Routledge
Taylor & Francis Group

Journal of Mental Health

ISSN: 0963-8237 (Print) 1360-0567 (Online) Journal homepage: http://www.tandfonline.com/loi/ijmh20

Reducing outcome measures in mental health: a systematic review of the methods

Wayne Smith, Anita Patel, Paul McCrone, Huajie Jin, Beatrice Osumili & Barbara Barrett

To cite this article: Wayne Smith, Anita Patel, Paul McCrone, Huajie Jin, Beatrice Osumili & Barbara Barrett (2015): Reducing outcome measures in mental health: a systematic review of the methods, Journal of Mental Health, DOI: 10.3109/09638237.2015.1101058

To link to this article: http://dx.doi.org/10.3109/09638237.2015.1101058

Published online: 04 Dec 2015.

Journal of MENTAL HEALTH

http://tandfonline.com/ijmh
ISSN: 0963-8237 (print), 1360-0567 (electronic)

J Ment Health, Early Online: 1–12
 DOI: 10.3109/09638237.2015.1101058

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

Reducing outcome measures in mental health: a systematic review of the methods

Wayne Smith[1], Anita Patel[1,2], Paul McCrone[1], Huajie Jin[1], Beatrice Osumili[1], and Barbara Barrett[1]

[1]*Institute of Psychiatry, Centre for the Economics of Mental and Physical Health, King's College London, London, UK and* [2]*Barts and The London School of Medicine and Dentistry, Centre for Primary Care and Public Health, Queen Mary University of London, London, UK*

Abstract

Background: Traditionally, classical test theory (CTT) has been used for instrument development and various methods have since been proposed for reducing outcome measures to shorter versions. These reduction methods have not previously been compared in mental or physical health.

Aim: To identify and compare the various methods used to develop brief versions of outcome measures from existing measures in mental health.

Method: A systematic review of the literature in Embase, Medline, PsychInfo and from a grey literature was done. Search strategies were developed in each database to identify all relevant literature based on the inclusion criteria. Each paper identified was briefly described and then assessed using a bespoke assessment checklist developed by the authors. Methods for reducing outcome measures found across all studies were compared.

Results: Ten papers were identified. Five methods were used for scale reduction: Rasch analysis (RA), exploratory factor analysis (EFA), graded response models (GRMs), all-subset regression, and regression. RA was the most widely used process.

Conclusion: The Rasch model (RM) is the only model where "'specific objectivity"' is a defining property of the model. This property is necessary for constructing scales in line with the fundamental principles of measurement.

Keywords

Outcome measures, outcomes research, questionnaire reduction, questionnaire development, mental health

History

Received 13 March 2015
Accepted 12 August 2015
Published online 25 November 2015

Introduction

This study aims at outlining the methods used to reduce existing outcome measures in mental health areas to shorter forms. It forms part of a larger doctoral project that aims at reducing the Health of the Nation Outcomes Scale (HoNOS) (Wing et al., 1998) to a shorter form for the purposes of economic evaluation in mental health and hence the review focuses on mental health outcomes. However, these techniques are also applicable to physical health. No previous review has been identified in the literature which compares item reduction methods for either physical or mental health outcome measures. Brazier et al. (2012) briefly summarises some reduction methods (in a review of the development of health state classification systems) as a first stage to developing preference-based outcome measures for use in economic evaluations, but his review contained only measures of physical health.

Outcome measures are important to assess the effectiveness of interventions by routinely recording changes in health and social care for people with mental illness.

Instrument development generally involves deriving a minimum number of items for use as an outcome measure which is suitably reliable and valid. However, in general, Health Related Quality of Life (HRQoL) instruments and outcome measures can be quite lengthy. Therefore, the development of short questionnaires has largely focused on reducing existing instruments (Prieto et al., 2003).

Classical test theory (CTT) methods (DeVellis, 2006) have traditionally been used to develop good measurement properties in scale development and reduction. CTT assumes that for each person, an observed score (O) represents a person's true score (T) and an error term (ε), where ($O = T + \varepsilon$). Thus tests or scales never produce a user's true score (T) but only an observed score (O). The standard deviation of these errors is known as the standard error of measurement (SEm) (Harvill 1991) and is therefore directly related to the reliability of a test. Reducing errors will increase the reliability of a test and lead to more true scores. CTT is associated with psychometrics at an overall test score level rather than at individual item level.

Correspondence: Wayne Smith, King's College London, Institute of Psychiatry, Centre for the Economics of Mental and Physical Health, London SE5 8AF, UK. Tel: +020 7848 0198, Fax: +020 7848 0458. E-mail: wayne.w.smith@kcl.ac.uk

This article was originally published with errors. This version has been corrected. Please see Corrigendum (http://dx.doi.org/10.3109/09638237.2016.1153246).

Table 1. Assumptions and properties of RMs.

Assumptions and properties of Rasch models
• Unidimensionality: usually a single underlying construct is measured.
• Local item independence: items should not be directly related to each other. An answer to one question should not affect how a respondent answers another.
• An Item Characteristic Curve (ICC) is the primary concept in IRT. It is a logistic regression line of ability (x-axis) and probability of a correct response (y-axis). The ICC shows the expected curve of the model. Observed data plotted against this curve is able to visualise any misfit to the model.
• Measurement invariance: Differential Item Functioning (DIF) (Clauser & Mazor, 1998) is assessed during Rasch Analysis. There should be no difference in item response between groups, at different occasions or under different conditions, for respondents with the same level of ability or latent trait.
• Ordered thresholds: disordered thresholds (Andrich, 2013) indicate that a classification system in not working as it should where increasing scores of a polytomous scale should represent an increase in a latent trait but it does not. For example, people with severe mobility problems indicating they have no problems with mobility. When this occurs it is usually an indicator that the item category responses are poorly worded or that respondents are not able to distinguish between the response levels. This is reflected graphically through the category probability curves.

Currently, latent trait models such as Item Response Theory (IRT) models or Rasch models (RMs) are used for scale development (Edelen & Reeve, 2007). The rationale of these approaches is that an individual's response to a particular test item is based on the characteristics of the individual (person parameters such as ability or any latent trait to be measured) and characteristics of the test items (item parameters such as test difficulty).

IRT models are logistic models which link trait (ability or disease severity) with item response probabilities. For dichotomous scales some IRT models include the 3-parameter logistic models (3-PL), 2-parameter logistic models (2-PL) or 1-parameter logistic models (1-PL). A 3-PL model shows the relationship between a respondent's ability and the probability of a correct response with three parameters (item difficulty, item discrimination and a guessing factor). Respondents with lower ability may tend to guess. When guessing is not a factor and is assumed to be zero then the 3-PL is reduced to a 2-PL model. When a second restriction is included (all items now have equal discrimination and guessing is not a factor) then the 2-PL is reduced to a 1-PL model. There is also a 4-PL model which has received less attention and this includes a factor which can be due to stress, tiredness or inattention for example (Magis, 2013).

Mathematically, the RM is identical to a one parameter logistic model in IRT. In IRT, an adequate fit of model to the data is expected for item analysis. However, a key difference with the RM is that where data do not conform to the model, the objective is not to find a more suitable model as in the paradigm of IRT, but to examine the fit of the data and the anomalies and adapt the data in order to create a more valid and reliable instrument (Bhakta et al., 2005). Both IRT models and RM share assumptions for model fit such as unidimensionality and local item independence (Chang & Reeve, 2005; Hays et al., 2000). However, the RM has additional assumptions which should be met (see Table 1).

For polytomous scales a number of models have been developed such as the Rasch based – Partial Credit Model and the Rating Scale Model along with other IRT models such as the Generalised Partial Credit Model, the Graded Response Model (GRM) for ordered response items and, the Nominal Model for items which are not ordered (Edelen & Reeve, 2007).

These are just a few of the methods and models which have been used for reducing outcome measures. It is therefore important to review reduction methods in order to establish an appropriate, standardised method which can be used to shorten pre-existing measures.

Method

Search strategy

A literature search was conducted in Embase, Medline and PsychInfo using the Ovid interface, plus OpenGrey to cover grey literature. Appendix 1 is an example of the search strategy developed for Embase. Appendix 2 is a prisma diagram showing the number of papers remaining at each stage. The selection of articles was double-checked blindly by two further reviewers (BO and JH) using the following inclusion criteria:

- All outcomes measures in mental health (MH); condition-specific MH or generic MH.
- Generic outcome measures such as EQ-5D (Longworth et al., 2014) were excluded.
- Patient population used in the analysis must include a MH group.
- Item reduction of health outcome questionnaires must be developed from an existing outcome instrument and not a combination of instruments.

Quality assessment

The quality of each paper was examined using a bespoke quality assessment checklist (see Appendix 3), which was developed specifically for this review since no assessment checklist can be identified from published literature.

Results

Two thousand four hundred and forty one abstracts were identified after which 436 duplicates and 1882 papers which did not fit the inclusion criteria were removed. The full text of 123 articles were examined for eligibility and 10 articles retained for review. In-text referencing and a grey literature search identified no additional papers for review.

The results are presented in two sections. The first section describes each included paper and highlights their key limitations. A summary of the characteristics of the included studies is reported in Table 2. The second section is a quality assessment of the papers found. Table 3 is a summary of the quality assessment.

Descriptive summary of studies

Bilker et al. (2003) developed the smallest possible subset of items in the schizophrenia quality of life scale (QLS) to

Table 2. Descriptive summary of outcome measures discussed.

Study	Full instrument	Mental health area for instrument use	Country/development sample	Scoring method	Reduction method	Short form	How used
Andresen et al. (2013)	(STORI) 50 items, Stages of recovery instrument	Schizophrenia, bipolar, other	Australian/232	Five stages, items in each stage rated (0–5)	Graded response model	STORI-30	Interview administered
Barkham et al. (2013)	CORE-OM, 34 items	People with common mental health illnesses	UK/1618 primary care patients for item selection	Items rated (0–4). 4 indicating worse QoL. Score range 0–136	Regression	CORE-10	Self-report
Bilker et al. (2003)	(QLS-21) 21 item QoL scale	Measure of functioning in schizophrenia	US study/198 patients with schizophrenia	Item score (0–6), severe impairment to high function	Predictive model approach, all possible subset analysis	QLS-7	Clinician reported outcome measure (CROM)
Boyer et al. (2010)	(S-QoL 41) QoL for people with schizophrenia	DSM-IV diagnosis (schizophrenia)	French/517 patients with schizophrenia	Dimensions scores (0–100). Total index score (0–100)	Rasch analysis	S-QoL 18	Self-report instrument
Las Hayas et al. (2010)	(HeRQoLED) Health related QoL for eating disorders questionnaire. 55 items	Eating disorders	Spain/324 patients with eating disorders	55 items covering five domains. Each domain converted to scores from 0 to 100	Rasch analysis	HeRQoLED-S	Self-report
Lovaglio & Monzani (2012)	(HoNOS-12) Health of the Nation outcome scale	Various diagnoses of mental ill health	Italy/1062 community mental health patients	Item score (0–4), no problems to severe problems	Rasch analysis and exploratory factor analysis	HoNOS-6D	Clinician reported outcome measure (CROM)
Mavranezouli et al. (2011)	(CORE-OM) Clinical Outcomes in Routine Evaluation-Outcome Measure, 34 items	People with common mental health illnesses	UK/400 primary care patients for Rasch analysis	Items rated (0–4), 4 indicating worse QoL. Score range 0–136	Rasch analysis	CORE-6D	Self-report
Mulhern et al. (2012)	(DEMQOL & DEMOQOL proxy) Health related QoL for people with dementia, 31 items	Dementia	UK/644-patient with dementia DEMQOL analysis 683 for proxy analysis	Items rated (1–4), 4 indicating better QoL. Score range 31–124	Rasch analysis	DEMQOL-U, five items & DEMQOL-proxy-U, four items	Patient and carer self-report
Ritsner et al. (2005a)	(Q-LES-Q) QoL Enjoyment and Satisfaction Questionnaire	DSM-IV diagnosis (schizophrenia, schizo-affective, and mood disorders)	Israel/339 patients with schizophrenia, schizo-affective and, mood disorders	Item score (0–5), 5 indicating better QoL	Predictive model approach, all possible subset analysis	Q-LES-Q-18	Self-report instrument
Ritsner et al. (2005b)	QLS-21	Measure of functioning in schizophrenia	Israel/133 patients with schizophrenia	Item score (0–6), severe impairment to high function	Predictive model approach, all possible subset analysis	QLS–5	CROM

Table 3. Quality assessment of studies.

Reference	Item/level structure outlined	Structure assess e.g. EFA/CFA	Sample size and population outlined	IRM used as alt. to RA	RA	Disordered thresholds examined	Fit statistics	DIF	Uni-dimensionality	Local dependence	Validation of short version	Alt. method also used or discussed	For CSPB valuation	Limitations and future application/ research
Andresen et al. (2013)	√	√	√	√							√			√
Barkham et al. (2013)	√		√								√			
Bilker et al. (2003)	√	√	√								√	√		√
Boyer et al. (2010)	√	√	√		√		√	√	√	√	√			√
Las Hayas et al. (2010)	√	√	√		√	√	√	√	√	√	√			√
Lovaglio & Monzani (2012)	√	√	√		√	√	√		√	√	√	√		√
Mavranezouli et al. (2011)	√	√	√		√	√	√	√	√	√	√		√	√
Mulhern et al. (2012)	√	√	√		√	√	√	√	√	√		√	√	√
Ritsner et al. (2005a)	√		√								√	√		√
Ritsner et al. (2005b)	√		√								√	√		√

Light shaded areas apply to Rasch analysis/IRT methods. Dark shaded areas indicate aspects that were not discussed in papers. √ Indicates description present in paper. *Abbreviations*: EFA/CFA, exploratory factor analysis/confirmatory factor analysis; IRM, Item response model; RA, Rasch analysis; DIF, Differential Item functioning; Alt, alternative; CSPB, Condition specific preference based.

predict the total score from the QLS-21. All subsets containing 1–10 items of the QLS were considered (1,048,575 models). Each predicted total QLS score from the subset models was assessed against the actual total QLS score using Pearson correlation coefficient. Models were validated using two validation datasets. The optimal model was a 7-item one which accurately predicted the QLS-21 and included all four theoretical constructs of the QLS.

Using a similar method, Ritsner et al. (2005b) applied a predictive model approach to reduce the length of QLS-21. A heuristic algorithm was used to select subsets that produced a maximum value of R-squared. The authors also compared the Pearson's product-moment correlations between the total scores of the selected QLS subsets and the original QLS-21. This procedure resulted in retaining five items to form the QLS-5. Psychometric properties of the QLS-5 were high and comparable to the QLS-21. This approach was also applied in Ritsner et al. (2005a) to reduce another 21-item outcome measure, the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q) to the Q-LES-Q-18. That work was based on a sample of 339 patients diagnosed with schizophrenia, schizoaffective or mood disorders to construct the model. The abbreviated version was subject to psychometric testing using CTT methods. In both studies (Ritsner et al., 2005a,b), validation was performed on separate samples.

Las Hayas et al. (2010) used Rasch analysis (RA) to produce a 20-item version from the 50-item HRQoL for eating disorders (ED) questionnaire. It also aimed at confirming the structure of the new version and examine its validity and reliability, using 324 patients with a diagnosis of ED. Confirmatory factor analysis (CFA) hypothesised two second-order latent traits. RMs were applied to both second order traits. Unidimensionality was assessed and items were examined for differential item functioning (DIF) (Clauser & Mazor, 1998) across three diagnosis subtypes. DIF examines whether items function differently across groups such as age and gender. Item residuals were examined for local dependency. RA was repeated throughout and item contents were examined by experts of the field before deciding on whether to remove items. One limitation highlighted in the method was that the short version was validated on the same patient sample at a different time frame rather than being validated in an independent sample.

Lovaglio & Monzani (2012) investigated the internal structure of the HoNOS-12 and proposed a shorter, one dimension 6-item version for use in community mental health services. They confirmed and tested the hypothesised factor structures of the HoNOS-12 found in the literature using CFA. The dimensionality of the HoNOS was explored using exploratory factor analysis (EFA) and parallel analysis (Ledesma & Valero-Mora, 2007). Dimensionality was also assessed following RA.

Two methods were used to reduce the HoNOS-12: RA and EFA. For the RA, assumptions for model fit such as ordered categories, item independence and unidimensionality were tested. Removal of items that misfit the RM resulted in a 6-item model which confirms the EFA version. Item that fits the RM was measured by the mean-square (MSQ) fit statistic (Smith, 1996).

Separate samples were used for training and to test validity. There was no mention of analysis for DIF for age

groups, gender or by diagnosis. One limitation identified was that RA was applied to a scale which was multidimensional. Rasch was not applied to each dimension of the scale.

Boyer et al. (2010) used a Rasch partial credit model, CTT and expert advice to reduce the S-QoL-41 (quality of life in patients with schizophrenia) and to validate a shorter version. Twenty three items were excluded from the original version. The reduced version was tested for construct validity and other psychometric properties. The construct validity was assessed using principal component analysis (PCA) to obtain the eight dimensions of the SQoL-41. Unidimensionality of each dimension was assessed using RA. DIF analyses were conducted for age, gender, educational level and clinical form. Dimension correlation was also explored. This study did not create a unidimensional questionnaire. RA was applied to each of the dimensions.

Andresen et al. (2013) developed a shorter version (STORI-30) from the Stages of Recovery Instrument (STORI). The structure and scoring method of the 50-item STORI was fully described. Data were divided into two groups, the first consisting of 232 participants from combined previous studies and the second consisting of 50 participants. The authors used a unidimensional GRM (Zhu & Stone, 2011) to identify six items for each of the five stages in the instrument thereby creating a 30-item model. The item selection process was not discussed in depth. Following item analysis using GRM on the first dataset, EFA was performed to determine whether the remaining 30 items were matched according to the stages to which they theoretically belonged. The second dataset was used for validation of the 30-item instrument. Internal reliability and correlation between subscales were examined. Correlation between the STORI-30 and the Recovery Assessment Scale (RAS) (Corrigan et al., 1999) was also investigated.

Mulhern et al. (2012) developed reduced versions of the self-report DEMQOL (HRQoL for people with dementia) and proxy (carer)-reported DEMQOL-Proxy using RA. It was the first stage of a study to develop a condition-specific preference-based measure which can be used directly in economic evaluations to generate quality-adjusted life years (QALYs) (Rowen et al., 2012). The structure and scoring method of the DEMQOL and its proxy were described. Two sources of data were used in this UK study, consisting of 644 patients in the DEMQOL analyses and 683 in the DEMQOL-Proxy analyses. The main objective of this study was to derive two brief measures, both amenable to eliciting preferences for health states using a time trade-off method (TTO; otherwise termed ''health state valuation''). The TTO and other valuation methods are described elsewhere in the literature (Drummond, 2005; Gudex, 1994). In the initial part of the study EFA was conducted to investigate the factor structure of the patient and proxy versions. Five-factor structures were derived in both groups. A separate RA was applied to each of the five factors. Item selection was based on the assumptions of model fit for RA: item level ordering, DIF for gender or age group, goodness of fit to the RM, unidimensionality and item independence. Based on these criteria, one item from each factor in the DEMQOL was chosen. The reduced version called the DEMQOL-U has five items with four health state levels. The reduced DEMQOL-Proxy version called the DEMQOL-Proxy-U, has four items each with four health state levels. The DEMQOL-U and the DEMQOL-Proxy-U result in 1024 (4^5) and 256 (4^4) health states respectively. The authors highlight that validation was not possible because of the limited sample size. Another limitation was the lack of analysis of DIF across dementia diagnosis groups. The authors also suggested that alternatives, such as advanced IRT models, could also be used for the item selection process.

Mavranezouli et al. (2011) used RA to form plausible health states amenable to valuation from the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM). The original CORE-OM structure, its validity and reliability, and its application in the UK are discussed with reference to previous articles. The authors also point out that valuing health states of the CORE-OM without applying item reduction would result in an unmanageable number of health states. They also give reference to a previous study (Evans et al., 2002) examining the structure of the CORE-OM using EFA. Data analysis for this study was based on 1500 patients. A random sample of 400 patients was used for the RA and another random sample of 400 patients was used for validation purposes. The authors justified the use of the smaller sample size for RA by citing Smith et al. (2008) who suggests higher type I errors (falsely rejecting items as misfitting) with increased sample size in polytomous data. The RA was fully described and all criteria for item exclusion, such as item threshold ordering, overall fit, item fit statistics and DIF for age, gender and ethnicity, were discussed. Additional exclusion criteria were applied to further reduce the instrument with an aim to develop a shorter and more manageable preference-based instrument. Statistical reduction methods were combined with advice from clinical expert opinion about which health states were plausible and which items should be retained or removed. Two of the limitations outlined were a low person separation index (Mallinson et al., 2004) and a resultant limited number of health states which does not capture the full range of plausible responses. Finally, the authors indicate that there are future opportunities to use this instrument to assess health care interventions for people with common mental health problems using a cost-utility analysis approach.

The final paper by Barkham et al. (2013) developed the CORE-10 from the 34-item CORE-OM, an outcome measure routinely used in psychological therapy. The key aim was to develop a brief and easy to use form. Data were obtained from primary care services with 5831 completed CORE-OMs. Item choices were based on set criteria such as: including two items each for depression and anxiety, one item each for trauma and physical problems, choosing items to reduce floor and ceiling effects, retaining items that cover certain domains and subdomains and dropping items due to high correlation. Items were retained in three steps. Firstly items were removed if they had low response rates. Secondly, selection of another item if the item in question was the only remaining item in a particular group. Thirdly, a regression analysis retained items with the highest R value that best predicted the original items on the CORE-OM. Psychometric properties such as reliability, convergent validity and acceptability of the shortened version were satisfactory.

Quality assessment

All included papers discussed the structure and scoring method of the original outcome measures as well as the sample sizes and sample population used in their analyses. Three papers did not explore the structure of the original outcome measures using approaches such as EFA or CFA (Barkham et al., 2013; Ritsner et al., 2005a, b).

Three articles (Bilker et al., 2003; Ritsner et al., 2005a, b) used variations of the ''all possible subset regression analysis'' method (Hocking & Leslie, 1967) which applies a predictive model approach to derive at a parsimonious subset of items which could accurately predict the total score of the QoL outcome measures. One paper used IRT GRMs to develop a 30-item version of the STORI from the original 50-item version (Andresen et al., 2013). However, this method was not fully described in the paper and reasons for retaining items of the original STORI were not clearly outlined. Only one paper used mostly regression analysis to retain items with the highest R squared values from the original outcome measure, and then used CTT methods to test the psychometric properties of the brief version (Barkham et al., 2013). One study compared RA and EFA as item reduction methods, producing the same items in both reduced versions of the outcome measure (Lovaglio & Monzani, 2012). The remaining 4 studies all used RA for item reduction (Boyer et al., 2010; Las Hayas et al., 2010; Mavranezouli et al., 2011; Mulhern et al., 2012). Wherever reported, the Rasch software, criteria for assessing unidimensionality, fit statistics and methods varied between some papers (Table 4).

Three studies (Las Hayas et al., 2010; Mavranezouli et al., 2011; Mulhern et al., 2012) of the five belonging to the Rasch group fully described the methods involved in item reduction and the assumptions behind the RM such as unidimensionality, local dependence, DIF, disordered thresholds and discussed fit statistics. There was no analysis of DIF by demographic characteristics such as age, group or gender in the RA performed by (Lovaglio & Monzani, 2012). Two papers briefly mentioned alternative methods to the method used in their analysis (Lovaglio & Monzani, 2012; Mulhern et al., 2012). All papers used validation samples during their analysis except for Mulhern et al. (2012) which indicated that the sample size was not sufficiently large to randomly subscribe patients to a validation and RA groups (Table 3).

Only two papers used item reduction methods with an aim at developing a condition-specific preference-based measure (Mavranezouli et al., 2011; Mulhern et al., 2012). Both papers applied RA to derive at a reduced health state classification system as a first step to develop a preference-based outcome measure for use in economic evaluations.

Discussion

Summary and critique of methods

In summary five processes were identified from these studies for item reduction: RA, IRT (GRM), EFA, all subset regression analyses and, in one paper, a procedure where the choice of items were driven by set criteria, completion rate assessment, item coverage and regression. The key aims of the papers were to develop reliable versions which were shorter and easier to use.

This review shows that RA is the most widely used procedure for item reduction. Papers discussed how the assumptions behind the RA such as unidimensionality, local independence, DIF and disordered item thresholds are addressed and explain the statistical methods behind the analysis. Mavranezouli et al. (2011) highlighted the ability of RA to develop health classification systems with independent dimensions as key since dependency can result in health states which do not make sense conceptually. However, none of these papers discussed why RA was preferable to other methods.

Rasch analysis

The RM outlines assumptions and properties which should be met in order to construct scales which are in line with the fundamental principles of measurement. Data which do not fit as expected are removed in order to create instruments which meet these assumptions.

As mentioned earlier, the theory behind the RA is that the probability of endorsing an item (question) is a log function of a person's ability (amount of underlying trait) and the difficulty of the item, where item difficulty is the only item parameter considered. Unlike CTT and IRT, the RM can produce sample free and test free measurement. This means that item difficulty estimates are the same regardless of who is included in the sample and person ability estimates are the same regardless of which items are used in a test. This is a unique property to RA called specific objectivity which allows for invariant measurement (Engelhard, 2012) where person and item parameters are separable and measured on the same invariant log scale. Analogous to specific objectivity is having a sufficient statistic with which to estimate parameters. Rasch (1966) and Rasch (1980) state that for data which fit the RM, the sum of the person's raw scores for all items is a sufficient statistic for the person parameters. Hagquist et al. (2009) cites Rasch and goes on to explain that conversely the sum of the item raw scores is a sufficient statistic for the item parameters. Instruments derived through RA allow ordered observations to be transformed into an interval scaled measure of the latent trait (Salzberger, 2010).

Item Response Theory (Graded response model)

IRT can use a number of additional item parameters (besides item difficulty) to describe data. Essentially the model that best describes the data is selected. Although these models are attractive in that they can better explain the variance in data, they are not developed for constructing measurement. The additional parameters of some IRTs (e.g. including a discrimination parameter) and how these models are used are not consistent with measurement theory since they violate the assumption of invariant measurement (Massof, 2002).

Although Rasch and IRT have often been grouped together in the literature under IRT (maybe because they have some overlap in their assumptions), they are separate theories. In fact proponents of each model often challenge each other in their application. Andrich (2004) discussed and compared the two test theories in detail.

Table 4. Summary of statistics used in Rasch papers.

	Rasch software/ method	Unidimensionality test statistics	Fit statistics	DIF	Local dependency	Reliability
Las Hayas et al. (2010)	Winsteps	PCA of residuals following RA, Violation additional factors with eigenvalues >3	Infit and outfit – mean square fit statistic (MSQ). Values between 0.7 and 1.3 are satisfactory	Performed	Correlation of >0.5 problematic	Person Separation Index (PSI) and Item Separation Index (ISI) values of >2 implies reliability comparable to 0.8 Cronbach's alpha
Lovaglio & Monzani (2012)	Not stated	PCA of Rasch residuals A violation implied additional factors with eigenvalues >1.5. Parallel analysis of Rasch residuals	Infit and Outfit-mean square fit statistic (MSQ). Misfit ≥1.3	Not discussed	Item correlation >0.3 indicate dependency	PSI and ISI
Boyer et al. (2010)	Winsteps	Construct validity – PCA with varimax rotation followed by RA of each dimension	Infit and outfit MSQ Values between 0.7 and 1.2	Performed	No formal threshold stated in paper	Cronbach's alpha
Mulhern et al. (2012)	RUMM2020	PCA of Rasch residuals	Respondents with fit residuals outside |2.5| and Chi-squared test sig level 0.01	Performed	≥0.3 within factors indicate dependency	PSI
Mavranezouli et al. (2011)	RUM2020	PCA of Rasch residuals 5% significance testing	Items with fit residuals beyond ±2.5 and/or significant Chi squared statistics (at the 0.01 level after Bonferroni adjustment) were excluded	Performed	No formal threshold stated in paper	PSI

Winsteps – Rasch analysis software which uses maximum likelihood estimation (Bond & Fox, 2013).
RUM2020 – Rasch analysis software which uses conditional pairwise maximum likelihood algorithm (van der Velde et al., 2009).
PCA, Principal Component Analysis.

Exploratory factor analysis

EFA uses various fitting procedures to explore the factor structure of outcome measures and determine which items load on particular domains of the outcome measure. For item reduction an *a priori* factor structure can be imposed. For example, a 1-factor model for unidimensionality and which then allows one to examine the individual items to see how well they load on that particular factor. Items which have low factor loadings can be dropped from the analysis. A wide range of statistical indices assess the goodness of fit to the proposed model, factor loadings and correlation among items (Fabrigar et al., 1999).

There are many different methods that can be used to conduct a factor analysis such as principal axis factor, maximum likelihood, generalized least squares and unweighted least squares. There are also various methods used in the rotation process such as orthogonal rotations, varimax and equimax for uncorrelated factors or promax for correlated factors. When factor structures are not imposed there are also various methods to determine how many factors to retain. The variation in the methods employed in an EFA can result in different models. Costello & Osborne (2005) fully discussed EFA and described the procedure as ''error prone even with very large sample sizes and optimal data''. Factor analysis does not have separable item or respondent properties therefore factor loadings are sample dependent. Factor analysis also assumes that raw scores have interval scale properties (Wright, 1996).

Predictive model approach/all possible subset analysis

The variant of ''all possible subset analysis'' was another popular method identified from this review. In these papers, the authors use a quick search algorithm to select subsets which maximise R-squared before comparing the total score on each relevant subset to the total score on the original instrument. Scrucca (2006) gives an example of a proposed algorithm for such cases which reduce the computing time for the analysis compared to an exhaustive search and analysis of all subsets. Although the resulting shortened instruments were validated, this analysis does not address the problems of using raw scores from ordinal data (Grimby et al., 2012).

Conclusion

Although there is a growing use of the RM for modern scale development and reduction it is also important to support this with evidence from substantial reviews. This is the first review which compares item reduction methods across mental health outcome measures. Similar reduction methods are also applicable in physical health. Considering some of the differences in relevant health-related outcomes in mental and physical health it is important to establish how these techniques have been applied to each area and then perhaps compare across them. This way, appropriate, consistent methods for producing shortened outcome measures across health can be established. As service provision and patient health (disease specific and quality of life) are measured using outcome instruments, improvements in measurement techniques could better inform service provision and clinical practice.

This study identified various methods used to reduce outcome measures. However, RA appears to be the only method that has been developed for constructing measurement. It is increasingly being applied in social science in the development of health outcome instruments. Scores produced by an instrument developed from the RM can be transformed to a scale with interval scoring properties whereas raw scores on outcome measures with polytomous scales which have not been developed in this way are not linear and therefore should not be treated as such. This linear assumption of raw data in EFA is a key reason why EFA is inappropriate for constructing measurement.

Further research should look at the various ways RA has been applied in deriving health outcome measures and reducing existing instruments. For example, for existing instruments shown to be multidimensional using alternative methods such as factor analysis, should Rasch then be applied to each dimension, applied to the overall instrument or should a multidimensional RM be used?

Also some researchers argued that factor analysis and RA methods are incompatible yet as evidenced they have been used or incorporated into studies which have purported the Rasch method. Perhaps clear guidance is needed to address this issue. Finally, many of the studies have been validated using techniques founded in CTT demonstrating its on-going importance in questionnaire development.

Declaration of interest

This study is funded by the Economic and Social Research Council (ESRC) Grant number ES/J500057/1 and South London and Maudsley NHS Trust.

References

Andresen R, Caputi P, Oades L. (2013). Development of a short measure of psychological recovery in serious mental illness: The STORI-30. Australas Psychiatry, 21, 267–70.

Andrich D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? Med Care, 42, I7–16.

Andrich D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any ''Threshold Disorder Controversy''. Educ Psychol Meas, 73, 78–124.

Barkham M, Bewick B, Mullin T, et al. (2013). The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. Couns Psychother Res, 13, 3–13.

Bhakta B, Tennant A, Horton M, et al. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. BMC Med Educ, 5, 9.

Bilker WB, Brensinger C, Kurtz MM, et al. (2003). Development of an abbreviated schizophrenia quality of life scale using a new method. Neuropsychopharmacology, 28, 773–7.

Bond TG, Fox CM. (2013). Applying the Rasch model: Fundamental measurement in the human sciences, Second edition. Oxon: Taylor & Francis.

Boyer L, Simeoni MC, Loundou A, et al. (2010). The development of the S-QoL, 18, A shortened quality of life questionnaire for patients with schizophrenia. Schizophr Res, 121, 241–50.

Brazier J, Rowen D, Mavranezouli I, et al. (2012). Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess, 16, 1–114.

Chang CH, Reeve BB. (2005). Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof, 28, 264–82.

Clauser BE, Mazor KM. (1998). Using statistical procedures to identify differentially functioning test items. Educ Meas, 17, 31–44.

Corrigan PW, Giffort D, Rashid F, et al. (1999). Recovery as a psychological construct. Community Ment Health J, 35, 231–9.
Costello A, Osborne J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pract Assess Res Eval, 10, 173–8.
DeVellis RF. (2006). Classical test theory. Med Care, 44, S50–9.
Drummond F. (2005). Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press, Incorporated.
Edelen MO, Reeve BB. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res, 16, 5–18.
Engelhard G. (2012). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. New York: Routledge.
Evans C, Connell J, Barkham M, et al. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. Br J Psychiatry, 180, 51–60.
Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods, 4, 272–99.
Grimby G, Tennant A, Tesio L. (2012). The use of raw scores from ordinal scales: Time to end malpractice? J Rehabil Med, 44, 97–8.
Gudex, C. (1994). Time trade-off user manual: Props and self-completion method [Online]. York: Centre for Health Economics, University of York. Available: http://www.york.ac.uk/che/pdf/op20.pdf.
Hagquist C, Bruce M, Gustavsson JP. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. International Journal of Nursing Studies, 46, 380–93. doi: 10.1016/j.ijnurstu.2008.10.007.
Harvill LM. (1991). Standard error of measurement. Educ Meas, 10, 9.
Hays RD, Morales LS, Reise SP. (2000). Item response theory and health outcomes measurement in the 21st century. Med Care, 38, II28–42.
Hocking R, Leslie R. (1967). Selection of the best subset in regression analysis. Technometrics, 9, 531–40.
Las Hayas C, Quintana JM, Padierna JA, et al. (2010). Use of Rasch methodology to develop a short version of the health related quality of life for eating disorders questionnaire: A prospective study. Health Qual Life Outcomes, 8, 1–12.
Ledesma RD, Valero-Mora P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. Pract Assess Res Eval, 12, 1–11.
Longworth L, Yang Y, Young T, et al. (2014). Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: A systematic review, statistical modelling and survey. Health Technol Assess, 18, 1–224.
Lovaglio PG, Monzani E. (2012). Health of the nation outcome scales evaluation in a community setting population. Qual Life Res, 21, 1643–53.
Magis D. (2013). A note on the item information function of the four-parameter logistic model. Appl Psychol Meas, 37, 304–15.
Mallinson T, Stelmack J, Velozo C. (2004). A comparison of the separation ratio and coefficient α in the creation of minimum item sets. Med Care, 42, I1–17.
Massof RW. (2002). The measurement of vision disability. Optom Vis Sci, 79, 516–52.
Mavranezouli I, Brazier JE, Young TA, Barkham M. (2011). Using Rasch analysis to form plausible health states amenable to valuation: The development of CORE-6D from a measure of common mental health problems (CORE-OM). Qual Life Res, 20, 321–33.
Mulhern B, Smith SC, Rowen D, et al. (2012). Improving the measurement of QALYs in dementia: Developing patient- and carer-reported health state classification systems using Rasch analysis. Value Health, 15, 323–33.
Prieto L, Alonso J, Lamarca R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. Health Qual Life Outcomes, 1, 1–13.
Rasch G. (1966). An individualistic approach to item analysis. In: Lazarsfeld PF, Henry NW, eds. Readings in mathematical social science (pp. 89–108). Chicago: Science Research Associates, 89–108.
Rasch G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.
Ritsner M, Kurs R, Gibel A, et al. (2005a). Validity of an abbreviated quality of life enjoyment and satisfaction questionnaire (Q-LES-Q-18) for schizophrenia, schizoaffective, and mood disorder patients. Qual Life Res, 14, 1693–703.
Ritsner M, Kurs R, Ratner Y, Gibel A. (2005b). Condensed version of the quality of life scale for schizophrenia for use in outcome studies. Psychiatry Res, 135, 65–75.
Rowen D, Mulhern B, Banerjee S, et al. (2012). Estimating preference-based single index measures for dementia using DEMQOL and DEMQOL-Proxy. Value Health, 15, 346–56.
Salzberger T. (2010). Does the Rasch model convert an ordinal scale into an interval scale? Rasch Meas Trans, 24, 2.
Scrucca L. (2006). Subset selection in dimension reduction methods. Perugia, Italy: Università di Perugia, Dipartimento Economia, Finanza e Statistica.
Smith A, Rush R, Fallowfield L, et al. (2008). Rasch fit statistics and sample size considerations for polytomous data. BMC Med Res Methodol, 8, 1–11.
Smith RM. (1996). Polytomous mean-square fit statistics. Rasch Meas Trans, 10, 516–17.
van der Velde G, Beaton D, Hogg-Johnston S, et al. (2009). Rasch analysis provides new insights into the measurement properties of the neck disability index. Arthritis Care Res 61:544–51.
Wing JK, Beevor AS, Curtis RH, et al. (1998). Health of the Nation Outcome Scales (HoNOS). Research and development. Br J Psychiatry 172:11–18.
Wright BD. (1996). Comparing Rasch measurement and factor analysis. Struct Equ Modeling Multidisciplin J, 3, 3–24.
Zhu X, Stone CA. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. J Educ Meas, 48, 81–97.

Appendix 1

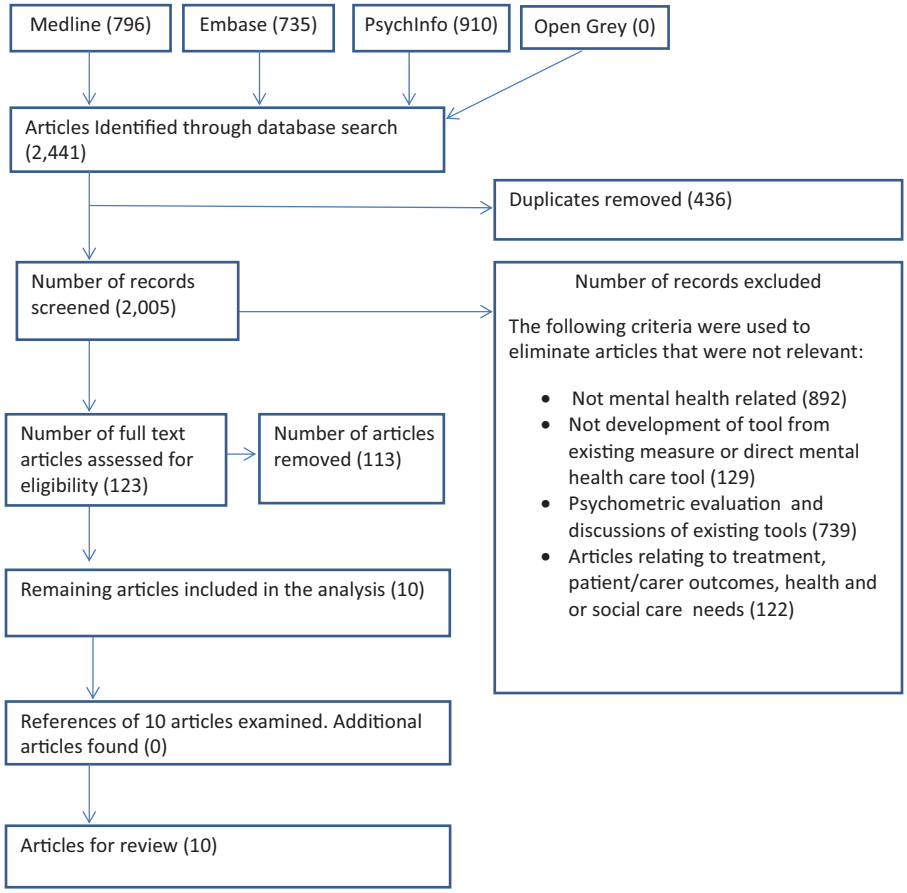
Embase Search Strategy

(1) exp Rasch analysis/
(2) exp factorial analysis/
(3) rasch analys$.mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(4) factor analys$.mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(5) (reduction adj3 theor$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(6) (latent adj3 theor$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(7) item reduction.mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(8) (item adj3 theor$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(9) (item adj3 method$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(10) exp *psychometry/
(11) exp *statistics/
(12) (preference$ adj3 index$).mp. [mp = title, abstract, subject headings, heading word, drug trade name,

original title, device manufacturer, drug manufacturer, device trade name, keyword]
(13) exp ‘‘quality of life"/
(14) (quality adj3 life$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(15) quality adjusted life year.mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(16) exp ‘‘quality of life index"/
(17) (preference$ adj3 index$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(18) (preference$ adj3 measure$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(19) exp outcome assessment/
(20) (health state adj2 system).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(21) instrument$.mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(22) questionnaire$.mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(23) exp questionnaire/
(24) exp mental disease/
(25) (mental adj3 dis$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(26) (mental adj3 problem$).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]
(27) "mavranezouli$".fc_auts. and ‘‘medical decision making$".fc_jour.
(28) "mavranezouli$".fc_auts. and ‘‘quality of life$".fc_jour.
(29) "lovaglio$".fc_auts. and ‘‘quality of life$".fc_jour.
(30) 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11
(31) 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19
(32) 20 or 21 or 22 or 23
(33) 24 or 25 or 26
(34) 30 and 31 and 32 and 33
(35) 27 or 28 or 29
(36) 34 and 35

Appendix 2

Prisma diagram showing the number of articles remaining at each stage of the selection process



Appendix 3

Quality assessment checklist

- Have the main aims and objectives been outlined?
- Was the place of study, sample size and population stated?
- Were the original scale and its structure discussed?
- Was the underlying structure of the scale explored in previous literature or in the current study e.g. Was exploratory or CFA used to explore the factor structure of the original scale?
- What method was used for item reduction? Was the method fully discussed?
- If Rasch analysis (RA) was used for item reduction then:
  - Has unidimensionality been assessed? This explores whether the responses to subset of items in the scale gives the same estimate of a person's ability. This is examined using a method called Principal Component Analysis (PCA). Following RA the main 'Rasch factor' is removed and residuals are left which do not comply with the RM. These residuals are examined to determine if there is a secondary dimension. A t-test is used in this case to check for unidimensionality between the positively and negatively loading items of the first PCA. The hypothesis test is there is no significant difference between the two subsets. This implies there is no secondary dimension.
  - Have the data been examined for disordered thresholds? This examines whether respondents are able to distinguish between adjacent levels in each item. For example, on an item with four levels, whether a respondent distinguish between two adjacent levels labelled 'I have some problems with washing and dressing' and 'I have a few problems with washing and dressing'.
  - Has overall fit, item fit and person fit been assessed? Overall fit of the model is assessed using the Chi-squared statistic. Item and person fit examines whether individuals and items are responding in the way that was expected. E.g. fit residual values which fall outside of ± 2.5 are an indication of deviation from the model. Statistics such as the Person Separation Index (PSI) or the Cronbach's alpha give an indication of how well the scale discriminates between respondents with different levels of an underlying trait.
  - Has local dependence been assessed? This can be explored by examining the correlation between the residuals of items.

  - Was DIF assessed? For example, were differences between age categories, ethnicity and gender explored?
  - Was RA repeated at each stage of item reduction? Removal of items is likely to affect the overall fit statistics and this must therefore be carefully monitored.
- Following item reduction methods, has the new questionnaire been validated?
- Were any alternative methods for item reduction explored?
- Was the ultimate purpose of item reduction for the development of a condition-specific preference-based measure?
- Was the final structure discussed?
- Were limitations of the study and future research implications discussed?