents are asked to select one of those options. There is one best (correct) option and other options that are not correct; so that each item is scored dichotomously as “right” or “wrong”. |
| Key Concepts | This is a list of the main ideas that people should understand and apply when assessing claims about the effects of treatments and making health choices, to avoid being misled and to enable well-informed choices (3). There are currently 34 Key Concepts, including concepts related to:* The basis for a claim and whether it is justified
* The comparisons (evidence) supporting a claim and whether they are fair and reliable
* Making informed choices
 |
| Test | We use this term generically and synonymous to questionnaire to refer to sets of items from the Claim Evaluation Tools database that are used for a specific purpose, including testing learners’ abilities, evaluating educational interventions, and surveys.  |
| Tool |  “Tools” in the name of the database is used to indicate the potential uses of sets of items taken from the database.To avoid confusion, we suggest referring to specific tools using terms that describe them (e.g. questionnaire or test), rather than as “a Claim evaluation tool”. |
| Treatments | Any action intended to improve the health or wellbeing of individuals or communities |
| The Informed Health Choices project (IHC) | The Informed Health Choices (IHC) project aims to develop learning resources to help children and adults recognise reliable and unreliable claims and make well-informed health choices. More information about the project can be found here: [www.informedhealthchoices.org](http://www.informedhealthchoices.org).  |
| Feedback from methodologists, teachers and end-users | Feedback from people with methodological (research) expertise (i.e. a deep understanding of the Key Concepts) or teachers is important for “face validity”, the extent to which a test is viewed as covering the concept it purports to measure. It refers to the transparency or relevance of a test as it appears to test participants. This includes their perceptions of the relevance of each item to the Key Concept that it addresses, their perceptions of the difficulty of each item, comments regarding terminology, problems with each item, and suggestions for improvements. For these purposes, it is also important to get feedback from end-users, including children and adults. This includes feedback on designs and instructions.  |
| Rasch analysis | Rasch analysis is a form of psychometric testing relying on Item Response Theory, and is a dynamic way of developing outcome measurement tools to achieve validity and reliability. The Rasch model relies on several assumptions; one of these is that the instrument should adhere to the Guttman pattern. Furthermore, the comparison of two people should be independent of which items are used within the set of items assessing the same variable (4-7). |
| Guttman pattern | A test that adhere to the Guttman pattern is one which a person succeeds on all the items up to a certain difficulty, and then fails on all the items above that difficulty. When individuals and items are ordered by raw score, this produces a data set with a "Guttman pattern".  |
| Unidimensionality and local dependency | The Rasch model requires that responses to any subset within a test should give the same estimate of ability. This is explored by testing for dimensionality, if more than one variable in the instrument guide people’s responses this may lead to measurement error. This requires that there is no local dependency among the items. Local dependence is the extent to which responses to one or more items within the data set relies on responses to other related items. |
| Discrimination and within item bias (DIF) | Through Rasch analysis, each item is also scrutinized for the extent to which it is able to discriminate among those with high and low ability (ability groups). Besides the item’s difficulty level, the only variable influencing people’s ability should be the latent trait (i.e. the ability to assess treatment claims). Therefore, as part of the Rasch analysis, we also look for signs of item bias such as Differential Item Functioning (DIF) based on relevant person factors such as gender or age. This means that it is undesirable that an item work differently for example for women than for men.  |
| Validity and reliability | An important aspect of validity is the trustworthiness of the score meaning and its interpretation (8). When the items fit to the Rasch model, ability is measured consistently with low measurement error (4). Therefore, the Rasch analysis also provides information about the reliability of the instrument tested. |

# **About the multiple-choice questions**

The core of the database consists of multiple-choice questions (hereafter referred to as [MCQ](#MCQ)). Multiple-choice questions are well suited for assessing application of knowledge, interpretation and judgements. In addition, they help problem-based learning and practical decision-making. Currently there are two types of MCQs in the database - single multiple-choice questions (addressing one concept), and multiple true-false items (addressing several concepts in the same item) (see figure 1 and 2). Each of the items we created opened with a [scenario](#Scenario) leading to a treatment claim and a [question](#Question), followed by a choice of response [options](#Options). We developed all items with “one-best answer” response options, the options being placed on a continuum, with one answer being unambiguously the “best” and the remaining options as “worse”.

Instead of a standard, fixed questionnaire, we have developed the Claim Evaluation [Tools](#Tools) database as a flexible battery of MCQs from which teachers, researchers and others can select those relevant for their purposes. This means that you can create your own test based on which Key Concepts you want to teach.

**Figure 1. Example of a single MCQ**



**Figure 2. Example of a multiple MCQ**



# **Who are the items suitable for?**

All items are developed to be used for children from the age of 10 and up, as well as for adults. This means that ideally the same items can be used for children and adults, including health professionals (although different sets of demographic items are recommended for children and adults) (1).

# **Have the items been validated?**

Most but not all the items in the Claim Evaluation Tools database have been and continue to be rigorously evaluated. Evaluation has included [feedback from experts, teachers and end-users](#Feedback), and [statistical testing using Rasch analysis](#Rasch).

Rasch analysis is a dynamic way of developing outcome measurement tools to achieve [validity and reliability](#Validity) (4, 5, 9). The Rasch model relies on several assumptions. One of these is that the instrument (questionnaire or test) should adhere to the [Guttmann pattern](#Guttman) (4, 9). Furthermore, the comparison of two people should be independent of which items are used within the set of items assessing the same variable in the test ([unidimensional](#Undimen)) (4, 9). This requires that there is little or no dependency between items; i.e. a person’s response to one item should not be dependent on other items ([local dependency](#Undimen)) (4, 9). An item should also work in the same way independent of person factors such as gender or age (4, 9). This means, for example, that, ideally, only one factor should guide peoples’ responses to the MCQs in the Claim Evaluation Tools database - their ability to assess treatment claims. When we test the MCQs in different settings, we test for such potential - [Differential Item Functioning (DIF)](#discrim). We are also conducting cross-cultural comparisons.

When data conform to the Rasch model, there is low risk of measurement error and ability is measured consistently (4, 9). To use a metaphor, one meter should be the same length when assessing the height of women, men, adults and children. The Rasch analysis thus provides information about the [reliability](#Validity) of the sample of items tested (4, 9).

We currently have items available in English (tested primarily in Uganda), Luganda, Spanish (tested in Mexico), Chinese, and Norwegian. A German translation is also underway. However, it should be noted that not all items have been tested using Rasch analysis, and that some items may not work in all settings. We have created an overview of the validation status for each item. This can be accessed by sending an email to asttrid.austvoll-dahlgren@fhi.no.

# **Who can access and use the items?**

The Claim Evaluation Tools database is open access and free for non-commercial use. However, we do not publish the items online to avoid learning effects, or “cheating”.

To access and use items from the Claim Evaluation Tools database, you must complete a form describing how you intend to use the items. We expect people who evaluate the items to share their findings with us, including any new items that are developed, the results of feedback from methodologists or teachers, findings from interviews with end-users, and findings from Rasch analyses. If you decide to lead a validation study in your context and to publish a paper on it, we can advise you on how to go about doing this and assist you with the analysis (see below). We strongly encourage open access publishing and aim to make all such evaluations accessible in the Claim Evaluation Tools database. Please see the form below:

****

# **How do I get started creating my own test?**

1. ***Introductory page***: We recommend that you include an introductory page in your test, explaining some of the terms used. In previous studies, we have found that people sometimes don’t understand what we mean by certain terms, and this may be a barrier to using the items. Examples of such terms include: “treatment”, “claim”, “study”, and “results”. A draft template for an introductory page can be found here:



1. ***Demographic items:*** Background information is needed for validating the items in your context, as well as for describing the participants in your sample. We suggest that you include demographic items about age and gender for all respondents. You may also consider including information about being trained to apply the Key Concepts. We have previously conceptualized this as training or participation in randomized trials, or training in evidence-based medicine, research methods, or medical statistics. Examples of demographic items used for children and adults can be found below:

********

1. ***Selecting MCQs***: When selecting MCQs you should start by looking at the Key Concepts list, and choose those that are most relevant to your purpose and target learners. For each Key Concept there are currently 1 to 6 corresponding MCQs from which to choose. We will provide you with a document including all items in your preferred language. Items are currently available in English, Luganda, Norwegian, Spanish (Mexico) and Chinese.



1. ***Consider including reading ability items***: In some contexts, people’s reading ability may be poor. Consequently, you may consider including items to test for reading ability as a person factor to be included in Rasch analysis for Differential Item Functioning. Reading ability items can be found below:

****

1. ***Consider including intended behaviour and attitude items:*** You may be interested in knowing more about learner’s intended behaviour, or their attitudes to assessing treatment claims. The database includes a set of such items that can be included in your test. These items can be found below:



1. ***Formatting all items***: The formats we use have been extensively user-tested and found to produce few missing or incomplete responses (1). Please note that if you change the format you may also change how people respond to the item. This may introduce measurement error or results that deviate from other tests using the format that we have tested.
2. ***Administration***: The test can be administered as paper-based written tests or electronic tests. In one setting, the test has also been validated as an oral test for a population with low literacy (10). In validation studies, we have tested for differential item function based on mode of administration. Results so far indicate that people respond similarly and independently of how the tests are administered.
3. ***Ordering of items:*** If the difficulty of the MCQs is known (based on validation in your setting), the ordering of the MCQs in the questionnaire or test should be from ‘easy’ to ‘difficult’.

# **How many items do I need?**

1. ***For teaching***: If you are preparing a test to evaluate how well your students performed after a lesson about one or more of the Key Concepts, you can pick whatever number of items you feel is appropriate within the available teaching timeframe. In our experience, children and adults spend about 30 minutes considering a set of 22 items.
2. ***For measuring outcomes in randomized trials, surveys or validation studies***:If you are conducting an evaluation of an educational resource, or you intend to measure people’s ability in a population survey, or you intend to perform a validation study, then reliability will be an issue. There is no gold standard for how many items are needed, but the general rule is that the more items you include, the higher the reliability. We have found that about 22 items provide sufficient reliability.

If you are conducting a validation study you also need to keep in mind that it is often necessary to remove or repair some items after validating subsets of them, so you need to plan for this. Accordingly, to allow for such revisions, we recommend that you include at least two items addressing each of the Key Concepts in which you are interested.

# **How should the tests be scored?**

Tests can be scored by calculating the overall proportion of correct responses across items. However, such scores can be difficult to interpret. To supplement this, we recommend using an absolute (criterion- referenced) standard to set a passing score (a cut score). It might also be desirable to use an absolute standard for a score that indicates mastery. Judgements about what the minimum score is for passing or to indicate mastery are inevitably pragmatic, and there are several ways of doing this (11-13). In two previous trials, evaluating the effects of educational resources in Uganda, we established criteria-referenced standards using a combination of Nedelsky’s and Angoff’s methods (14, 15). For an example of this approach, see below:



### ***How should the data be entered if I want to do a validation study?***

If you would like to contribute to validating the items in your setting, we can assist you with Rasch analysis. The data for this should be entered or exported into an Excel file. A draft template can be found below:



# **What sample size is needed?**

1. ***For teaching***: If you are preparing a test to evaluate how well your students performed after a lesson about one or more of the Key Concepts, then the sample size is simply the number of students who attended your lesson.
2. ***For randomized trials or surveys***: If you are conducting an evaluation of an educational resource in a randomized trial or intend to measure people’s ability in a population survey, then you will need to use standard approaches to sample size calculation. We will include data that can be used in these calculations in the Claim Evaluation Tools database as they become available.
3. ***For validation studies:*** If you are validating items in your setting, then the sample size depends on how many items you include and on how many person factors you intend to explore for Differential Item Functioning in the Rasch analysis. Based on previous validation studies including >22 MCQs, we recommend including at least 150 people per person factor (such as age or gender). Please also make sure that people with training in the Key Concepts are represented in your sample. Examples include training in epidemiology, statistics, evidence-based practice, research methodology, and critical thinking. Having been trained as a health professional is not a reliable indication of ability to understand and apply the [Key Concepts](http://www.testingtreatments.org/key-concepts-for-assessing-claims-about-treatment-effects/). Please see Table 2 for suggested sample sizes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample** | **Females** | **Adults** | **Total** |
| **Children** | 75+ | 75+ | 150+ |
| **Adults** | 75+ | 75+ | 150+ |
|  |  |  | 300+ |

**Table 2. Suggested sample size for validation studies**

# **How is the Claim Evaluation Tools database maintained?**

Maintenance and revisions of the Claim Evaluation Tools database will reflect changes in the list of Key Concepts. Both the Key Concepts list and the Claim Evaluation Tools database will continue to evolve. The database is managed by a working group consisting of Astrid Austvoll-Dahlgren, Andy Oxman and Iain Chalmers. Development and testing of the initial items were coordinated by the Key Concepts working group (AA, AO, IC), in consultation with other members of the Informed Health Choices (IHC) project group.

# **Would you like to give us your feedback?**

If you have any comments or suggestions you would like to give us on this manual or the Claim Evaluation Tools database, we would be delighted to hear it. Please send us an email to astrid.austvoll-dahlgren@fhi.no.

# **References**

1. Austvoll-Dahlgren A, Semakula D, Nsangi A, Oxman A, Chalmers I, Rosenbaum S, et al. Measuring ability to assess claims about treatment effects: The development of the “Claim Evaluation Tools”. Accepted manuscript BMJ open. 2016.

2. Austvoll-Dahlgren A, Nsangi A, Semakula D. Key concepts people need to understand to assess claims about treatment effects: a systematic mapping review of interventions and evaluation tools. Accepted paper BMC. 2016.

3. Austvoll-Dahlgren A, Oxman AD, Chalmers I, Nsangi A, Glenton C, Lewin S, et al. Key concepts that people need to understand to assess claims about treatment effects. Journal of Evidence-Based Medicine. 2015;8(3):112-25.

4. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007;57(8):1358-62.

5. Andrich D. Rasch models for measurement. Beverly Hills: Sage Publications; 1988.

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

7. Rasch G. Probabilistic Models for Some Intelligence and Achivement Tests. Copenhagen: Danish Institue for Educational Research. Expanded Edition 1983: Chicago: MESA Press; 1960.

8. Baghaei P. The Rasch Model as a Construct Validation Tool. Rasch Measurement Transactions, 2008, 22:1 p. 1145-6.

9. Rasch analysis. <http://www.rasch-analysis.com/>. Accessed 2016.

10. Nsangi A, Semakula D, Austvoll-Dahlgren A, Guttersrud Ø, Oxman A, Group. TI. Measuring ability to assess claims about treatment effects in Luganda: testing of a two sub-sets of the “Claim Evaluation Tools” using Rasch modelling. Manuscript in preparation. 2016.

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

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

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

14. Nsangi A., Semakula D., Oxman M., Austvoll-Dahlgren A., Rosenbaum S., Kaseje M., et al. Evaluation of resources to teach children in low income countries to assess claims about treatment effects. Protocol for a randomized trial. Submitted manuscript. 2016.

15. Semakula D., Nsangi A., Oxman M., Austvoll-Dahlgren A., Rosenbaum S., Kaseje M., et al. Can an educational podcast improve the ability of parents of primary school children to assess claims about the benefits and harms of treatments? Protocol for a randomized trial Submitted manuscript. 2016.