Appendix 5. Psychometric analysis methods

The psychometric analyses and the report below were completed by Sarah C Smith (PhD), Associate Professor, London School of Hygiene & Tropical Medicine, and Jolijn Hendriks (PhD), Assistant Professor, London School of Hygiene & Tropical Medicine, UK.

Psychometric Principles

Psychometrics is the scientific field concerned with the measurement of subjective judgements using numerical scales and the evaluation of the measurement properties of such scales (e.g. reliability, validity, responsiveness).[1] In general so-called "modern psychometrics" can be divided into two types of methods: Item Response Theory (IRT) proposed by Frederic Lord;[2-4] and *RMT* measurement[5-8] proposed by Georg Rasch.[9] Both of these types of methods are becoming more widespread in health outcome research.

Modern psychometric methods examine the differences between observed and predicted item responses to determine the extent to which the data are consistent with ('fit') a mathematical model(s). When data fit the model, the estimates derived from the model are considered robust because the measurement theory is supported by the data. When data do not fit the model, the IRT approach is to question the appropriateness of the mathematical model (the IRT approach); the RMT approach is to question the appropriateness of the data. The RMT approach was used in this evaluation because it can provide a wide range of diagnostic information that help us to understand how the AMR awareness scale can be improved

RMT analysis indicates the extent to which rigorous measurement is achieved by examining the difference (or 'fit') between the observed scores (persons' responses to items) and the expected values predicted from the data by the Rasch model. The criteria for measurement in RMT analysis are evaluated interactively using a single mathematical model, the Rasch model[5-8]. Thus, a range of evidence is used to evaluate each individual item in the scale. This evidence is then used to make a judgment about the overall quality of the scale.

Fundamentally, this method differs from traditional psychometric methods (based on Classical Test Theory) as their focus is the relationship between a person's measurement and their probability of responding to an item, rather than the relationship between a person's measurement and their observed scale total score. The main advantages of RMT analysis are that it:

1. provides measurements of people that are independent of the sampling distribution of the items used, and locates items in each scale independent of the sampling distribution of the people in whom they are derived;
2. improves the potential to diagnose item-level psychometric problems;
3. allows for a more accurate picture of individual person measurements derived from questionnaire instruments.

Key RMT parameters evaluated

Item fit validity

The items of the scale must work together (fit) as a conformable set both clinically and statistically. Otherwise, it is inappropriate to sum item responses to reach a total score. When items do not work together (misfit) in this way, the validity of a scale is questioned. The fit of each item to the Rasch model was evaluated both statistically (fit residuals >+/-2.5), chi square (Bonferroni corrected significance levels) and graphically (visual inspection of the item characteristic curves). No single piece of information can confirm the fit of an item to the model and it is important therefore to consider all the evidence together.

The response categories of each item should work the way they are intended and should reflect an ordered continuum (e.g. 0,1,2,3). Although these response categories may appear to be clear and easily understood, they must also work when the items are combined into a scale. We evaluate this statistically and graphically by considering threshold locations and plots. We would expect the threshold values between adjacent pairs of response options to be ordered by magnitude (e.g. less to more, worse to better etc). Disordered thresholds can indicate where respondents have misunderstood or been unable to use response categories consistently.

Targeting

Scale-to-sample targeting describes the extent to which the distribution of person estimates matches the distribution of item estimates and is evaluated by comparing the spread of person and item locations. Targeting gives us information about whether the items (questions) are capturing information at the correct level for the skill level of the people in the sample (and also how suitable the sample is for evaluating the questionnaire). When an instrument has better targeting it gives us greater confidence in the rest of the psychometric data.

Item dependency

The responses to one item should not influence the responses to another item. The extent to which this happens is evaluated by considering item dependency. Item dependency is determined by examining the residual correlations between items after the Rasch factor is partialled out (r >0.30 indicates potential dependency).

Reliability

Reliability is assessed with the Person Separation Index (PSI), a reliability statistic comparable to Cronbach's alpha, which quantifies how well the scale discriminates between people with different levels of the measured construct. Higher PSI values indicate better reliability (>0.70 indicates adequate reliability).

Stability of items

Item stability is evaluated by considering Differential Item Functioning (DIF). This uses ANOVA to evaluate whether for a given level of the construct, the item performs in the same way for different groups within the sample. Uniform DIF is indicated by a significant main effect for the group and non-uniform DIF is indicated by significant interaction between the group and the class intervals (groupings of people with approximately the same amount of the construct). The presence of uniform DIF can be corrected by calibrating problem items separately for each level of the group (known as "splitting" items). Items showing non-uniform DIF may need to be investigated and/or removed from the item set. We evaluated DIF for groups defined by gender, age and profession of the respondent and also country.

References

1. Stevens S. Mathematics, measurement and psychophysics. In: Stevens S, editor. Handbook of Experimental Psychology. New York: Wiley; 1951.
2. Lord F. A theory of test scores. Psychometric monographs. 1952; **No. 7**.
3. Lord FM. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika. 1952; **17**(2): 181-94.
4. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley; 1968.
5. Wright BD. Solving measurement problems with the Rasch model. Journal of Educational Measurement. 1977; **14**(2): 97-116.
6. Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978; **43**: 561-73.
7. Wright BD, Stone MH. Best test design: rasch measurement. Chicago: MESA; 1979.
8. Wright BD, Masters G. Rating scale analysis: Rasch measurement. Chicago: MESA; 1982.
9. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research (Expanded edition (1980) with foreword and afterword by B.D. Wright, Chicago: The University of Chicago Press, 1980. Reprinted Chicago: MESA Press, 1993. Available from www.rasch.org/books.htm); 1960.