Health Technology Assessment 2009; Vol. 13: No. 12

Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods

J Hobart and S Cano

February 2009
DOI: 10.3310/hta13120

Health Technology Assessment
NIHR HTA Programme
www.hta.ac.uk

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can be purchased only for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods

J Hobart* and S Cano

Neurological Outcome Measures Unit, Department of Clinical Neurosciences, Peninsula College of Medicine and Dentistry, Plymouth, UK, and Institute of Neurology, London, UK

*Corresponding author

Declared competing interests of authors: none

Published February 2009

DOI: 10.3310/hta13120

This report should be referenced as follows:

Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009;**13**(12).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (*SciSearch*®) and *Current Contents*®/Clinical Medicine.

NIHR Health Technology Assessment Programme

The Health Technology Assessment (HTA) Programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The research findings from the HTA Programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA Programme is needs led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, from the public and consumer groups and from professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA Programme then commissions the research by competitive tender.

Second, the HTA Programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Third, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer reviewed by a number of independent expert referees before publication in the widely read journal series *Health Technology Assessment*.

Criteria for inclusion in the HTA journal series

Reports are published in the HTA journal series if (1) they have resulted from work for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this issue of the journal was commissioned by the National Coordinating Centre for Research Methodology (NCCRM), and was formally transferred to the HTA Programme in April 2007 under the newly established NIHR Methodology Panel. The HTA Programme project number is 95/01/05. The contractual start date was in February 2005. The draft report began editorial review in January 2007 and was accepted for publication in March 2008. The commissioning brief was devised by the NCCRM who specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme or the Department of Health.

ISSN 1366-5278

Published by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk), on behalf of NCCHTA.

Printed on acid-free paper in the UK by the Charlesworth Group. MR

Abstract

Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods

J Hobart* and S Cano

Neurological Outcome Measures Unit, Department of Clinical Neurosciences, Peninsula College of Medicine and Dentistry, Plymouth, UK, and Institute of Neurology, London, UK

*Corresponding author

Objectives: In this monograph we examine the added value of new psychometric methods (Rasch measurement and Item Response Theory) over traditional psychometric approaches by comparing and contrasting their psychometric evaluations of existing sets of rating scale data. We have concentrated on Rasch measurement rather than Item Response Theory because we believe that it is the more advantageous method for health measurement from a conceptual, theoretical and practical perspective. Our intention is to provide an authoritative document that describes the principles of Rasch measurement and the practice of Rasch analysis in a clear, detailed, non-technical form that is accurate and accessible to clinicians and researchers in health measurement.

Review methods: A comparison was undertaken of traditional and new psychometric methods in five large sets of rating scale data: (1) evaluation of the Rivermead Mobility Index (RMI) in data from 666 participants in the Cannabis in Multiple Sclerosis (CAMS) study; (2) evaluation of the Multiple Sclerosis Impact Scale (MSIS-29) in data from 1725 people with multiple sclerosis; (3) evaluation of test–retest reliability of MSIS-29 in data from 150 people with multiple sclerosis; (4) examination of the use of Rasch analysis to equate scales purporting to measure the same health construct in 585 people with multiple sclerosis; and (5) comparison of relative responsiveness of the Barthel Index and Functional Independence Measure in data from 1400 people undergoing neurorehabilitation.

Results: Both Rasch measurement and Item Response Theory are conceptually and theoretically superior to traditional psychometric methods. Findings from each of the five studies show that Rasch analysis is empirically superior to traditional psychometric methods for evaluating rating scales, developing rating scales, analysing rating scale data, understanding and measuring stability and change, and understanding the health constructs we seek to quantify.

Conclusions: There is considerable added value in using Rasch analysis rather than traditional psychometric methods in health measurement. Future research directions include the need to reproduce our findings in a range of clinical populations, detailed head-to-head comparisons of Rasch analysis and Item Response Theory, and the application of Rasch analysis to clinical practice.

Contents

List of abbreviations

1-PL	one-parameter logistic model
2-PL	two-parameter logistic model
3-PL	three-parameter logistic model
ANOVA	analysis of variance
BI	Barthel Index
CAMS study	Cannabis and Multiple Sclerosis study
CI	confidence interval
CPC	category probability curve
D_i	discrimination index
DIF	differential item functioning
EF	endorsement frequency
FAMS	functional assessment of multiple sclerosis
FEW	FAMS emotional well-being scale
FIM	Functional Independence Measure
GHQ	General Health Questionnaire
ICC	item characteristic curve
IRT	Item Response Theory
pADL	personal activities of daily living
PCA	principal components analysis
PSI	Person Separation Index
RMI	Rivermead Mobility Index
MS	multiple sclerosis
MSIS-29	Multiple Sclerosis Impact Scale
SD	standard deviation
SE	standard error
SED	standard error of the difference
SEM	standard error of measurement
SF-36	Medical Outcomes Study 36-item Short Form Health Survey
SLM	simple logistic model
SRM	standardised response mean
T1	time 1
T2	time 2
UKNDS	UK Neurological Disability Scale

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices, in which case the abbreviation is defined in the figure legend or in the notes at the end of the table.

Executive summary

Background

Rating scales are used increasingly as measurement instruments in clinical trials, clinical studies, clinical audit and clinical practice. The results of these studies influence the care of individual people, the making of health policy and the direction of future research. The inferences made from these studies are based on the analysis of numbers generated by the rating scales they use as outcome measures. If clinically meaningful interpretations are to be made from these studies, it is a requirement that the rating scales used are rigorous measures of the variables (aspects of health) they claim to quantify.

This report concerns psychometric methods: these are methods for developing and evaluating rating scales, and for analysing their data. There are many different psychometric methods for evaluating scales in health measurement. Each uses a different type of evidence to determine the extent to which a scale has achieved its goal of generating measurements. This monograph concerns three psychometric methods: traditional psychometric methods, Rasch measurement and Item Response Theory (IRT).

Objective

We evaluate the added value of the new psychometric methods over existing 'traditional' psychometric methods. The report is in two parts. Chapters 1–3 concern theory. Chapters 4–8 are practical demonstrations using existing sets of rating scale data. The report is aimed at clinicians and researchers working in health measurement and tries to provide clear, detailed, non-technical explanations, and a link into the existing but somewhat inaccessible and abstruse literature. The practical demonstrations are comprehensive with full explanations and extensive visual illustrations. There is repetition across chapters to ensure that the basic principles are conveyed.

Methods

The first part of this monograph (Chapters 1–3) presents reviews of the existing literature. Chapter 1 concerns the role of rating scales and the theory and practice of traditional psychometric methods. Chapter 2 outlines the impetus behind the new psychometric methods (Item Response Theory and Rasch measurement), charts their development, and explains their similarities and differences. In this chapter, we provide the case underpinning the reasons why the rest of the monograph focuses on Rasch measurement and not on Item Response Theory. Chapter 3 describes the theory behind Rasch measurement, the development of the Rasch measurement model, the properties of the model and how it 'works' in practice.

The second part of this monograph (Chapters 4–8) presents five practical head-to-head comparisons of Rasch analysis and traditional psychometric methods based on data sets produced from a variety of settings. Chapter 4 compares evaluations of the Rivermead Mobility Index (RMI) in 666 people with multiple sclerosis (MS). Chapter 5 compares evaluations of the Multiple Sclerosis Impact Scale (MSIS-29) in 1725 people with MS. Chapter 6 compares evaluations of test–retest reliability of the MSIS-29 in 150 people with MS. Chapter 7 demonstrates the use of Rasch measurement to equate four scales measuring physical functioning and four scales measuring psychological functioning. Chapter 8 compares the evaluation of relative responsiveness of the Barthel Index and Functional Independence Measure motor scale in 1400 people admitted to a neurorehabilitation unit.

Results

Our reviews of the health measurement literature reveal that: (1) the dominant traditional paradigm for the construction, evaluation and analysis of scales (traditional psychometric methods) is based

on a weak theory; (2) new psychometric methods (Rasch measurement and Item Response Theory) represent a concerted attempt to bring theory and structure to an inherently weak field; and (3) Rasch measurement and Item Response Theory are fundamentally very different approaches.

In the second half of the monograph we focus on worked examples comparing Rasch measurement with traditional psychometric methods. In Chapters 4 and 5, our comprehensive evaluations of the Rivermead Mobility Index (RMI) and the Multiple Sclerosis Impact Scale (MSIS-29) reveal the limitations of traditional psychometric methods and demonstrate the advantages of Rasch measurement. In Chapter 6 we demonstrate the use of different data designs to answer the various components of complex problems and the examination of differential item functioning in test–retest reliability. In Chapter 7 we demonstrate the use of equating tables that enable users of different scales to compare their results. Finally, in Chapter 8 we find that group-based statistics may mislead, and highlight the value and importance of being able to examine change data at the individual person level.

Conclusions and recommendations

We believe that when taken together the arguments and demonstrations in this monograph, both theoretical and empirical, illustrate that Rasch measurement is vastly superior to traditional psychometric methods. Although we have highlighted the value of Rasch measurement in the context of only a limited number of scales for people with MS, we feel that it has much to offer all health measurement, state-of-the-art clinical trials and, most importantly, the individual patients treated by clinicians.

There are a number of future research directions. As next steps, we recommend: (1) that other researchers and clinicians reproduce our findings in a range of clinical populations; (2) detailed head-to-head comparisons of Rasch measurement and Item Response Theory; (3) work to determine further sample size requirements for adequate person and item estimations; and (4) exploration of the application of Rasch measurement to clinical practice in areas including prioritising problems, facilitation of communication, screening potential problems, identifying preferences, monitoring changes or responses to treatment, training new staff and clinical audit.

Chapter 1

Rating scales and traditional psychometric methods

Overview

The aim of this chapter is to explain what is meant by traditional psychometric methods and to document their limitations. This will act as the basis for understanding the reasoning that led to the need for, and development of, new psychometric methods. We start the chapter by looking at two typical rating scales used in health measurement to explain what they are trying to achieve, how they 'work' and how they are most commonly evaluated. We then discuss the theory that underpins those methods and the limitations of that theory. Finally, we discuss the limitations of traditional psychometric methods for rating scale evaluation, development and handling of data.

Rating scales and how they work

Some things that we wish to quantify can be measured directly using devices or machines. Examples include weight, height and protein levels in the cerebrospinal fluid. Other things that we wish to measure cannot be measured directly. Examples include health variables such as disability, anxiety, fatigue and quality of life. These variables must be measured indirectly through their observable manifestations. For example, we can only measure the physical disability of a person by engaging the person in physical tasks.

In fact, all measurements, whether direct or indirect, are of this nature. The weight of a person can only be determined by engaging their weight with an instrument which reacts to it; the height of a person can only be measured by engaging it with an instrument against which we can read off its length. Likewise, we infer the extent to which a person is depressed from the symptoms of this health variable that the person manifests, either through observation or by formalising some questions. In order to stress this inferred aspect of health variables, they are typically referred to in the rating scale literature as latent traits.

Appendices 1 and 2 show two typical health rating scales, the Rivermead Mobility Index (RMI)[1] and the Multiple Sclerosis Impact Scale (MSIS-29).[2]

Consider first the RMI (see Appendix 1). This is a rating scale purporting to measure mobility. It has 15 questions (items). Each item concerns a mobility-related task that has two response categories: 'no' – I am unable to do this task; 'yes' – I am able to do this task. The RMI is scored by clinicians from patient interview and/or observation. Scores for people are constructed by counting (summing) the number of 'yes' responses. Higher scores indicate greater ability and less disability. By convention, the scale scores achieved by summing item scores in this way are called total scores, raw scores or summed scores. These three terms are often used interchangeably.

Items with two response categories are called dichotomous items. It is more common for the items of health measures to have three or more response categories that are ordered from less to more (or vice versa). These are called polytomous items. An example is the MSIS-29 (see Appendix 2.1), which purports to measure the physical and psychological impact of multiple sclerosis (MS). Each item has five response categories. However, the same scoring process applies. Sequential integers (1, 2, 3, 4 or 5) are assigned to the ordered response categories and scores for people are constructed by counting the integer responses across the items. The response categories and their integer scoring system are sometimes called the item scoring function.

The purpose of most health rating scales is to measure people. More correctly, they measure an aspect of people, such as their mobility (RMI) or the physical and psychological impact of MS (MSIS-29). In the light of this fact, consider the aims of the RMI and the MSIS-29. For the RMI mobility is being thought of (conceptualised) as a quantitative variable in the sense that it reflects a property that can have a range of values from 'less' to 'more'. The 15 RMI items attempt to map out this idea so that responses to the 15 RMI items

can be seen as indicators of the level of mobility. Essentially, the RMI seeks to map out mobility as a line (continuum) varying from more to less on which people can be located. Their location is determined by their total score. Thus, the 15 RMI items make operational (operationalise) the idea of the mobility variable. Because mobility is observed by a variety of manifestations, rather than directly, it is considered to be a latent (hidden) property. The words 'trait', 'construct' and 'aspect' are typically used instead of the word property.

Figure 1 represents graphically, and simply, this conceptualisation of the mobility variable by the RMI, and the idea of measurement by locating a person on that line mapped out by the 15 items of the RMI. The notion of locating people (or something) on a line (or continuum) is an important one, is common to all measurement and is often termed 'scaling'.

This conceptualisation implies that each item represents a mark on the 'ruler' of mobility mapped out by the RMI. More specifically, the mark defined by each item represents the transition point of the score from 0 to 1, i.e. the point at which a person moves from scoring '0' (unable to do) to '1' (able to do), and the point below which he or she scores '0' and above which he or she scores '1'.

The MSIS-29 is slightly more complex. First, the RMI generates one total score whilst the MSIS-29 generates two total scores. Items 1–20 are summed to generate the total score for the physical impact subscale; items 21–29 are summed to generate the total score for the psychological impact subscale. Second, each item provides multiple (four) marks on the continuum, because each mark represents the transition between two adjacent response categories (i.e. 'not at all'/'a little'; 'a little'/'moderately'; 'moderately'/'quite a bit'; 'quite a bit'/'extremely'). Finally, note that the ordering implied by the response categories goes in different directions. For the RMI high scores indicate more ability; for the MSIS-29 high scores indicate more disability. This simply means that the continua implied by the two scales run in different directions.

Psychometrics is defined as the study of methods for measuring psychological variables. The field has broadened to include many circumstances where rating scales are used as measurement instruments. Essentially, when evaluating a scale such as the RMI, the aim of a psychometric analysis is to determine whether this idea (conceptualisation) of mobility as a variable mapped out by the 15 items of the RMI has been achieved. In measurement speak, the purpose of a psychometric analysis is to establish the extent to which a quantitative conceptualisation has been operationalised successfully.[3] To achieve that goal a range of evidence is used.

Traditional psychometric methods for evaluating scales

There are many different psychometric methods. This monograph concerns three: traditional psychometric methods, Rasch analysis and Item Response Theory (IRT). Other methods include Thurstone's method of paired comparisons,[4] Thurstone and Chave's method of equal-appearing intervals,[5] Likert's method of summated ratings[6] and Guttman's scalogram analysis.[7] There are many more. Edwards' text[8] gives an excellent account of some of the earlier methods.

Different psychometric methods use different ranges of evidence to determine the extent to which a quantitative conceptualisation has been operationalised successfully. Traditional

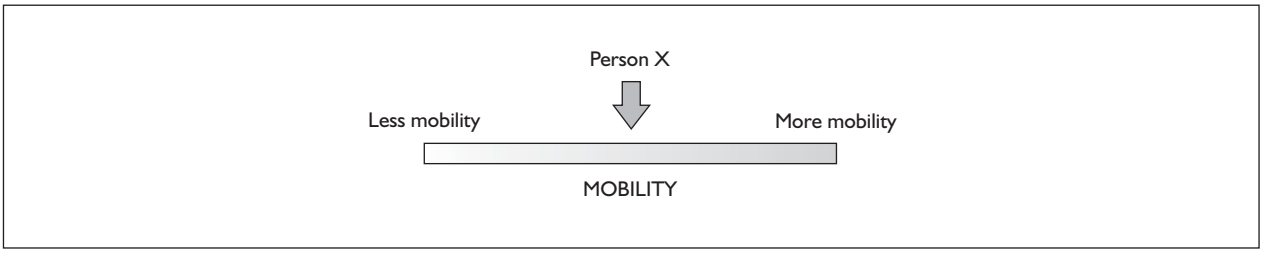


FIGURE 1 *Conceptualisation of the Rivermead Mobility Index. This shows that a variable, here mobility, can be represented as a line, or continuum, ranging from less mobility to more mobility and also illustrates the location of a person on the variable and indicates the amount of mobility that person has.*

psychometric methods, the commonest psychometric method for evaluating rating scales for nearly a century, use evidence predominantly from correlations and descriptive statistics.

Typically, traditional psychometric methods consider evaluating rating scales in terms of three main properties: reliability, validity and responsiveness. However, there is no consensus in the literature, although there are guidelines[9] as to how the results of psychometric analyses should be reported. Hence, they are documented in the literature in a variety of ways. We have previously recommended that a comprehensive scale evaluation using traditional psychometric methods should involve the examination of six psychometric properties: data quality, scaling assumptions, targeting, reliability, validity and responsiveness.[2] Data quality concerns the extent to which a scale can be administered successfully in the target sample.[10] Tests of scaling assumptions examine whether it is legitimate to sum item scores to generate a single scale score.[11] Targeting concerns the extent to which the distribution of disability in the sample matches the range of disability measured by the scale.[12] Reliability describes the extent to which scale scores are free from random error.[13] Validity refers to the extent to which a scale measures what it purports to measure.[13] Responsiveness is the ability of an instrument to detect change accurately when it has occurred.[14] These methods are fully documented elsewhere in a manner accessible for clinicians[2,12,14–16] and are described more fully in the Chapter 4, which documents the evaluation of the RMI.

Classical Test Theory: the theory underpinning traditional psychometric methods

Traditional psychometric methods are underpinned by a theory called Classical Test Theory. It has also been termed Classical True Score Theory, Weak True Score Theory[17] or the Classical Test Model.[18] A test theory, or test model, is a mathematical representation of the factors influencing scores generated by a rating scale, and is described by its assumptions.[19] The reason that the word 'test' (which often confuses clinicians) is used is that the early work concerning the use of rating scales as measurement instruments was in the fields of education and psychology where the aim of measurement was often testing.

Classical Test Theory is a theory of how errors of measurement can influence the scores achieved on rating scales. The theory assumes certain conditions to be true. If these assumptions are reasonable, then the conclusions derived from the model are reasonable. If the assumptions are not reasonable then use of the model can lead to faulty conclusions. Classical Test Theory has seven main and fairly straightforward assumptions. It is interesting to note that a few simple assumptions expand into a set of principles for developing rating scales and for testing their reliability and validity.[19]

The assumptions underpinning Classical Test Theory

The seven main assumptions that underpin Classical Test Theory are discussed at length in a variety of texts[13,17–22] to varying degrees of sophistication. They are summarised and presented in a clinically friendly form below. Assumptions 1–5 present Classical Test Theory's definition of error of measurement. Essentially, according to Classical Test Theory, measurement error is a random (unsystematic) deviation of a person's observed score from a theoretically expected observed score (the true score). In Classical Test Theory, systematic errors are not called errors of measurement. Assumption 6 is the definition of parallel scales. Two scales are parallel if for each person the true score (T) and the error variance are the same. Assumption 7 states the definition of τ-equivalent scales. Two scales are τ equivalent when the true scores for each individual are the same except for an additive constant. Assumptions 6 and 7 do not concern us in this monograph but, for completeness, they have implications for some types of reliability testing.[17,19]

Assumption 1

Classical Test Theory postulates the idea of an observed score, a true score and an error score. The observed score (O) is the score that a person actually achieves on a scale. The true score (T) is the person's real score. This is unobservable (a theoretical value) because of the associated measurement error – the error score (E). Classical Test Theory then postulates that the observed score (O) is the sum of the true score (T) and the error score (E). Thus, it assumes that the relationship between the true score and the error score is additive rather than anything else (e.g. multiplicative). Thus:

$$O = T + E$$

The true score is, of course, a theoretical value. For any person and scale, the true score is assumed to be constant (and unobservable) at any one point in time. However, the error and thus the observed score vary. So, if a person was measured with a rating scale on multiple occasions at the same point in time (say 100 times – an obviously fictitious situation) there would be a distribution of observed scores and a distribution of errors.

Assumption 2
When a scale is administered to a person on multiple occasions, as suggested above, the mean of their observed scores is equal to their true score. Thus, the true score is the theoretical mean of the distribution of observed scores that would be found in repeated independent testing of the same person, at the same point in time. Note that assumption 2 concerns a theoretical distribution of observed scores over different administrations of *one person and one scale*. It further assumes that over these multiple administrations there is no influence of one administration on any other, i.e. each is an independent measurement.

Assumption 3
The error scores and the true scores obtained by a population of people on one scale are not correlated. That is, errors of measurement are not related to the observed score.

Assumption 4
The error scores associated with two scales are uncorrelated. This means that if one person completes two rating scales the errors on the two scales are not related.

Assumption 5
The error scores on one scale are uncorrelated with the true score on another scale.

Therefore, to summarise, Classical Test Theory is a theory which postulates the existence of a true score, that errors are uncorrelated with each other and with true scores, and that observed, true and error scores are linearly related.[23] A number of mathematical conclusions (proofs) can be derived from the seven assumption of Classical Test Theory. These conclusions concern the behaviour of observed, true and error scores, and underpin methods for computing scale reliability and the error associated with scores [the standard error of measurement (SEM)]. Thus, if the assumptions are considered reasonable, the conclusions can be considered true. Likewise, if the assumptions are considered unreasonable, the conclusions cannot be considered true. The conclusions drawn from the assumptions of Classical Test Theory and their mathematical proofs are detailed in various texts,[17,19–21] and many of the basic equations underpinning Classical Test Theory are reported in Spearman's early papers.[24–28]

Limitations in the theory of Classical Test Theory

Classical Test Theory is a useful model that has served developers of rating scales well for many decades.[18] Nevertheless, it has some fundamental problems. Perhaps the most important of these is that the values of the true score (T) or the error score (E) cannot be determined. They are unobservable 'theoretical' variables, the assumptions underpinning the theory cannot be tested and we can only surmise when they would be appropriate.[19] For exactly this reason, most of its equations are unlikely to be contradicted by data.[20] Lord[20] demonstrates that the other equations of Classical Test Theory (which arise from or are related to the basic $O = T + E$ and are subject to the same core limitations) cannot be falsified because they are tautologies.These facts, combined with the fact that Classical Test Theory fails to define mathematically the functional form of the observed score, true score or error score distributions,[23] underpin why Classical Test Theory is also called Weak True Score Theory. In essence, because the assumptions cannot be tested they can be considered met by most test data sets. Therefore, the model can, and is, widely applied.[18] Unfortunately, weak assumptions lead to weak conclusions, in this case about the performance of scales and the measurement of people.

Limitations of traditional psychometric methods

Irrespective of these concerns about the theory underpinning Classical Test Theory, there are significant problems with traditional psychometric methods of evaluating scales, handling their data and constructing scales. The most clinically important of these are discussed below.

Ordered counts are not interval measures

Most health rating scales construct scores by counting answers (responses) to a predetermined set of questions (items). Essentially then, most health rating scales are structured interviews in

which scores are constructed as ordered counts. One important limitation of ordered counts is that they are not interval-level measurement systems. An interval-level measurement system is one in which the unit of measurement implied by the data is equal across the whole range of the continuum.

Two assumptions are made when sequential integers are assigned to ordered response categories and when integer responses to items are summed. The first assumption is that the 'distance' between response categories is consistent within and across items. By distance we mean the real but unobservable distance in interval-level units implied by the score. For example, in the MSIS-29, the distance between 'not at all' and 'a little' is the same as the distance between 'quite a bit' and 'extremely' because they are allocated sequential integers. The second assumption is that the distance between summed (total or raw) scores is consistent across the range of the continuum represented by scale. For example, a difference or change in disability of 5 points has the same meaning (i.e. consistent implications) across the range of the scales. These assumptions have important implications for the measurement of change over time or associated with treatments, the measurement of difference between people at one point in time and the use of statistical tests for analysing data.

It is fairly logical that rating scales such as the RMI and MSIS-29 generate ordinal-level data at both the item score and total score levels. Therefore, it might seem surprising that more is not made of this issue. The reason for this is that psychometricians argue that the ordinal scores generated by rating scales adequately approximate interval-level measurements. The evidence for this is said to have originated from Likert's work in the early 1930s.[6,29,30] When Likert was undertaking his research into rating scales, he was in competition with Thurstone, who believed, amongst other things, that interval-level measurement was a requirement if rating scales were to be valid.[31] He (Thurstone) developed ways of approximating this.[4,5] Likert viewed Thurstone's methods as too cumbersome and complex and proposed a simple method of scoring Thurstone's scales. This has become known as the method of summated ratings.[6] In this method, ordered response categories are allocated sequential integer values, and item scores are simply summed without weighting or standardisation to give total scores. Likert went on to demonstrate high correlations between his total scores and the interval-level measurements generated by Thurstone's methods, as well as equivalent reliability. The simplicity of Likert's approach is one of the reasons why it has been and still is very popular.[32]

Unfortunately, that is not the whole story. Massof [33] demonstrates that Likert's conclusions can be considered to rest on a 'mathematical sleight of hand'. Also, as we will see later in this monograph, the fact that ordinal score and intervalised measurement are highly correlated does not mean that ordinal scores approximate interval measures.

Results for scales are sample dependent

Another important limitation of traditional psychometric methods is that the performance of a scale is dependent on the sample in which it is assessed.

However, the problem is deeper and really concerns the characteristics of the items. Consider the RMI. *Table 1* shows the items' endorsement frequencies (EF = the proportion of people who responded 'yes' to each item) for three samples of people from the Cannabinoids for treatment of spasticity and other symptoms related to multiple sclerosis (CAMS) study.[34] The three samples were everyone (EF1, $N = 666$); the less disabled half of this sample (EF2, $n = 321$ people above the median RMI total score of 6); and the more disabled half of this sample (EF3, $n = 306$ people below the median RMI total score of 6). The results for each item differ notably.

This is important because the frequency of endorsement for an item indicates the 'difficulty' of the item. Consider RMI item 1 (turning over in bed). The endorsement frequency (EF) in the total sample is high (EF1 = 0.73), indicating that most people in this sample can do this task (*Table 1*, column 2). This indicates that it is an easy item for this sample. Compared with the other items, item 1 has the highest EF. This indicates that more people responded 'yes' to item 1 than to all the other items except one. This in turn indicates that item 1 is easier than 13 of the other 14 items. *Table 1* shows that the EF depends on the distribution of disability in the sample, except for item 15. It shows that the endorsement frequencies are sample dependent, and also the relative endorsement frequencies between items are sample dependent. The consequence is that decisions on the adequacy of items depend on the sample in which they are examined.

TABLE 1 *Rivermead Mobility Index (CAMS study visit 1): endorsement frequency (EF) and discrimination index (D_i) in three samples*

Item	EF1	EF2	EF3	D_i1	D_i2	D_i3
1. Turning over in bed	0.73	0.97	0.46	0.53	0.09	0.62
2. Lying to sitting	0.68	0.98	0.34	0.65	0.05	0.74
3. Sitting balance	0.73	0.94	0.51	0.45	0.06	0.55
4. Sitting to standing	0.61	0.91	0.28	0.64	0.13	0.52
5. Standing supported	0.46	0.85	0.05	0.80	0.33	0.10
6. Transfer	0.66	0.98	0.30	0.69	0.08	0.65
7. Stairs	0.33	0.64	0.02	0.62	0.54	0.05
8. Walking inside, with aid if needed	0.54	0.95	0.10	0.85	0.15	0.21
9. Walking outside (even ground)	0.33	0.66	0.01	0.65	0.61	0.02
10. Walking inside, no aid	0.14	0.29	0.01	0.28	0.54	0.02
11. Picking off floor	0.29	0.61	0.003	0.60	0.70	0.01
12. Walking outside (uneven ground)	0.14	0.29	0.003	0.29	0.56	0.01
13. Bathing	0.49	0.84	0.13	0.72	0.25	0.28
14. Up and down four steps	0.13	0.27	0.0	0.26	0.47	0.00
15. Running	0.01	0.02	0.0	0.02	0.02	0.00

D_i1, discrimination index in total sample; D_i2, discrimination index in more able subsample; D_i3, discrimination index in less able subsample; EF1, total sample ($n = 666$; median RMI = 6); EF2, more able sample ($n = 321$, median RMI = 10); EF3, less able sample ($n = 306$, median RMI = 2).
NB For $n = 39$, RMI = 6.

Table 1 shows another index of the utility of an item – its discrimination ability (D_i).[35] This is related to EF, and indicates the likelihood that a person who has a high RMI total score (i.e. is less disabled) is more likely to have responded 'yes' and vice versa. For example, RMI item 8 (walking inside, with aid if needed) has the highest D_i in the total sample ($D_i1 = 0.85$) indicating that most people in this sample are more likely to do this task. In contrast, RMI item 15 (running) has the lowest D_i in the total sample ($D_i1 = 0.02$) indicating that most people in this sample are less likely to do this task (*Table 1*, column 5).

One computation for the discrimination ability is:

$$D_i = \frac{U - L}{n}$$

where:

U = number of people above the median score who score 'yes' on the item
L = number of people below the median score who score 'yes' on the item
n = number of people above (or below) the median (i.e. half the sample).

Table 1 shows the D_i values for each of the three samples described above. Like the endorsement frequencies, discrimination indices and the relative differences in D_i are sample dependent. This is because we get three scores for any one individual. If a person has moderate disability then the likelihood of getting a high score is greater in D_i2 (more able sample) than in D_i3 (less able sample). The discrimination index is related to EF in that items which discriminate best among people have endorsement frequencies of 0.5 near the cut-off point (typically the median score) of the scale. The index differs from endorsement frequencies in that it looks at the items in relation to all the others in the scale and not just in isolation.[35] The reason that this is important is that in traditional psychometric methods, item selection and item performance are based on item discrimination. It is also fundamentally important because it means that the performance of our measurement scales is sample dependent. That is, the measurement scale is affected by the sample it is supposed to be measuring. Logically, this seems incorrect. It is certainly undesirable as investigators need their scales to be stable.

Results for samples are scale dependent

Another important limitation of traditional approaches is that people's measurement on a health trait is dependent on the scale on which they are measured and the sample within which they are measured. For example, consider the measurement of physical function in MS. If we use three recognised scales for measuring physical function, e.g. the Medical Outcomes Study 36-item Short Form Health Survey physical functioning dimension (SF-36PF),[36] the Barthel Index (BI)[37] and the MSIS-29 physical subscale, we will get three different total scores. People who have moderate disability would probably get a high score on the BI, a low score on the SF-36PF and a middle-range score on the MSIS-29. If, however, they are measured within a sample of people with severe disability, they will be in a high percentile of that sample, and if they are measured along with a group of mildly disabled people they will be in a low percentile of that sample. As Wright[38] states, a person's level of ability depends on the scale used and the company they keep.

Thus, a problem with using traditional psychometric methods is that people's level of function cannot be measured independently of the scale, or more correctly the set of items used. Logically, this seems incorrect. At any point in time a person's level of disability is a fixed value (albeit unknown). That value should be the same (within statistical reason) regardless of the items he or she completes, provided those items measure a common variable and are appropriately targeted to the person.

Missing data

A dilemma posed by rating scales is how to handle missing data. More specifically, what should be done when a respondent fails to give a response to one or more items of a scale? There is a literature on this subject and a discussion of it is beyond this monograph. However, one widely used approach in traditional psychometric methods, purported to be scientifically valid,[36] but for which we can find no supportive empirical evidence, is to 'impute' for missing data, i.e. to replace the missing item score(s) with the person-specific mean score (the mean score of the items that have been answered by that individual). This is considered a legitimate process when up to 50% of the items in a scale are missing.

Replacing missing data with the person-specific mean score raises a number of concerns. First, it makes an assumption, albeit based on the average response, as to how a person would have responded to an item. Second, it is appropriate only if the items have the same difficulty, i.e. they share the same location on the continuum. A more scientific alternative is to use only the answered items to generate a score. In either case, it is questionable whether an accurate measurement can be made if only half of the items have been completed.

Standard error of measurement

Using traditional psychometric methods, the error around an individual person's score (SEM) is a constant value regardless of the person's location on the continuum. It is computed from the reliability and standard deviation (SD) of scores in the sample using the following formula:[19]

$$\text{SEM} = \text{SD} \times \sqrt{(1 - \text{reliability})}$$

This has a number of implications. First, it seems illogical that people at the floor and ceiling of a scale, the people whose scores are logically the least precise, have the same level of precision as those who score in the centre of the scale (the most precise). Second, because the computation involves the standard deviation it is dependent on the distribution of the sample. Third, typically the SEM is large; thus, the precision of measurement is limited. For these reasons rating scales are not usually recommended for individual patient decision making.

Scaling items

Before anything can be measured, the variable along which the measures are to be made must be marked out.[39] Traditional psychometric methods do not 'scale' items. That is, they do not give items values that locate them on the measurement continuum. Thus, they neither map out the variable on which people can be measured nor mark out the continuum along which people can be located.

Summary

Health rating scales are measurement instruments whose aim is usually to quantify characteristics of people that cannot be measured directly. They are constructed to map out variables as a line on which

people can be located. Psychometric methods concern the construction and evaluation of rating scales in order to determine their success at making the variable operational and locating people.

The most widely used method for constructing and evaluating rating scales is termed traditional psychometric methods. Underpinning these methods is Classical Test Theory, which postulates the existence of a true score, that error scores are uncorrelated with each other and with the true scores, and that observed true and error scores are linearly related. However, because true scores and error scores cannot be determined, the appropriateness of the assumptions cannot be verified and we can only surmise that they are met.

Most health rating scales use Likert's method of summated ratings. Here, sequentially ordered response options are allocated sequentially ordered integers, and item scores are summed to give total scores. Whilst this method is nearly ubiquitous in health measurement, the belief that ordinal-level total scores approximate interval-level measurements is not well founded. Other problems arising from the use of total scores and traditional psychometric methods are that evaluations of scales are sample dependent and the measurement of people is scale dependent. Even at a purely logical level these undermine a belief in the interpretation of total scores as measurements.

Chapter 2
The new psychometric methods

Overview

Chapter 1 identified that most rating scales in health care use Likert's method of summated ratings as their format, are constructed and evaluated using traditional psychometric methods (if any psychometric methods are used) which are underpinned by Classical Test Theory and use total scores as the basis for their analyses. It then highlighted the limitations of these approaches, specifically that the ordinal total scores generated by Likert's method of summated rating are not interval-level measurements, that the theory of Classical Test Theory cannot be verified in data and that the analysis of total scores as measurements is problematic.

This chapter concerns developments that have been undertaken to overcome the limitations of Classical Test Theory, traditional psychometric methods and total score analysis. The main outcomes of these developments are two 'new' psychometric methods: Item Response Theory (IRT) and Rasch measurement. Essentially, like Classical Test Theory, the new psychometric methods postulate theories of how the scores generated by rating scales relate to measurements of the variables they seek to estimate. However, unlike Classical Test Theory, they provide mathematical realisations of these theories (mathematical models) that enable their verification in data. That is, they can be formally and rigorously tested.

This chapter charts the development of the two new psychometric methods, highlighting that they come from different origins and represent different research perspectives, despite clear similarities. Perhaps not surprisingly, this has led to confusion and misunderstandings about their similarities and differences. We explore why this might have occurred.

The fact that IRT and Rasch measurement are fundamentally different means that two groups have evolved: proponents of IRT and proponents of Rasch measurement. We examine why each group has chosen to adopt their approach and not the other. However, because of the similarities between IRT and Rasch measurement, and the resultant confusion that has arisen, it is our experience that there is a third group: those who do not know that there is a difference between the two.

Background to the new psychometric methods

The term 'new', or 'modern', psychometric methods refers, mainly, to two methods for constructing and evaluating rating scales, or analysing rating scale data. These methods are called Item Response Theory (IRT) and Rasch measurement. The new psychometric methods represent a logical progression from Classical Test Theory, because they attempt to improve the scientific quality of the theory underpinning rating scales. For Classical Test Theory the underpinning theory is weak, and by definition the conclusions about the performance of scales as measurement instruments and the measurement of people cannot be strong.[19] The logical next step was to try to strengthen that theory and, therefore, the scientific quality of the conclusions that could be drawn.

A theory is basically an idea about something. In medicine, theories are proposed about the causes of disease and evidence is gathered which either supports or refutes the theory. In this way theories become refined and developed. Eventually, usually aided by these theories and a bit of serendipity, the cause of a disease is often discovered, and we move on. However, this is not always the case; for example, although various theories have been proposed, the cause of multiple sclerosis remains elusive.

Other theories are ideas about the relationship between variables and entities. These theories serve to describe, explain or predict some behaviour or occurrence. Structure is brought to theories by postulating formal models. These are explications of theories, or parts of theories, that connect the theory to observable events. Their function is to permit logical deduction of relationships that have not been demonstrated but may be demonstrable.

One method of bringing a concrete structure, or framework, to a theory is to express it as formal logical (mathematical) relations between variables or events. The articulation of a theory as a mathematical model (i.e. equation or formula) has three main advantages. First, because equations predict the relationships between variables, they enable careful and sophisticated checks on the consistency of the data to determine how observations (observed data) satisfy the 'fit' of the predictions of a model, i.e. is the extent to which the data satisfy the theory, or alternatively the extent to which the theory is supported by the data. Second, they enable predictions to be made for the future.

The third value of mathematical models comes from the analysis of deviations of the observed data from the predictions of the mathematical model. These analyses can lead to important developments because they allow theories and models to be explored in two directions. One direction is to question whether the model is an adequate representation of the theory. This can lead to developments in our understandings of the theory and changes in the models used to formalise them. However, this direction requires confidence in the observed data. The other direction in exploring discrepancies between observed data and the model prediction is to question whether the data are adequate. This can also lead to developments in our understandings of the theory, but is more likely to lead to developments in understandings of the data themselves and how they might be collected. This direction requires confidence that the models used are adequate representations of the theory.

If sets of items are to be used as a measurement instrument, a theory is needed to underpin the development of scales, evaluation of scales and analysis of rating scale data. Otherwise any collection of items could be used in the belief that they represent a measurement instrument. We outlined earlier that Classical Test Theory is a theory of measurement error. To recap, it postulates that, when using a rating scale, a person has a 'true score'. A person's true score is the score a person would get on a rating scale if there were no measurement error. Thus, it is unobservable because of measurement error. The only observable measurement is the actual score on the scale – the 'observed score'. The theory then postulates that there is an additive relationship between these three scores (true = error + observed) and expresses this relationship as a mathematical formula (model or equation):

$$T = O + E$$

The theory also postulates that the size of the error is not correlated with the size of either the true or the observed score, and that the size of the error on one scale is not correlated with the true or observed scores on another scale. Finally, the theory postulates that the true, observed and error scores are linearly related, i.e. they have a straight line relationship. Thus, a fixed unit of change in one scale means a fixed unit of change in the other. It does not mean, necessarily, that the unit of change is the same. We have seen that from these statements postulated by Classical Test Theory, and its mathematical model, we are able to draw a number of conclusions and derive further equations for testing scale reliability and validity.

The main problem with Classical Test Theory, as we have seen, is that it is a weak theory. This is because the postulated statements constituting the theory, and the derived mathematical equations, could not be tested, verified or refuted. Therefore, it is an easy theory to satisfy. The new psychometric methods are an attempt to raise the scientific bar of the theories underpinning rating scales. Like Classical Test Theory, the new psychometric methods concern the relationship between a person's unobservable true measurement and his or her observed score. Unlike Classical Test Theory, the new psychometric methods focus on the relationship between a person's unobservable measurement on the underlying trait and the probability of responding to one of the response categories of a scale item.

There are a number of points to note about the final sentence of the last paragraph. The term 'person's unobservable measurement' is used because the new psychometric methods recognise the limitations of total scores. The new methods are interested in someone's true measurement on the construct being measured by the scale, i.e. his or her location on an interval-level continuum, rather than the true total score, which we know is ordinal in nature. The second point to note is that a person's true interval-level location governs his or her response to an item, i.e. the location on the construct should determine which item response category a person chooses on an item. This is logical. We would expect a person who is severely disabled to be unable to do many

tasks of the Rivermead Mobility Index (RMI). We expect such a person to answer 'no' to most items. Thus, the response to any item is a function of someone's level of disability. Also logically, a person's response to an item is related to the difficulty of the tasks: running (RMI item 15) is usually a more difficult task than turning over in bed (RMI item 1) if a person is physically disabled. The third point to note is the use of the phrase 'probability of responding to one of the response categories of a scale item'. This recognises that it is unreasonable to try to predict exactly how someone will respond to an item. We can only say that he or she is likely to choose a specific response category. Thus, the relationship between true interval-level measurement and response to an item is best formalised in terms of probabilities. The final point to note is that the focus has changed from the total score level in Classical Test Theory to the item score level in the new psychometric methods.

It seemed reasonable, therefore, to consider developing mathematical models that would relate the probability of a response to an item to the person's location on the continuum measured by the items and to some characteristics (parameters) of the items. The challenge then was to find, or derive, a mathematical model that would link these variables together, and give the information required, using the responses of a sample of people to a set of items. Essentially, the mathematical model must use this information to allow us to determine if a set of items forms a reliable and valid rating scale on which people can be located, preferably in interval-level units.

At this stage it is worth noting some issues relating to terminology. The term 'Item Response Theory' is used, not surprisingly, because both of the new psychometric methods concern theories about responses to items. Confusion has arisen because some people use the term to encompass both branches of new psychometric methods (IRT and Rasch measurement), while others use it to mean only the IRT branch of new psychometric methods. Other synonyms for the new psychometric methods are Person–Item Response Theory, Strong True Score Theory (to differentiate from Weak True Score Theory), Latent Trait Theory (to emphasise the fact that the constructs being measured are unobservable or latent) and modern test theory.

As pointed out in the introduction to this chapter, the two new psychometric methods have different origins and directions and represent different perspectives. Thus, they are fundamentally different. However, to complicate matters and ensure confusion, they also have important similarities.

Development of Item Response Theory

Louis Thurstone, at the University of Chicago, was the first person to attempt to bring strong theoretical and mathematical underpinnings to rating scales. In the 1920s, he articulated a set of requirements for measuring 'social' variables, as he called them, with rating scales.[4,5,31,40] Essentially, Thurstone stated that:

- Rating scale items should define a continuum and be located on the continuum as landmarks of different levels of construct of interest.
- Rating scales should measure clearly defined single aspects of things or people.
- Rating scales should measure that entity on an equal-interval (interval-level) scale.
- The performance of the rating scale should not be affected by the sample.
- The measurement of a person should not depend on the scale (items) used.

Thurstone formalised these requirements for measurement using mathematical models. He argued that it should be possible to 'rationalise the problem and establish the functions that underlie the data' and that using correlations was an 'acknowledgement of failure' to do this.[40] He developed two methods for constructing rating scales: the method of paired comparisons[4] and the method of equal appearing intervals.[5]

Soon after Thurstone published these methods, Likert proffered his method of summated ratings.[6] Likert argued that Thurstone's approach was 'exceedingly laborious' and complex. He questioned whether Thurstone's scales 'worked better than simpler scales' and whether it is possible to 'construct equally reliable scales without making unnecessary statistical assumptions'. In his paper, Likert[6] demonstrated that, when one of Thurstone's equal-appearing interval scales was scored using his (Likert's) simpler method, the two versions of the scale correlated highly and had similar reliability coefficients. Likert took this as evidence that his simpler scoring methods were just as good as Thurstone's more complex approach. Likert also demonstrated that when one of his own scales was scored using both his simple scoring method and a method based on approximating equal-appearing intervals, these two versions of

the scale correlated highly. He took this to mean that the sequential integer scoring of response categories was an adequate approximation of interval-level measurement.

The application of Likert's method of summated ratings then took off. Today it is an almost ubiquitous measurement method in the social sciences.[41] This is not surprising; Likert's approach was easy and simple to understand and his arguments appeared convincing. In contrast, Thurstone's work was more complex and abstract. Soon after Likert's publication, Thurstone published his lecture notes on reliability and validity testing,[42] then focused his research energies on factor analysis, which he pioneered and saw as a method for determining the dimensions that underpin complex multidimensional variables.[43]

After Thurstone's change of focus, others continued to develop the field of mathematical models for rating scales. Important contributions were made by Richardson,[44] Ferguson,[45–47] Lawley,[48,49] Tucker,[50] Brogden[51] and Lazarsfeld.[52] However, progress was slow, probably due to some of the mathematical complexities of the area and the lack of computational facilities.[18] In addition, during this time there were some important methodological developments in terms of traditional psychometric methods for reliability[53–55] and validity[56–58] testing that almost certainly helped to strengthen the position of traditional psychometric methods as the dominant paradigm for developing and evaluating measurement scales.

However, the work of Frederic Lord[59,60] seems widely regarded to be the birth of IRT. He is considered to be one of the first people to develop a mathematical model describing the relationship between a person's level of (dis)ability and his or her response to the items of a rating scale, and to apply this model successfully to real data.[18] Lord, along with Novick and Birnbaum, went on to publish a landmark book on the area in 1968[17] and then another on his own in 1980.[20]

Early work in IRT focused on trying to develop mathematical models that formalised Classical Test Theory. Specifically, this work related a person's measurement on the construct to the features (parameters) of items that were seen to be important in traditional psychometric item analyses: item difficulty and item discrimination. In traditional psychometric methods, the 'difficulty' of an item is indicated by its endorsement frequency. As we saw in the previous chapter the EF is the proportion of people (p) responding in each item response category. On dichotomous items, as on those of the RMI, the proportion of people responding 'yes' is an indicator of the item difficulty. On the RMI items, a higher proportion of people saying 'yes' indicates an item that more people can do, i.e. an easier item, compared with one having a low proportion of people responding 'yes'. *Table 2* gives these p-values for the RMI items (column 1). They differ, indicating different endorsement frequencies and thus different item difficulties (locations on the continuum).

The item discrimination index, discussed in Chapter 1, is the likelihood that a person who has a high RMI total score (i.e. is less disabled) is more likely to have responded 'yes' (can do this task). It indicates the ability of the items to discriminate between people who attain different levels on the construct. The higher the discrimination index, the better the discriminant ability, the better the item. These are also shown in *Table 2*.

Mathematical models relating the probability of a response to an item to the person's location, the item's difficulty and the item's discrimination are known as two-parameter (2P) models. For some reason, the person location is not considered in the number of parameters used to name the model. Examinations of these models did not necessarily account adequately for observed data sets. This led researchers to consider adding other item and person characteristics (parameters) to the basic 2P model so that the data might be better explained. In educational testing, where much of the early work was done, guessing the responses to items was felt to be important, so models including an item guessing parameter,[61] a person guessing parameter[62] and a person discrimination parameter[63] were developed. Of these 'multiparameter' models, those used mostly are the basic 2P model and the three-parameter (3P) model in which the third parameter is item guessing.

The general approach in the development of IRT was to try to develop mathematical models that explained the observed rating scale data. Essentially, models were postulated and examined in data. When the observed data were not adequately explained by the mathematical model, i.e. when the data did not fit the chosen model closely enough, another model was tried. Thus, the justification for model selection was empirical evidence of its suitability.[23] The choice of one model over another indicated that it accounted better for the data.[64] The data were considered given. It is important to note that most

TABLE 2 *Rivermead Mobility Index (CAMS study visit 1): endorsement frequency and discrimination index* (n = *666)*

Item	EF	U	L	*n*	$D_i = U - L/n$
1. Turning over in bed	0.73	311	142	321	0.53
2. Lying to sitting	0.68	315	105	321	0.65
3. Sitting balance	0.73	303	157	321	0.45
4. Sitting to standing	0.61	291	87	321	0.64
5. Standing supported	0.46	272	14	321	0.80
6. Transfer	0.66	314	91	321	0.69
7. Stairs	0.33	206	7	321	0.62
8. Walking inside, with aid if needed	0.54	305	31	321	0.85
9. Walking outside (even ground)	0.33	213	3	321	0.65
10. Walking inside, no aid	0.14	92	2	321	0.28
11. Picking off the floor	0.29	195	1	321	0.60
12. Walking outside (uneven ground)	0.14	94	1	321	0.29
13. Bathing	0.49	270	40	321	0.72
14. Up and down four steps	0.13	85	0	321	0.26
15. Running	0.01	5	0	321	0.02

D_i, discrimination index; EF, endorsement frequency; *L*, number of people below median[a] scoring 'yes'; *n*, number of people above median; *U*, number of people above median scoring 'yes'.
a Median RMI for sample = 6.

circumstances of 'modelling data' involve finding a model that fits the data. In addition, the finding that proposed models did not fit observed data means that model development is also justified empirically.

Development of Rasch measurement

At the same time that Frederic Lord was developing his theories at the Educational Testing Service in the US, Georg Rasch was developing his theories at the Danish Institute of Educational Research in Copenhagen.[65] It is important to note that Rasch's work was independent of Lord's (and vice versa), although Lord discusses Rasch's work in his landmark text published 8 years later.[17] In addition, to our knowledge, Rasch's work was independent of influence from any other work into psychometric methods because, it is said, he rarely read research papers.

At the beginning, Rasch took a similar approach to Lord's in that he tried to model the data. He was working with reading test data, in particular students' ability to read texts of varying difficulty, and the distribution of errors. He chose to work with a Poisson model as this is generally used as an error count model. However, to make the Poisson 'work' he needed to have a parameter for each person's ability and each reading text difficulty. Thus, Rasch treated each student in his data set as an individual. This contrasts with more conventional mathematical modelling in which people tend to treat groups of individuals as random replications of each other so that the error includes the variation among individuals. He fixed those, and there was randomness only in the response process of a person to an individual text.

This work gave Rasch a very elegant model in which he could eliminate the person parameter to carry out tests of fit and to get the relative difficulty of the texts. He showed this to Ragmar Frisch (Norwegian economist and Nobel Prize winner), who noted that the person parameter dropped out. This observation made Rasch think more closely about the mathematics of the model and the class of all such models. From this he derived the specific model for dichotomous responses independently of any data of the kind. He then analysed two existing tests according to this model. One set of data fitted the model, the other did not.[66,67]

Over time, Rasch increasingly noticed the implications of a property of the model he

had deduced. The property was that the item locations and person locations could be estimated independently of each other. This will be discussed and demonstrated later. However, it is the implication of this property that is important. It means that the measurement of people can be made independently of the sampling distribution of the items used, and the location of items on the continuum can be made independently of the sampling distribution of the people from whom they are derived. Put another way, this means that the relative locations of any two people does not depend on the items they take, and that the relative locations of any two items does not depend on the people from whom the estimates are made. This is Rasch's criterion of invariance – or stability.[66]

Of course there is one essential proviso: that the data fit the Rasch model. This is because the property of invariance is a consequence of the model, not of the data or just the *application* of the model. Thus, for invariance to be a consequence of the data, the data must fit the Rasch model within statistical reason. However, this means that when the data do fit the Rasch model, different subsets of items will give equivalent person estimates and different subsets of persons will give equivalent item estimates.

Rasch realised that this property was a fundamental property of a very important class of models. From then on he shifted his focus from describing data sets to studying a class of models with a unique property. The implications are discussed further below.

Rasch developed his model for dichotomous variables. His work was developed at the University of Chicago by Wright,[68] with Stone[69] and Masters,[39] and by Andrich, who generalised the Rasch model for use in rating scales with polytomous response categories.[70]

In the Rasch paradigm, the mathematical model is given primacy. This is not because it describes any set of data. The case rests on the inherent properties of the model. Essentially, it provides the optimum criterion for fundamental measurement.

The general approach in Rasch measurement is to use only Rasch models to construct scales, evaluate scales and analyse rating scale data. The use of other item response models is not typically considered. When the data do not fit the Rasch model, another model is not chosen. Instead, the finding invokes an examination of the data to determine why, for example, a set of items hypothesised to be a measurement instrument are not performing as such. Thus, the justification for model selection is theoretical evidence of its suitability.[23] The data are not considered given. It is important to note that this is not the typical approach to using mathematical models with data. In addition, the finding that observed data do not fit the chosen model means that developments in construct theory, and the items used to operationalise them, is justified empirically.

Item Response Theory and Rasch measurement: similarities and differences

Similarities

There are a number of important similarities between IRT and Rasch measurement. This only serves to increase the potential for confusion. The first similarity is that both families are item response models for constructing and evaluating rating scales and analysing their data. The second similarity is that both appeared at around the same time. The first written reports about both IRT and Rasch measurement were written in the 1950s[71] and their landmark texts were published in the 1960s.[17,65]

Perhaps the greatest similarity between IRT and Rasch measurement is the structure of the mathematical models.[33] *Figure 2* shows the mathematical models of the Rasch model and the two-parameter and three-parameter models. All three are 'logistic' models. This simply means that central to them is the expression (e/1+e). Hence, the Rasch model is known as the simple logistic model (SLM) and the two IRT models as the two- and three-parameter logistic models (2-PL, 3-PL). Second, the models are such that:

- if $C = 0$, the 3-PL becomes the 2-PL
- if $a = 1$, the 2-PL becomes the SLM
- if $a = 1$ and $C = 0$, the 3-PL becomes the SLM

where C = constant, describing the lower asymptote due to guessing and a = discrimination.

This has led many people to consider the Rasch model as nothing more than a 'special case' of the 2- and 3-PL models with convenient properties,[72,73] and to describe it as a one-parameter logistic model (1-PL). That perspective and terminology is reasonable if the aim is to find the model that best fits the data. However, it ignores the fundamental reasoning behind the development

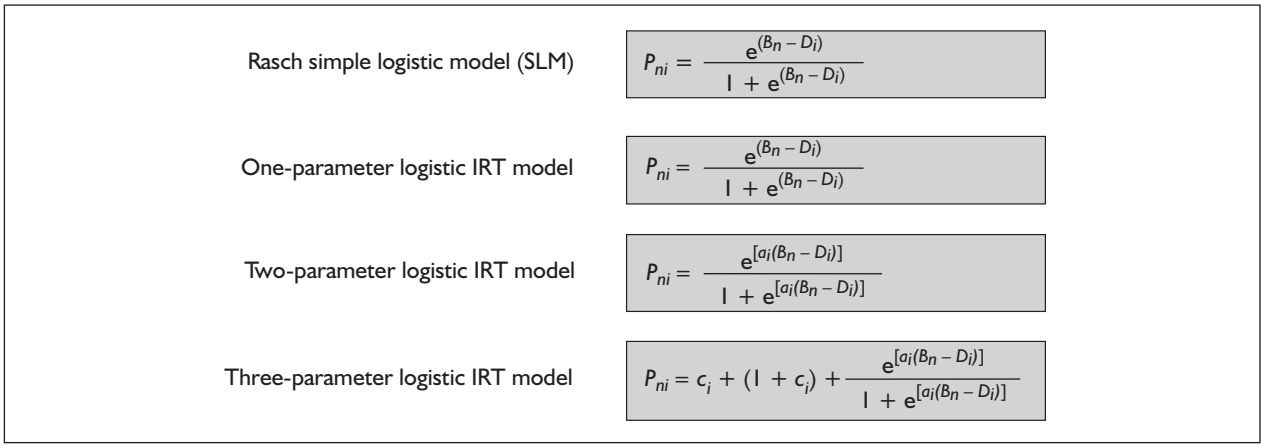


FIGURE 2 *Formulae for the Rasch model, and for the one- two- and three-parameter logistic models. c_i = constant, describing the lower asymptote due to guessing for item i; a_i = slope of item i (discrimination); D_i = difficulty of item i; and the one-person parameter is B_n = ability of person n. Note. Some refer to the Rasch model as the 'one-parameter model'. This is the case here to demonstrate the mathematical similarity of the models. However, as explained in the text, the conceptions of the Rasch model and the 1-PL IRT model were very different.*

of the Rasch model and its value. That perspective is not reasonable if the reasoning behind the development of the Rasch model is taken into account. This is because the addition of parameters to the Rasch SLM prevents the separate estimation of item and person parameters and thus the property of invariance. Massof[33] demonstrates this formally.

There are two more similarities between IRT and Rasch measurement that are noteworthy. The first is that proponents of both have some common goals despite their different agendas. Both approaches are concerned with the development of better measurement methods, improved evaluation of scales and greater understanding of the constructs that we seek to measure.[64,74] The second is that both approaches often acknowledge the importance of Thurstone's work, but for different reasons. Proponents of IRT acknowledge Thurstone for being the first person to apply mathematical models to the field of 'social' measurement. Thus, IRT is in keeping with their approach. Proponents of Rasch measurement acknowledge Thurstone's work because he laid down a series of *requirements* for measurement, which he argued rating scales should satisfy to be considered valid (and which Likert misinterpreted as assumptions[75]). Thus, Thurstone is advocating the primacy of the theory and the model.

Differences

The fundamental difference between IRT and Rasch measurement is the approach to the problem or the research agenda.[66] IRT aims to find the item response model that best explains the data. Rasch measurement will use only the Rasch model, and if data do not fit this researchers will seek to understand why and, if necessary, remove data, recollect data or reconceptualise the construct. This is because proponents of IRT prioritise the data. In contrast, proponents of Rasch measurement prioritise the model.

Another difference is that proponents of IRT may well use Rasch models but proponents of Rasch measurement are highly unlikely to use other item response models. This is because proponents of IRT, in the process of trying to find the model that best explains the observed data, are highly likely to examine the extent to which the simplest item response models fit the data. The simplest item response models are those with the fewest parameters, i.e. Rasch models. However, as models with more parameters often achieve better fit of the model to the data, Rasch models are rarely the models ultimately chosen for describing any data set. Conversely, proponents of Rasch measurement are unlikely to use other item response models because the presence of additional parameters in these mathematical models destroys the very property (invariance) they hold with primacy.

Other differences between IRT and Rasch measurement arise from their different approaches for example how certain findings within the data are considered, interpreted and managed. This applies particularly to features such as item discrimination and guessing. Both the 2-PL and 3-PL item response models include an item discrimination parameter. This means that the discrimination of each item is estimated. The Rasch model does not have an item discrimination parameter. Thus, if items have empirically different discriminations this is shown up as misfit of

the data in the Rasch model. More specifically, different item discrimination shows up as misfit for individual *items* that requires further qualitative exploration and explanation.

Similarly, the 3-PL model contains a guessing parameter for items. Thus, this is estimated in the analysis. The Rasch model does not contain a guessing parameter, and 'guessing' shows up as misfit for individual *people*. It is important to note that the analysis does not tell us that these individuals were guessing. It identifies their response pattern across the items as inconsistent with expectation, hence misfitting. The cause of the misfit requires further qualitative exploration of the individuals.

Why proponents of Item Response Theory favour their approach

Any opinion necessarily consists of arguments for it and counter-opinions against it. Proponents of IRT give primacy to the data and an attempt to best explain it. Thus, it is easy to understand why one might seek to find the item response model with the best fit. Another reason people favour the IRT approach is concern about using an item response model that does not include an estimate of item discrimination, given that there is empirical evidence to demonstrate that items have different discriminatory abilities.[76] A further reason supporting the IRT approach is concern about the concept of finding data to fit the model. This has been felt by some to disturb construct validity.[76] However, giving primacy to the data does imply confidence that the data are not fallible. This might raise cause for concern given the ambiguities of items and the fact that many items could be placed in scales measuring a range of constructs.

Why proponents of Rasch measurement favour their approach

The reason that proponents of Rasch measurement favour their approach, and give primacy to the choice of model, lies in the inherent properties of the model. In essence, it offers the optimum criterion for fundamental measurement.[66] It is appropriate to explore this statement further because it goes right back to the meaning of measurement itself.

Many psychometric textbooks quote the definition of measurement formalised by Stanley Smith Stevens,[77] but which Stevens,[77] and others,[78,79] attribute to Campbell.[80] He defined measurement as 'the assigning of numbers to things according to rules'. Stevens proceeded not to define the rules for assigning numbers to things but rather to offer a four-level classification of scales (nominal, ordinal, interval, ratio) and the statistical tests that ought, or ought not, to be applied to them. Although there have been other variants (and combinations) of these, Stevens' classification of scales as ordinal, interval and ratio has been widely adopted as the basic scales of measurement and has had a major influence on psychometric theory and its development.

It is perhaps interesting to note that many standard psychometric texts over the years[19,79,81–88] offer little discussion of the topic of the nature of measurement, despite the tomes devoted to it.[81] In fact, measurement is a very important and advanced concept.[66] It has been defined as the *sine qua non* of science,[89] Helmholtz[90] is reputed to have stated that all science is measurement,and it is widely acknowledged that advances in measurement methods have underpinned the progress of physical sciences.[66] Measurement in the physical sciences, which was called fundamental measurement by Campbell in the 1920s,[91] has been studied and debated at length by mathematicians, philosophers and social scientists in order to articulate the characteristics of measurement, against which systems purporting to generate measurements (for example, rating scales) can be tested.

Underpinning most physical measurement systems, e.g. weight and height, there is an empirical combining (concatenating) process. As Perline *et al*.[92] state: 'It is easy to show that two lumps of clay joined into one is equal to the sum of the weights of the individual lumps.' Campbell[91] deduced that the ability to concatenate was the fundamental property on which physical measurement (i.e. measurement in physics) was based, coined the term 'fundamental measurement' and argued that there could be no fundamental measures in psychology because concatenation of psychological properties seemed impossible.

This challenge motivated the theory of conjoint measurement, which aimed to determine the conditions required for rating scales to achieve fundamental measurement. Luce and Tukey[93]

made an important contribution to the field in 1964.Essentially, for an item response model to produce fundamental measurements it must satisfy the mathematical requirements of a non-interactive conjoint structure.[94] These mathematical requirements concern relationships between the three key variables in the model: the person location and item location estimated by the model and the observed responses to items. A non-interactive conjoint structure is one that exhibits additive relationships, double cancellation, solvability, the Archimedean axiom and the independent effects that the item locations and person locations have on the observed responses.

The Rasch model satisfies these five conditions. Other item response models do not.[66] This underpins why the Rasch model offers the optimum potential for constructing fundamental measurements from rating scale data,[66] why the status of other item response models as measurement models has been challenged[33] and why proponents of Rasch measurement favour their models and do not tend to use others.

Limitations of new psychometric methods

The main limitation of the new psychometric methods is that they require a complex, more advanced level of mathematical understanding, and investment in training in the use of new software. For these reasons, Rasch analysis appears complicated and is not widely used, and there are few clinicians and researchers trained in its use and interpretation. Thus, the key question is 'Do the clinical advantages of Rasch analysis outweigh the necessity for specialised knowledge and software?' We believe that the new methods offer clinically meaningful, scientific advantages that far outweigh concerns about the necessity for specialised knowledge and software. In particular, the benefits of Rasch analysis provide a substantial leap from traditional methods in measuring patient outcomes. These benefits are that Rasch analysis: (1) offers the ability to construct interval-level measurements from ordinal-level rating scale data, thereby addressing a major concern of using rating scales as outcome measures; (2) enables us to obtain estimates suitable for individual person analyses rather than only for group comparison studies; (3) enables us to use subsets of items from each subscale rather than all items from the scale, yet still allows us to compare scores using different sets of items; and (4) allows for missing item data to be handled scientifically, rather than on the basis of assumption, because Rasch analysis computes an estimate from the available data rather than requiring missing data to be imputed.

We hope that this monograph will help to illustrate some of the reasons behind our belief in the relative advantages of Rasch measurement.

Summary

The aim of this chapter was to try to introduce the new psychometric methods – IRT and Rasch measurement. Yet, they are not so new; both approaches were available 40 years ago. However, they are complicated and there are few non-technical accounts that are accessible to people from a clinical background. To that end, we have tried to provide some insights to help interested people understand some of the fundamental issues and enable them to access what is undoubtedly a complicated literature.

Classical Test Theory is weak and the field of measurement using rating scales needed strong theories to underpin the science for it to be taken seriously. Two independent fields evolved, but they had many similarities that were bound to cause confusion, especially when coupled with some non-specific terminology.

The fundamental difference between Item Response Theory and Rasch measurement is substantial but subtle: one gives primacy to the data, the other gives primacy to the mathematical model. This results in a different approach to addressing measurement problems, and different approaches to the management of the findings. It is hardly surprising that Andrich[66] argues that they are incompatible paradigms with proponents who are most likely to agree to disagree. There have been many misunderstandings between the two fields in the past, but it is interesting to note how infrequently the true differences of the two approaches are acknowledged publicly. As a consequence, our experience that few people are aware of the differences and similarities comes as no surprise.

The purpose of attempting to compare and contrast the two approaches is to help health-care professionals make up their own minds as to which they find most suited to their needs. In this chapter

we have not argued that one way is right and the other is wrong. Certainly, both are superior to Classical Test Theory from a scientific perspective. We are proponents of the Rasch approach, the reason for our perspective being the opportunities offered by the Rasch model for advancing measurement in health care and, as a consequence, improving patient care.

Chapter 3

The Rasch measurement model

Overview

In Chapter 2, we discussed that the need for strong theories, articulated in the form of mathematical models, underpinned the new psychometric methods. We identified that there were two main directions within the new psychometric methods – Item Response Theory and Rasch measurement – and that these methods had fundamental differences.

This chapter, and the rest of the monograph, focuses on Rasch measurement. We start by discussing the theory behind the Rasch model and showing how it is derived. Next, we explain and demonstrate two important properties inherent to the mathematics of the Rasch model: the fact that item and person estimates can be generated independently of the sample distributions of each other; and the fact that total scores for persons and items are 'sufficient statistics' for locating items and people on the continuum. Finally, we try to bring it all together and explain how a set of 'yes'/'no' responses of a sample of people to a set of items can be used to generate estimates of items and persons that, when the data fit the Rasch model, are stable linear estimates. In this chapter we also need to introduce some of the mathematics.

Theory

There are two main components to the theory of Rasch measurement. The first component is that a person's response to an item is governed by only two factors: the location of the person and the location of the item on the shared continuum. The second component is Rasch's criterion of invariance, specifically that the relative location of any two persons on the continuum should be independent of the items used to make that comparison. The symmetrical aspect of this criterion is that the relative location of any two items on the continuum should be independent of the persons used to make that comparison.

It is valuable to recap, and develop, some of the issues. When a set of items is used as a measurement instrument, the aim of the items is to map out a continuum onto which people can be measured (located). It is more correct, as Thurstone[31] pointed out, to say 'on which aspects of people' can be located. This is because we measure something about people, for example their height, weight, mobility, disability or quality of life. As the items mark out the measurement continuum, their locations define it. Thus, there is a common continuum on which the aspect of the people that is to be measured and the items for doing the measuring are located. Thus, when we use a set of items to measure people there is a simultaneous interaction between people and items.

A simply analogy may help to make this idea of simultaneous interactions concrete. This analogy concerns the measurement of strength and it seems to have been used first by Edwards[8] in explaining Thurstone's method of paired comparisons. A similar analogy was used to illustrate the Rasch measurement model by Wright in his lectures, and by Wright and Stone in their publications.[95,96]

Let us imagine that an investigator wants to determine the weights of a set of objects which range from very light to very heavy. Unfortunately, he does not have a physical scale for measuring weight. One alternative is to present the objects to a sample of people and ask them to make a judgement about the weights. For example, people could be asked to arrange the objects in order from lightest to heaviest, or they could be presented with all possible pairs of objects and asked to judge which object in each pair was heaviest. In fact, these two approaches underpinned Thurstone's method of equal-appearing intervals[5] and the method of paired comparison.[4] Another approach is simply to ask each person to lift each object and record whether he or she can or cannot lift it.

The analogy of people lifting objects highlights that every observation involves a simultaneous interaction of multiple forces.[96] Here, the simultaneous interaction is between the person's strength and the object's weight, and the shared variable is of strength/weight. However, does a person lift an object because that object is light or because the person is strong? Similarly, does a person fail to lift an object because that object is heavy or the person is weak? Clearly, any interpretations of these observations will be confounded until there is a way of making separate estimates of the different forces (here, person strength and object weight). This, as we have seen, is a problem with traditional psychometric methods – the measurement of people depends on the scale, and the properties of the scale depend on the sample. They cannot be separated.

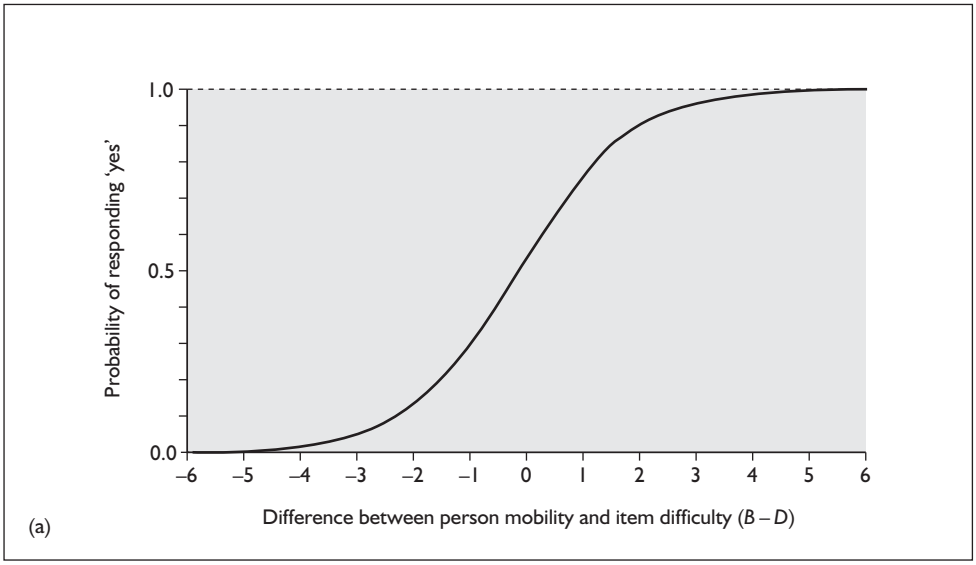
Rasch's criterion of invariance warrants a further comment. When two people are measured they are located on the continuum of whatever is being measured, e.g. mobility using the RMI. Rasch's criterion of invariance is that the relative locations of these two people should be the same regardless of the RMI items chosen to measure their mobility. Similarly, the relative location of any two items on the continuum should be the same regardless of the people chosen to generate those estimates. Note that this refers to the *relative* locations, not the absolute locations of persons and items. The relative locations refer to the distances between the locations of people or items.

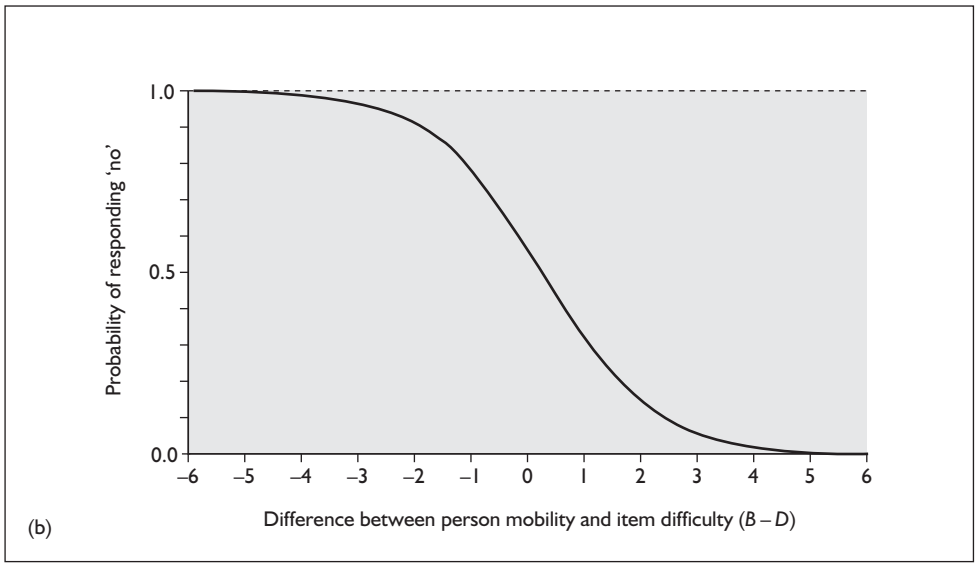
Mathematics

We will start by pursuing the people lifting objects analogy. Logically, when any person in the study sample (standard notation = n) attempts to lift any of the study objects (standard notation = i), the thing that governs the likelihood (probability) of a successful lift is the difference between a person's strength (standard notation = B) and the object's weight (standard notation = D). Thus, for person n lifting object i, the probability of that person lifting that object is governed by the difference between the strength of person n (B_n) and the weight of item i (D_i), i.e. ($B_n - D_i$).

When person n is stronger than an object i is heavy, this difference is greater than zero ($B_n - D_i > 0$), and thus the probability of a successful lift is more than 50% (> 0.5); and the more B_n exceeds D_i,the greater the probability of a lift. Similarly, when person n is weaker than an object i is heavy this difference is less than zero ($B_n - D_i < 0$), and the probability of a lift is less than 50% (< 0.5); and the more D_i exceeds B_n the greater the probability of failing to lift. Logically, therefore, when a person is as strong as an object is heavy, the strength/weight difference is zero ($B_n - D_i = 0$), and the probability of a lift is 50% ($= 0.5$).

Now consider this logic in relation to any person with MS from the CAMS study completing the RMI. The actual response to any of the 15 items is determined





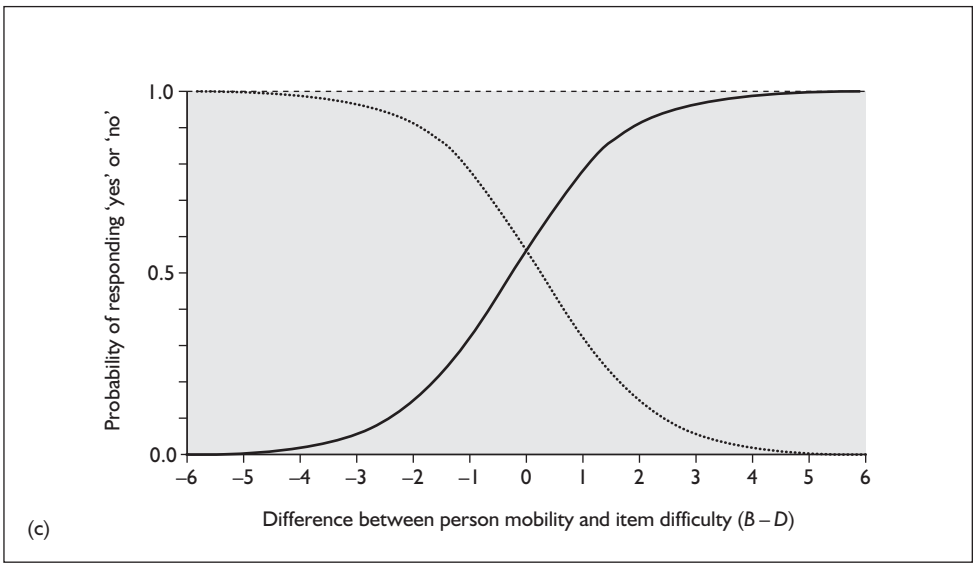


FIGURE 3 *(a) Probability of responding 'yes' to an item versus difference between person and item locations. (b) Probability of responding 'no' to an item versus difference between person and item locations. (c) Probability of responding 'yes' or 'no' to an item versus difference between person and item locations.*

by the person's mobility level and the item's difficulty. Also, the probability that the person will choose one of the two response categories ('No, I can't do this task' or 'Yes, I can do this task') is governed by the difference between the his or her location on the mobility continuum and the item's location on the same continuum. Thus, when person n, whose location is B_n, answers RMI item i, whose location is D_i, the probability of responding 'yes' or 'no' is governed by the value $(B_n - D_i)$.

The general relationship between the probability of any person responding 'yes' to any item of the RMI, and the difference between that person's and the item's locations $(B_n - D_i)$ on the mobility variable can be shown graphically (*Figure 3a*). The resultant curve is S-shaped (also known as an ogive). Three things about this curve are worth noting:

1. The probability of responding 'yes' (y-axis) never reaches 0 or 1 – as there is always a chance, however small, of a person being unable to do a task. That is the nature of probabilities.
2. The difference between B_n and D_i on the mobility variable (x-axis) ranges from minus to plus infinity ($-\infty$ to $+\infty$), i.e. the range is unbounded.
3. The graph relating the probability of responding 'no' to any item and the difference $(B_n - D_i)$ on the mobility variable is the reciprocal curve, and is shown in *Figure 3b*.

When *Figure 3a* and *b* are combined to give *Figure 3c* we have the type of graph that enables us to predict, for any difference between person location and item location on the mobility variable, the probability of responding 'yes' and 'no'. From *Figure 3c* note that:

- The curves cross at probability = 0.5 (y-axis) and $B_n - D_i = 0$ (x-axis).
- The sum of the probabilities of responding 'yes' and 'no' at any point on the x-axis is 1.0.
- These graphs tell us neither how able an individual is nor how difficult an item is – they merely give the probabilities of responding 'yes' or 'no' governed by the difference between person mobility and item difficulty.

So far we have postulated a theory that the difference between the person's location and the item's location on the mobility continuum governs the probability of the outcome of any person's response to any item. This theoretical relationship for the 'yes' response is an ogive (see *Figure 3a*). If this theory can be articulated as a formula, we would have a mathematical model relating the probability of a 'yes' response to the difference between the person location and the item location on the shared mobility continuum. If such a formula existed we might be able to deduce the mobility of people and the difficulty of the RMI items from a matrix of responses of people to the RMI (how is explained later).

Essentially, we need to be able to relate bounded probabilities $(0 \leq p \leq 1)$ to unbounded differences between B_n and D_i $[-\infty \leq (B_n - D_i) \leq +\infty]$. For the probability of a response to be used to estimate a person's mobility and an item's difficulty both sides of the equation must have the same boundaries: either 0–1 or $-\infty$ to $+\infty$.

The value $(B_n - D_i)$ can be made to have the range of 0–1 by using a two-step transformation.

First, if $(B_n - D_i)$ is expressed using natural logarithms the expression will have limits of zero and infinity:

$$0 \leq e^{(B_n - D_i)} \leq \infty$$

Second, if this value $[e^{(B_n - D_i)}]$ is expressed as an odds ratio, the expression has the limits of zero and 1 and could be a formula for the probability of a correct response:

$$0 \leq \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} \leq 1$$

If we take this formula as an estimate of the probability of person n responding 'yes' to RMI item i the relationship is:

Probability of person n responding 'yes' to item i =

$$\frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

And, therefore:

Probability of person n responding 'no' to item i =

$$1 - \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

because:

Probability of responding 'yes' + Probability responding 'no' = 1

Thus, for completeness:

Probability of person n responding 'no' to item i =

1 – Probability of person 'n' responding 'no' to item 'i'

Probability of person n responding 'no' to item i =

$$1 - \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

Probability of person n responding 'no' to item i =

$$\frac{1 + e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} - \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

Probability of person n responding 'no' to item i =

$$\frac{1}{1 + e^{(B_n - D_i)}}$$

Thus, for dichotomously scored items a general equation can be written:

$$\text{Probability of person } n \text{ getting outcome } x \text{ on item } i = \frac{e^{x(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

where the outcome x can be either the response 'yes' ($x = 1$) or the response 'no' ($x = 0$), and $e^0 = 1$. This is the Rasch model for dichotomous items.

Note. The fact that there is an x in the numerator is important when it comes to items with more than two response options (polytomous) as x can be 0, 1, 2, etc.

Andrich[97] offers another way of deriving the Rasch model which exploits the relationships between odds and probabilities: the odds of a 'yes' response = probability of a 'yes' response/probability of a 'no' response or for short:

$$\frac{\text{Probability 'yes'}}{\text{Probability 'no'}}$$

But, as

$$\text{Probability 'yes'} + \text{Probability 'no'} = 1, \text{ then}$$

$$\text{Probability 'no'} = 1 - \text{Probability 'yes'}$$

Thus, the odds of a 'yes' response =

$$\frac{\text{Probability 'yes'}}{1 - \text{Probability 'yes'}}$$

From before:

$$\text{Probability 'yes'} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

$$\text{Probability 'no'} = \frac{1}{1 + e^{(B_n - D_i)}}$$

and thus also, the odds of a 'yes' response = $\dfrac{\dfrac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}}{\dfrac{1}{1 + e^{(B_n - D_i)}}}$

Thus, the odds of a 'yes' response also = $e^{(B_n - D_i)}$. Hence, the odds of a 'yes' response:

$$\frac{\text{Probability 'yes'}}{1 - \text{Probability 'yes'}} = e^{(B_n - D_i)}$$

And, by taking logs:

$$\text{Log of the odds of success} = \ln\left(\frac{\text{Probability 'yes'}}{\text{1–Probability 'yes'}}\right) = (B_n - D_i)$$

where ln = natural logarithm. Thus, estimates of B_n and D_i are additive, and in log-odds units (logits). Taking the logarithm of a ratio scale turns it into an interval scale. Thus, the units of the person and item estimates generated by the Rasch model are interval level.

Important properties of the Rasch model

The above theory about the relationship between probability and values is logical and sensible. The mathematics seems reasonable. The important question is how this relates to a set of data, such as patients' responses (either 0 or 1) to the items of the RMI. Before this can be explained, two fundamental properties of the Rasch model need to be demonstrated because they are required to understand the computation of item and person estimates. The first property, mentioned in the previous chapter, is that the parameters can be estimated separately.[97] The second property is that the total score is a sufficient statistic.[97] Both are explained below.

The item and person parameters can be estimated separately (parameter separation)

In Chapter 2, it was stated that a fundamental property of the Rasch model is that the item locations can be estimated independently of the distribution of the person's location, and the person locations can be estimated independently of the distribution of the item's location. This is significant as it is a requirement for Rasch's condition of invariance, and is demonstrated below. It is important to recognise that this is demonstrated mathematically. If the data fit the Rasch model then it becomes a conclusion that can be drawn from that data set as well. It is not a conclusion that arises just because the Rasch model is used in any data set.

Consider one person (n) responding to two RMI items (1 and 2). Consistent with our previous notation the person has location B_n, and the items have locations $D1$ and $D2$. Each item has two response categories, 'no' = 0 and 'yes' = 1. There are four possible response outcomes:

Response to item 1	Response to item 2
0 (No to item 1)	0 (No to item 2)
1 (Yes to item 1)	0 (No to item 2)
0 (No to item 1)	1 (Yes to item 2)
1 (Yes to item 1)	1 (Yes to item 2)

Only two of these four possible response outcomes are useful in terms of the relative information they offer about the difficulty of the items. These are the ones where the responses are different, i.e. 1,0 and 0,1. Response outcomes 0,0 and 1,1 do not have a distinction in outcomes and therefore no information in terms of the relative difficulties of the items. (This is akin to the ceiling and floor effect on any scale: we don't know how far above the ceiling or below the floor a person is, and we cannot determine relative differences between people.) We can now

focus on these two outcomes and replace the outcome with the probability of that outcome in terms of the Rasch model.

Probability of response to item 1	Probability of response to item 2
$p\,(x=1)=\dfrac{e^{(B_n-D_1)}}{1+e^{(B_n-D_1)}}$	$p\,(x=0)=\dfrac{1}{1+e^{(B_n-D_2)}}$
$p\,(x=0)=\dfrac{1}{1+e^{(B_n-D_1)}}$	$p\,(x=1)=\dfrac{e^{(B_n-D_2)}}{1+e^{(B_n-D_2)}}$

We can denote the expressions that are common to each item using:

$$Y=\frac{1}{1+e^{(B_n-D_1)}}$$

$$Z=\frac{1}{1+e^{(B_n-D_2)}}$$

Thus, the table above can be simplified to:

Probability of response to item 1	Probability of response to item 2
$p\,(x=1)=Y\,e^{(B_n-D_1)}$	$p\,(x=0\,)=Z$
$p\,(x=0)=Y$	$p\,(x=1)=Z\,e^{(B_n-D_2)}$

We can now compute the probability for each of the two outcomes (1,0 and 0,1) given that the outcome is either 1,0 or 0,1. Thus, this is a conditional probability.

First, the probability of outcome 1,0 given that the outcome is either 1,0 or 0,1:

$$\frac{Ye^{(B_n-D_1)}\times Z}{[Ye^{(B_n-D_1)}\times Z]+[Y\times Ze^{(B_n-D_2)}]}$$

which simplifies to:

$$\frac{e^{(B_n-D_1)}}{e^{(B_n-D_1)}+e^{(B_n-D_2)}}$$

from which the person parameter (B_n) is eliminated to give:

$$\frac{e^{-D_1}}{e^{-D_1}+e^{-D_2}}$$

This equation means that the probability of the first item receiving a 'yes' response, when only one of the two items receives a 'yes' response, depends only on the relative locations of the items and not on the location of the person.

Second, the probability of outcome 0,1 given that the outcome is either 1,0 or 0,1 is:

$$\frac{Y \times Ze^{(B_n - D_2)}}{[Ye^{(B_n - D_1)} \times Z] + [Y \times Ze^{(B_n - D_2)}]}$$

which simplifies to:

$$\frac{e^{(B_n - D_2)}}{e^{(B_n - D_1)} + e^{(B_n - D_2)}}$$

from which the person parameter (B_n) is eliminated to give:

$$\frac{e^{-D_2}}{e^{-D_1} + e^{-D_2}}$$

This equation means that the probability of the second item receiving a 'yes' response, when only one of the two items receives a 'yes' response, depends only on the relative locations of the items and not on the location of the person.

The same issue applies when two people with locations $B1$ and $B2$ answer the same item D_i.

There are four possible response outcomes:

Response of person 1	Response of person 2
0 (No by person 1)	0 (No by person 2)
1 (Yes by person 1)	0 (No by person 2)
0 (No by person 1)	1 (Yes by person 2)
1 (Yes by person 1)	1 (Yes by person 2)

We are interested in the two outcomes where the responses are different (1,0 and 0,1). The probabilities of these two outcomes are:

Probability of response of person 1	Probability of response of person 2
$p\,(x=1) = \frac{e^{(B_1 - D_i)}}{1 + e^{(B_1 - D_i)}}$	$p\,(x=0) = \frac{1}{1 + e^{(B_2 - D_i)}}$
$p\,(x=0) = \frac{1}{1 + e^{(B_1 - D_i)}}$	$p\,(x=1) = \frac{e^{(B_2 - D_i)}}{1 + e^{(B_2 - D_i)}}$

We denote the expressions that are common to each item using:

$$Q = \frac{1}{1 + e^{(B_1 - D_i)}}$$

$$R = \frac{1}{1 + e^{(B_2 - D_i)}}$$

Thus, the table above can be simplified to:

Probability of response of person 1	Probability of response of person 2
$p\ (x = 1) = Qe^{(B_1 - D_i)}$	$p\ (x = 1) = R$
$p\ (x = 1) = Q$	$p\ (x = 1) = Re^{(B_2 - D_i)}$

We can now compute the probability for each of the two outcomes (1,0 and 0,1) given that the outcome is either 1,0 or 0,1.

First, the probability of outcome 1,0 given that the outcome is either 1,0 or 0,1:

$$\frac{Qe^{(B_1 - D_i)} \times R}{[Qe^{(B_1 - D_i)} \times R] + [Q \times Re^{(B_2 - D_i)}]}$$

which simplifies to:

$$\frac{e^{(B_1 - D_i)}}{e^{(B_1 - D_i)} + e^{(B_2 - D_i)}}$$

from which the item parameter (D_i) is eliminated to give:

$$\frac{e^{B_1}}{e^{B_1} + e^{B_2}}$$

This equation means that the probability of person 1 responding 'yes' when only one of the people responds 'yes' depends only on the relative locations of the persons and not on the location of the item.

Second, the probability of outcome 0,1 given that the outcome is either 1,0 or 0,1 is:

$$\frac{Q \times Re^{(B_2 - D_i)}}{[Qe^{(B_1 - D_i)} \times R] + [Q \times Re^{(B_2 - D_i)}]}$$

which simplifies to:

$$\frac{e^{(B_2 - D_i)}}{e^{(B_1 - D_i)} + e^{(B_2 - D_i)}}$$

from which the item parameter (D_i) is eliminated to give:

$$\frac{e^{B_2}}{e^{B_1}+e^{B_2}}$$

This equation means that the probability of person 2 responding 'yes' when only one of the people responds 'yes' depends only on the relative locations of the persons and not on the location of the item.

The total score is a sufficient statistic

The second important property of the Rasch model is that the total score for a person is a sufficient statistic. What this means is explained below. This is a consequence of the model and directly of the ability to separate the parameters.

Reconsider the example above of one person responding to two items. There were four possible response outcomes. This time the total score has been added:

Response to item 1	Response to item 2	Total score
0	0	0
1	0	1
0	1	1
1	1	2

In the previous section, we noted that no information about the relative item locations was given by the two response outcomes where the items scores were the same (0,0 and 1,1). We also noted that the probability of the other two outcomes (1,0 and 0,1) depended only on the relative locations of the two items. It did not depend on the person location because this parameter was eliminated.

Thus, the probability of outcome 1,0 (given that the outcome is either 1,0 or 0,1) was:

$$\frac{e^{-D_1}}{e^{-D_1}+e^{-D_2}}$$

and the probability of outcome 0,1 (given that the outcome is either 1,0 or 0,1) was:

$$\frac{e^{-D_2}}{e^{-D_1}+e^{-D_2}}$$

The absence of the person's location parameter (B_n) from these two equations means that when the total score is 1 the *response pattern* contains no information about the person location estimate. Therefore, all the information needed to estimate the person location is contained in the total score. Thus, the person's total score, obtained by summing across his or her item scores, is the sufficient statistic for estimating the person location. No further information is contained in the response pattern (i.e. in the data) for estimating the person location, because it is all absorbed in the total score.

There is a symmetrical argument for the items. The total score for an item, computed by summing the scores across people, contains all the information for estimating the item location. No further information for estimating the item location is contained within the response pattern of the people to the items.

Thus, in the Rasch model the total score is a sufficient statistic. The importance of sufficient statistics was determined by Fisher,[98] and is discussed by Rasch and Andrich.[65,66,97]

The direct implication of this for the Rasch model is that all persons with the same total scores will get the same location estimate irrespective of response pattern. This does not mean that the response pattern is not important – it is. Differences in response patterns are reflected in the person fit statistics. In contrast, in Item Response Theory differences in response patterns are reflected in different location estimates.

It goes without saying that the parameter separation and sufficient statistics are inherent properties of the Rasch model. They are realised only in data when the data fit the Rasch model, within statistical reason.

Putting it all together: relating theory and mathematics to real data

The theory that the relative locations of an item and a person govern the probability of a response is logical and sensible. The mathematics articulates this theory and has some unique properties. However, how does this relate to a set of data? How can we use the simple matrix of patients' responses (0 or 1) to the 15 items of the RMI to generate estimates of item locations, person locations and probabilities of responses, and then go on to test the fit of the observed data to the expectations of the model?

In practice, this process is complex and requires the use of computer software. However, the general principles are relatively straightforward. Although it is best understood by considering a dichotomous scale, such as the RMI, where the item response options are either 0 or 1, the process generalises to items with more than two response options.[70,99]

We know from the Rasch model that the total score for each person, i.e. the sum of one person's responses to the RMI items, is the sufficient statistic for computing the person location. It contains all the information required to estimate the person location. Similarly, we know that the total score for each item, the sum of the responses of all people in the sample to one item, is the sufficient statistic for computing the item location. It contains all the information required to estimate the item location.

We also know from the Rasch model that the probability of a 'yes' response (i.e. a score of 1) to any RMI item is given by the formula:

$$\text{Probability of person } n \text{ responding 'yes' (score = 1) to item } i = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

Likewise, the probability of a 'no' response (i.e. a score of 0) to any RMI item is given by the formula:

$$\text{Probability of person } n \text{ responding 'no' (score = 0) to item } i = \frac{1}{1+e^{(B_n - D_i)}}$$

The matrix of 'yes' (score = 1) and 'no' (score = 0) responses to the RMI items allows us to compute the *observed proportion* of people who responded 'yes' and 'no' to each item. The observed proportion of people in a sample who respond 'yes' and 'no' to an item is an estimate of the *theoretical proportion* of people who respond 'yes' and 'no' to each item in the population we are studying. A theoretical proportion is a probability, because the definition of a probability of an outcome is the theoretical proportion of times that we expect that outcome to occur on a very large number of replications.

The next stage extrapolates Bernoulli's work (1700s) on probabilities and random variables. Essentially, if a coin is tossed multiple times the total number of heads is the sum of the theoretical average (probability) that a head will occur at each toss. Thus, conceptually, the sum of the theoretical averages (probabilities) of the number of times each item would be answered correctly should be equal to the number of items that are answered correctly. Thus, the total score for each person is the sum of the probabilities of getting each item correct and the total score for each item is the sum of the probabilities of each person getting that item correct. The probability of a person getting each item correct is given by the equation above.

Rasch analysis computer programs, such as RUMM2020, apply this logic in an iterative approach to generate estimates. Essentially, initial estimates of item and person locations are proposed. These are put into the equation above so that the probability of responding 'yes' can be computed for each item–person interaction (i.e. each cell in the data matrix). The probabilities are summed and this value is compared with the total score. Based on the differences between the total score and the sum of the probabilities, the initial estimates are refined. These refined estimates are put into the equation and new probabilities are computed, summed and compared with the total scores. This process continues until the sum of the probabilities is close enough to the total score. The computer program sets the criterion for 'close', typically within 0.001 unit difference.

This process of iteration results in best estimates of item and person locations given the total scores for each person and item, derived from the Rasch model. The process now works backwards. From these best estimates the computer program can determine the expected value for each item–person interaction. These expected values, computed from the Rasch model, are compared with the observed scores generated by people completing items. The difference between the observed scores and the expected values is determined and indicates the extent to which the observed data fit the predictions of the Rasch model. If the data fit the model, within statistical reason, then the properties of the model hold in the data, the total scores are sufficient statistics computing item and person estimates and the estimates generated are invariant (stable) linear estimates. If the data do not fit the model, the properties of the Rasch model do not hold in the data. This provokes an investigation of items and persons to determine why.

There are many ways of evaluating the fit of the data to the model. No one method is necessary or sufficient to summarise it. More details on tests of fit are discussed in the Chapter 4.

Summary

This chapter has attempted to demonstrate the theory and mathematics of the Rasch model for dichotomous variables. The theory on which the model is grounded is logical and relatively straightforward. The mathematics articulates the theory. However, it is the properties of the mathematical model that are important. The ability to estimate the item and person locations independently of each other is a feature only of Rasch models. It is not a feature of other item response models because the additional parameters prevent it mathematically. Massof[33] demonstrates this. By removing the influence of persons on item estimates, and items on person estimates, the Rasch model is able to deliver mathematically stable linear measurements of these parameters. This enables invariant comparisons of people and items to be made when the data fit the Rasch model. This is considered a requirement for fundamental measurement and is the reason why proponents of the Rasch model seek to fit data to that model, rather than try to explain observed data.

The Rasch model for items with more than two ordered response categories is a generalisation of the model for dichotomous response options.[70] The mathematics will not be discussed here. That job is best left to Andrich, who developed the model.[70,74,99] The rest of this monograph focuses on demonstrations of the application of the Rasch model to health measurement situations using the Rasch Unidimensional Measurement Models computer program.[100]

Chapter 4

The Rivermead Mobility Index

An evaluation using traditional and Rasch psychometric methods

Overview

The aim of this chapter is to illustrate the Rasch analysis of an existing scale, and to compare and contrast scale evaluation using traditional and Rasch psychometric methods. This chapter uses data from the CAMS study.[34] Specifically, we analyse data for the Rivermead Mobility Index (RMI). This scale has been chosen because each item has only two response categories ('yes' and 'no'); that is, it consists of dichotomously scored items. It is valuable to illustrate the Rasch model in a dichotomous scale before moving on to examine its use in scales with polytomous (more than two) response options. We do this in Chapter 5.

This chapter has five sections. First, the scale is presented. Second, the sample is described. Third, we report the results of a comprehensive evaluation of the RMI using traditional psychometric statistical methods. Fourth, we report a comprehensive evaluation of the RMI using Rasch analysis. Finally, we compare and contrast the information derived from the two analyses and the inferences made about the RMI.

The Rivermead Mobility Index

Appendix 1 shows the RMI. It is a rating scale purporting to measure mobility and has 15 items. Each item concerns a mobility-related task that has two response options: 'no' – I am unable to do this task; 'yes' – I am able to do this task. The RMI is scored by clinicians from patient interview and observation. Typically, item scores are summed to give a total score that ranges from 0 (all 'no' responses to items) to 15 (all 'yes' responses to items).

Now let us consider the aims of the RMI in more detail. Mobility is being thought of (conceptualised) as a quantitative variable in the sense that it reflects a property that can have a range of values from 'less' to 'more'. The RMI items attempt to map out this idea so that responses to its 15 items can be seen as indicators of the level of mobility. Essentially, the RMI seeks to map out the mobility as a line varying from less to more on which people can be located. Thus, the 15 RMI items make operational (operationalise) the idea of the mobility variable. Because mobility is observed through a variety of manifestations, rather than directly, it is considered to be a latent (hidden) property. The words 'trait' and 'construct' are typically used instead of the word 'property'. *Figure 1* represents graphically this conceptualisation of the mobility variable by the RMI, and the idea of measurement by locating a person on that line.

This conceptualisation implies that each item represents a mark on the 'ruler' of mobility mapped out by the RMI. More specifically, the mark defined by the item represents the transition point of the score from 0 to 1, i.e. the point at which a person moves from scoring '0' (unable to do) to '1' (able to do). The aim of a psychometric analysis is to determine, using a range of evidence, the extent to which this conceptualisation of mobility by the RMI has been achieved. In measurement terminology, the purpose of a psychometric analysis is to establish the extent to which a quantitative conceptualisation has been operationalised successfully.[3]

Sample

This illustration uses data from the CAMS study.[34] The CAMS study was designed to test the notion that cannabinoids have a beneficial effect on spasticity and other symptoms related to MS. It was a multicentre, randomised, placebo-controlled clinical trial at 33 UK neurology and rehabilitation centres.

A range of outcomes were measured. The primary outcome was the Ashworth scale, a clinician-rated measure of overall spasticity. Secondary outcome measures included the RMI, Barthel Index (BI), 30-item version of the General Health Questionnaire (GHQ) and the UK Neurological Disability Scale (UKNDS).

Each patient attended eight clinic visits over 15 weeks. There were two pre-treatment pre-randomisation visits, a 5-week dose titration phase (visits 3 and 4) and an 8-week plateau phase during which people remained on a stable dose (visits 5, 6, 7). Dose reduction was carried out during week 14, and people were medication free during week 15. The final assessment (visit 8) was conducted at the end of week 15.

A total of 667 people were enrolled. Of these, 630 were treated with oral cannabis extract (Cannador, $n = 211$), synthetic Δ^9-tetrahydrocannabinol (Marinol, $n = 206$) or placebo ($n = 213$). A total of 611 people were followed up for the primary end point. Treatment with cannabinoids did not have a significant effect on spasticity as assessed by the Ashworth scale.

In this illustration we used the RMI data as they have dichotomously scored items. We report the analysis of RMI data from visit 1 (including a total of $n = 667$ people). One person was determined to be ineligible before the RMI assessment was undertaken; hence, the data set consists of the responses for $n = 666$ people with MS.

Traditional psychometric evaluation of the RMI

Methods

Overview

There is no consensus as to how the results of traditional psychometric analyses should be reported. Hence they are documented in a variety of ways. We have previously recommended that a comprehensive scale evaluation using traditional psychometric methods should involve the examination of six psychometric properties: data quality, scaling assumptions, targeting, reliability, validity and responsiveness.[2] Data quality concerns the extent to which a scale can be administered successfully in the target sample.[10] Tests of scaling assumptions examine whether it is legitimate to sum item scores to generate a single scale score.[11] Targeting concerns the extent to which the distribution of disability in the sample matches the range of disability measured by the scale.[12] Reliability describes the extent to which scale scores are free from random error.[13] Validity refers to the extent to which a scale measures what it purports to measure.[13] Responsiveness is the ability of an instrument to detect accurately change when it has occurred.[14] These methods are fully documented elsewhere.[2,12,14–16,101,102]

Purists might consider these subheadings a little false, as some analyses provide evidence for more than one property, and there is overlap among properties. However, we have found it valuable to use these subheadings in order to help clinicians organise their thoughts about evaluating scales and scale performance.

In this chapter, the evaluation of the RMI focuses on within-scale analyses, i.e. the evaluations that can be undertaken on the responses from a single administration of the RMI to a sample of people. Thus, we did not examine test–retest reliability validity in terms of correlations with other scales or responsiveness. These issues are covered in other chapters [specifically, Chapter 6 describing the comparison of test–retest reproducibility methods based on the scales of the MSIS-29, Chapter 5 comparing psychometric evaluations of the MSIS-29 and Chapter 8 comparing responsiveness methods using the BI and Functional Independence Measure (FIM)]. The following analyses were undertaken.

Data quality

Indicators of data quality are the percentage of missing item responses and the percentage of the sample for whom total scores could be calculated.[11] It has been suggested that when responses are missing, a total score can be calculated if at least 50% of the items (i.e. $n \geq 8$ RMI items) have been completed. Under these circumstances each missing item is replaced with an imputed score, the patient-specific mean score, which is the mean score across completed items for that individual.[36]

Scaling assumptions

It has been proposed by others,[6,13] and it is generally accepted,[9] that a series of criteria should be satisfied for a set of items to be summed, legitimately, to form a single 'total' score. The RMI was tested against these criteria, which are:

1. Items should be roughly parallel, i.e. measure at the same point on the scale and have similar variance, otherwise they do not contribute equally to the variance of the total score and

should be standardised before combination.[11] A set of items is considered parallel when their item response option frequency distributions and their item mean scores and standard deviations are roughly similar.[6] When items have similar variances it has been agued that these do not need to be standardised before items are combined.[103]

2. Items should measure the same underlying construct (trait or property), otherwise it is not appropriate to combine them to generate a total score, i.e. the items of a scale should be internally consistent. A set of items is considered to be measuring the same construct when each item's corrected item–total correlation,[104] which is the correlation between each item and the total score computed from the remaining items in that scale, exceeds a recommended criterion. Three such correlations have been suggested: 0.20,[35] 0.30[105] and 0.40.[11]
3. Items in the scale should contain a similar proportion of information concerning the construct being measured. This criterion is considered satisfied when the corrected item–total correlations exceed 0.30.103 It is argued that when these are roughly equal there is no need to weight the items for differences in factor content before they are summed.

It should be noted that Likert's original test of internal consistency was to order the sample by their total score on a scale, take the subsample representing the top and bottom 10% (approximately), compute the mean item scores for those two subsamples and rank order the items by the magnitude of the difference in mean score.[6] The greater the difference between the mean score of the top and bottom 10%, the better the item. This approach was not used as it has been replaced by other methods of assessing internal consistency.

It should also be noted that others[106] have suggested an additional scaling assumptions criterion when evaluating a rating scale with multiple subscales measuring different constructs. Such scales measure multiple traits, so the test has become known as 'multitrait scaling'.[103] Examples of multitrait scales are the MSIS-29, which measures two constructs (physical and psychological functioning), and the SF-36, which measures eight constructs. Multitrait scaling tests evaluate convergent and discriminant validity,[106] which simply extends the concept of internal consistency by examining the relationships (correlations) between each item and all the subscales of an instrument. Clearly, items should correlate more highly with the total score of the subscale in which they are hypothesised to exist than with the total scores of the other subscales. That is, the item–own scale correlation should exceed the item–other scale correlations. Ideally, these differences should be significant. This item-level analysis follows the logic of the scale-level analyses' multitrait–multimethod approach to examining convergent and discriminant construct proposed by Campbell and Fiske.[57] As the RMI generates only one score, tests of item convergent and discriminant validity are not required and not possible.

Targeting

Targeting refers to the match between the distribution of problems in the sample and the range of problems measured by the scale. The better this match, the greater the potential for precise measurement. Targeting of the RMI to the CAMS sample at time 1 (T1) was evaluated by examining score distributions, skewness statistics and floor and ceiling effects. Floor effects are the percentage of patients scoring 15 (greatest impact on walking) and ceiling effects are the percentage of patients scoring 0 (least impact on walking). Two criteria have been proposed as upper limits for floor and ceiling effects: 15%[107] and 20%.[108]

Reliability

The reliability of a scale is defined as the extent to which its scores are associated with random error.[54] The most widely used estimate of reliability is Cronbach's coefficient alpha,[55] which determines the error associated with scores from the intercorrelations (internal consistency) among the items. There are good explanations of the theory behind Cronbach's alpha,[13] and some misconceptions and misunderstandings.[109] A range of minimum values has been suggested, the lowest being 0.50,[85] but it is generally accepted that Cronbach's alpha should exceed 0.70,[9] and preferably 0.80.[22]

One well-known limitation of Cronbach's alpha is that it is dependent on the number of items in a scale: the larger the number of items, the higher the alpha.[109] The implication of this is that the relationships among items might be 'hidden' by a high alpha when the number of items is relatively large. One way to address this problem is to also report the mean item–item correlation (the homogeneity coefficient[110]). As the RMI has 15 items, which is considered relatively large, reporting of the homogeneity coefficient is helpful.

Another useful reliability index is the standard error of measurement (SEM). It is computed as SEM = SD×√(1−reliability).[105] One value of the SEM is that it makes the reliability of a scale more tangible and clinically meaningful. It does this by enabling the computation of the 95% confidence intervals around individual person scores. This is computed as ±1.96 SEM.[105]

Likert's original method of testing reliability was the split-half method – split the scale's items into two halves, compute the correlation between the two and correct the value for the number of items using the Spearman–Brown prophecy formula (because reliability is related to scale length). This has now been largely replaced by Cronbach's alpha,[55] which can be considered as the mean of all split-half reliabilities,[13] at least under certain circumstances.[109]

Validity

The validity of a scale is defined as the extent to which it measures the construct(s) it purports to measure. As Fitzpatrick *et al*.[101] and others[111] have stated, this is far from simple.

Determining the validity of a scale involves bringing together pieces of evidence from various sources. First, there is non-statistical evidence, including the process of scale development and the extent to which the items form a clinically meaningful and conformable set.[111] Second, there is the empirical evidence such as within-scale analyses (termed by some as internal construct validity[89]), correlations with other scales (convergent and discriminate construct validity[56]), examination of group differences[112] and hypothesis testing.[15]

This example focuses on validity evidence generated by within-scale analyses. In traditional methods, a range of within-scale analyses has been used to provide evidence of validity. These include item–total correlations, alpha coefficients and homogeneity coefficients, which indicate the internal consistency of items and thus are taken to be evidence that they measure a common construct. Another within-scale analysis used to support validity is an exploratory factor analysis of an item set. The aim of the factor analysis is to identify clusters of items that intercorrelate but do not correlate with other clusters of items. Clusters of items identified by a factor analysis are potential candidates for scales and subscales. Support for the validity of the RMI would be provided by a factor analysis implying that the 15 items are best considered as a single cluster.

Results

Tables 3–6 show the results of these analyses on the RMI data from the CAMS study. *Table 5* summarises the results.

Data quality

There were no missing data and total scores could be computed for everyone. These findings imply good data quality.

Scaling assumptions

Table 3 shows the distribution of responses to each item given as a percentage of the total sample, item mean scores, item standard deviations (which indicate the variance) and corrected item–total correlations. The item mean scores (range 0.01–0.73) and variances (range 0.09–0.50) are in keeping with the differences in response distributions. However, as the only response options are 0 and 1, the means and variances for many of the 15 items appear to be similar.

Item–total correlations, corrected for overlap, range from 0.10 to 0.76. All items except running (item 15) have values that exceed the range of published requirements of 0.20,[35] 0.30[22] and 0.40.[11] In addition, these values satisfy the suggested criterion of 0.30[103] which indicates 'equivalence' of item–total correlations.

Table 4 shows the same data as *Table 3*, reordered in terms of increasing proportions of people responding 'no'. That is, the first item ('turning over in bed') is the item to which most people answered 'yes' (i.e. able to do). The last item ('running') is the one to which most people responded 'no' (i.e. unable to do). Put another way, the items are ordered in terms of increasing difficulty.

Targeting

Table 5 shows the distribution of RMI total scores. The proportion (percentage) of people responding across the different items covers the complete scale range of 0–15, and the sample mean score (6.3) is near the scale mid-point (7.5). Although there is a floor effect (11.9), this is below the recommended upper limit.

Reliability

Table 5 shows that the alpha is high (0.91), indicating good reliability. The mean item–item correlation (0.38) exceeds the recommended criterion of 0.30.[113] The confidence interval around the RMI score for any individual is ±2.53 points, indicating that the score lies somewhere within a

TABLE 3 *Rivermead Mobility Index (n = 666): scaling assumptions*

Item	Percentage responding 'no' (= 0)	Percentage responding 'yes' (= 1)	Mean score	SD	Corrected item–total correlation
1. Turning over in bed	27	73	0.73	0.44	0.57
2. Lying to sitting	32	68	0.68	0.47	0.69
3. Sitting balance	27	73	0.73	0.45	0.48
4. Sitting to standing	39	61	0.61	0.49	0.61
5. Standing supported	54	46	0.46	0.50	0.74
6. Transfer	34	66	0.66	0.48	0.71
7. Stairs	67	33	0.33	0.47	0.66
8. Walking inside, with aid if needed	46	54	0.55	0.50	0.76
9. Walking outside (even ground)	66	33	0.33	0.47	0.70
10. Walking inside, no aid	86	14	0.14	0.35	0.48
11. Picking off the floor	71	29	0.29	0.46	0.70
12. Walking outside (uneven ground)	86	14	0.14	0.35	0.50
13. Bathing	51	49	0.49	0.50	0.66
14. Up and down four steps	87	13	0.13	0.34	0.47
15. Running	99	01	0.01	0.09	0.10

SD, standard deviation.

TABLE 4 *Rivermead Mobility Index items reordered by mean score (n = 666)*

Item	Percentage responding 'no' (= 0)	Percentage responding 'yes' (= 1)	Mean score	SD	Corrected item–total correlation
1. Turning over in bed	27	73	0.73	0.44	0.57
3. Sitting balance	27	73	0.73	0.45	0.48
2. Lying to sitting	32	68	0.68	0.47	0.69
6. Transfer	34	66	0.66	0.48	0.71
4. Sitting to standing	39	61	0.61	0.49	0.61
8. Walking inside, with aid if needed	46	54	0.55	0.50	0.76
13. Bathing	51	49	0.49	0.50	0.66
5. Standing supported	54	46	0.46	0.50	0.74
7. Stairs	67	33	0.33	0.47	0.66
9. Walking outside (even ground)	66	33	0.33	0.47	0.70
11. Picking off the floor	71	29	0.29	0.46	0.70
10. Walking inside, no aid	86	14	0.14	0.35	0.48
12. Walking outside (uneven ground)	86	14	0.14	0.35	0.50
14. Up and down four steps	87	13	0.13	0.34	0.47
15. Running	99	01	0.01	0.09	0.10

SD, standard deviation.

TABLE 5 *Rivermead Mobility Index (RMI) summary of evaluation using traditional psychometric methods (n = 666)*

Psychometric property	Values
Scaling assumptions	
Item mean scores [mean (SD); range]	0.42 (0.44), 0.008–0.73
Item variances [mean (SD); (range)]	0.18, 0.008–0.25
Corrected item–total correlations (range)	0.10–0.76
Targeting	
RMI scale (mid-point; range)	7.5; 0–15
RMI observed scores [mean (SD); range]	6.3 (4.3); 0–15
Observed score (range)	0–15
Scale score (range)	0–15
Floor effect (% scoring 0)	11.9%
Ceiling effect (% scoring 15)	0.3%
Reliability	
Cronbach's alpha	0.91
Inter-item correlation (mean; range)	0.38 (0.01–0.75)
SEM [SD×(1−alpha)]	1.29
95% CI around individual person scores	±2.53
Validity (within-scale analyses)	
Corrected item–total correlations (range)	0.10–0.76
Cronbach's alpha	0.91
Inter-item correlation (mean; range)	0.38 (0.01–0.75)

CI, confidence interval; SD, standard deviation; SEM, standard error of measurement.

TABLE 6 *Principal components analysis (varimix rotation with Kaiser normalisation) of the Rivermead Mobility Index (n = 666)*

	Component matrix when one component solution requested	**Rotated component matrix when extraction set as eigenvalues > 1.0**		
Item		**1**	**2**	**3**
1. Turning over in bed	0.64	0.74	0.12	–0.03
3. Sitting balance	0.74	0.85	0.15	0.03
2. Lying to sitting	0.54	0.61	0.13	0.07
6. Transfer	0.67	0.69	0.24	–0.03
4. Sitting to standing	0.79	0.60	0.51	0.01
8. Walking inside, with aid if needed	0.76	0.83	0.20	0.03
13. Bathing	0.72	0.38	0.67	–0.14
5. Standing supported	0.81	0.64	0.50	–0.01
7. Stairs	0.76	0.40	0.70	0.08
9. Walking outside (even ground)	0.54	0.14	0.67	–0.05
11. Picking off the floor	0.76	0.36	0.74	–0.05
10. Walking inside, no aid	0.57	0.12	0.72	0.21
12. Walking outside (uneven ground)	0.71	0.64	0.36	0.04
14. Up and down four steps	0.53	0.07	0.72	0.10
15. Running	0.12	0.04	0.08	0.97

range of 5 points, which is larger than the standard deviation of scale scores (4.3).

Validity

The item–total correlations (except for item 15), alpha coefficient and homogeneity coefficient shown in *Table 5* provide evidence supporting the internal construct validity of the RMI.

Factor analysis (more correctly, principal components analysis – PCA) provides some support for the 15 RMI items as a statistically conformable set (*Table 6*). When a one-component solution was requested (varimax rotation), a total of 44.3% of the variance was explained. Component loadings range from 0.12 (item 15 – 'running') to 0.81 (item 8 – 'walking inside'), with 14 items having values above 0.53. Other recommended criteria to determine the number of components produce different results: there are three components with eigenvalues exceeding 1.0,[114] four components explaining more than 5% of the variance,[115] and the scree plot supports either a two- or three-component solution depending where the 'elbow' is judged to lie.[116]

Solutions with two, three and four components were examined. The two-component solution explained 56% of the variance and had six items that crossloaded > 0.30 (of which five items cross-loaded > 0.40) onto the other component and one item that did not load onto either component. The three-component solution explained 62% of the variance and also had six items that crossloaded > 0.30 (three crossloading > 0.40) and one component with only one item loading onto it. The four-component solution explained 67% of the variance and had seven items that crossloaded > 0.30 (four crossloading > 0.40), and two components with only one item loading onto each of them. Thus, from a statistical perspective the one-component model was the best. Cross-validation of these results using principal axis factoring produced similar findings.

Table 6 shows the component loadings for the one-component and three-component models.

Summary of traditional psychometric analysis of the RMI

Data quality was high, implying that there were no problems in using the RMI in this large sample. Scaling assumptions were partially satisfied. The 15 items had variable mean scores and standard deviations implying that they were not parallel. Nevertheless, corrected item–total correlations, with the exception of the 'running' item, exceeded 0.30, implying that 14 of the items measured a common underlying construct and satisfying the criterion for summation without weighting. Scale-to-sample targeting was good but the floor effect implied that a cohort of people with greater mobility problems were not measured well. Reliability was high. There was evidence to support the validity of the RMI from the within-scale analyses. Although the PCA suggested that the one-component solution was the most satisfactory from a statistical perspective, this solution explained only 44% of the total variance.

Thus, the RMI satisfies most traditional criteria for rigorous measurement. Some problems were demonstrated that suggest improvements could be made if the scale were considered for revision, specifically consideration of removing the running item and extension of the measurement range in the more disabled range.

Evaluation of the RMI using Rasch analysis

Methods

Overview

The purpose of a psychometric analysis, as stated succinctly by Andrich and Styles,[3] is to establish whether a quantitative conceptualisation has been operationalised successfully. In the case of the RMI, which is typical of many rating scales in that successive integers are assigned to the successive categories of the response options for each item (0 = no = can't do; 1 = yes = can do), item scores are summed to give a single value representing the amount of the trait. Thus, the role of a psychometric analysis is to determine if this process of summing is legitimate.

Different psychometric methods use a different range of evidence to achieve that goal. As we have seen, in traditional psychometric methods the range of evidence comes, predominantly, from correlation-based analyses. In a Rasch analysis, the range of evidence stems from a mathematical conceptualisation of the conditions of measurement that permit the summation of integer scores. Essentially, then, the observed data should, within reason, fit this model for measurement to be considered to have been achieved. In circumstances where the data fit the model, two fundamental inferences can be made. First, the measurement of the persons can be considered to be on a linear

scale. Second, these measurements are invariant across designated groups for which the fit has been confirmed.

Like all psychometric analyses, the Rasch analysis of an existing scale consists of gathering and integrating the evidence from a series of analyses. Typically, these analyses are not reported in the literature under the same subheadings that we and others have used for traditional methods. Although some have compared and contrasted the reliability and validity evidence from traditional and Rasch analyses,[117] it might be more clinically meaningful to build on the approach documented by Wright and Masters.[39] This is because a Rasch analysis gives us explicit and separate information about the scale and the sample. It seems to us clinically meaningful to think about scale and sample separately after first considering, in general terms, the suitability of the sample for evaluating the scale and the suitability of the scale for measuring the sample. With this in mind we recommend that consideration be given to reporting Rasch analyses by means of three main questions:

1. Is the scale-to-sample targeting adequate for making judgements about the performance of the scale and the measurement of people?
2. Has a measurement ruler been constructed successfully?
3. Have the people been measured successfully?

Is the scale-to-sample targeting adequate for making judgements about the performance of the scale and the measurement of people?

This is an important question that influences the interpretation of the results about the scale and the sample. A simple examination of the relative distributions of the item and person locations, their basic summary statistics and the power of the tests of fit (explained in the results) provides a frame of reference for interpreting the other results.

Has a measurement ruler been constructed successfully?

This question has five components:

Do the items map out a discernible line of increasing intensity?

Before anything can be measured, the variable along which measurements are to be made must be marked out. Rating scales such as the RMI define the variable they intend to measure using a set of items. Therefore, for the RMI to define a mobility variable along which measures can be interpreted, the items must be located at different points so that the direction and meaning of the variable can be identified. This question is answered by examining the item locations, their range, how they are spread, their proximity to each other and their precision (standard error).

Is the location of items along this line reasonable?

The location of items places them at points along a possible line. Thus, the ordering of item locations provides a description of the reach and hierarchy of the variable. If the ordering is consistent with clinical expectation, it provides evidence towards the construct validity of the variable. Departures from expectation require investigation and explanation, and can occur when the items are ambiguous, misleading or poorly worded.

Do the items work together to define a single variable?

The items of a scale must work together as a conformable set. Thus, examining the responses to each item for their consistency is important to determine if the items define a single continuum. More specifically, the responses to items should be in general agreement with the ordering of persons implied by the majority of items. When this is not the case, the validity of the items is suspect. These ideas are examined formally using fit statistics, i.e. fit of the data to the model. They are discussed in the results section.

Does the response to one item directly influence the response to another?

The response to one item should not bias the response to another. The technical term is that items should be locally *independent*. When items are locally *dependent,* measurement is artificially inflated (pushed up) or deflated (pushed down) relative to the true measurement (i.e. biased) depending on the nature of the dependency. In addition, reliability is artificially elevated.

The concept of local independence can be confusing because at one level the answer to one item is related to another – but this relationship is probabilistic; the fact that a person is unable to do a task makes it more probable that he or she will be unable to do another task relative to a person who is less disabled. Local independence is subtly different. A simple example of two locally dependent items is the following:

What is 2+1?

$A = 3$; $B = 5$

What is 15/(2+1)?

$A = 5$; $B = 3$

People who answer *A* to item 1 will answer *A* to item 2. Similarly, those answering *B* to item 1 will answer *B* to item 2. Hence these items are locally dependent.

Local dependence within health measures can be less obvious and the difference between items being locally dependent or probabilistically related is subtle, but important, and not necessarily immediately apparent. Hence it needs to be actively sought. For example, a potential instance of local dependence in a mobility questionnaire might be three 'yes'/'no' items: 'I can walk 10 metres'; 'I can walk 20 metres'; 'I can walk 30 metres'. A person who cannot walk 10 metres is likely to answer 'no' to the other two items; although we would expect some disagreement as the Rasch model is a probabilistic model. However, it might be that the answer to one item determined the answer to another. Such a problem is better addressed with a single item ('Can you walk?') with three response options (10 metres, 20 metres, 30 metres). Local independence can be looked for by examining the correlations among the residuals and by performing a subtest analysis. These approaches are illustrated using examples in the results section.

Are the locations of the items stable across clinically important groups?
When the rulers mapped out by rating scale items are stable, the measurements generated by them can be used to make meaningful comparisons. Thus, we need our items to perform similarly across important groups that we might wish to study and compare (e.g. men and women, different levels of disability, different types of MS and groups undergoing different treatments in a clinical trial).

When items do not perform similarly across important groups, the technical term is that they demonstrate differential item functioning (DIF), the measurement ruler is not stable across circumstances and measurement is affected to an unknown degree.

DIF can, and should, be examined in clinically important groups. The concept of DIF and methods used to examine it are discussed in Chapter 6.

Have the people in the sample been measured successfully?

This question has three components.

Are the persons in the sample separated along the line defined by the items?
It is important to examine the extent to which a scale detects differences between people in the sample under study, as that is often our aim in measurement. In Rasch analysis, the separation of people can be examined and quantified in terms of the Person Separation Index (PSI).

Do individual placements on the variable make sense?
This constitutes a check on the reasonableness of the measurements made.[39] This can be achieved by means of a comparison with measurements made using other scales, examination of group differences and hypothesis testing. In addition, for example, therapists can be asked to order people by their clinical evaluation of level of mobility. If that ordering of the persons is supported by the ordering of their measurements, there is evidence to support their validity.

How valid is each person's measurement?
When we measure a person we want to verify that the individual, or the person who has measured that individual, has used the items in the way expected, i.e. consistent with the idea that the items map out a variable along which the items have a unique order. This can be determined by examining the extent to which the responses for an individual person are in general agreement with the ordering of items implied by the majority of persons. If they do not agree, the validity of that person's measurement is questionable.

Results

Rasch analyses of the RMI data were undertaken using the software program Rasch Unidimensional Measurement Models (RUMM2020).[100] Results are reported and interpreted for each of the questions identified in the methods section, then interactively at the end.

In a Rasch analysis, people at the extremes of the scale range, those at the floor and ceiling, are excluded from the estimation of item statistics as they offer no comparison across the items to facilitate the examination of relative item difficulty. This is because people at the extremes of the scale range achieve the same score on all the items of

the scale. In the CAMS sample there were $n = 81$ people at the extremes; thus, the item statistics were computed from $n = 585$.

A similar logic applies to the estimation of person locations. People at the extremes of the scale range are problematic because it is impossible to estimate their location accurately. Essentially, we do not know how far above the ceiling, or below the floor, of the scale they are. However, people at the floor and ceiling are given person location estimates in a Rasch analysis as these people are an important part of the sample and need to be included to provide a complete picture of the range of scores obtained. These estimates are achieved by extrapolation. If many extreme scores are present, this will influence the variability across the range of the test and can influence the targeting.

Is the scale-to-sample targeting adequate for making judgements about the performance of the scale and the measurement of people?

Figure 4 shows the targeting of the patient sample (top histogram) to the items (bottom histogram). The scale range has been set from −8 to +8 units (logits) for symmetry. The histogram bars represent the relative location of the item(s) and people on the same variable. At this stage we are simply trying to answer a specific question so we can consider this graph relatively superficially and return to it later.

The most obvious finding is that the item locations are covered by the people, but the person locations are not covered by the items. Thus, we could infer that this is a reasonable sample to examine the scale but a suboptimal scale for measuring the sample. Hence, we would expect the scale to provide limited information about people at the extremes of the sample distribution. This is formalised by examining the information function for the scale, which is the inverse of the standard error associated with a measurement provided by the scale, at every location on the continuum: the higher the information value, the less the standard error and the greater the precision of measurement at that location. The greater the precision, the greater the information imparted by a measurement at that location. This is represented by the curve on the graph. Essentially, it informs us where on the continuum the scale performs best. This shows us that the performance of the scale is better in the centre (range −3 to +1.5 logits) and worse at the extremes. *Figure 4* demonstrates that many people in this sample are located outside the best functioning of the scale.

RUMM2020 also evaluates the 'power' of the tests of fit as 'excellent', 'good', 'reasonable', 'low' or 'too low'. For the RMI, this is considered excellent. This evaluation refers to the *power* in detecting the extent to which the data do not fit the model.[118] It does not imply that the fit of the data to the model is excellent. The power in detecting the extent to which the data do not fit the model is influenced by the targeting of the sample to the items, the PSI (explained later) and the variability of the sample. It is clearly important to interpret the fit statistics in the light of the power of the tests of fit.

The power of the tests of fit is intimately related to the PSI. If the PSI is low due to limited sample variability, i.e. people have similar locations and are not spread across the continuum, the power of the tests of fit is low. This is because lack of variability in people's locations makes it impossible to determine whether people with higher locations tend to get higher scores on items (the nubbins of the Rasch model). Under these circumstances, the fit statistics may appear to be very good but the power of the tests of fit will be poor.

Has a measurement ruler been constructed successfully?

Do the items map out a discernible line of increasing intensity?

Table 7 shows the 15 RMI items ordered by their location (also called calibration), in ascending order (from the most negative to the most positive). They range from about −3 to +6 logits, which is a wide range. Thus, the items define a line of increasing intensity, a continuum, rather than just a point.

Items have both negative and positive values because the mean item location is always set to 0 to give an arbitrary origin. It is used as a constraint because Rasch analysis estimates the locations of the items and people relative to each other, and not their absolute locations. That is, in the case of any two items, only the difference in their locations, not an independent value for each item, can be estimated. Thus, to give the items and people absolute values a constraint is required – the convention is that the mean of the item locations is set to 0. This explains why Rasch (and IRT) analyses typically have the curious finding of values ranging from negative to positive.

Table 7 also shows the standard error associated with each item location estimate. These vary; specifically, standard errors are smallest at the centre and largest at the extremes. They are also influenced by the size of the study sample, and

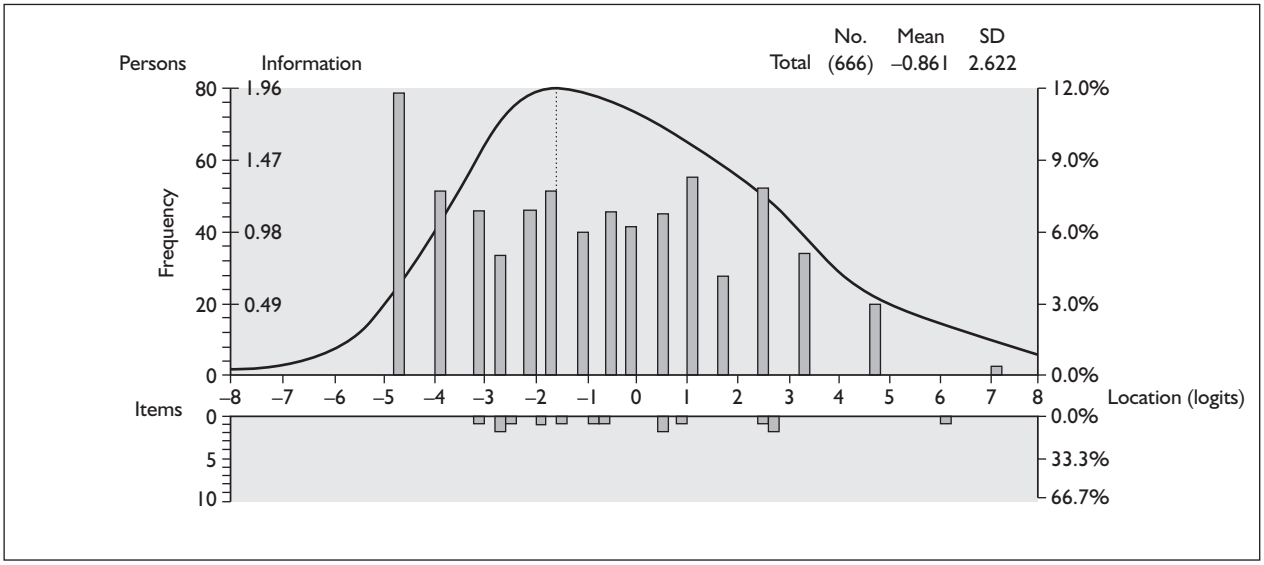


FIGURE 4 *Targeting of the patient sample (top) to the items (bottom). Person–item threshold distribution (grouping set to interval length of 0.20 making 80 groups).*

the targeting of the scale to the sample. This is because the formula for the standard error (SE) is given as [SE = 1/√sum of $p(1-p)$ for each person in the sample]. At the heart of this equation is $1/p(1-p)$. This is smallest when $p = 0.5$, i.e. when the proportion of people responding 'yes' = proportion of people answering 'no' = 50%. As p moves away from 0.5, either down towards 0 or up towards unity, the value of $1/p(1-p)$ gets larger.

Figure 5 complements *Table 7* by showing the item locations graphically. It shows that the 15 RMI items are not evenly spread. They are bunched in two ways. First, items are bunched such that at three places on the continuum multiple items have the same location. These are marked with thin arrows. *Table 7* confirms the items with the very similar locations: 'sitting balance' (−2.781) and 'lying to sitting' (−2.707); 'walking outside on even

TABLE 7 *Item locations in ascending order (n = 585; 666 with 81 extremes excluded)*

RMI item	Item location	Location SE
1. Turning over in bed	–3.032	0.130
3. Sitting balance	–2.781	0.126
2. Lying to sitting	–2.707	0.125
6. Transfer	–2.423	0.122
4. Sitting to standing	–1.863	0.117
8. Walking inside, with aid if needed	–1.568	0.116
13. Bathing	–0.872	0.115
5. Standing supported	–0.619	0.115
9. Walking outside (even ground)	0.444	0.119
7. Stairs	0.515	0.119
11. Picking off the floor	0.940	0.122
10. Walking inside, no aid	2.566	0.144
12. Walking outside (uneven ground)	2.641	0.146
14. Up and down four steps	2.715	0.148
15. Running	6.042	0.356

RMI, Rivermead Mobility Index; SE = standard error.

ground' (+0.444) and 'stairs' (+0.515); and 'walking outside on uneven ground' (+2.641) and 'up and down four steps' (+2.715). Items with similar locations raise the possibility of one of the items being redundant.

Items are also bunched such that there are notable gaps in the continuum mapped out by the items between +3 and +6 units, +1 and + 2.5 units, and −0.5 and +0.5 units. These are marked with thick arrows. Gaps imply limited measurement at those areas on the continuum they attempt to map out, despite the fact that they spread over a reasonable range (−3 to +6 units).

Figure 5 provides a clear representation of the measurement ruler mapped out by the 15 RMI items. Its adequacy and limitations are explicit. This figure acts as an evidence base for improvement of the RMI. For example, it would benefit from including items that are more difficult than going up and down four steps (location +2.715) but not as difficult as running 10 metres (location +6.042). It is clear to see how such a figure could be invaluable during scale construction.

Is the location of items along this line reasonable?

The continuum mapped out by the items warrants some explanation. On this continuum, items to the left indicate those items that are 'easy' mobility tasks to do whilst items to the right indicate those that are 'hard' mobility tasks to do. This is because high scores on the RMI indicate people who answer more 'yes' more times (i.e. can do more tasks) and thus have relatively lower disability. In contrast, low RMI scores (more 'no' responses) indicate higher disability. Thus, as people become progressively more disabled they move along the disability continuum from right to left, and if people improve they move from left to right.

In contrast, a high score for an item indicates more people who answer 'yes' (can do), and thus means the item is relatively easy. A low score for an item indicates more people who answer 'no' and a difficult item. Thus, the continuum for items runs from the easiest task to do on the left ('turning over in bed') to the most difficult task to do on the right ('running'). This makes sense: as people become progressively more disabled their RMI score falls and they move from right to left on this continuum, and as they become disabled the first task they find difficult is the most difficult ('running'). If the scoring had been the other way round (0 = can do; 1 = can't do), or for scales on which high scores indicate greater disability (e.g. MSIS-29), the direction of the continuum would be reversed.

The ordering of the 15 RMI items is clinically reasonable. Of the 15 items, 'running' (location +6.042) is predicted to be the most difficult and 'turning over in bed' (location −3.032) the easiest. Also, for example, the relative ordering of items as 'walking inside with an aid' (location −1.568) before 'walking inside without an aid' (location +2.566),

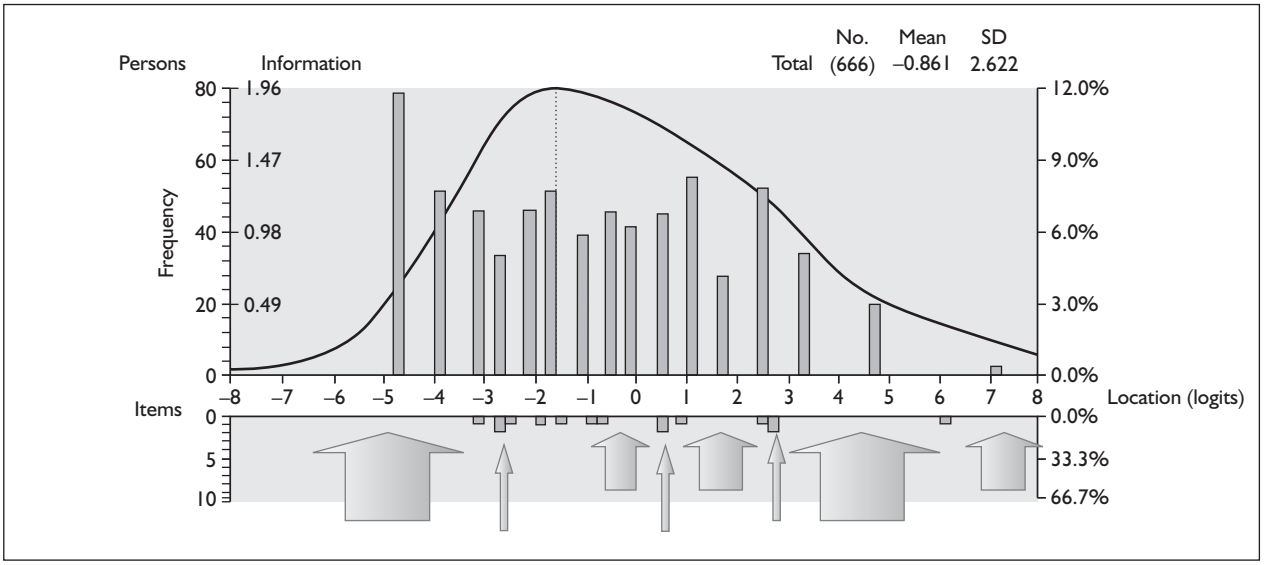


FIGURE 5 *Targeting of the patient sample providing a representation of the measurement ruler mapped out by the 15 RMI items. Person–item threshold distribution (grouping set to interval length of 0.20, making 80 groups).*

and 'walking outside (even ground)' (location +0.444) before 'walking outside (uneven ground)' (location +2.641), makes clinical sense.

Do the items work together to define a single variable?

A central theme of a Rasch analysis is an examination of the 'fit' of the observed data to the expectations of the mathematical model. By fit we mean the extent to which the observed responses of persons to items are predicted by, or recovered from, the mathematical model.

To recap, the Rasch model is a mathematical expression of the requirements that rating scale data must meet to generate values that can be considered measurements. This means that the Rasch model defines the relationships between item locations and person locations that should exist if a set of items is to deliver on the key tenets of measurement: invariance, unidimensionality and interval-level estimates.

What happens in practice? The Rasch model uses the total score for people (achieved by summing the scores of the items each person responds to) and the total scores for items (achieved by summing the scores of the persons who responded to each item) to derive best estimates of person mobility and item difficulty. As we have seen, this is legitimate, mathematically, because the total score can be proven to be the sufficient statistic for estimating person mobility and item difficulty. Having derived these estimates, the analysis now works backwards. It uses these estimates to derive the predicted responses that should have occurred for the items and persons to satisfy the Rasch model. All that remains is to compare the observed responses of the patients to the RMI items with the predicted responses derived from the model. Fit indicators essentially summarise the extent of the difference between the observed and expected responses.

In the Rasch paradigm, there is no one indicator of fit of the observed data to the mathematical model that is necessary and sufficient to summarise fit. There are multiple methods. Each addresses a different aspect of fit, or fit from a different perspective. For this reason, all available fit indicators should be examined; but their interpretation should be simultaneous and interactive rather than singular and in isolation of each other. However, in presenting the data for the RMI below it is necessary to consider each indicator separately before interpreting them interactively.

For dichotomously scored items RUMM2020 provides two numerical (fit residual and chi-squared) and one graphical (item characteristic curve, ICC) indicators of fit.

Fit residual

The fit residual (sometimes called the log residual) evaluates the fit of the observed data to the Rasch model from the perspective of the items.[118] For each item, the fit residual summarises the interaction between that item and all the persons for whom there is a response to that item. More specifically, the fit residual for an item is a summary of the differences between observed and expected values from each and every response to it (item–person interaction).

For every item–person interaction there is an observed score (0 or 1 for the RMI) and an expected score derived from the Rasch model (any value between 0 and 1). The difference between the observed and expected scores is called a residual (observed – expected = residual). For each item of the RMI, the residuals from the interactions with each of the $n = 585$ people in the sample are squared, summed and transformed to give a summary value (the fit residual) with possible range of $-\infty$ to $+\infty$.

Fit residuals are standardised to approximate a standard normal deviate. This means that the fit residuals are expected to be normally distributed with mean = 0 and SD = 1. This is based on the hypothesis that if the data fit the model the deviations (residuals) between the observed responses and the model-derived expected values should be no more than random errors. Therefore, if the data fit the Rasch model the mean fit residual, across all the items should be close to 0, the SD close to 1, and the individual values for the items should be distributed in the approximate range −2.5 to +2.5 (more specifically, 99.5% of values in the range −2.5 to +2.5, and 99.9% in the range −3.0 to +3.0).

For individual items, an observed fit residual of 0 indicates no difference between observed and predicted scores (perfect fit). The greater the departure from 0 (regardless of accompanying + or – sign), the greater the discrepancy between observed and predicted responses, and thus the greater the misfit of the observed data to the model. The sign (+ or –) associated with the fit residual value alludes to the type of misfit. Negative values indicate overdiscriminating items relative to the model, while positive values indicate

underdiscriminating values relative to the model. This will be explained in more detail later.

Table 8 shows the fit residuals for the 15 RMI items in ascending order. The mean item fit residual is −0.8, the SD 1.7 and the range −3.775 to +2.188. This indicates there is some misfit of the observed data to the Rasch model that needs to be explored and explained. Four items (11, 5, 9 and 8) have fit residuals outside the range of −2.5 to +2.5, indicating that the observed responses to these four items are not consistent with those predicted by the Rasch model. It is noteworthy that fit residuals for three of these four items (5, 9 and 8) lie within the range −3.0 to +3.0 and thus only a small amount outside the recommended range. Hence, only item 11 ('picking off the floor') fails notably this one criterion of fit. The implications of this will be explored later.

Chi-squared value and its probability

The chi-squared value is an indicator of the interaction between the individual item and the trait measured by the set of items[118] (here mobility measured by the 15 RMI items). It is much easier to understand the chi-squared test of fit by seeing both the numerical value and a visual representation. This will be done by referring to *Table 9a*, which shows the 15 RMI items ordered by increasing chi-squared value, and *Figure 6*, which is the item characteristic curve (ICC) (explained below) for item 1, whose chi-squared value in *Table 9a* is 4.557.

The chi-squared value in *Table 9a* is a summary statistic. It is computed by summing the chi-squared values for a series of class intervals. A series of class intervals is achieved by dividing the sample into a number of similarly sized groups based on their level of disability. For each class interval, the mean location of the people and their mean score on each of the 15 items are computed. Then, for each of the 15 items, the mean observed scores for the class intervals are compared with the scores for those items predicted by the Rasch model at the mean location of the class interval. The chi-squared value for each item is the sum of the chi-squared values computed for each of the six class intervals. The associated chi-squared probability is the probability that the discrepancy between the observed mean and the expected value is large relative to chance. If the chi-squared value is significant ($p < 0.05$ or 0.01) the item should be examined.

Figure 6 shows this process graphically. The S-shaped curve is the ICC for item 1. This indicates the *expected* score (*y*-axis) on item 1 for each possible location on the mobility continuum (*x*-axis). Note that the expected score can be any value between 0 and 1, unlike the observed responses, which can only be 0 or 1. The six small vertical marks on the *x*-axis indicate the mean person locations for each of the six class intervals. The six black dots indicate the *observed* mean score on item 1 for each of these six class intervals. The chi-squared values in *Table 9a* (4.577 for item 1) summarise the coherence between the observed responses (black dots) and expected responses (ICC) at the six points on the continuum. They are computed by summing the component chi-squared values for each class interval (*Table 9b*).

It is important to note that the number of class intervals can be altered. The RUMM program will select an appropriate number of class intervals (possible range 2–10) based on the size of the study sample. The analyst can override this default. Altering the number of class intervals will alter the chi-squared values, but is less likely to alter the associated probability and inferences. We find it informative to examine the implications of changing the number of class intervals.

A number of facts need to be considered when interpreting the chi-squared values. Chi-squared values increase with sample size. Hence, the developers of RUMM2020 recommend that when using large samples the sample size is amended to $n = 500$ to compute the values.[118] Chi-squared values are affected by the number of class intervals chosen, as discussed above. Chi-squared values only approximate a chi-squared statistic, and are inflated when the estimated probabilities are close to 0 or 1.

For these reasons, Andrich *et al*.[118] suggest that it is best to use the chi-squared statistic as an order statistic (i.e. order of degree of misfit) to see which items show much greater values than others, and to examine the ICC (see below). Consider *Table 9a* in which the items are ordered by ascending chi-squared values. They range from 3.971 (item 14, 'up and down four steps') to 51.566 (item 3, 'sitting balance'). Andrich *et al*. recommend examination of the values and how they change sequentially across items. For the RMI, there is a gradual increase over the first 10 items from item 14 ($\chi^2 = 3.971$) up to item 4 ($\chi^2 = 17.230$), then there is a larger step (item 11 = 25.877), followed by a fairly gradual increase for the next three items ($\chi^2 = 28.581$–31.761) before a large step to the final item (item 3, $\chi^2 = 51.566$). The chi-squared probabilities are significant at the 0.01 level for the last six items (4,

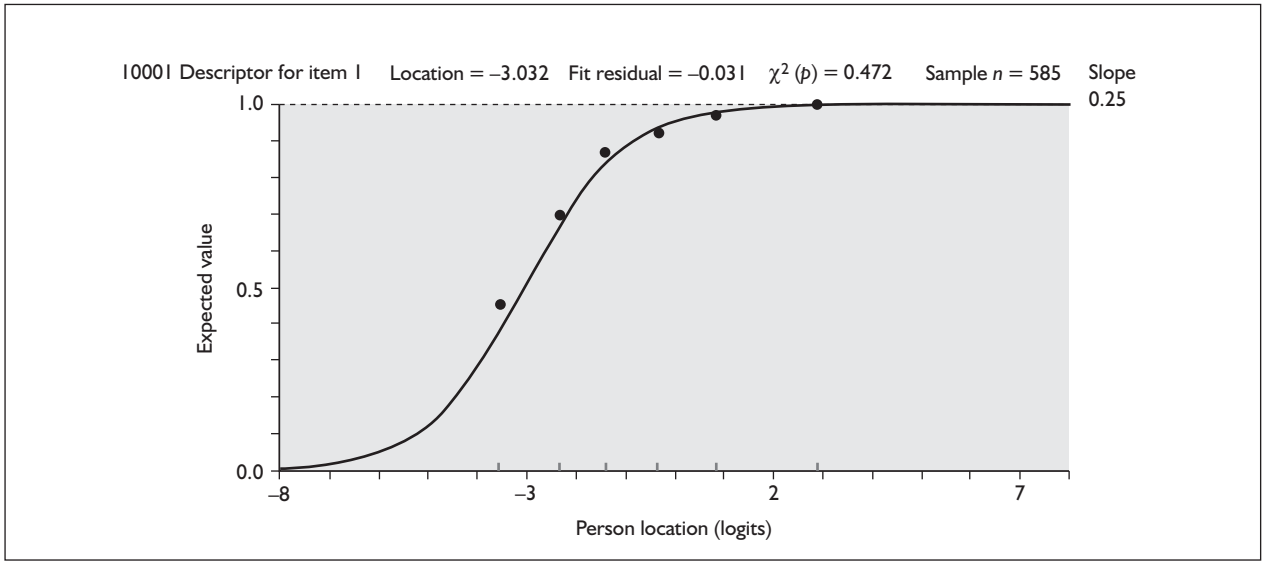


FIGURE 6 *Item characteristic curve for item 1 including class intervals.*

11, 8, 6, 2, 3) in *Table 9a*. Of these items, item 3 is notably the most wayward.

Item characteristic curve

The third indicator of observed data-to-Rasch model fit generated by RUMM2020 is the ICC. This is a graphical rather than a numerical indicator of fit and it aids the interpretation of the two fit statistics detailed above.

The ICC is a graph for an individual item. It plots the expected response (predicted from the model) to an item at each and every level of the measurement continuum. Clearly, the only available responses are 0 (no) and 1 (yes). However, the potential responses are continuous and vary from 0 to 1.

TABLE 8 *Items in fit residual order (n = 585; 666 with 81 extremes excluded)*

RMI item	Item location	Location SE	Fit residual
11. Picking off the floor	0.940	0.122	–3.775
5. Standing supported	–0.619	0.115	–2.850
9. Walking outside (even ground)	0.444	0.119	–2.750
8. Walking inside, with aid if needed	–1.568	0.116	–2.729
6. Transfer	–2.423	0.122	–1.514
7. Stairs	0.515	0.119	–1.143
2. Lying to sitting	–2.707	0.125	–0.803
14. Up and down four steps	2.715	0.148	–0.759
12. Walking outside (uneven ground)	2.641	0.146	–0.503
15. Running	6.042	0.356	–0.202
1. Turning over in bed	–3.032	0.130	–0.031
10. Walking inside, no aid	2.566	0.144	0.088
13. Bathing	–0.872	0.115	1.217
3. Sitting balance	–2.781	0.126	1.626
4. Sitting to standing	–1.863	0.117	2.188

RMI, Rivermead Mobility Index; SE, standard error.

TABLE 9a *Rivermead Mobility Index (RMI) items in chi-squared probability order (*n = *585; 666 with 81 extremes excluded; six class intervals)*

RMI item	Item location	Location SE	Fit residual	χ^2 value	χ^2 probability
14. Up and down four steps	2.715	0.148	–0.759	3.971	0.5536
1. Turning over in bed	–3.032	0.130	–0.031	4.557	0.4722
12. Walking outside (uneven ground)	2.641	0.146	–0.503	4.760	0.4459
10. Walking inside, no aid	2.566	0.144	0.088	7.322	0.1978
13. Bathing	–0.872	0.115	1.217	8.770	0.1186
15. Running	6.042	0.356	–0.202	10.973	0.0519
9. Walking outside (even ground)	0.444	0.119	–2.750	11.453	0.0431
7. Stairs	0.515	0.119	–1.143	11.585	0.0409
5. Standing supported	–0.619	0.115	–2.850	13.742	0.0173
4. Sitting to standing	–1.863	0.117	2.188	17.230	0.0041
11. Picking off the floor	0.940	0.122	–3.775	25.877	0.0001
8. Walking inside, with aid if needed	–1.568	0.116	–2.729	28.581	0.0000
6. Transfer	–2.423	0.122	–1.514	29.498	0.0000
2. Lying to sitting	–2.707	0.125	–0.803	31.761	0.0000
3. Sitting balance	–2.781	0.126	1.626	51.566	0.0000

SE = standard error.

TABLE 9b *Item 1: component chi-squared values for six class intervals*

Class interval		Location		Component			Category responses	
No.	Size	Max	Mean	Residual	χ^2		0	1
1	97	–3.170	–3.544	1.555	2.148	Obs. *p*	0.55	0.45
						Est. *p*	0.63	0.37
[OM = 0.45; EV = 0.38; OM–EV = 0.08; ES = 0.16]						Obs. *t*		0.45
2	79	–2.097	–2.310	0.486	0.237	Obs. *p*	0.30	0.70
						Est. *p*	0.33	0.67
[OM = 0.70; EV = 0.67; OM–EV = 0.03; ES = 0.05]						Obs. *t*		0.70
3	90	–1.105	–1.387	0.804	0.646	Obs. *p*	0.13	0.87
						Est. *p*	0.16	0.84
[OM = 0.87; EV = 0.84; OM–EV = 0.03; ES = 0.08]						Obs. *t*		0.87
4	86	–0.053	–0.334	–0.620	0.385	Obs. *p*	0.08	0.92
						Est. *p*	0.06	0.94
[OM = 0.92; EV = 0.94; OM–EV = –0.02; ES = –0.07]						Obs. *t*		0.92
5	100	1.119	0.846	–0.615	0.378	Obs. *p*	0.03	0.97
						Est. *p*	0.02	0.98
[OM = 0.97; EV = 0.98; OM–EV = –0.01; ES = –0.06]						Obs. *t*		0.97
6	133	4.716	2.896	0.703	0.494	Obs. *p*	0.00	1.00
						Est. *p*	0.00	1.00
[OM = 1.00; EV = 1.00; OM–EV = 0.00; ES = 0.06]						Obs. *t*		1.00

ES, expected score; EV, expected value; OM observed mean.
Item: $\chi^2 = 4.557$ (df = 9; $p = 0.472249$).

Figure 7 shows the ICC for item 1 ('turning over in bed'). The ICC plots the expected values for the item (predicted by the Rasch model) on the *y*-axis against the location on the RMI-measured mobility continuum. Essentially, as a person moves from left to right on the *x*-axis, i.e. from more disabled (low RMI score) to less disabled (high RMI score), his or her expected value on item 1 (in fact any of the items) increases.

When the class intervals are added to the plot they appear as black dots (see *Figure 6*). In this example we have chosen six class intervals. Each dot represents the intersection between the item mean score for the people in the class interval (*y*-axis) and the mean person location on the mobility continuum for the class interval (*x*-axis). The closer the dots follow the ICC, the better the fit of the observed data to the predictions of the Rasch model. As we have discussed, the chi-squared statistics in *Tables 9a* and *9b* give the numerical values.

For item 1 the dots representing the six class intervals follow the ICC very well. Hence the chi-squared values are small. However, it is noteworthy that in *Figure 6* there is only one class interval whose mean is in the lower half of the continuum covered by the item. This situation is not altered substantially by increasing the numbers of class intervals. Thus, this sample is not a stringent test of item 1's performance across the range of the continuum. The best test of an item is a sample that has class intervals across the item's range (e.g. item 5; *Figure 8*).

The importance of the graphical indicator of fit is that it gives perspective to the numerical values in *Tables 8* and *9a*. Chi-squared values and their associated probabilities can appear daunting and worrying when considered in isolation of the graph, which visualises the extent to which each class interval departs from expectation. However, this plot is rich with other information. It shows the suitability of the sample for examining the items. For example, the statistics may be excellent because the sample is poorly targeted to the item (*Figure 9*; the ICC for item 15).

The plot of the ICC and class intervals also helps to interpret the + or – sign associated with fit residual values. Positive fit residuals occur when the slope of the line from the class intervals is flatter than the predicted ICC – this is interpreted as underdiscriminating (*Figure 10*; ICC with class interval for item 3). When the slope of the line from the class intervals is steeper than the predicted ICC, this means that the item is overdiscriminating (*Figure 11*; ICC with class interval for item 7).

General comments about fit indicators

In addition to the specific comments associated with each fit indicator the interpretation of fit statistics needs to consider a number of facts:

1. The predictions of a mathematical model are predictions of perfection. Observed data are associated with limitations of various origins. Therefore, misfit is to be expected. The key issue is what does it mean and does it matter?

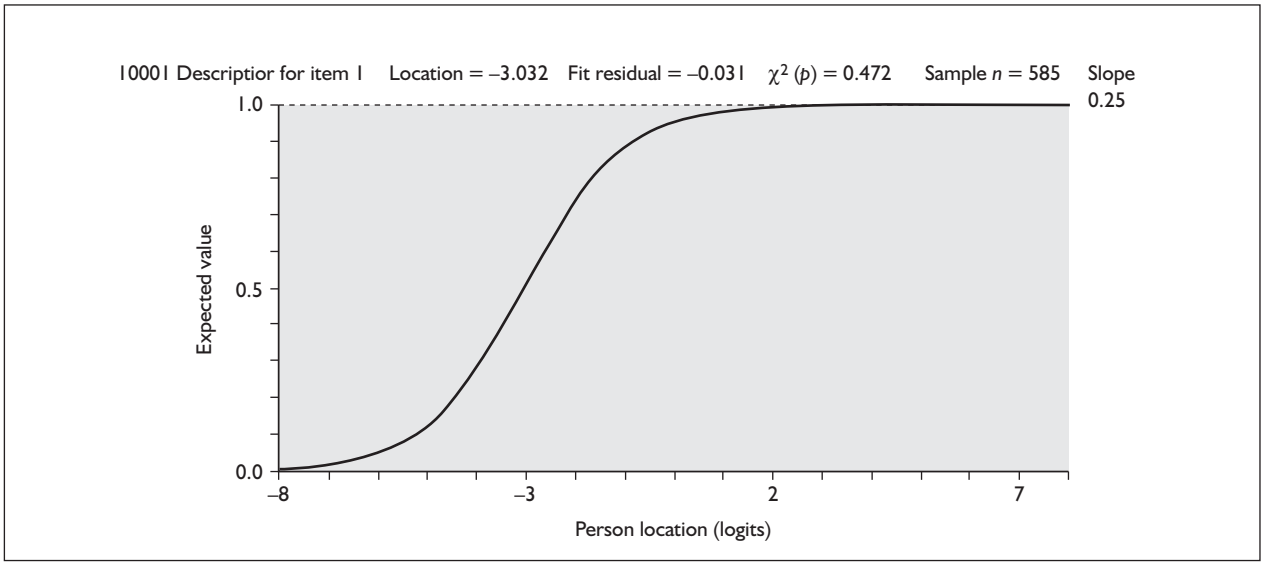


FIGURE 7 *Item characteristic curve for item 1 excluding class intervals.*

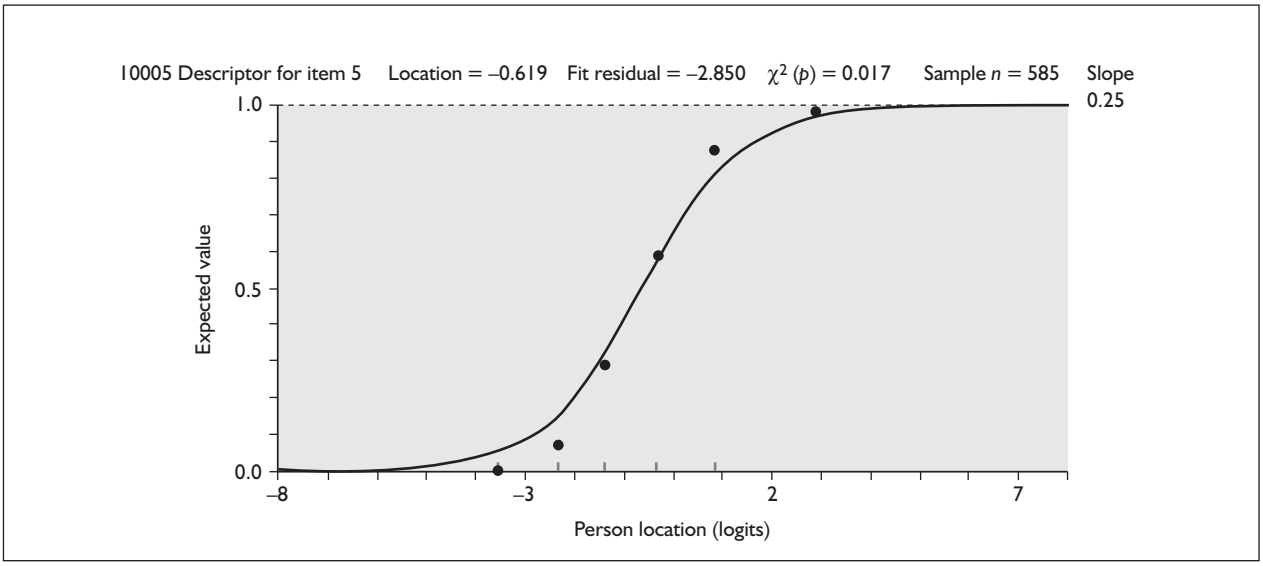


FIGURE 8 *Item characteristic curve for item 5 including class intervals.*

Thus, the degree of fit must be interpreted in the light of the goals of measurement and evaluation. For example, many established scales were developed using either traditional psychometric or no psychometric methods. Rasch analysis is more likely to identify problems with those scales than with scales constructed using Rasch methods (when applied appropriately). The immediate response might be to change existing scales – but that might not be the best way forward.

2. No one fit indicator is necessary and sufficient to summarise fit. Andrich recommends that the results of fit indicators are interpreted interactively and together, rather than sequentially and in isolation (much like the results of a clinical assessment).[119]
3. Fit statistics inform us that the responses to items are not as predicted. They do not diagnose the cause of the misfit. It is the analyst's job to try and explain why misfit occurs. The knee jerk response of removing items that misfit is to be avoided.
4. Fit statistics, like all Rasch statistics, give an indication of fit within the frame of reference of the item set. Thus, removing and adding items may change the values achieved.
5. Changing the number of class intervals will change the chi-squared values and the black dots on the ICC (but not the ICC itself). It is likely to have less effect on the associated probability and inferences made. Changing the number of class intervals will not affect the fit residual value.
6. Items that anchor the extremes of the scale range might have somewhat erroneous fit statistics.
7. Poor targeting may under- or overestimate fit statistics.

Interactive interpretation of fit indicators for the RMI

What does this all mean for the 15 RMI items? *Table 10* summarises the results of the three indicators of fit in a very literal sense (pass/minor fail/major fail). This table shows that six items (1, 10 and 12–15) pass all three fit criteria, two items (8 and 11) fail all three criteria, two items (2 and 3) fail two criteria and the remaining five items (4–7 and 9) fail one criterion. Only three of the 15 criterion failures are really notable: item 11's fit residual at −3.775 and item 3's chi-squared value (51.566) and adherence to the ICC. The remaining departures from expectation are relatively small.

The implication of misfit is that it undermines the inferences made from the data. Essentially, for the total score to be a sufficient statistic, and for the estimates of items and persons generated to be invariant, and on an equal interval scale, the data need to fit the Rasch model. The question then becomes to what extent does the misfit associated with a specific analysis disrupt this process?

One common approach to the problem of misfitting items is to remove them in order to create a modified RMI with better fit statistics. A better approach is to try to diagnose why misfit has occurred in some items given that the 15 items of

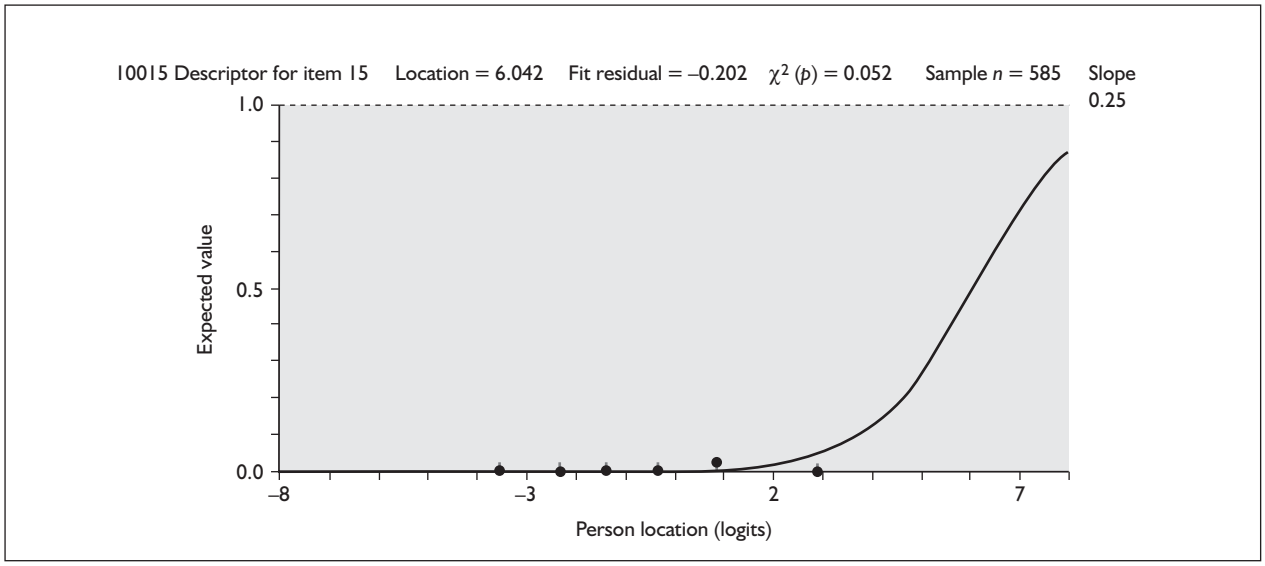


FIGURE 9 *Item characteristic curve for item 15 including class intervals.*

the RMI are conceptually related to mobility – the construct the RMI seeks to measure.

Item 11 ('picking off the floor') failed all three item fit criteria. The fit residual is negative, indicating an overdiscriminating item relative to the model. The ICC (*Figure 12*) shows this – the slope of the item (rate of change of item score across the continuum) is steeper than the model requirement. The class interval dots tend to be below the line at the left-hand end and above the line at the right-hand end. Thus, among people with greater levels of disability, more people than predicted were *unable* to do this task, and, among people with lesser levels of disability, more people than predicted were *able* to do the task. One explanation of the misfit is that this item involves more than mobility, and is less related to mobility than the other items. Another explanation is that there may been some ambiguity, or that the items may not be a common task for people to do. So, there may be some estimating by people as to their ability to do the task – or the interviewers may set different standards for success and failure.

Item 7 ('stairs') also failed all three item fit criteria and demonstrated a similar pattern of fit results to item 11. As it stands, item 7 is ambiguous in that the number of stairs constituting a flight varies widely. Some people might not climb their flight of stairs as there are too many steps. So, they would report unable when they are able to do

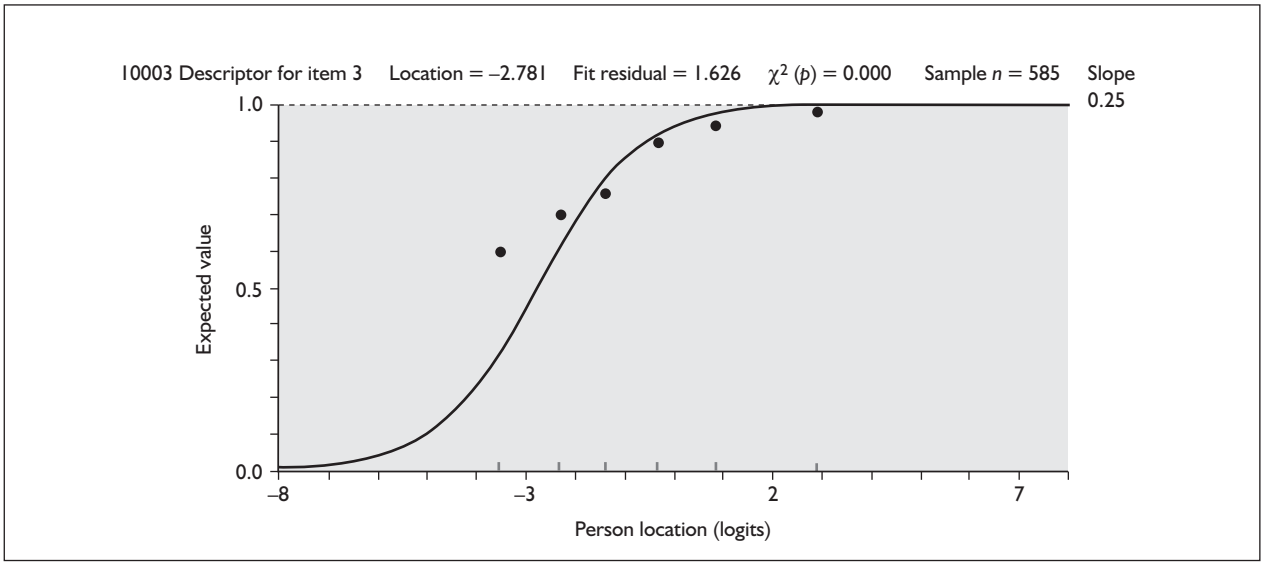


FIGURE 10 *Item characteristic curve for item 3 including class intervals.*

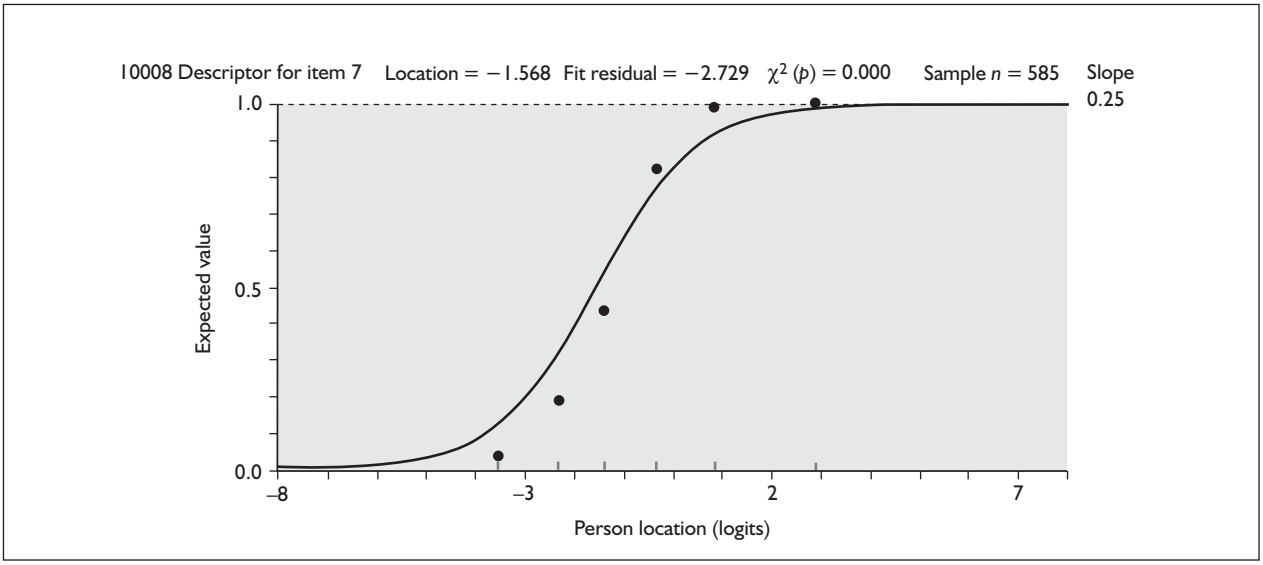


FIGURE 11 *Item characteristic curve for item 7 including class intervals.*

a few steps – which for others might constitute a flight. In addition, there is no clarification of 'help', which can be in many forms (handrail, verbal encouragement, hands-on help from another person). Thus, there are many explanations for misfit, and were one to modify the RMI it would be appropriate to address these issues before removing items. It is perhaps surprising to note that managing a flight of stairs (location = −1.568) appears to be considerably easier than going up and down four steps (location = +2.715).

Items 3, 7 and 11 came out worst. All three items have some degree of ambiguity. Item 3 involves a judgement of time and is not a regular task. It is uncertain why 10 seconds was used as the criterion. Item 7 does not define what constitutes a 'flight of stairs' or 'help'. However, it is interesting that item

TABLE 10 *Summary of item fit statistics*

Item	Fit residual	χ^2	ICC	F+	F++
1	P	P	P	0	0
2	P	F+	F+	2	0
3	P	F++	F++	0	2
4	P	F+	P	1	0
5	F+	P	P	1	0
6	P	P	F+	1	0
7	P	P	F+	1	0
8	F+	F+	F+	3	0
9	F+	P	P	1	0
10	P	P	P	0	0
11	F++	F+	F+	2	1
12	P	P	P	0	0
13	P	P	P	0	0
14	P	P	P	0	0
15	P	P	P	0	0

F+, minor fail; F++, major fail; ICC, item characteristic curve; P, pass.

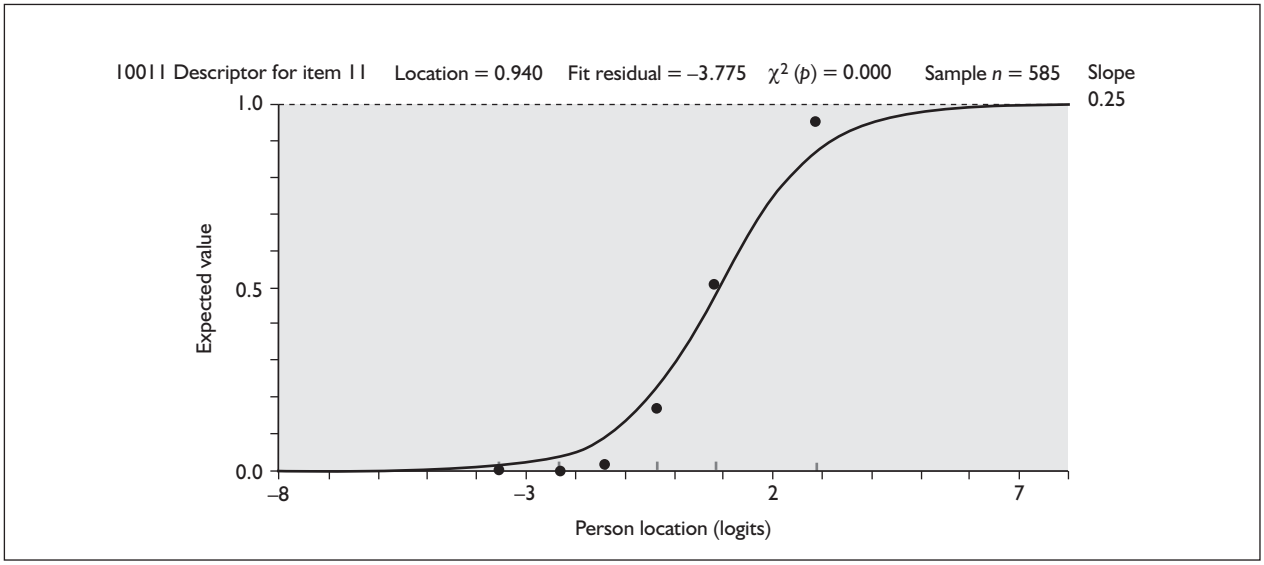


FIGURE 12 *Item characteristic curve for item 11 including class intervals.*

3 ('sitting balance') failed two item fit criteria. It had the largest chi-squared values and therefore chi-squared probability. The ICC (see *Figure 10*) shows that the slope of the observed responses is flatter than the ICC. Thus, the item is less discriminating than the model requirements. This explains the positive fit residual. Thus, at the more disabled end of the continuum, more people are able to do this task than predicted. Also, people in four successive class intervals, and by definition with different levels of disability, tend to get the same mean score on the item. One explanation is that item 3 may be difficult to interpret.

Does the response to one item directly influence the response to another?

The Rasch model requires that the items are locally independent, i.e. the response to one item in a scale is independent of the response to another in the scale. In a Rasch analysis, one way of studying local independence stems from the theory that the residuals (observed–expected values) should represent random error. Therefore, the residuals for the items should not be correlated. If they are correlated the implication is that the answer to one item is dependent on another. Thus, the correlations among the residuals should be low. The criterion of < 0.30[118] has probably been chosen as this represents 10% shared variance. For the RMI, in this sample, none of the residual correlations exceeded 0.30. This implies that the responses of the items are independent of each other and that the items are locally independent.

Andrich has recently developed another way of examining for local independence. This method is not yet freely available or published. Essentially, this method involves combining items that might be dependent to form a single item with more response categories (he has named these subtests), and re-examining the modified scale. If the PSI of the modified scale falls notably relative to the original scale this implies that the reliability of the original scale was inflated by the item dependency. The fact that item dependency inflates reliability estimates has long been known. The importance of this method lies in the fact that the residual correlations do not always identify item dependency. It is therefore superior both conceptually and empirically.

We examined for local dependency using this new method. Specifically, we examined the implications of combining two pairs of items that could be dependent: 'walking outside (even ground)' with 'walking outside (uneven ground)'; and 'walking inside with an aid if needed' with 'walking inside with no aid'. Each pair of items was combined to form one item with four response options using the subtest facility within RUMM2020. The PSI was unaffected, implying no significant local dependency.

Are the locations of the items stable across clinically important groups?

In this example we examined item functioning across the three treatment arms of the study (placebo, Cannador, Marinol). The three groups are clinically important because they represent three different subsamples of the study. Whether these randomly generated samples are statistically equivalent in terms of their clinical characteristics

is an empirical question. Whether these samples 'handle' the scale in the same way is a further empirical question. There was no evidence that the three groups handled any of the items differently. Thus, we have evidence that item performance across the three treatment groups is stable and that the three groups can be measured on a common ruler. There is a discussion of DIF and how to examine and interpret it in Chapter 6.

Have the people in the sample been measured successfully?

Are the persons in the sample separated along the line defined by the items?

Figure 4 shows the distribution of person measurements (locations) relative to the item locations. The sample is well spread with values ranging from around −4.75 to+7 logits. The mean is −0.861 (SD 2.622), indicating that the sample is off-centre of the items (as the mean of the item locations is always 0).

Figure 4 is a graphical indicator that the items of the scale have been successful in separating this sample of people with MS. One numerical indicator of the degree of separation is the PSI.[120] This is computed from the person location estimates as the variation among person locations relative to the error of estimate for each person. Thus, it is consistent with the traditional definition of reliability of a scale, i.e. how reliably the scale distinguishes between the responders. The PSI tells us how much of the variation of person estimates can be attributed to error variances, i.e. the extent to which scores are associated with random error. Thus, the PSI is a reliability indicator, and like most reliability indicators ranges from 0 (all error) to 1 (no error). The fact that it focuses (in both name and computation) on the separation of the persons indicates that the PSI is not a property of the scale but a property of the scale in relation to the specific sample of persons measured. In contrast, the analogous reliability statistic of traditional psychometric analysis, Cronbach's alpha, has the same formulaic structure as the PSI, but relates the variance among persons to the variance of the items.

It is interesting to note that the PSI and Cronbach's alpha often produce near identical values. Indeed, RUMM2020 reports Cronbach's alpha (partly to satisfy those people who appeared to want the reassurance of at least seeing this index, perhaps as it is the main reliability estimate for traditional psychometric analyses[121]). However, there are fundamental differences between the PSI and coefficient alpha. First, the PSI is expressed entirely in terms of person locations and so meets the true definition of reliability. Second, the PSI is computed from linear measurements rather than raw scores. Third, the PSI can be computed when data are incomplete (i.e. there are missing item responses) whereas alpha requires complete data (i.e. it is computable only on the subsample with complete data). This is because, in Rasch analysis, missing responses affect the standard error of a person location not the ability to generate an estimate. The PSI will, however, decrease. Further details about the PSI can be found in Andrich's paper.[120]

In the RMI data set, both PSI and alpha are reported as there are no missing data. The values are essentially identical at 0.91 indicating good reliability. It is important to note that the PSI, like alpha, is sample dependent because it is computed from the person locations. Essentially, it indicates the ability of a set of items to separate the study sample. Thus, the PSI is a function of the data, not an independent function of the scale.

There are some notable features about the distribution of the people in the sample. The sample is not normally distributed. This is neither expected nor wanted, as the distribution of the sample is an empirical finding rather than a requirement. However, it does have implications by suggesting it would be advantageous not to make assumptions about the distribution of samples and traits in populations. The largest frequency of patients ($n \approx 80$, ≈12%) is at the floor of the scale range, and about 50% are within the lower third of the scale range. This provides further evidence of suboptimal targeting of the RMI to the study sample, especially in the context of a clinical trial, when the ability to detect change is paramount.

Figure 13 shows the plot of RMI total scores against the intervalised measurements they imply. The curve is S-shaped, although not substantially so. *Table 11* shows the measurement implied by each RMI total score. These are the best estimates and can be used when people have complete data. Also tabulated are the changes in interval measurement units implied by each single-point change score. These vary 4.7-fold across the range of the scale.

Table 11 also gives the standard error associated with each of the 16 possible RMI person locations. This is a function of the number of items, the relative relationship of the item locations to the person locations (targeting) and the person's

total score. The greater the number of items a person responds to, and the better the targeting, the smaller the standard error. The reason, as for the standard error of item estimates, is that the formula for computing the standard error of a person's location estimate is {SE = 1/√[sum of $p(1-p)$ across items answered]}.

Figure 14 plots the SE against each RMI location when there are complete data (i.e. all items have been answered). The curve is U-shaped and indicates that standard errors vary threefold across the range of the scale, and are greatest at the extremes and smallest at the centre of the scale range. This is logical. People who score at the floor and the ceiling are those for whom we have the least confidence about their estimate – we do not know how far above or below the ceiling or floor they really lie.

There are two implications of the U-shaped curve in *Figure 14*. The first is that the most precise measurement occurs in the centre of the range covered by a scale. The second implication concerns the measurement of change. The statistical significance of change is influenced by a person's location at each measurement time point (e.g. pre and post treatment) as well as the size of the change score.

It is noteworthy that in *Table 11* each RMI raw score implies one linear measurement. Some are concerned by this as it implies that no regard is given to the combination of items that are 'passed' or 'failed'. Consider an RMI raw score of 7. In practice, this can be achieved by scoring 1 for any seven RMI items. The extreme situation would be two people who attain an RMI total score of 7: person A scores 1 for each of the seven easiest items (i.e. items 1, 3, 2, 6, 4, 7, 8 and 13), while person B scores 1 for each of the seven hardest items (i.e. items 9, 7, 11, 10, 12, 14 and 15). These two people achieve the same RMI total score, the same RMI linear measure and the same standard error. This concerns some people, who argue that a person's measurement should be influenced by which specific items they 'passed' and not simply by their aggregate score. They would expect person B to have a higher location (i.e. better mobility) than person A, and so might choose another item response model to analyse their rating scale data. In the Rasch model, the reason that persons A and B get the same location results from one of the mathematical properties of the model: that the total score is the sufficient statistic for estimating a person's location. This was explained in Chapter 3. This means that there is no more information in the response pattern for determining their location. This does not mean that the pattern of responses isn't important. On the contrary, it is fundamentally important and accounted for in the person fit statistic – the extent to which a particular individual's responses to the 15 RMI items accord with the expectations of the measurement model (explained below). Person B would be identified as a marked 'misfit' that would require explanation before his or her measurement was considered for further analysis. From a Rasch perspective, the pattern of responses invokes enquiry as to why such an unlikely pattern of responses occurs, rather than invoking a change in the estimate of that person's location.

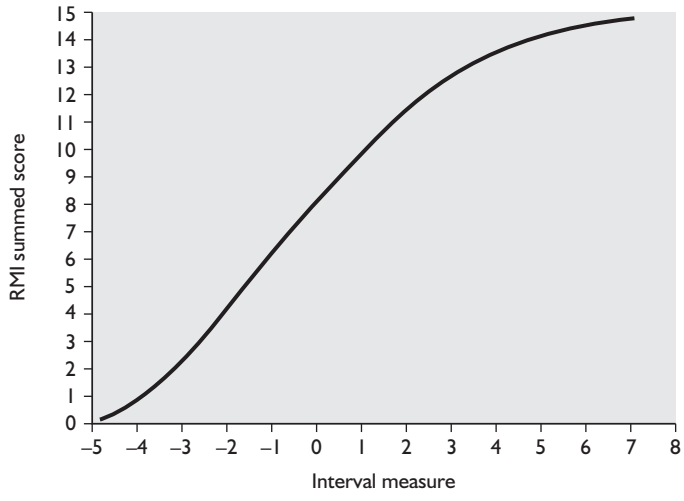


FIGURE 13 *Plot of RMI total scores against the intervalised measurements they imply.*

TABLE 11 *Rivermead Mobility index (RMI) raw scores and the interval measures they imply*

RMI raw score	Interval measure	Standard error	Change implied by 1-point difference[a]
0	–4.780	1.358	NA
1	–3.882	0.984	0.898*
2	–3.170	0.824	0.712
3	–2.607	0.755	0.563
4	–2.097	0.725	0.510
5	–1.603	0.715	0.494
6	–1.105	0.718	0.498
7	–0.590	0.729	0.515
8	–0.053	0.746	0.537
9	0.513	0.768	0.566
10	1.119	0.798	0.606
11	1.779	0.837	0.660
12	2.496	0.896	0.717
13	3.323	1.018	0.827
14	4.713	1.374	1.390
15	7.047	2.406	2.334

NA, not applicable.
a This column represents the change in interval-level measurement implied by a 1-point difference in RMI raw score. For example, a change in RMI raw score from 0 to 1 implies a change in interval-level measure of 0.898 logits (*).

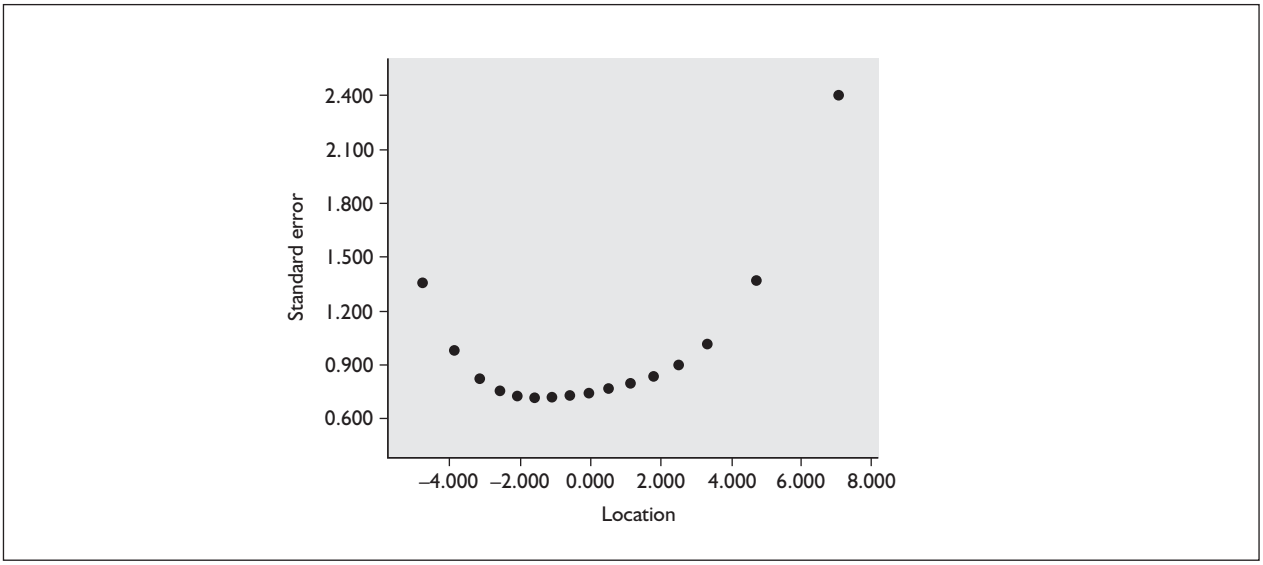


FIGURE 14 *Plot of standard error against location.*

Also, note that *Table 11* is for complete data only. That is, the person location implied by each total score (e.g. a total score of 7 implies a location of −0.590) is applicable only for a person who has responded to all 15 items. It is important to be certain about what happens when there are missing item responses (i.e. incomplete data). Clearly, tables cannot be produced for all eventualities. For example, consider person C, who achieved a total score of 7 but responded only to 10 items; the other items where left blank. This person's location would differ from that of a person with complete data and a total score of 7, and the standard error would differ (as this is related to the number of items). The key issue here is that the Rasch analysis uses the available data to generate an estimate and does not make assumptions about the missing data.

Do individual placements of people on the variable make sense?

This analysis did not examine correlations with other measures, group differences, hypothesis testing and clinical ordering of persons in terms of their relative locations.

How valid is each person's measurement?

The item fit residual, discussed above, summarises the extent to which responses to each individual item are consistent with those expected by the Rasch model. This value is achieved for each of the 15 RMI items by summarising the residuals arising from 585 patients' responses to that item. Similarly, we can achieve a value for each of the 667 patients by summarising the residuals from each person's responses to the 15 RMI items. This is called the person fit residual; it summarises the extent to which responses by each person are consistent with those expected by the Rasch model, and is used to identify misfitting individuals.

As before, the residual is produced by subtracting the expected score from the observed score. The predicted scores are calculated from the Rasch model using the estimates of the person and item locations. The residual is standardised by dividing it by the square root of the variance, which is computed from the expected value (EV) using the following formula: variance = $EV - EV^2$. This process leads to a standardised fit residual for each person's response to each item and these are transformed and summarised to form the person fit residual which approximates a standard normal deviate. Thus, for each person the fit residual should ideally lie within the range −2.5 to +2.5.

In this sample, person fit residuals ranged from −1.042 to +1.969 (mean −0.224; SD 0.421). Thus, none outside the range −2.5 to +2.5. *Tables 12* and *13* show the observed and expected values for one of the better and one of the worse fitting persons. Note that the fit residual for the better fitting person (*Table 12*) is −1.042, despite a perfect response 'pattern'. This is because the model expects some departures from perfection – which is seen as overfitting.

Summary of results of the Rasch analysis

The sample was adequate for examining the scale, but the scale was suboptimal for measuring the sample. This implies that any group-level changes detected were underestimates.

The items of the RMI mapped out a variable of increasing intensity, but the item locations indicated areas on the continuum within the range measured by the RMI where measurement could be improved. The ordering of items along the variable was clinically sensible, except for the location of the 'stairs' item. This appeared to be easier than predicted. This needs further qualitative examination as the reasons for these findings are not immediately apparent.

The fit statistics highlighted items for which the observed responses did not fit the expectations of the measurement model. One of these items was the 'stairs' item, whose location has already been questioned. The other notably misfitting items were 'picking off the floor' (item 11), 'sitting balance' (item 3) and 'lying to sitting' (item 2). These require further examination.

There did not appear to be any dependence among the items in terms of residual correlations or Andrich's subtest analysis. None of the items demonstrated differential functioning across the three treatment arms. To recap, the three arms were placebo, Cannador and Marinol. These were considered clinically important because they represent three different subsamples of the study.

The 15 RMI items separated the sample well. The responses of all people were within boundaries of expectation, indicating no misfitting persons. A 1-point change in RMI total scores implied a variable change, up to almost fivefold, in interval-level measurements across the range of the continuum.

TABLE 12 *Observed and expected responses for one of the better fitting persons (n = 585; 666 with 81 extremes excluded)*

RMI item	Item location	Location SE	Observed score	Expected value	Fit residual
1. Turning over in bed	–3.032	0.130	1	0.992	0.090
3. Sitting balance	–2.781	0.126	1	0.990	0.102
2. Lying to sitting	–2.707	0.125	1	0.989	0.106
6. Transfer	–2.423	0.122	1	0.985	0.122
4. Sitting to standing	–1.863	0.117	1	0.974	0.162
8. Walking inside, with aid if needed	–1.568	0.116	1	0.966	0.188
13. Bathing	–0.872	0.115	1	0.934	0.266
5. Standing supported	–0.619	0.115	1	0.917	0.301
9. Walking outside (even ground)	0.444	0.119	1	0.792	0.513
7. Stairs	0.515	0.119	1	0.780	0.532
11. Picking off the floor	0.940	0.122	1	0.698	0.657
10. Walking inside, no aid	2.566	0.144	0	0.313	–0.675
12. Walking outside (uneven ground)	2.641	0.146	0	0.297	–0.650
14. Up and down four steps	2.715	0.148	0	0.282	–0.626
15. Running	6.042	0.356	0	0.014	–0.119

RMI, Rivermead Mobility Index; SE, standard error.
Person location = 1.779; SE = 0.837; fit residual = –1.042.

TABLE 13 *Observed and expected responses for one of the worse fitting people (n = 585; 666 with 81 extremes excluded)*

RMI item	Item location	Location SE	Observed score	Expected value	Fit residual
1. Turning over in bed	–3.032	0.130	0	0.984	–7.966
3. Sitting balance	–2.781	0.126	1	0.980	0.142
2. Lying to sitting	–2.707	0.125	1	0.979	0.148
6. Transfer	–2.423	0.122	1	0.972	0.170
4. Sitting to standing	–1.863	0.117	1	0.952	0.225
8. Walking inside, with aid if needed	–1.568	0.116	1	0.936	0.261
13. Bathing	–0.872	0.115	1	0.880	0.370
5. Standing supported	–0.619	0.115	1	0.850	0.419
9. Walking outside (even ground)	0.444	0.119	1	0.662	0.714
7. Stairs	0.515	0.119	0	0.646	–1.352
11. Picking off the floor	0.940	0.122	0	0.545	–1.093
10. Walking inside, no aid	2.566	0.144	0	0.190	–0.485
12. Walking outside (uneven ground)	2.641	0.146	1	0.179	2.141
14. Up and down four steps	2.715	0.148	0	0.169	–0.450
15. Running	6.042	0.356	1	0.007	11.725

RMI, Rivermead Mobility Index; SE, standard error.
Person location = 1.119; SE = 0.798; fit residual = 1.969.

Standard errors also varied threefold across the range of the scale, indicating that the significance of individual person change is dependent on a person's location pre and post treatment.

Summary

This analysis of RMI data using both traditional and Rasch psychometric analyses offers the opportunity to compare and contrast the two approaches and highlight some of the similarities and differences between them. Andrich and Styles[3] suggest that the Rasch model may be considered as a refinement of, or advance on, traditional analyses. An important similarity is that both methods view the total score produced by summing the item scores as a key statistic. However, the reasons underpinning this are different: in traditional methods the total score is important because the analyses use total scores; in Rasch analysis total scores are important because a property of the Rasch model is that the total score is the sufficient statistic from which accurate estimates can be derived.

This study has demonstrated a number of refinements that result from using the Rasch model to analyse the RMI data. These refinements can be considered, separately, in terms of the construction of the scale and the measurement of people.

In Rasch analysis, the 15 items of the RMI have been located, relative to each other, on an interval-level continuum. Moreover, these estimates are independent of the distributional properties of the sample from which the estimates were made, and we have an estimate of the error associated with the location. The fact that these estimates are freed up from the sample distributional properties is a fundamental requirement if we are to determine the stability of these locations, and thus the stability of the ruler they imply, across clinically different samples. Traditional psychometric methods do not generate estimates of item locations. This does not mean we can't get an idea of their potential locations – this can be achieved by examining the item mean scores. However, these values are not on an interval-level scale, and are dependent on the distribution of the sample from which they were derived. Thus, inferences about the potential stability of items (even if they could be made) could not be undertaken meaningfully.

In addition to generating numerical location estimates for the items, RUMM2020 generates a graph that allows these estimates to be seen. This plot enables investigators to determine, immediately, the extent to which a set of items maps out the intended variable. The breadth of measurement, coverage across the continuum and gaps in the measurement process are explicit. This provides the empirical basis of improved measurement by, for example, plugging gaps, extending the continuum and identifying redundant items for removal. In traditional methods, item redundancy is decided based on the correlations between pairs of items and overlapping content. The correlation between two items does not tell us about their place on the continuum.

In Rasch analysis, the extent to which a set of items works together to define a single variable, their internal consistency, is determined rigorously: the fit of the observed responses to a mathematical model. In traditional methods, the internal consistency of a set of items is determined by correlational analyses: corrected item–total correlations, Cronbach's alpha and homogeneity coefficients. Correlations, as Thurstone pointed out nearly 80 years ago,[40] are limited as 'they constitute an acknowledgement of failure to rationalise the problem and to establish the functions that underlie the data'.

Traditional methods and Rasch analysis came to different conclusions about the extent to which the 15 RMI items are a conformable set. Traditional methods implied the items were a cohesive set but identified item 15 ('running') to be poorly related to the others. Rasch analysis identified problems with a number of items that warranted further explanation. Certainly, attention needs to be paid to the wording of some items and their descriptions. Item 15 was fine, and the reasons for its poor item–total correlation are its distant location from the other items and its relative targeting to the sample. This highlights a poorly understood limitation of correlations.[122]

Finally, in terms of examining items, Rasch analysis enabled a formal examination of dependency among items and differential performance of items. This was not possible with traditional methods.

Traditional and Rasch analysis provide different information about the measurement of persons and the inferences than can be made from them. The first difference is that traditional methods generate total scores and 'measure' people on an ordinal scale, whilst Rasch analysis generates

measurements and measures people on an interval scale. Moreover, this is the same interval-level scale on which the items are located. Raw scores, which increase in successive integers, are non-linear because they have a non-linear relationship to the underlying trait (here mobility) that they seek to measure. In contrast, Rasch-derived person locations (and item locations) are linear measures because they have a linear relationship to the underlying trait that they seek to measure.

We have seen that the change in measurement implied by a 1-point change in RMI raw score varies across the range of the scale from 0.494 logits (score change in RMI 4 to 5) to 2.334 (score change in RMI 14 to 15). Thus, a change of 1 RMI raw score varies 4.7-fold across the scale range. This has critical implications for clinical trials in which accurate measurement of change underpins the inferences made. It also has implications for performing basic statistical tests on RMI raw scores. It questions the legitimacy of adding (which underpins means and SDs), subtracting (which underpins change score analysis), division and multiplication and all the statistical tests that handle data in this way. In contrast, it is legitimate to undertake these statistical analyses on Rasch-derived measurements because they are on an interval scale.

Person locations generated by a Rasch analysis are freed up from the distributional properties of the items from which they were generated. This is fundamental to examining the stability of person measurements, and thus the whole idea of equating scales on the same metric, and using different combinations of items to measure people in the knowledge that their measurements can be referred back to a common metric (item banking).

Another advantage of Rasch analysis is that it generates a standard error for each person location. In contrast, traditional analyses generate a single estimate of the standard error that is considered applicable across the whole range of the scale. The problem with a single estimate of error is that it is illogical – it is clear that the people for whom we have the least confidence about their measurement are those at the extremes of the scale range.

The availability of individualised standard errors makes measurement at the individual person level a legitimate process. This means that rating scale data from clinical trials and other studies could be analysed for individual person clinical decision making as well as group comparison study. The advantages of this are illustrated in Chapter 8. In contrast, raw scores are not considered suitable for individual person-level analysis.[107]

The demonstration that variable standard errors are associated with different person location estimates is important for another reason. It demonstrates that significant change for any one individual is not simply a function of the magnitude of their change. It also depends on their location on the continuum at the measurement time points. This important fact is not accounted for in group-based analyses of change. This is discussed further in Chapter 8.

Traditional psychometric analyses do not give us information about individuals' responses to items. Rasch analysis does, and quantifies this as the person fit statistic. This information is clinically valuable for identifying individuals for whom the pattern of responses is unlikely. Once identified, steps can be taken to diagnose why these people gave these responses. In addition, individuals identified in this manner might be excluded from the analysis as they could confound the results and inferences made.

Last, let us say a few words on missing responses to items. This occurs in many studies, although not in the RMI data from the CAMS study. As we have discussed, the traditional psychometric approach to missing item response data is to replace missing responses with a person-specific mean score. Thus, the missing item responses are replaced with the mean score of the completed items (provided that 50% of the items have been completed). This method is reputed to be psychometrically sound,[36] although to our knowledge it has not been tested formally. It might be reasonable if all the items were measured at the same point on the scale (one of Likert's original scaling criteria[6]). However, as demonstrated here for the RMI, this is clearly not the case. It does not appear scientifically sound to make assumptions as to how people might have responded to items, especially as the basis for making those assumptions is weak. Rasch analysis takes a more scientific different approach to missing data. It simply uses the available data to generate a best estimate. Missing data are accounted for in terms of the confidence of the estimate – the standard error is increased.

From the above discussion, it seems clear that Rasch analysis offers substantial clinical and scientific advantages over traditional psychometric methods in the development and evaluation of rating scales, and in the analysis of rating

scale data. Nevertheless, there have been a few publications that have concluded that the benefits are modest.[123] This has led a number of people to argue that traditional psychometric methods are just as good as new psychometric methods. However, these studies simply examined the impact of inferences on a single study of using Rasch transformed summed scores versus summed scores and therefore there are a number of reasons why the findings are not surprising. First, as we have seen in *Figure 13*, summed scores are monotonically related to the interval measurements they imply. Second, group-based and individual person-based analyses can come to different conclusions. Third, neither of these studies examined differences at the individual person level. Fourth, the impact of inferences will be study dependent; thus, findings from a single study will not generalise and it will be impossible to determine in advance the situations in which a difference might be found.

More importantly, we think that the main problem with these studies comparing Rasch with traditional scoring is that they have missed the point. As we have seen, the ability to construct interval measurements from ordinal rating scale data is one of a series of advantages that Rasch analysis offers over traditional psychometric methods. Their major role will be in the development of newer scales,[119] recommended modifications of scales developed using traditional psychometric approaches and informing about the constructs being measured.[66] As rating scales are increasingly the central dependent variable on which treatment decisions are made about patients we treat, we do not feel that there can be any compromise in the efforts made to achieve rigorous measurements in clinical studies.

Chapter 5

A re-evaluation of the MSIS-29

The lessons of Rasch analysis

Overview

The aim of this chapter is to report a comprehensive evaluation of the Multiple Sclerosis Impact Scale (MSIS-29; see Appendix 2.1) using Rasch analysis, and to use this as another vehicle for comparing and contrasting the traditional and Rasch-based evaluations of rating scales. This chapter builds on the previous chapter. The RMI, the subject of Chapter 4, has items with two response categories. In contrast, the MSIS-29 has items with five response categories.

The MSIS-29 was originally developed using traditional psychometric methods. In doing so, it has fully satisfied the requirements of traditional psychometric evaluations. However, since we developed the scale, there has been increasing interest in the application of new psychometric methods (Rasch analysis and Item Response Theory). Thus, the aim of this analysis was to examine how a scale developed using traditional psychometric methods (i.e. the MSIS-29) stands up to evaluation using new psychometric methods, in this case Rasch analysis.

Although Rasch analysis provides support for the MSIS-29, it also highlighted problems with the MSIS-29 that were not detected by traditional psychometric analyses. Specifically, there were problems with the five category response options and the extent to which the items of the two subscales were conformable sets. In addition, Rasch analysis indicated how these problems may be solved to produce a more reliable, valid and responsive scale. Results also have implications for scale development. They highlight the advantages of using Rasch analysis, building scales on a strong conceptual basis; establishing the response options early in the process; and measuring over a wide range of the continuum.

The clinical problem

Rating scales are increasingly used as outcome measures in clinical studies. As such, they are the central dependent variables on which clinical decisions are made.[124] This important role necessitates that every effort be made to maximise the scientific quality of the rating scales used in clinical research and practice.

There are two main approaches to rating scale development and evaluation. The most widely used approach applies traditional methods. There is, however, increasing interest in the application of new psychometric methods. Both approaches have the same goal: to determine if it is legitimate to generate a total score by combining the integer scores from a group of items, and if that total score is reliable and valid. The two approaches differ in the evidence used to achieve that goal. Traditional psychometric analyses are based on Classical Test Theory and are embodied in the work of Likert[6,29] and others. Their evidence comes mainly from analyses of correlations. New psychometric methods stem from the work of Thurstone,[125] Lord and Novick,[17] and Rasch.[65] Their evidence comes from checking the observed data against an a priori explicit mathematical model of how a set of items must behave to permit the summation of items to generate a reliable and valid total score.[3]

New psychometric methods offer potential scientific and clinical advantages over traditional methods. From a scientific perspective they enable more sophisticated evaluations of scales. From a clinical perspective they offer, amongst other things, the ability to transform ordinal scores into interval measures, and legitimate individual person measurement. However, there are limited comparisons in the literature to determine the added value of using the new psychometric methods.

The Multiple Sclerosis Impact Scale (MSIS-29)

The MSIS-29 (see Appendix 2) is a 29-item self-report rating scale for measuring the impact of MS. It has two subscales – a 20-item physical impact scale and a nine-item psychological impact scale – and seeks to measure the impact of MS on physical and psychological functioning. All items have five response categories ('not at all'; 'a little'; 'moderately'; 'quite a bit'; 'extremely') that are assigned sequential integers (1, 2, 3, 4, 5). Total scores for the two scales are achieved by summing the item scores for the 20 physical items (items 1–20) and the nine psychological impact items (items 21–29). When some items have not been endorsed by an individual, a total score can still be computed provided that at least 50% of the items have been answered (i.e. 10 or more physical impact items; five or more psychological impact items). Under these circumstances investigators can replace each missing item score with the person-specific item mean score, i.e. the mean score of the items completed by that individual. This process is called imputing and is considered to be psychometrically sound.[36]

The MSIS-29 was generated to be suitable for use as an MS outcome measure in appropriate clinical trials, epidemiological studies, audit and routine clinical practice. It is increasingly widely used and has been officially translated into around 20 different languages.

The MSIS-29 was developed using traditional psychometric methods for scale development. Full details of its development and evaluation are described elsewhere.[2,126] Briefly, some 3000 statements concerning the impact of MS were generated from 30 tape-recorded one-to-one patient interviews, literature review and expert opinion. These statements were examined for their content, overlap and redundancy. From the statements 141 potential scale items were written. These items were administered as a questionnaire to a large, randomly selected and geographically stratified sample of people with MS ($N = 1250$) from the membership database of the Multiple Sclerosis Society of Great Britain and Northern Ireland.

Data analysis focused on 129 items. This was because 12 items concerning walking were excluded as they were appropriate only to a limited number of people with MS. These 12 items formed the MS Walking Scale,[127] which has since been genericised for neurological conditions.[128,129]

Full details of the item reduction process, i.e. the methods used to select the final 29 items from the original 129, are given elsewhere.[126] First, 36 items were removed due to high item–item correlations and thus presumed redundancy. Next, 51 items were removed on the basis of psychometric performance (either high item-level floor/ceiling effects or poor endorsement profiles). The remaining 42 items were entered into an exploratory factor analysis. A number of possible factor solutions were examined. The two-factor solution was the most clinically and statistically appropriate, but three items were removed as they loaded similarly on both factors. The other 39 items were grouped into two scales (26 items and 13 items) whose content concerned the physical and psychological impact of MS. Refinement of the two scales using tests of item convergent and discriminate validity identified items that might confound measurement and thus were removed. The final instrument had two scales: a 20-item physical impact scale and a nine-item psychological impact scale.

The reliability and validity of the MSIS-29 were examined in a second large independent survey. We examined data quality, scaling assumptions, targeting, reliability (internal consistency and test–retest reliability) and validity (convergent and discriminate construct validity, group differences, hypothesis testing), and undertook a provisional responsiveness study. A comprehensive responsiveness study, in which the MSIS-29 was compared with a range of other scales, was also undertaken and reported.[14] Across the range of health outcomes measurement literature, it is extremely rare to find examples which include the extent of scale evaluation in all the areas we have examined in the case of the MSIS-29. Subsequent evaluations,[130–133] also using traditional psychometric methods, have supported the reliability, validity and responsiveness of the scale.

Sample

MSIS-29 data from a total of 1725 people with MS were analysed. These data were generated by the two field tests of the MS Society membership databases ($n = 768$ and $n = 712$) undertaken during the development of the MSIS-29[2,126] and a study of its responsiveness ($n = 245$).[14] In this responsiveness study we had three samples of people with MS: $n = 64$ patients admitted for inpatient rehabilitation at the National Hospital for Neurology and Neurosurgery (NHNN) in London; $n = 77$ patients admitted to the NHNN for steroid

treatment of MS relapses; and $n = 104$ people with primary progressive MS from the NHNN database. *Table 14* shows the sample characteristics.

Methods

Overview

We have used the same template for examining the MSIS-29 as we did in Chapter 4.

Physical and psychological scales are examined and their results are reported separately. There are, however, two main differences between this chapter and Chapter 4. The first difference arises out of the fact that the items of the MSIS-29 have more than two response options. Thus, one important aspect of the analysis is to examine whether these response options work as intended. The second difference is that we have a large sample size. When the sample size analysed substantially exceeds 500, Andrich recommends amending the sample size to $n = 500$ for the computation of the chi-squared fit statistics. The reason for this is that the chi-squared values and probabilities are sample size dependent. The larger the sample size, the greater the values and the greater the apparent misfit. Amending the sample size to 500 for the computation of fit statistics provides the analyst with a better feel for the behaviour of the data and gives the fit statistics a chance to reveal any misfit to the model. (Random selection was used to generate the sample of $n = 500$ in our study.) It is important to remember that, for any fit statistic, no one fit indicator is necessary and sufficient, and this chi-squared estimate is a guide only to give a feel for the behaviour.[134] It is important also to note that making this adjustment does not affect other aspects of the analysis.

Is the scale-to-sample targeting adequate for making judgements about the performance of the scale and the measurement of people?

The relative distributions of the item locations and the sample locations were examined. The distribution of the item locations for polytomous items, such as the MSIS-29, differs from that of dichotomous items, such as the RMI studied in Chapter 4. They differ because polytomous items have multiple locations on the continuum (correctly termed thresholds and explained below), whereas dichotomous items have only one location on the continuum. For polytomous items RUMM2020, and other Rasch analysis programs, usually report threshold values and a single item location, the mean of the threshold values.

TABLE 14 *Sample characteristics (n = 1725)*

Characteristic	Numerical value
Sex (percentage female)	68.2%
Age [mean (SD); range]	50.5 (12.2); 18–87
Duration of MS [mean (SD)]	
Since onset	17.4 (10.8)
Since diagnosis	11.3 (7.5)
Employed	19.8%
Retired as a result of MS	54.5%
Married	65.8%
Educated after minimum school leaving age	57.1%
Degree or professional qualification	35%
Mobility level	
Walk unaided	19.8%
Walk with an aid	54.5%
Wheelchair user	25.7%

MS, multiple sclerosis; SD, standard deviation.

Has a measurement ruler been constructed successfully?

Do the item response categories work as intended?

Each item of the MSIS-29 has five response categories. Consistent with Likert-type rating scales, these response categories are ordered to imply a continuum of increasing impact, from less ('not at all') to more ('extremely'). This continuum of increasing impact is implied further by assigning sequential integers to the response categories – 1 = not at all to 5 = extremely – which divide the continuum into five accordingly. It is, of course, an assumption that these response categories work as intended.

The response categories of dichotomous items also imply ordering. For the RMI, the ordering was 0 = 'unable to do the task' and 1 = 'able to do the task'. For the items of the RMI, the location is the point (threshold) on the mobility continuum where each of the responses has a 50% chance of occurring, i.e. the probability of scoring either 0 or 1 is the same (50%).

Conceptually, and mathematically, polytomous items are simply an extension of dichotomous items. Essentially, the RMI items have two ordered response categories and the MSIS-29 items have five ordered response categories. Thus, for items of the MSIS-29 which have five response categories there are four thresholds (standard mathematical nomenclature uses the Greek symbol 'τ' (tau) instead of 'T' to signify threshold):

- τ1 – where the probability of scoring either 1 ('not at all') or 2 ('a little') is the same
- τ2 – where the probability of scoring either 2 ('a little') or 3 ('moderately') is the same
- τ3 – where the probability of scoring either 3 ('moderately') or 4 ('quite a bit') is the same
- τ4 – where the probability of scoring 4 ('quite a bit') or 5 ('extremely') is the same.

The fact that item response category scoring begins with either 0 or 1 does not matter conceptually (as what comes after is sequential) or mathematically (the RUMM2020 program changes the integers so that they always begin with 0).

The generalisation of the Rasch model, which was developed for dichotomous items, to polytomous items was developed by Andrich.[70] RUMM2020 estimates the threshold values for polytomous items, and thus provides an empirical check on the extent to which the ordered categories are working as intended. This analysis was undertaken for the MSIS-29.

Disordered thresholds, that is the finding that the locations of the thresholds are not ordered sequentially (τ1, τ2, τ3, τ4), imply item scoring functions (the technical phrase for response categories) that are not working as intended.[70] There are three main reasons for this. First, responders cannot use the response categories consistently. This creates noise in the data and impacts on the reliability of an item. Second, the categories do not characterise the intended meaning of what it takes to reflect more of the property within an item. This impacts on the validity of the items. Third, the item does not measure the same underlying trait as the other items. When item categories are not ordered as intended it is possible to combine categories post hoc and investigate the kinds of categorisation that might work.

Do the items map out a discernible line of increasing intensity?

To answer this question we examined the items' locations, their range, how they are spread, their proximity to each other and the precision of these estimates' standard error.

Is the location of items along this line reasonable?

This question was answered by examining the ordering of the items to determine the extent to which it is consistent with clinical expectation.

Do the items work together to define a single variable?

This question was answered by examining three fit statistics: the item–person fit residuals; item–trait chi-squared values and probabilities; and item characteristic curves.

Does the response to one item directly influence the response to another?

This question was answered by examining the correlations between residuals, and using Andrich's subtest analysis for items that on clinical grounds might be locally dependent.

Are the locations of the items stable across clinically important groups?

We examined the extent to which item locations were stable across multiple groups: men and women; people with shorter and longer durations of MS; and people with different levels of mobility.

Have the people in the sample been measured successfully?

Are the persons in the sample separated along the line defined by the items?

This question was addressed by examining the distribution statistics of the sample and the PSI.

Do individual placements on the variable make sense?

This question was addressed by examining the relationship with other measures and group differences.

How valid is each person's measurement?

This question was addressed by examining the person–item fit residuals.

Results I: MSIS-29 physical impact subscale

Is the scale-to-sample targeting adequate for making judgements about the performance of the scale and the measurement of people?

Figure 15 shows the targeting of the patient sample to the location's 20 items of the physical impact subscale. Item locations range from about −1 to +1 logits. Person locations range from about −4.5 logits to +4.5 logits. Thus, the sample covers the scale well, indicating that this is a good sample for examining the performance of the scale. In contrast, the scale does not cover the full range of the sample, implying that the scale is suboptimal for measuring the people in the sample. However, each item location represents the mean of four thresholds. *Figure 16* shows the targeting of the patient sample to the thresholds of the 20 items of the physical impact subscale. Item locations now spread from about −2 logits to +2 logits but still do not cover the range of person locations in the sample.

Has a measurement ruler been constructed successfully?

Do the item response categories work as intended?

Table 15 shows the threshold estimates and the location estimates (mean of the four threshold estimates) for the 20 items of the physical impact subscale. Nine items had reversed thresholds (1, 4, 12, 14, 15 and 17–20), i.e. the threshold estimates were not ordered sequentially. For example, for item 1, the estimate for $\tau 3$ (−0.861) is less than that for $\tau 2$ (−0.649). This means that the estimated point on the continuum at which the probability of scoring either 3 ('moderately') or 4 ('quite a bit') is the same is lower than the point on the continuum at which the probability of scoring either 2 ('a little') or 3 ('moderately') is the same. Clearly, this does not make sense and indicates that the response categories are not working as intended.

All nine items with reversed thresholds have the same pattern of reversal, i.e. threshold $\tau 3$ is less than threshold $\tau 2$. This finding implies that people are not reliably discriminating between the three middle response categories: 'a little', 'moderately' and 'quite a bit'. This consistent finding implies

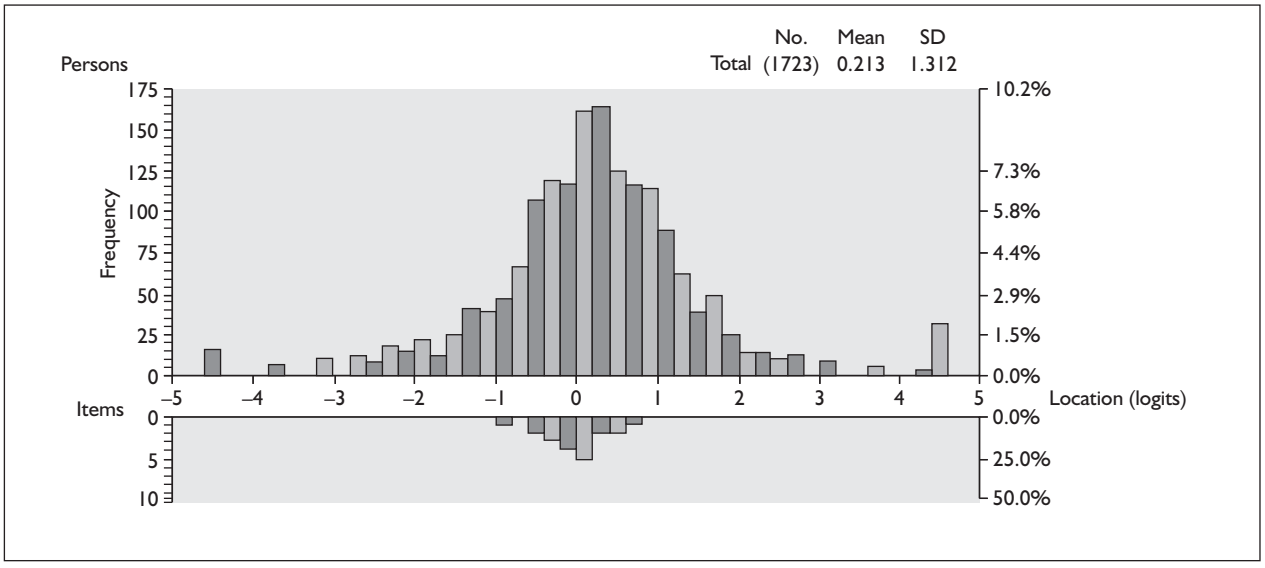


FIGURE 15 *MSIS-29 physical subscale – targeting of sample to item locations. Person–item location distribution (grouping set to interval length of 0.20, making 50 groups).*

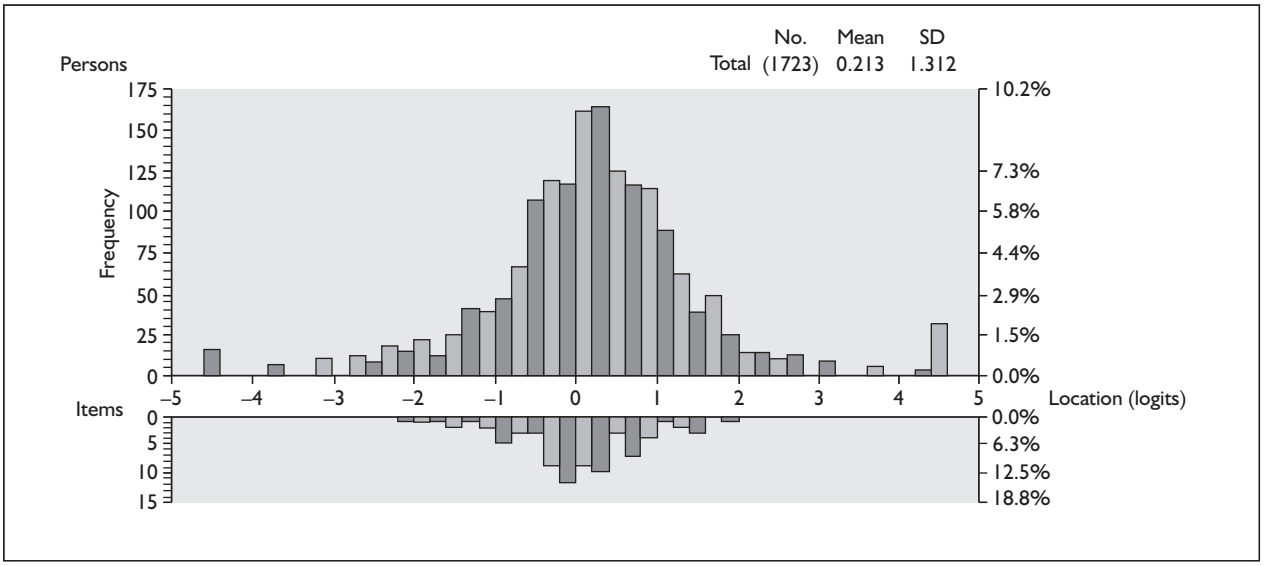


FIGURE 16 *MSIS-29 physical subscale – targeting of sample to item locations. Person–item threshold distribution (grouping set to interval length of 0.20, making 50 groups).*

TABLE 15 *MSIS-29 physical impact subscale: item thresholds and location estimates*

		Threshold estimates				
Item	**Location**	**τ1**	**τ2**	**τ3**	**τ4**	**RT**
1	–0.880	–2.165	–0.649	–0.861	0.155	X
2	0.443	–0.363	0.299	0.644	1.194	
3	–0.190	–1.011	–0.319	–0.029	0.598	
4	–0.427	–1.589	–0.353	–0.423	0.658	X
5	0.022	–0.983	–0.168	0.334	0.903	
6	0.107	–1.632	0.183	0.313	1.565	
7	0.106	–0.916	–0.112	0.101	1.352	
8	–0.240	–1.248	–0.289	–0.257	0.833	
9	0.764	0.033	0.366	0.702	1.955	
10	0.513	–0.336	0.371	0.503	1.513	
11	–0.239	–1.465	–0.149	–0.133	0.791	
12	–0.091	–1.106	0.170	–0.192	0.766	X
13	0.291	–0.618	0.187	0.355	1.239	
14	0.080	–0.276	0.036	–0.105	0.664	X
15	0.392	–0.587	0.406	0.299	1.450	X
16	–0.030	–0.822	–0.142	0.041	0.802	
17	0.179	–0.097	0.245	0.219	0.351	X
18	–0.443	–1.906	–0.284	–0.409	0.829	X
19	–0.229	–0.725	–0.115	–0.204	0.130	X
20	–0.129	–0.892	–0.112	–0.144	0.632	X

Location, mean of thresholds; RT, reversed threshold; X, item with reversed thresholds.
τ1, point at which probability of responding 'not at all' and 'a little' is the same.
τ2, point at which probability of responding 'a little' and 'moderately' is the same.
τ3, point at which probability of responding 'moderately' and 'quite a bit' is the same.
τ4 = point at which probability of responding 'quite a bit' and 'extremely' is the same.

that there may be too many response categories or that the wording attached to the response categories is difficult for people to relate to in practice.

The category probability curve (CPC) for each item shows these relationships graphically. *Figure 17* is the CPC for item 5. This graph plots the probability of a response (*y*-axis) against a person's location on the physical functioning continuum mapped out by the 20 items of the MSIS-29 physical impact subscale. Each of the four lines, labelled 0–4, represents the probability of responding to category 0–4. (Note that the RUMM2020 progam has renumbered the categories 1–5 as categories 0–4.)

For the MSIS-29, high scores indicate more disability (the reverse of the RMI). Thus, as one moves from left to right along the *x*-axis, people become more physically disabled. Logically, we would expect that, as a person's level of physical disability increases, the probability of that person responding 'not at all' (the curve labelled 0 on the graph) falls and the probability of responding 'a little' (the curve labelled 1 on the graph) increases and becomes the most likely response category chosen. As disability increases, the probability of responding 'a little' increases, then decreases, and the probability of responding 'moderately' (the curve labelled 2 on the graph) becomes the most likely response category chosen. As the person becomes more disabled, the probability of responding 'quite a bit' (the curve labelled 3 on the graph) increases, becomes the most likely chosen category and is finally replaced by 'extremely' as the most likely response category. Thus, the response categories of item 5 are working as intended. The intersections between adjacent curves (0 and 1; 1 and 2; 2 and 3; 3 and 4) represent the points at which the probability of responding to either of a pair of adjacent response categories (e.g. 0 and 1) are the same. These points mark the four thresholds for item 5, they are ordered sequentially along the continuum (−0.983; −0.168; +0.334; +0.903), and so the response categories of item 5 are working as intended (*Figure 17*).

Figure 18 shows the CPC for item 1, one of the items with reversed thresholds. This graph shows that there is no point on the continuum at which response categories 2 and 3 have the highest probability of being chosen. Thus, the intersection of response categories 1 and 2 ($\tau 2 = -0.649$) is above the intersection of response category 2 ($\tau 3 = -0.861$). Hence the thresholds are disordered and the response categories are not working as intended.

Examination of the CPCs of all the nine items with disordered thresholds shows a similar pattern to *Figure 18*: the category labelled 2 ('moderately') is never the category with the highest probability of being endorsed. Thus, people whose disability location is at the relevant range of the continuum are more likely to choose either of the two flanking categories ('a little' or 'quite a bit'). It is important

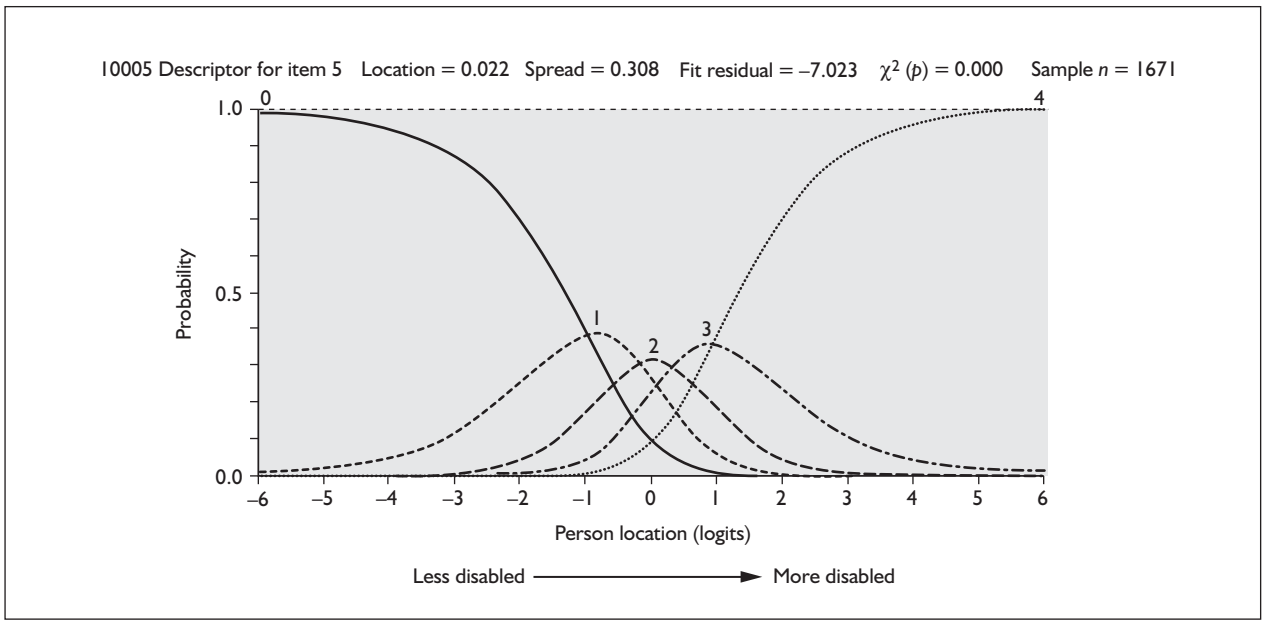


FIGURE 17 *MSIS-29 physical subscale – category probability curve for item 5.*

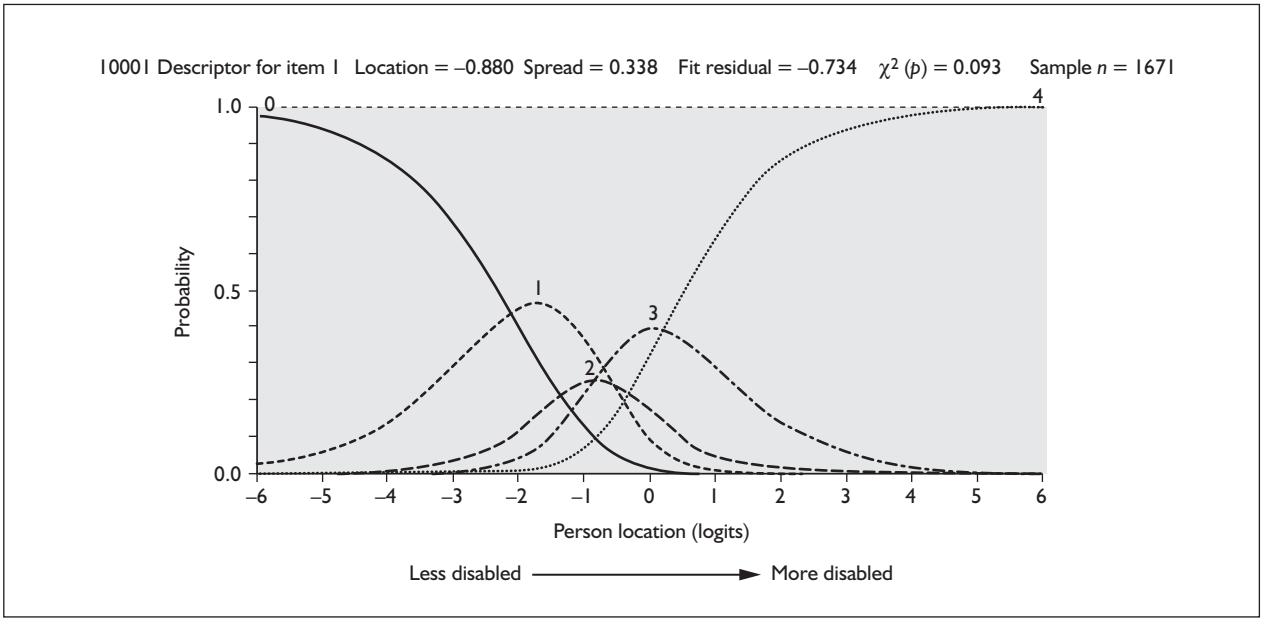


FIGURE 18 *MSIS-29 physical subscale – category probability curve for item 1.*

to recognise that this does not mean that the category has not been endorsed. Indeed, 13% of the sample ($n = 206$) endorsed 'moderately' for item 1.

Do the items map out a discernible line of increasing intensity?

Table 15 shows that the item locations (mean of four thresholds) range from −0.880 (item 1) to +0.764 (item 9), and that the item thresholds range from −2.165 (item 1 τ1) to +1.955 (item 9 τ4). Thus, the items spread out and map out a continuum. However, the spread of item locations is less than the item range for the RMI (−3.0 to +6.0).

Figure 19 shows the continuum mapped out by the 20 items. Items tend to be bunched towards the centre of the scale range. There are notable gaps in the continuum towards the extremes, indicating areas where there is suboptimal measurement and pointing to sites for improvement.

Is the location of items along this line reasonable?

The ordering of items along the continuum mapped out by the MSIS-29 physical impact subscale makes clinical sense. The items with the most negative locations, i.e. those items that are first to become problematic when people develop disability, are 'doing physically demanding tasks', 'taking longer to do things' and 'problems with balance'. Similarly, the items at the other end of continuum, which are the items that typically present a problem at greater levels of disability, are 'difficulties using hands', 'gripping things' and 'muscle spasms'.

It is, however, easier to judge the suitability of item ordering when items have dichotomous response options. This is because they have a single location. In contrast, when items have multiple response options, the single location estimates are, by definition, a summary of their multiple thresholds.

Do the items work together to define a single variable?

Item–person fit residuals

Table 16 shows the fit statistics in the order the items appear on the questionnaire. *Table 17* shows them ordered by fit residual magnitude. *Figure 20* shows the values diagrammatically with boundary lines drawn at −2.5 and +2.5.

Nine items fall within the recommended range (1, 3, 4, 6, 13, 14, 15, 17 and 19). A further two items (2 and 16) are near the boundary lines. One item (item 20, 'needing to go to the toilet urgently') is very misfitting, and eight items are notably misfitting.

Chi-squared values and probabilities

Table 16 shows the fit statistics in item sequence order. *Table 18* shows them ordered by chi-squared values. *Figure 21* shows the values diagrammatically.

Figure 21 shows that item 20 misfits substantially compared with the rest. Three further items (5, 12 and 18) are more misfitting than the main group.

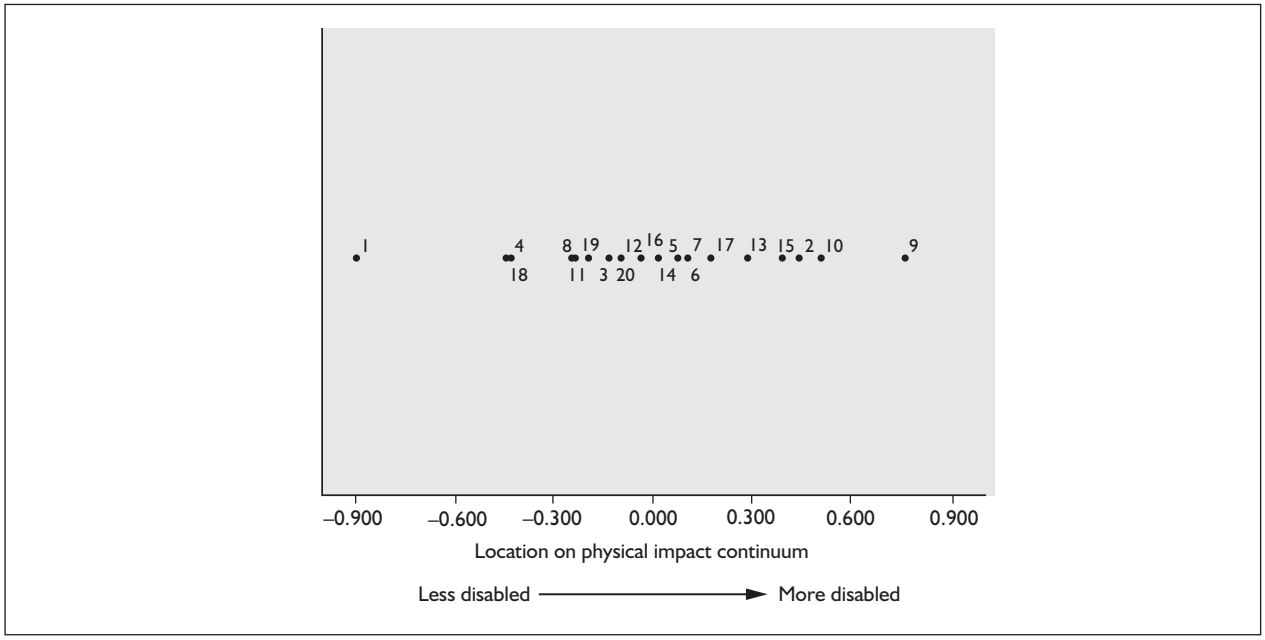


FIGURE 19 *MSIS-29 physical subscale – continuum mapped out by the 20 physical impact items.*

TABLE 16 *MSIS-29 physical impact subscale: item locations, standard errors and fit statistics [*n = *1671 (adjusted to* n = *500 for calculation of* χ^2*); 10 class intervals]*

Item				Fit statistics		
Number	**Label**	**Location**	**Standard error**	**Fit residual**	χ^2	χ^2 **probability**
1	Do physically demanding tasks	–0.880	0.029	–0.734	4.468	0.878
2	Grip things tightly	0.443	0.025	3.714	4.095	0.905
3	Carry things	–0.190	0.026	–1.851	5.795	0.760
4	Problems with your balance	–0.427	0.027	1.501	4.014	0.911
5	Difficulties moving about indoors	0.022	0.026	–7.023	26.168	0.002
6	Being clumsy	0.107	0.027	–0.307	5.478	0.791
7	Stiffness	0.106	0.026	6.002	10.882	0.284
8	Heavy arms and/or legs	–0.240	0.026	5.728	7.816	0.553
9	Tremor of your arms or legs	0.764	0.026	5.789	12.350	0.194
10	Spasms in your limbs	0.513	0.025	4.901	17.612	0.040
11	Your body not doing what you want it to do	–0.239	0.026	–5.300	15.268	0.084
12	Having to depend on others to do things for you	–0.091	0.025	–8.122	28.297	0.001
13	Limitations in social and leisure activities at home	0.291	0.025	0.657	3.981	0.913
14	Being stuck at home more than you would like	0.080	0.024	1.358	2.409	0.983
15	Difficulties using your hands in everyday tasks	0.392	0.025	–1.842	6.086	0.731
16	Having to cut down time spent on work/daily activities	–0.030	0.025	3.084	9.160	0.423
17	Problems using transport	0.179	0.023	1.079	3.807	0.924
18	Taking longer to do things	–0.443	0.027	–7.887	29.895	0.000
19	Difficulty doing things spontaneously	–0.229	0.024	–0.880	6.074	0.733
20	Needing to go to the toilet urgently	–0.129	0.025	15.845	73.537	0.000

TABLE 17 *MSIS-29 physical impact subscale: items ordered by increasing item–person fit residual [n = 1671 (adjusted to n = 500 for calculation of χ^2); 10 class intervals]*

Item				Fit statistics		
Number	Label	Location	Standard error	Fit residual	χ^2	χ^2 probability
12	Having to depend on others to do things for you	–0.091	0.025	–8.122	28.297	0.001
18	Taking longer to do things	–0.443	0.027	–7.887	29.895	0.000
5	Difficulties moving about indoors	0.022	0.026	–7.023	26.168	0.002
11	Your body not doing what you want it to do	–0.239	0.026	–5.300	15.268	0.084
3	Carry things	–0.190	0.026	–1.851	5.795	0.760
15	Difficulties using your hands in everyday tasks	0.392	0.025	–1.842	6.086	0.731
19	Difficulty doing things spontaneously	–0.229	0.024	–0.880	6.074	0.733
1	Do physically demanding tasks	–0.880	0.029	–0.734	4.468	0.878
6	Being clumsy	0.107	0.027	–0.307	5.478	0.791
13	Limitations in social and leisure activities at home	0.291	0.025	0.657	3.981	0.913
17	Problems using transport	0.179	0.023	1.079	3.807	0.924
14	Being stuck at home more than you would like	0.080	0.024	1.358	2.409	0.983
4	Problems with your balance	–0.427	0.027	1.501	4.014	0.911
16	Having to cut down time spent on work/ daily activities	–0.030	0.025	3.084	9.160	0.423
2	Grip things tightly	0.443	0.025	3.714	4.095	0.905
10	Spasms in your limbs	0.513	0.025	4.901	17.612	0.040
8	Heavy arms and/or legs	–0.240	0.026	5.728	7.816	0.553
9	Tremor of your arms or legs	0.764	0.026	5.789	12.350	0.194
7	Stiffness	0.106	0.026	6.002	10.882	0.284
20	Needing to go to the toilet urgently	–0.129	0.025	15.845	73.537	0.000

The chi-squared probability for these four items is < 0.01, indicating that the difference between observed scores and expected values is large relative to chance.

The remaining 16 items have more similar-level chi-squared values and *Table 18* shows that there is a gradual increase in chi-squared value from 2.4 (item 14) to 17.6 (item 10) before there is a notable step increase in value (item 5; chi-squared value = 26.2).

Item characteristic curves

Figure 22 shows the ICCs for the four items that 'failed' both the item–person fit residual and the item–trait chi-squared value tests of fit. All four items show reasonable adherence between the curve defined by the 10 class intervals (black dots) representing the observed data and the ICC curve representing the expected values derived from the Rasch model. *Figure 22* implies less concern about the degree of misfit of these items than is implied by the values reported in Tables 16–18 and by *Figures 20* and *21*.

Figures 23 and *24* show the ICCs for the six items (2, 7, 8, 9, 10 and 11) for which fit residuals were outside the recommended range (−2.5 to +2.5). For all these items the visual discrepancy between observed scores and expected values across 10 class intervals is small. For comparison, *Figure 25* shows the ICCs for two items (1 and 17) that passed both the item–person fit residual and item–trait chi-squared tests of fit.

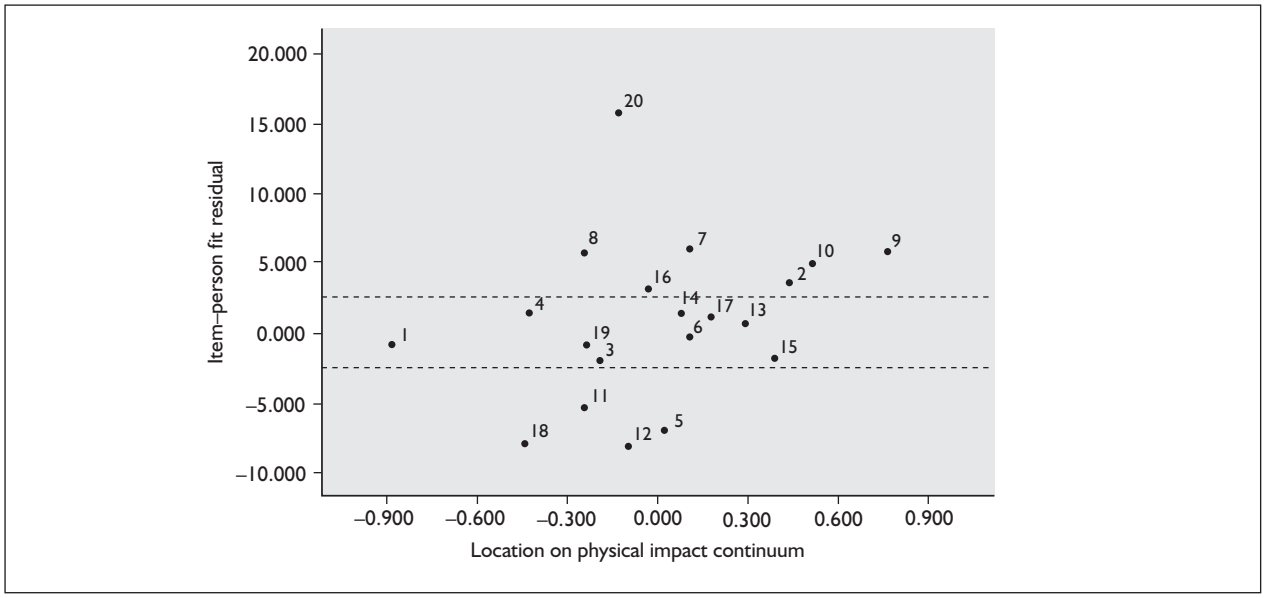


FIGURE 20 *MSIS-29 physical subscale – plot of item–person fit residuals against location.*

TABLE 18 *MSIS-29 physical impact subscale: items ordered by increasing chi-squared value [*n = *1671 (adjusted to* n = *500 for calculation of* χ^2*); 10 class intervals]*

Item				Fit statistics		
Number	**Label**	**Location**	**Standard error**	**Fit residual**	χ^2	χ^2 **probability**
14	Being stuck at home more than you would like	0.080	0.024	1.358	2.409	0.983
17	Problems using transport	0.179	0.023	1.079	3.807	0.924
13	Limitations in social and leisure activities at home	0.291	0.025	0.657	3.981	0.913
4	Problems with your balance	–0.427	0.027	1.501	4.014	0.911
2	Grip things tightly	0.443	0.025	3.714	4.095	0.905
1	Do physically demanding tasks	–0.880	0.029	–0.734	4.468	0.878
6	Being clumsy	0.107	0.027	–0.307	5.478	0.791
3	Carry things	–0.190	0.026	–1.851	5.795	0.760
19	Difficulty doing things spontaneously	–0.229	0.024	–0.880	6.074	0.733
15	Difficulties using your hands in everyday tasks	0.392	0.025	–1.842	6.086	0.731
8	Heavy arms and/or legs	–0.240	0.026	5.728	7.816	0.553
16	Having to cut down time spent on work/daily activities	–0.030	0.025	3.084	9.160	0.423
7	Stiffness	0.106	0.026	6.002	10.882	0.284
9	Tremor of your arms or legs	0.764	0.026	5.789	12.350	0.194
11	Your body not doing what you want it to do	–0.239	0.026	–5.300	15.268	0.084
10	Spasms in your limbs	0.513	0.025	4.901	17.612	0.040
5	Difficulties moving about indoors	0.022	0.026	–7.023	26.168	0.002
12	Having to depend on others to do things for you	–0.091	0.025	–8.122	28.297	0.001
18	Taking longer to do things	–0.443	0.027	–7.887	29.895	0.000
20	Needing to go to the toilet urgently	–0.129	0.025	15.845	73.537	0.000

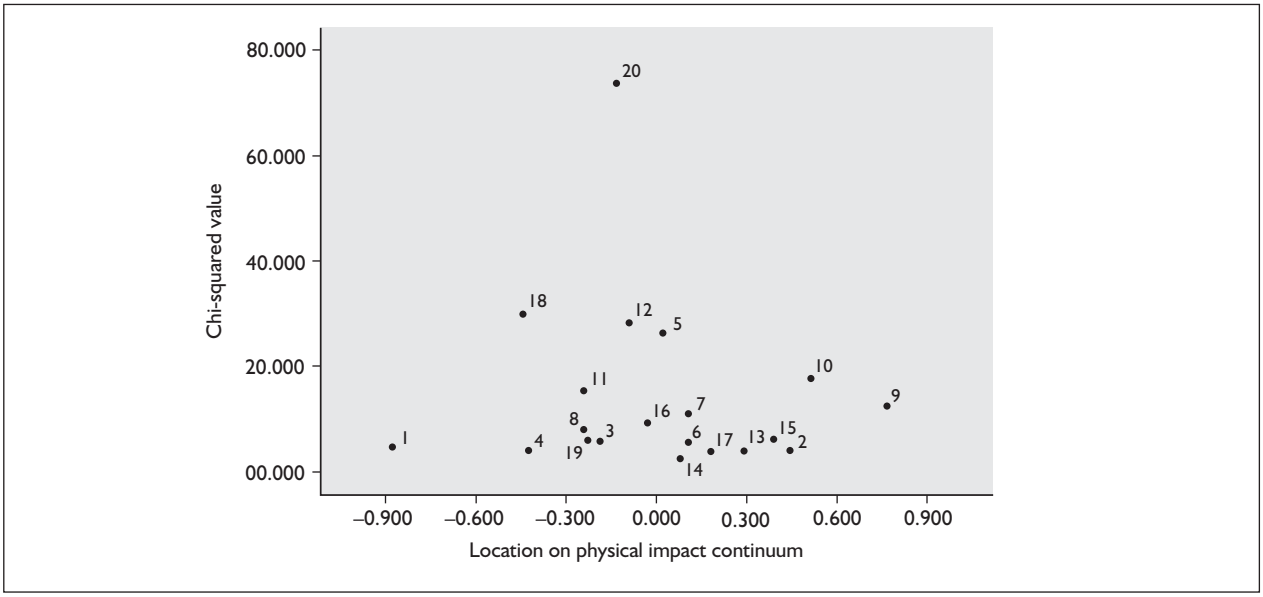


FIGURE 21 *MSIS-29 physical subscale – plot of chi-squared values against location.*

Note that *Table 19* shows the locations, errors and fit statistics with the sample *not* adjusted to $n = 500$ to demonstrate that only the chi-squared values and chi-squared probabilities are affected.

Does the response to one item directly influence the response to another?
In general, the correlations between the residuals were low. Only four of the 190 correlations exceeded 0.30, and none exceeded 0.40. The pairs of items whose residual correlations exceeded 0.30, and their values were:

Items 2 and 3 = 0.32

Items 2 and 15 = 0.37

Items 9 and 10 = 0.38

Items 13 and 14 = 0.33

Are the locations of the items stable across clinically important groups?
We examined differential item functioning (DIF) in relation to five clinically important subgroupings: sex (men and women); mobility level (unaided; with an aid; in a wheelchair); MS type (relapsing–remitting MS; secondary progressive MS; primary progressive MS); educational level (higher degree/qualification or not); sample (rehabilitation; steroids; first postal survey; second postal survey).

In the context of these subgroupings, DIF was only demonstrated in one circumstance (mobility level). Here, three items (5, 12 and 20) demonstrated statistically significant DIF. However, a number of class intervals had very small numbers (approximately 15). Nevertheless, re-evaluation of DIF using two class intervals produced the same results.

Have the people in the sample been measured successfully?

Are the persons in the sample separated along the line defined by the items?
Figures 15 and *16* show the distribution of person locations compared with the item locations (see *Figure 15*) and item thresholds (see *Figure 16*). It is clear from these two figures that the sample locations are spread across the continuum. This is reflected in the PSI, which is 0.955, indicating that the 20 physical impact items are able to distinguish reliably between responders on the trait they measure.

Figure 26 is a plot of raw scores against the interval measures they imply. The relationship is S-shaped (an ogive) rather than linear (a straight linear). This indicates that equal changes in one metric are not associated with equal changes in the other. The correlation between raw scores and interval measures is 0.952.

Table 20 gives the interval-level location values that correspond with each raw score. Note that a change from 0 to 1 raw score point corresponds to a change in interval measure of 0.81 logits. A change from 42 to 43 raw score points corresponds to a

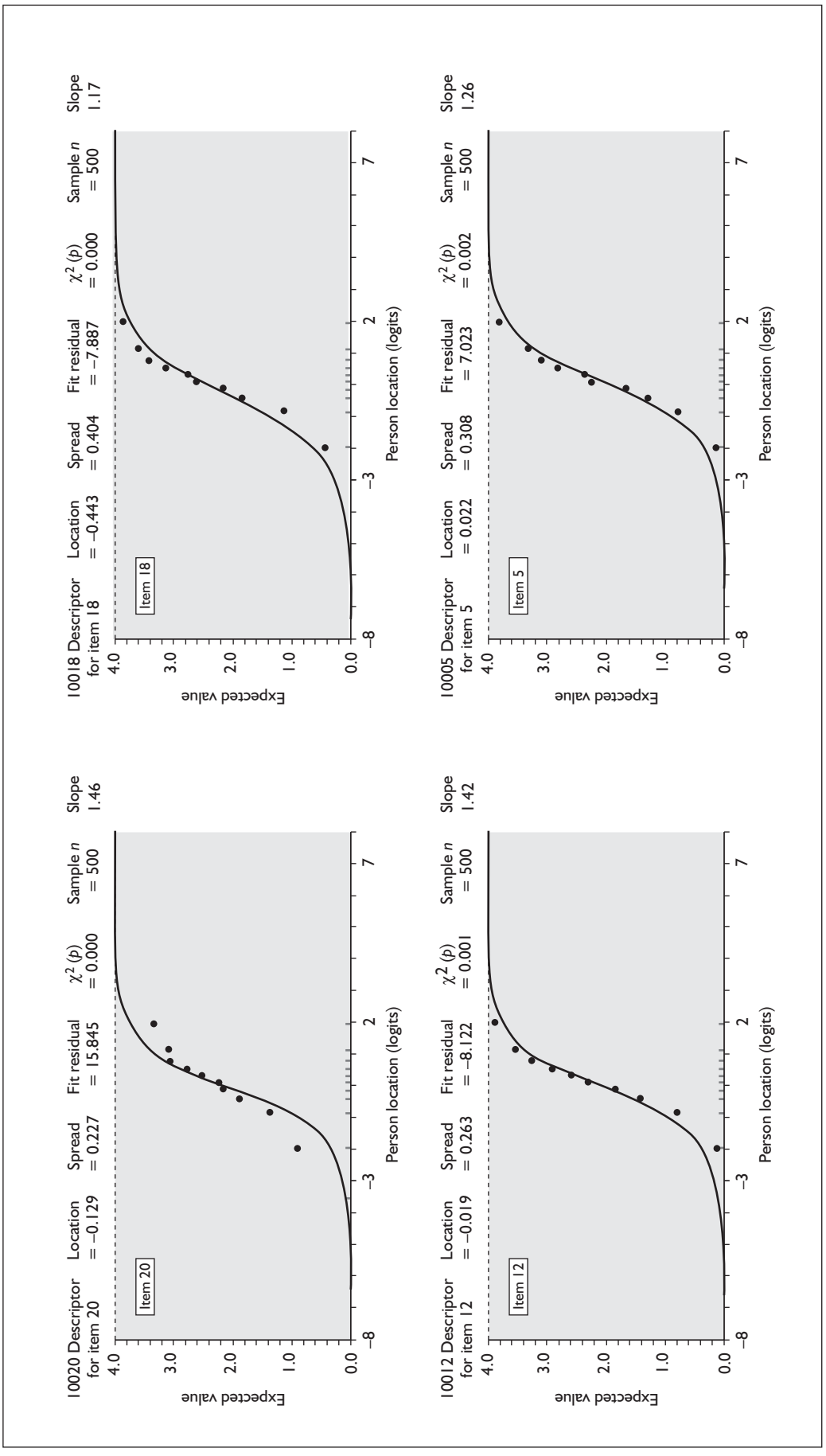


FIGURE 22 *MSIS-29 physical subscale – item characteristic curves for items 20, 18, 12 and 5.*

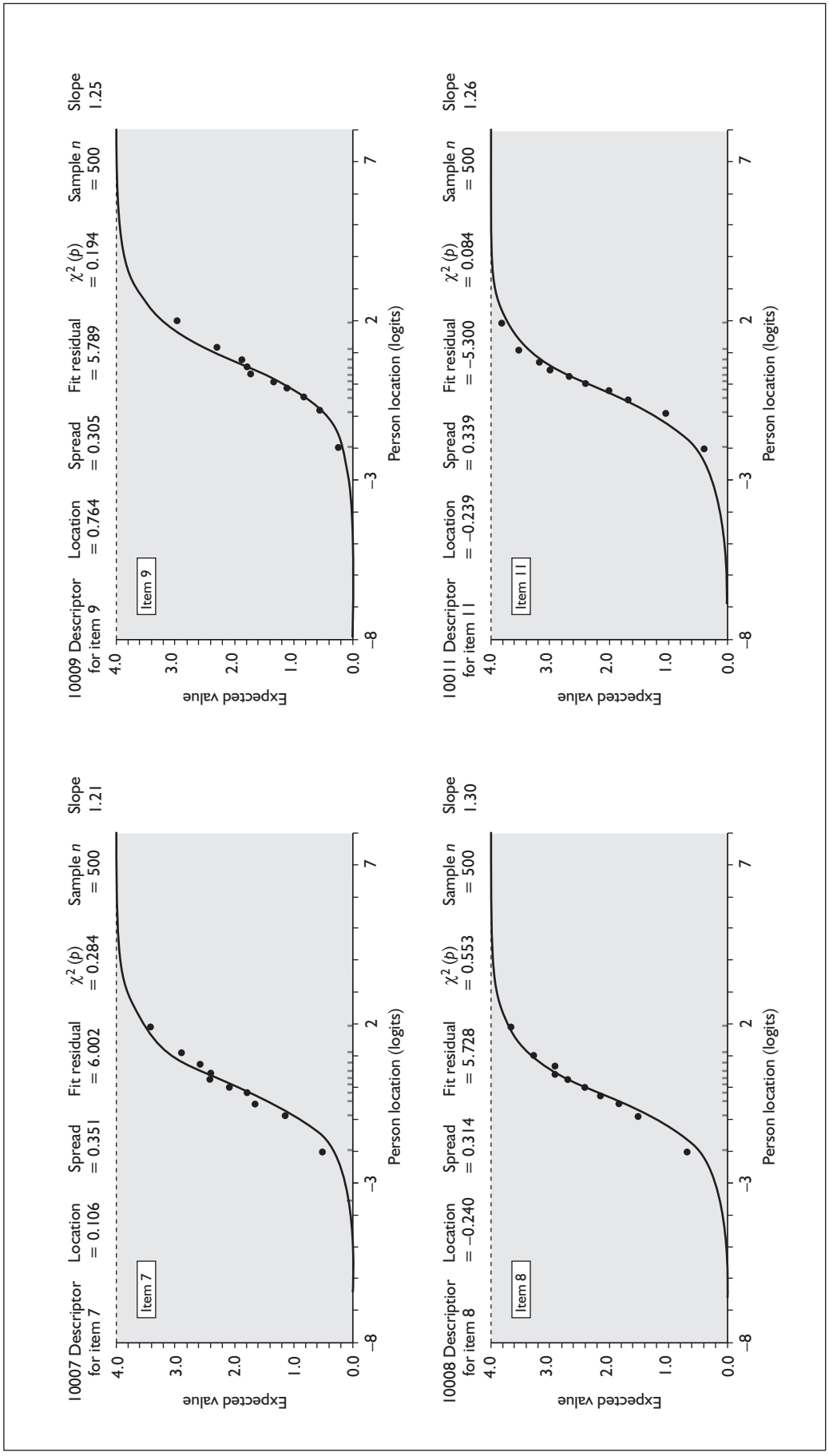


FIGURE 23 *MSIS-29 physical subscale – item characteristic curves for items 7, 9, 8 and 11.*

DOI: 10.3310/hta13120

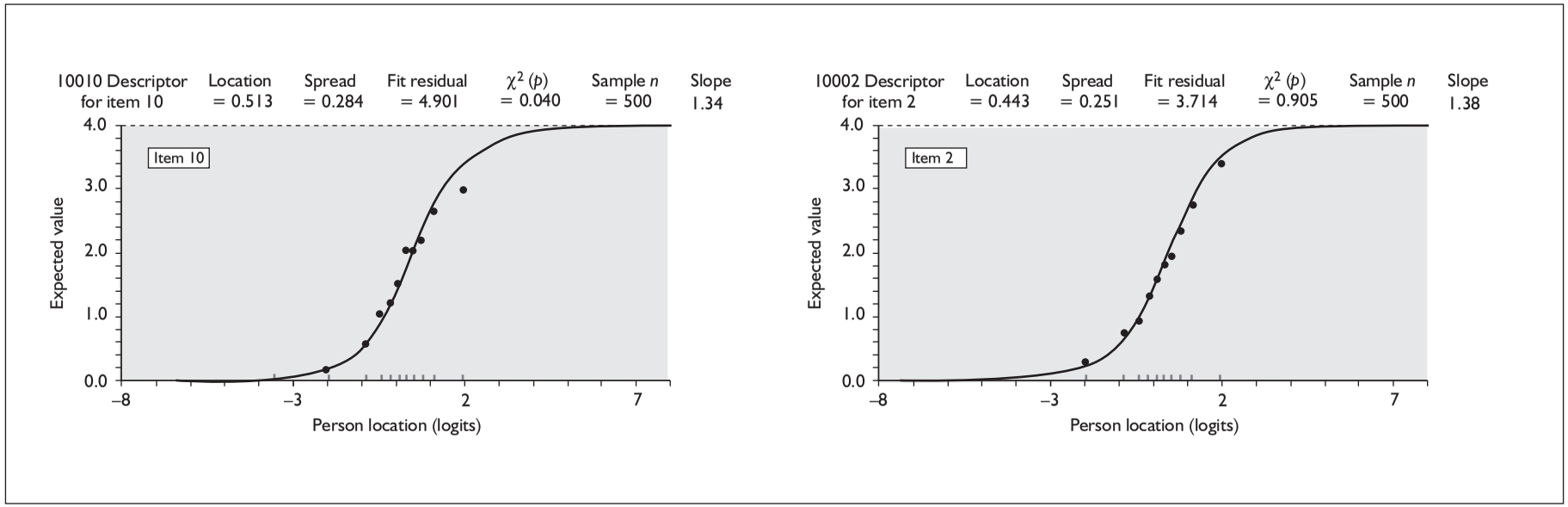


FIGURE 24 *MSIS-29 physical subscale – item characteristic curves for items 10 and 2.*

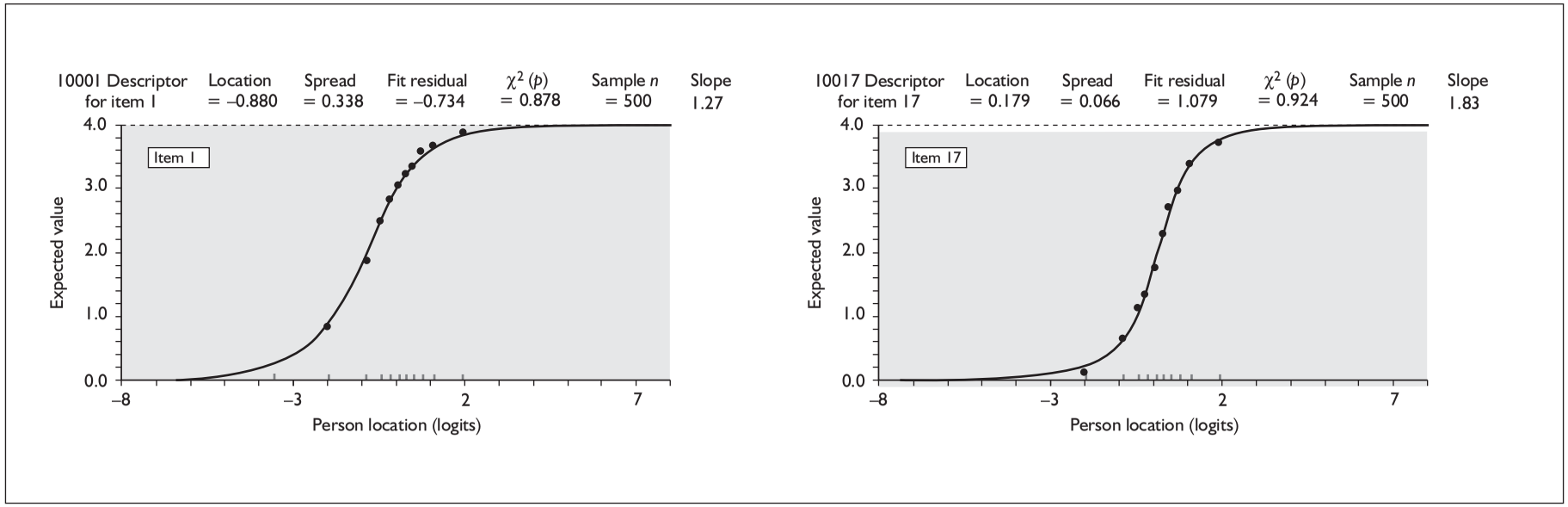


FIGURE 25 *MSIS-29 physical subscale – item characteristic curves for items 1 and 17.*

TABLE 19 *MSIS-29 physical impact subscale: item locations, standard errors and fit statistics [n = 1671 (sample size not adjusted for χ^2 calculation); 10 class intervals]*

Item				Fit statistics		
Number	Label	Location	Standard error	Fit residual	χ^2	χ^2 probability
1	Do physically demanding tasks	–0.880	0.029	–0.734	14.931	0.093
2	Grip things tightly	0.443	0.025	3.714	13.684	0.134
3	Carry things	–0.190	0.026	–1.851	19.366	0.022
4	Problems with your balance	–0.427	0.027	1.501	13.414	0.145
5	Difficulties moving about indoors	0.022	0.026	–7.023	87.452	0.000
6	Being clumsy	0.107	0.027	–0.307	18.306	0.032
7	Stiffness	0.106	0.026	6.002	36.369	0.000
8	Heavy arms and/or legs	–0.240	0.026	5.728	26.120	0.002
9	Tremor of your arms or legs	0.764	0.026	5.789	41.273	0.000
10	Spasms in your limbs	0.513	0.025	4.901	58.860	0.000
11	Your body not doing what you want it to do	–0.239	0.026	–5.300	51.027	0.000
12	Having to depend on others to do things for you	–0.091	0.025	–8.122	94.570	0.000
13	Limitations in social and leisure activities at home	0.291	0.025	0.657	13.306	0.149
14	Being stuck at home more than you would like	0.080	0.024	1.358	8.052	0.529
15	Difficulties using your hands in everyday tasks	0.392	0.025	–1.842	20.339	0.016
16	Having to cut down time spent on work/ daily activities	–0.030	0.025	3.084	30.614	0.000
17	Problems using transport	0.179	0.023	1.079	12.722	0.176
18	Taking longer to do things	–0.443	0.027	–7.887	99.910	0.000
19	Difficulty doing things spontaneously	–0.229	0.024	–0.880	20.298	0.016
20	Needing to go to the toilet urgently	–0.129	0.025	15.845	245.762	0.000

change in interval measure of 0.03 logits. Thus, the implication of a 1-point change in raw score varies up to 27 times across the range of the scale.

Table 20 also includes the standard errors associated with each location. These vary across the range of the scale, being largest at the extremes and smallest in the centre of the scale range. *Figure 27* shows the relationship between standard error and location on the continuum. The relationship is U-shaped and shows how much the precision of measurement falls as locations move away from the centre of the scale. *Figure 28* also shows this relationship, superimposed on the sample and item distribution. Essentially, *Table 20* and *Figures 27* and *28* show that good measurement occurs only at the centre of the scale range.

It is important to note that the raw scores, interval locations and standard errors shown in *Table 20* are for complete data. By complete data we mean people who have endorsed a response category for all 20 items. The location estimates for people who have not completed all 20 items will depend on the items they endorsed and the total score they achieved. This could be overcome by imputing for missing data using a person-specific mean score (as discussed earlier).

Another point to note is that the location estimate and standard error corresponding to each raw score is the same irrespective of the pattern of item responses (provided that all 20 items have been endorsed). This does not mean that the pattern of responses is irrelevant. On the contrary, in Rasch

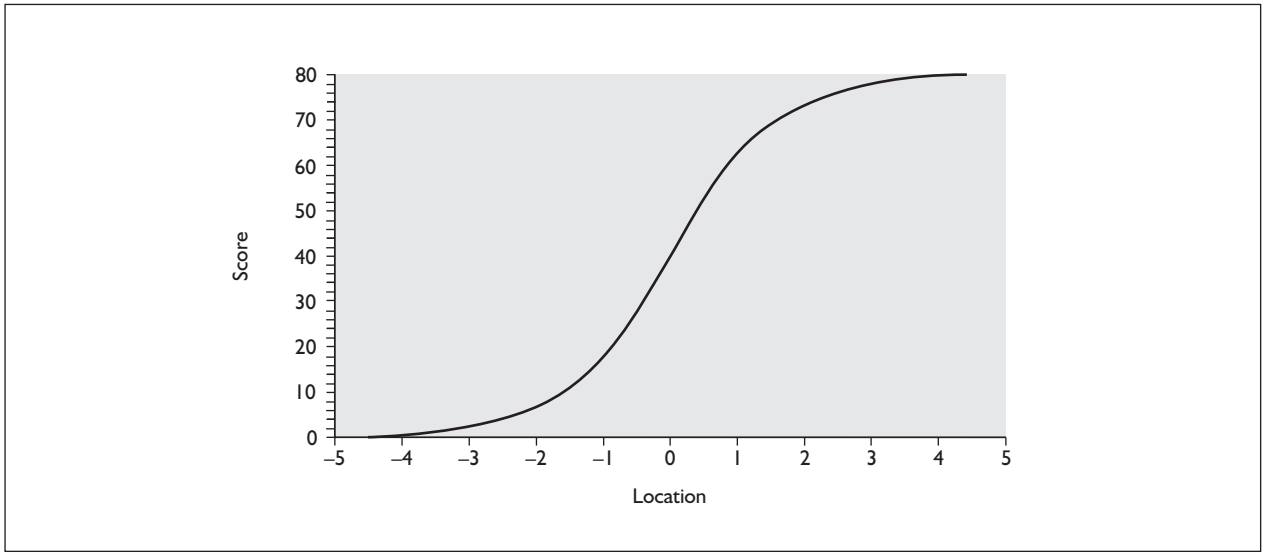


FIGURE 26 *MSIS-29 physical subscale – plot of raw scores against interval measures.*

TABLE 20 *MSIS-29 physical impact subscale: interval-level locations implied by each ordinal-level raw score*

Raw score	Interval measure	Standard error
0	–4.490[a]	1.180
1	–3.684	0.840
2	–3.144	0.650
3	–2.783	0.560
4	–2.510	0.490
5	–2.291	0.450
6	–2.108	0.410
7	–1.950	0.380
8	–1.813	0.360
9	–1.691	0.340
10	–1.582	0.320
11	–1.482	0.310
12	–1.391	0.300
13	–1.307	0.290
14	–1.229	0.280
15	–1.156	0.270
16	–1.088	0.260
17	–1.023	0.250
18	–0.961	0.250
19	–0.903	0.240
20	–0.847	0.240
21	–0.793	0.230
22	–0.741	0.230
23	–0.691	0.220

continued

TABLE 20 *MSIS-29 physical impact subscale: interval-level locations implied by each ordinal-level raw score*

Raw score	Interval measure	Standard error
24	–0.643	0.220
25	–0.596	0.220
26	–0.551	0.210
27	–0.507	0.210
28	–0.463	0.210
29	–0.421	0.210
30	–0.380	0.210
31	–0.339	0.200
32	–0.298	0.200
33	–0.259	0.200
34	–0.220	0.200
35	–0.181	0.200
36	–0.142	0.200
37	–0.104	0.200
38	–0.066	0.200
39	–0.028	0.200
40	0.010	0.200
41	0.047	0.200
42	0.085	0.200
43	0.123	0.200
44	0.161	0.200
45	0.199	0.200
46	0.238	0.200
47	0.277	0.200
48	0.316	0.200
49	0.356	0.200
50	0.397	0.200
51	0.438	0.210
52	0.480	0.210
53	0.523	0.210
54	0.566	0.210
55	0.611	0.220
56	0.657	0.220
57	0.705	0.220
58	0.754	0.230
59	0.805	0.230
60	0.857	0.230
61	0.912	0.240
62	0.970	0.240
63	1.030	0.250
64	1.094	0.260
65	1.161	0.260

TABLE 20 *MSIS-29 physical impact subscale: interval-level locations implied by each ordinal-level raw score*

Raw score	Interval measure	Standard error
66	1.232	0.270
67	1.308	0.280
68	1.390	0.290
69	1.478	0.300
70	1.575	0.320
71	1.681	0.330
72	1.799	0.350
73	1.931	0.380
74	2.083	0.400
75	2.260	0.440
76	2.471	0.480
77	2.735	0.550
78	3.086	0.650
79	3.613	0.830
80	4.400[a]	1.170

a Note that the measurements for people at the extremes of the scale range (floor and ceiling) are extrapolation based on actual estimates. For these people no person–item fit residual can be computed.

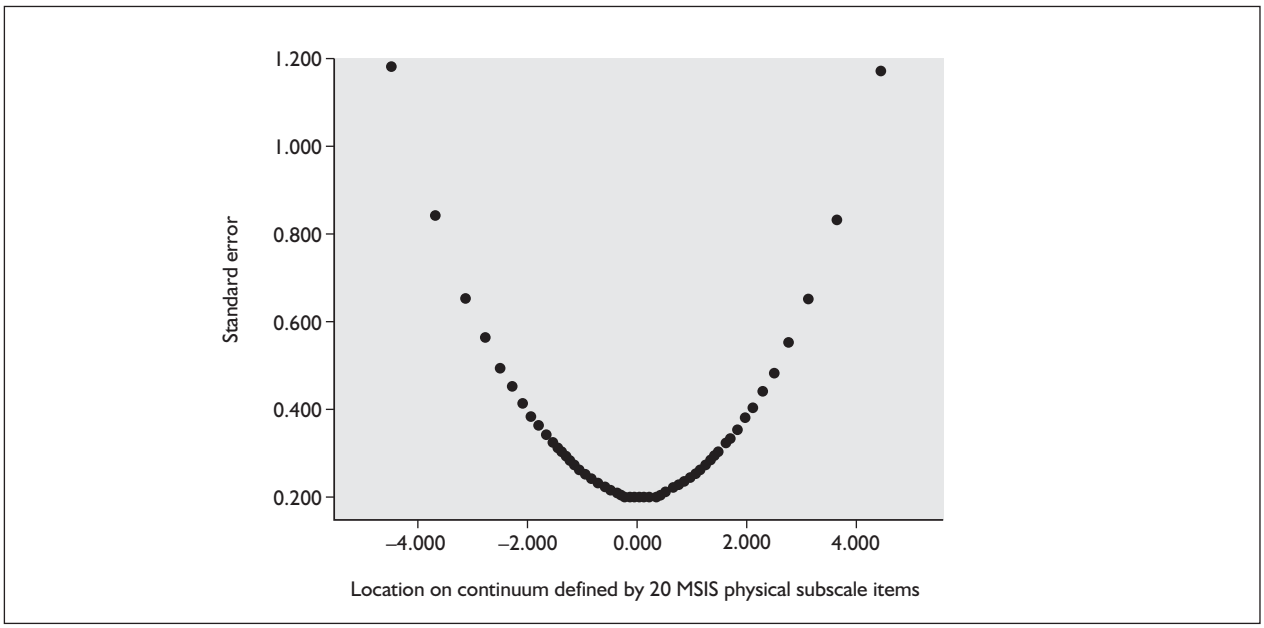


FIGURE 27 *MSIS-29 physical subscale – plot of standard error against location.*

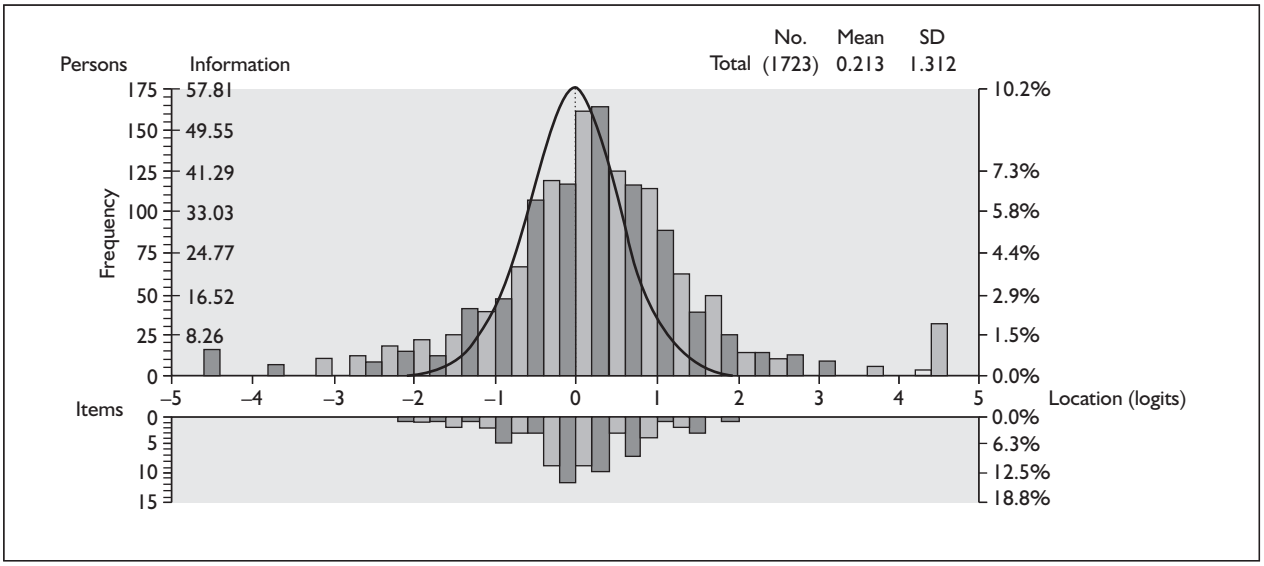


FIGURE 28 *MSIS-29 physical subscale – information from scale relative to sample. Person–item threshold distribution (grouping set to interval length of 0.20, making 50 groups).*

analysis differences in the pattern of item responses are reflected in the person–item fit statistic (discussed below).

Finally, as noted before, the range of the total scores in *Table 20* is 0–80. The response categories on the questionnaire are labelled 1–5, and thus the total score for the 20 physical impact items ranges from 20 to100. The reason for the difference is that in the RUMM2020 program the first item response option is labelled 0. To convert from questionnaire-determined total score to RUMM2020-determined total score simply subtract 20.

Do individual placements on the variable make sense?
The placements of individuals on the physical function continuum mapped out by the items was largely consistent with expectation. For example, people who walked without an aid had worse physical function than people who used a walking aid, and people who used a walking aid had worse physical function than people who used a wheelchair.

How valid is each person's measurement?
Person–item fit residuals indicate the extent to which the responses of a specific individual are consistent with expectation based on the Rasch model. They are analogous to the item–person fit residuals that indicate the extent to which the responses to a specific item are consistent with expectation based on the Rasch model. It is recommended that person–item fit residuals lie in the range −2.5 to +2.5.

In our sample of 1723 people, person–item fit residuals could be computed for 1671, as they could not be computed for the 52 people who scored at the extremes. Values were outside the recommended range for 189 people (11% of the sample). These exceeded −3.0 to +3.0 in only 16 people (1%).

Summary of results of Rasch analysis of the MSIS-29 physical impact subscale

Rasch analysis of the 20-item MSIS-29 physical subscale identified a number of problems. The five-category scoring function for the items did not work as intended for half of the items. The variable mapped out by the items was not very wide, and there was bunching of thresholds at the centre of the scale. The items did not work as well together to map out a variable as implied by traditional psychometric analyses. There was a small amount of differential item functioning. These limitations were not detected by traditional psychometric methods.

Results II: MSIS-29 psychological impact subscale

Is the scale-to-sample targeting adequate for making judgements about the performance of the scale and the measurement of people?

Figure 29 shows the targeting of the item locations to the sample. *Figure 30* shows the targeting of the item thresholds to the sample. The sample covers the scale well, implying that this is a good sample in which to study the scale. However, the scale does not cover the sample well, implying that the scale is suboptimal for measuring the people. This is discussed further below.

Has a measurement ruler been constructed successfully?

Do the item response categories work as intended?

Only one item (item 22) has reversed thresholds, indicating that the item scoring function for this item was not working as anticipated. *Figure 31* shows the CPC for this item. However, a close look at the CPCs for the other eight items, and their threshold values in *Table 21*, indicates that thresholds τ3 and τ4 for all items occur at very similar locations on the continuum. *Figure 32* shows the CPCs for items 24–27. As with the physical subscale items, this finding implies that there may be too many response options in the psychological subscale items.

Do the items map out a discernible line of increasing intensity?

Table 21 shows that item locations range from −0.48 ('feeling mentally fatigued') to + 0.17 ('feeling depressed'), less than one logit, while item thresholds vary from −1.62 to +1.50. Thus, although the items spread out they do not spread to define a wide continuum. *Figure 33* shows the psychological impact continuum mapped out by the nine items. There are clear gaps and there is bunching around the centre of the scale range. *Figure 30*, the plot of item thresholds, shows 10 of a total of 36 thresholds lying between 0.0 and 0.2 logits, a very narrow range.

Is the location of items along this line reasonable?

Table 22 shows the nine items in location order. The continuum implies that the first of these nine problems to be manifest as a person moves from left to right down the psychological impact continuum is likely to be 'feeling mentally fatigued'. As the psychological impact becomes greater there are likely to be problems with concentrating and anxiety. Last is the likelihood of a perception of depression.

Do the items work together to define a single variable?

Item–person fit residuals

Table 23 shows the item–person fit residuals. *Figure 34* shows these values diagrammatically. Three items lie outside the recommended range; one

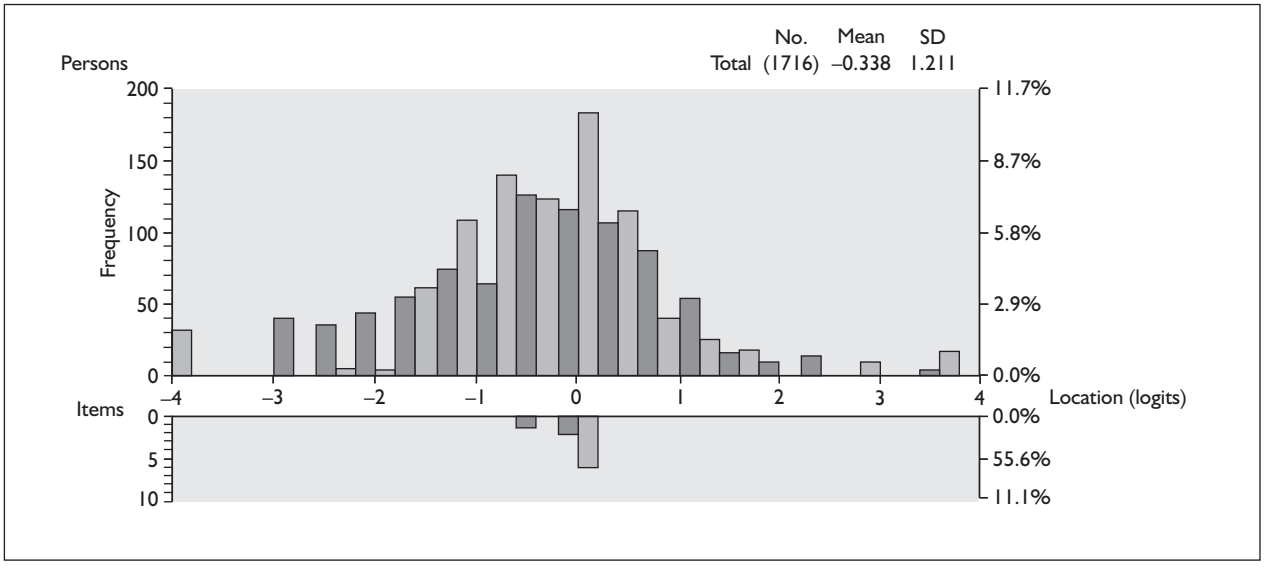


FIGURE 29 *MSIS-29 psychological subscale – targeting of sample to item locations. Person–item location distribution (grouping set to interval length of 0.20, making 40 groups).*

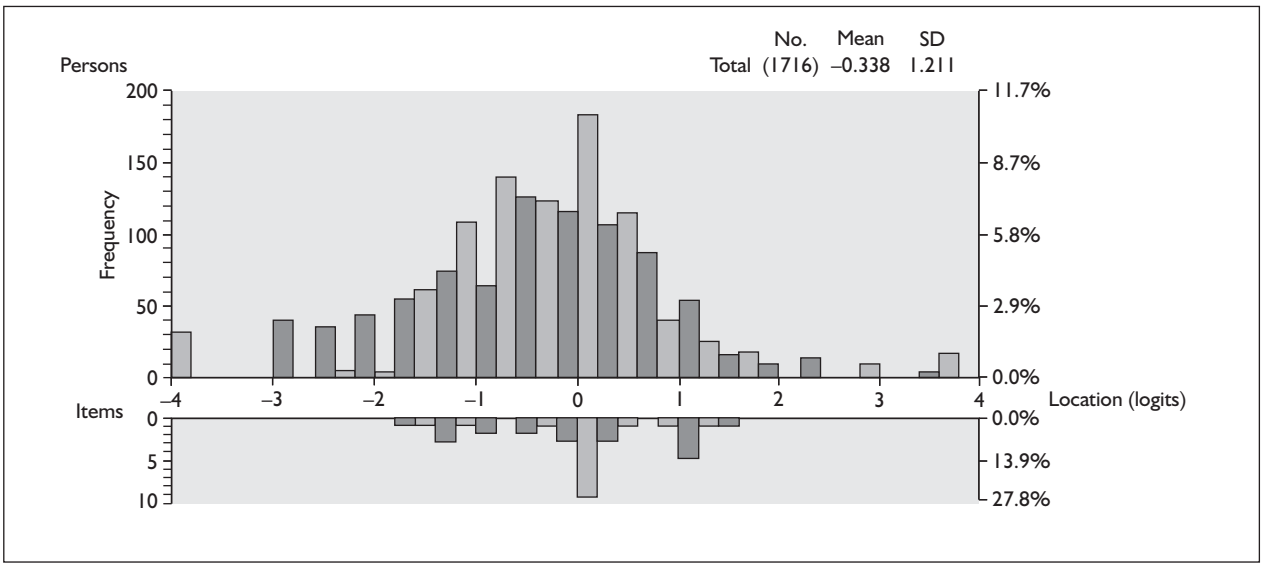


FIGURE 30 *MSIS-29 psychological subscale – targeting of sample to item thresholds. Person–item threshold distribution (grouping set to interval length of 0.20, making 40 groups).*

item (item 22) misfits substantially more than the other two (items 29 and 25).

Item–trait chi-squared values

Table 23 shows the chi-squared values and *Figure 35* plots these values diagrammatically. The same three items with misfitting fit residuals have more extreme chi-squared values relative to the others. For two items (items 22 and 25) the chi-squared values are statistically significant, indicating that the discrepancies between observed scores and expected values are larger than expected by chance.

Item characteristic curves

All ICCs were examined. *Figure 36* shows the ICCs for the three worst-fitting items (22, 25 and 29). Although the fit statistics raise concerns about these three items, the curve defined by the observed scores has an increasing monotonic relationship

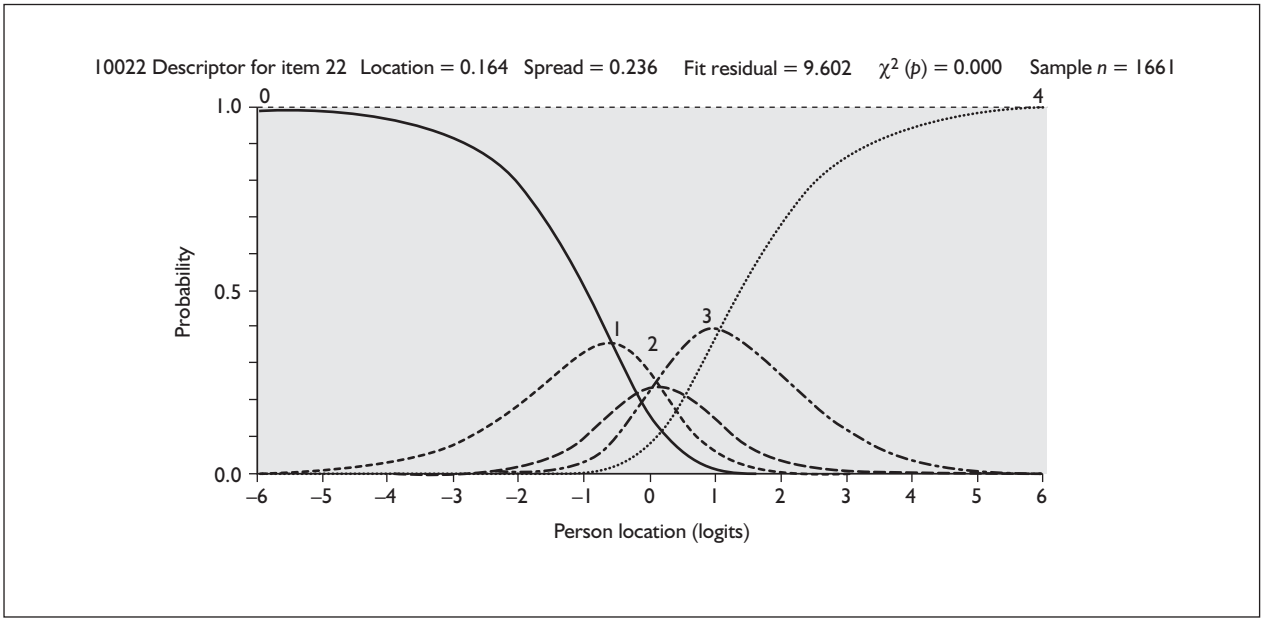


FIGURE 31 *MSIS-29 psychological subscale – category probability curve for item 22.*

TABLE 21 *MSIS-29 psychological impact subscale: item thresholds and location estimates*

		Threshold estimates				
Item	**Location**	**τ1**	**τ2**	**τ3**	**τ4**	**RT**
21	0.097	–1.175	0.021	0.356	1.185	
22	0.164	–0.579	0.178	0.001	1.055	X
23	–0.476	–1.623	–0.594	–0.249	0.561	
24	–0.031	–1.323	0.105	0.107	0.986	
25	0.053	–1.361	–0.077	0.157	1.495	
26	–0.024	–1.452	0.019	0.144	1.194	
27	0.036	–1.323	0.033	0.170	1.263	
28	0.008	–0.853	–0.088	–0.058	1.029	
29	0.173	–0.922	0.243	0.262	1.110	

Location, mean of thresholds; RT, reversed threshold; X, item with reversed thresholds.
τ1, point at which probability of responding 'not at all' and 'a little' is the same.
τ2, point at which probability of responding 'a little' and 'moderately' is the same.
τ3, point at which probability of responding 'moderately' and 'quite a bit' is the same.
τ4, point at which probability of responding 'quite a bit' and 'extremely' is the same.

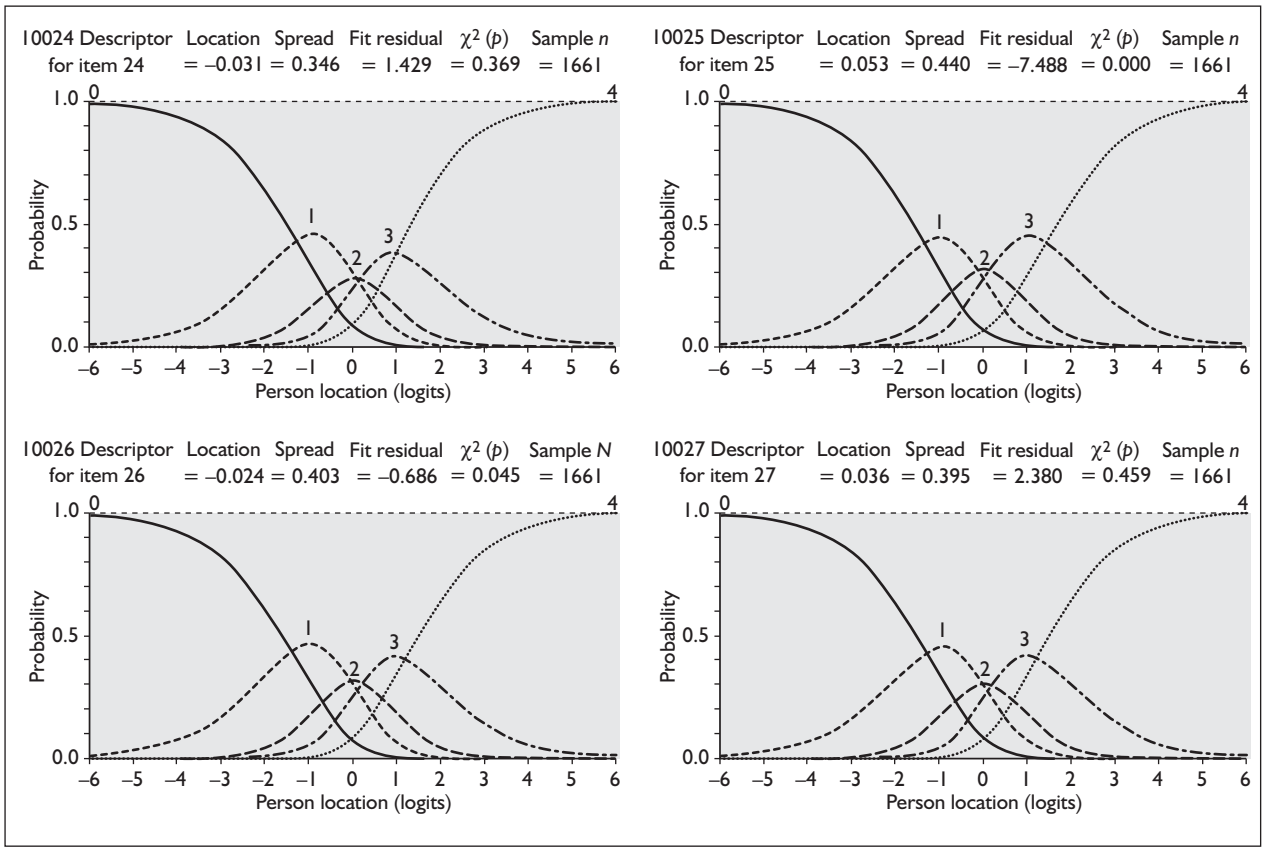


FIGURE 32 *MSIS-29 psychological subscale – category probability curve for items 24, 25, 26 and 27.*

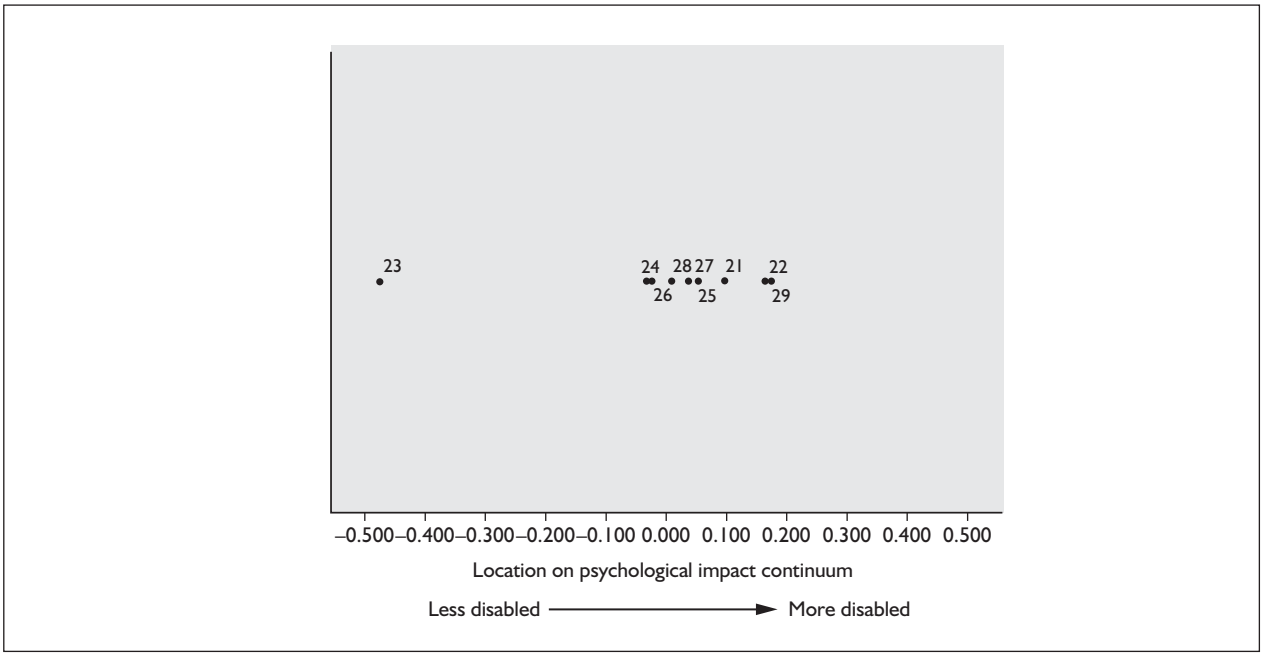


FIGURE 33 *MSIS-29 psychological subscale – continuum mapped out by the nine psychological impact items.*

with location on the continuum and follows the expected values fairly well. These findings imply less concern about the extent to which these items misfit.

Does the response to one item directly influence the response to another?
Correlations between residuals were low. Only two of the 36 correlations exceeded 0.30, and none exceeded 0.40. The pairs of items whose residual correlations exceeded 0.30, and their values, were:

Items 23 and 24 = 0.33

Items 24 and 27 = 0.31

Are the locations of the items stable across clinically important groups?
DIF was tested for sex (men and women); mobility level (unaided; with an aid; in a wheelchair); MS type (relapsing–remitting MS; secondary progressive MS; primary progressive MS); educational level (higher degree/qualification or not); and sample (rehabilitation; steroids; first

TABLE 22 *MSIS-29 psychological impact subscale: item locations, standard errors and fit statistics ordered by increasing item location [n = 1661 (adjusted to n = 500 for calculation of χ^2); 10 class intervals]*

Item				Fit statistics		
Number	Label	Location	Standard error	Fit residual	χ^2	χ^2 probability
23	Feeling mentally fatigued	–0.476	0.026	0.916	4.813	0.850
24	Worries about your MS	–0.031	0.026	1.429	2.111	0.990
26	Feeling irritable, impatient or short-tempered	–0.024	0.027	–0.686	3.506	0.941
28	Lack of confidence	0.008	0.025	–1.698	6.669	0.672
27	Problems concentrating	0.036	0.026	2.380	3.380	0.947
25	Feeling anxious or tense	0.053	0.027	–7.488	31.894	0.000
21	Feeling unwell	0.097	0.027	–0.711	4.185	0.899
22	Problems sleeping	0.164	0.025	9.602	40.791	0.000
29	Feeling depressed	0.173	0.026	–4.993	17.991	0.035

TABLE 23 *MSIS-29 psychological impact subscale – item locations, standard errors and fit statistics [n = 1661 (adjusted to n = 500 for calculation of χ^2); 10 class intervals]*

Item				Fit statistics		
Number	Label	Location	Standard error	Fit residual	χ^2	χ^2 probability
21	Feeling unwell	0.097	0.027	–0.711	4.185	0.899
22	Problems sleeping	0.164	0.025	9.602	40.791	0.000
23	Feeling mentally fatigued	–0.476	0.026	0.916	4.813	0.850
24	Worries about your MS	–0.031	0.026	1.429	2.111	0.990
25	Feeling anxious or tense	0.053	0.027	–7.488	31.894	0.000
26	Feeling irritable, impatient or short-tempered	–0.024	0.027	–0.686	3.506	0.941
27	Problems concentrating	0.036	0.026	2.380	3.380	0.947
28	Lack of confidence	0.008	0.025	–1.698	6.669	0.672
29	Feeling depressed	0.173	0.026	–4.993	17.991	0.035

postal survey; second postal survey). No items demonstrated differential item functioning for age, gender, sample or mobility level. Note that DIF is discussed at length in Chapter 6.

Have the people in the sample been measured successfully?

Are the persons in the sample separated along the line defined by the items?

Figure 29 demonstrates that the person locations in this sample ranged from about −4.0 to +4.0 logits. Thus, despite the limited range of item locations, the scale seems to detect important differences. The PSI is 0.892, indicating that the nine-item scale is able to distinguish reliably between the responders on the trait that it measures.

Table 24 shows the interval-level locations implied by each psychological subscale total score. *Figure 37* plots the values. The correlation is 0.965, but the graph is S-shaped. *Table 24* indicates that a change of 1 raw score point at the scale extreme implies a change in location of 0.80 logits, whereas in the centre of the scale a change of 1 raw score point implies a change of 0.08 logits. Thus, a change of 1 point implies a variable change in interval-level location of up to 10-fold depending on where in the scale range the change occurs.

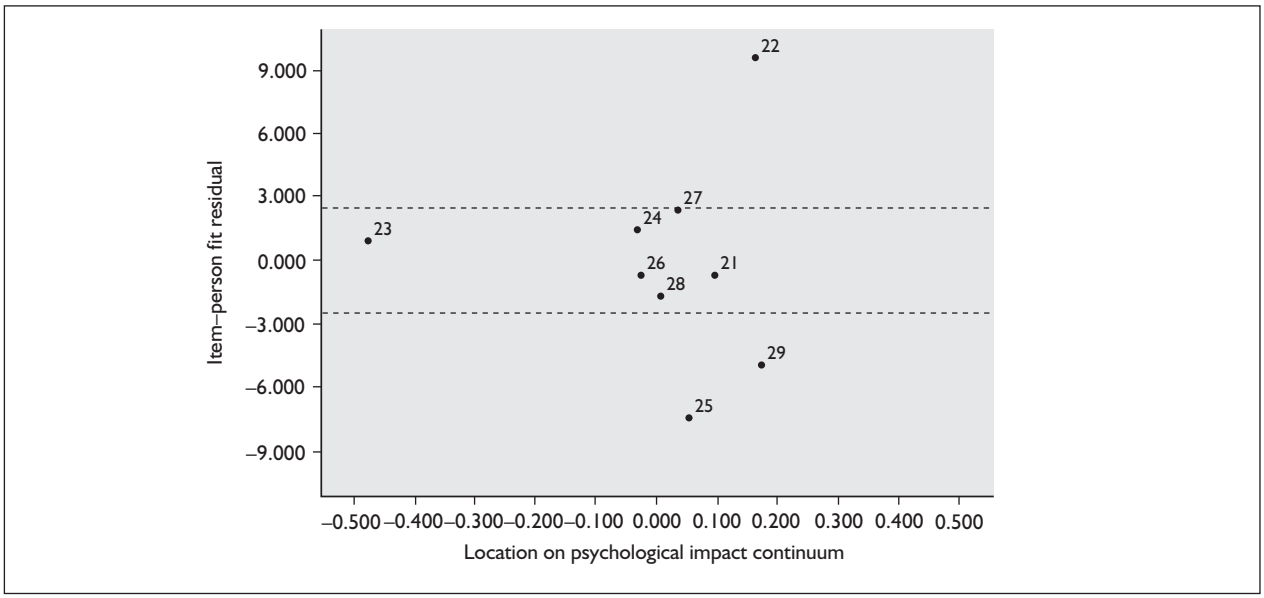


FIGURE 34 *MSIS-29 psychological subscale – plot of item–person fit residuals against location.*

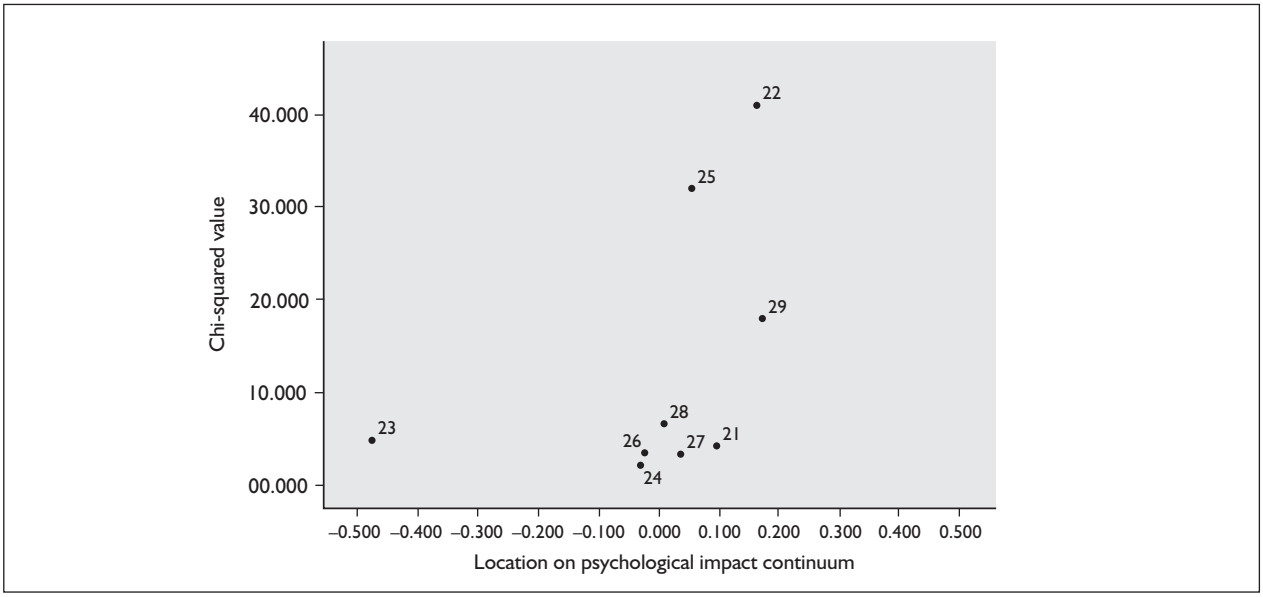


FIGURE 35 *MSIS-29 psychological subscale – plot of chi-squared values against location.*

Table 24 also reports the standard error associated with each person location estimate. These vary fourfold across the range of the scale and are greater at the extremes. *Figure 38* plots the standard error against the location. This shows that the scale is at its most precise in the central range. *Figure 39* plots the information function for the scale in relation to the study sample and the item thresholds. Many people lie outside the range in which this scale provides precise measurements.

Do individual placements on the variable make sense?

How valid is each person's measurement?

A total of 1716 people were measured in the sample. Of these, 50 scored at the floor ($n = 34$) or ceiling ($n = 16$) and did not have person–item fit residual estimates. Of the 1666 people who had estimates of their person–item fit residuals, 116 people (7%) had values that were outside the range −2.5 to +2.5, and 83 (5% of the sample) had fit residuals outside the range −3.0 to +3.0. These findings indicate that 93% of people gave responses that were consistent with expectation.

Summary of results of Rasch analysis of the MSIS-29 psychological impact subscale

Scale-to-sample targeting indicated that this was a good sample for studying the psychometric properties of the nine-item MSIS-29 psychological impact scale. The five-category item scoring function for the scale was adequate, but there was evidence that there were too many categories. The nine items mapped out a variable of increasing intensity, and the ordering of items along this variable made clinical sense with respect to its intended goal of quantifying psychological impact. However, the range of item thresholds and locations was narrow, and there were gaps in the continuum and bunching of thresholds. Three of the items failed the two numerical tests of fit, implying that responses to them were not adequately consistent with expectation. This raises questions about their placement within the nine-item set. However, the graphic test of fit implied less concern about these items. Responses to the remaining six items were consistent with expectation. There was no evidence to suggest that responses to all items were not biased by responses to another, and no evidence of significant DIF across clinically important groups.

The scale proved to be a reliable indicator in this sample and there was evidence that the pattern of item responses for more than 90% of people in the sample was consistent with expectation.

Summary

The aim of this chapter was to examine, with new psychometric methods, a scale that had been developed using traditional psychometric methods. The purpose of this examination was to compare and contrast the two approaches, to determine the added value (if any) of using a more sophisticated method, and whether any advantages justified the

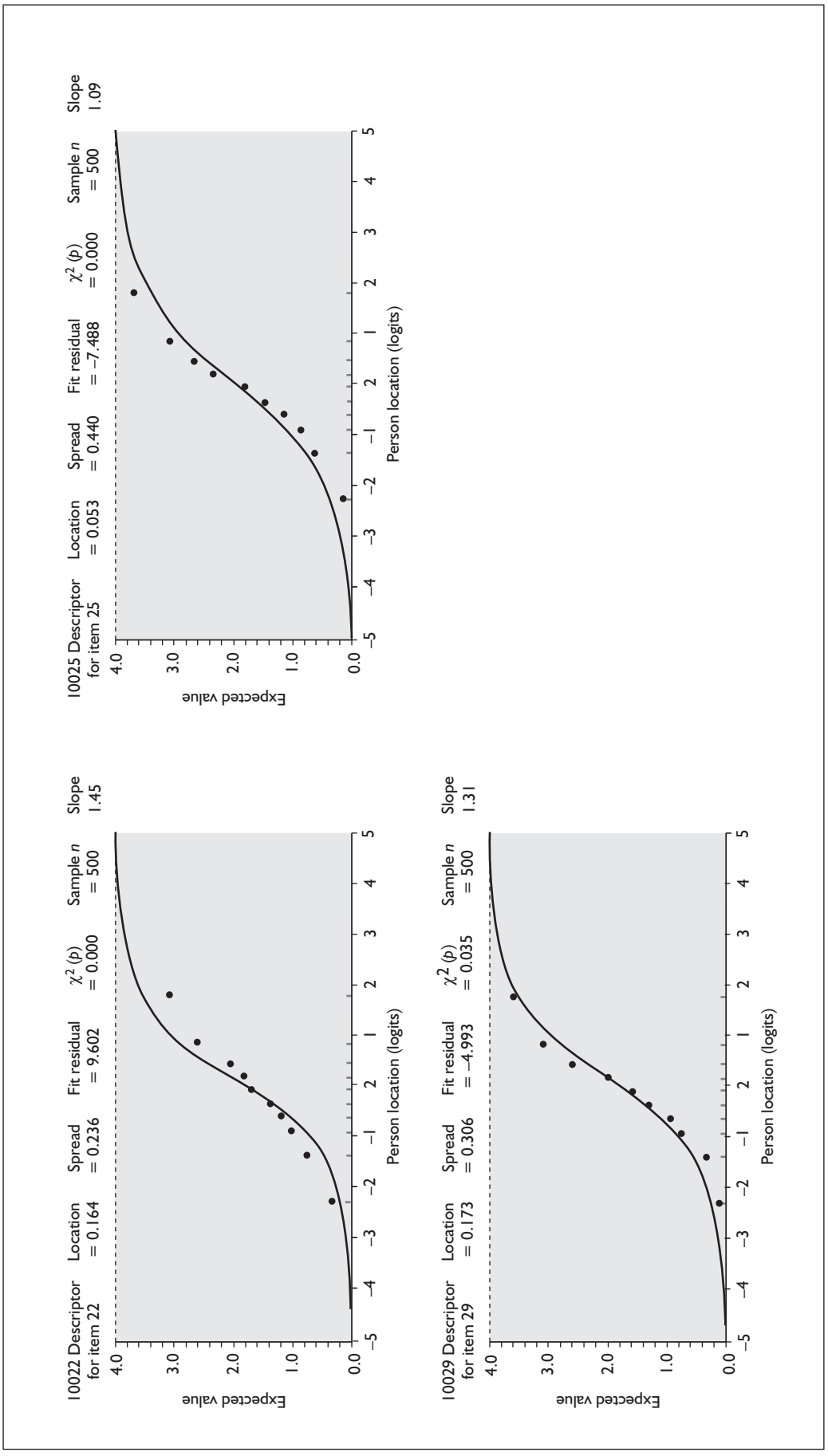


FIGURE 36 *MSIS-29 psychological subscale – item characteristic curves for items 22, 25 and 29.*

TABLE 24 *MSIS-29 psychological impact subscale: interval-level locations implied by each ordinal-level raw score*

Raw score	Interval measure	Standard error
0	–3.810[a]	1.190
1	–2.990	0.860
2	–2.430	0.670
3	–2.050	0.570
4	–1.760	0.510
5	–1.530	0.460
6	–1.330	0.430
7	–1.160	0.400
8	–1.010	0.380
9	–0.880	0.360
10	–0.750	0.350
11	–0.640	0.340
12	–0.530	0.330
13	–0.430	0.320
14	–0.330	0.310
15	–0.240	0.310
16	–0.150	0.310
17	–0.060	0.300
18	0.020	0.300
19	0.110	0.300
20	0.200	0.300
21	0.290	0.310
22	0.380	0.310
23	0.470	0.310
24	0.570	0.320
25	0.670	0.330
26	0.780	0.340
27	0.890	0.350
28	1.020	0.370
29	1.160	0.390
30	1.320	0.420
31	1.500	0.450
32	1.720	0.490
33	1.990	0.560
34	2.350	0.650
35	2.890	0.840
36	3.700[a]	1.180

a Note that the measurements for people at the extremes of the scale range (floor and ceiling) are extrapolation based on actual estimates. For these people no person–item fit residual can be computed.

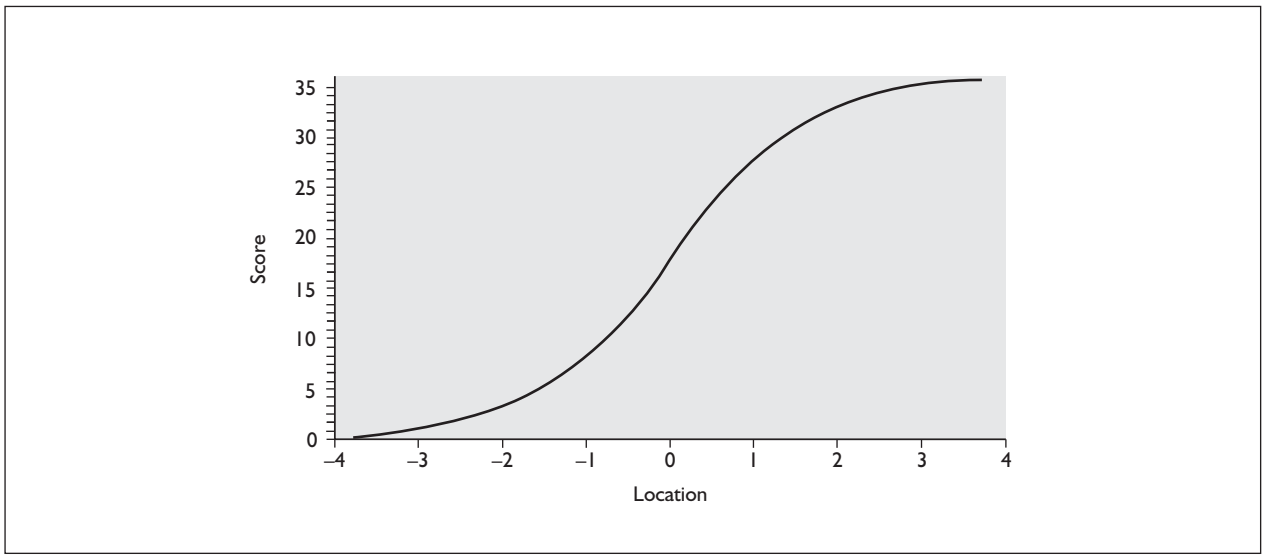


FIGURE 37 *MSIS-29 psychological subscale – plot of raw scores against interval measures.*

additional expertise required to undertake such analysis.

Implications for the MSIS-29

The MSIS-29 was developed using traditional psychometric methods. Those of us involved in its development would like to think that we applied those methods rigorously. Others think so.[135] We followed the well-established three-stage approach: item generation; item reduction and scale formation; and scale validation. People with MS were intimately involved at all stages and the study was in collaboration with the MS Society of Great Britain and Northern Ireland. To generate a pool of items, we interviewed people with MS, representing the full range of the condition, until redundancy, canvassed expert opinion from a range of health professionals and undertook a comprehensive literature review. This process generated in the region of 3000 statements concerning the health impact of MS. A preliminary questionnaire was prepared from these statements, and pre-tested in a sample of people with MS, before being sent to a large sample of people with MS across the UK.

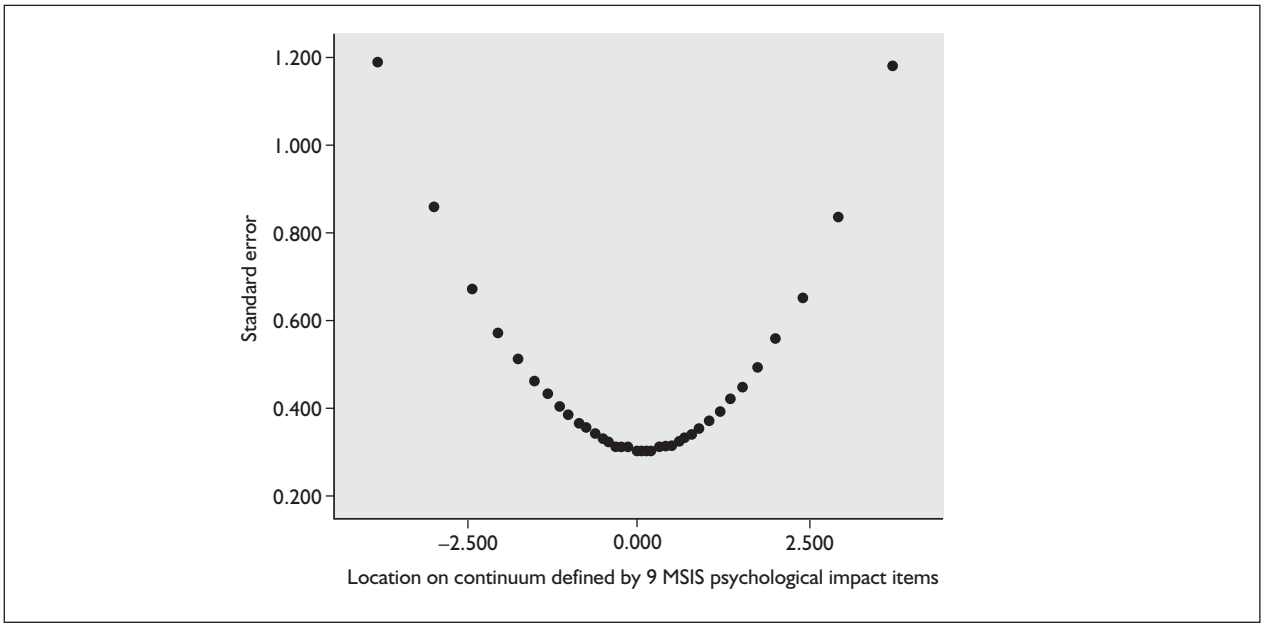


FIGURE 38 *MSIS-29 psychological subscale – plot of standard error against location.*

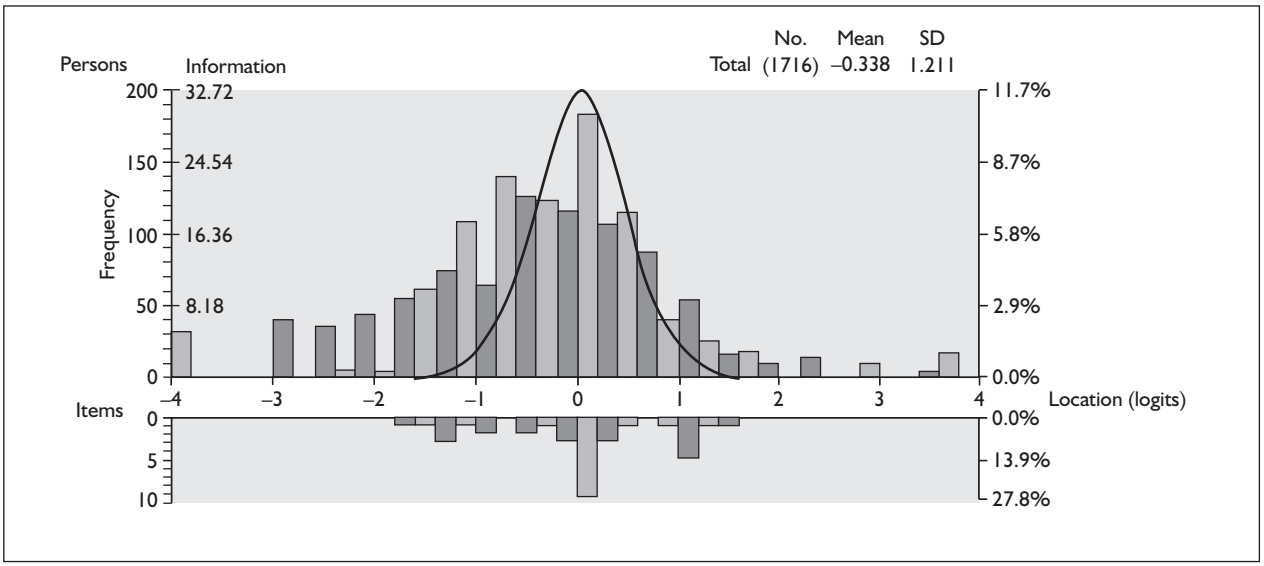


FIGURE 39 *MSIS-29 psychological subscale – plot of information from scale relative to sample. Person–item threshold distribution (grouping set to interval length of 0.20, making 40 groups).*

Many approaches have been used to reduce a pool of items to form a rating scale, but there are no consensus guidelines. On account of this, we identified the many criteria used by others, and applied these whenever possible. The result was the 29-item Multiple Sclerosis Impact Scale (MSIS-29).

The MSIS-29 was validated, comprehensively, in a large, independent sample of people with MS. A range of other widely used self-report scales were used as validation instruments. The MSIS-29 satisfied all criteria that we tested, and outperformed competing scales. Others have found similar results.

Rasch analysis of the MSIS-29 provided evidence to support the findings of the traditional psychometric evaluations. However, Rasch analysis also demonstrated important limitations in the scale and in traditional psychometric methods, and provoked thinking into the constructs measured and the whole process of scale development.

Evidence in support of the measurement properties of the MSIS-29 came from a demonstration that the items of both subscales map out continua of increasing intensity; are located along those continua in a clinically sensible order; work reasonably well together to define single variables; consist of items that are locally independent; and do not exhibit differential item functioning. Further support for the use of the subscales came from the findings that both subscales were able to separate the sample reliably and that people's patterns of responses were consistent with expectation.

However, Rasch analysis detected important limitations of the MSIS-29 which were not identified by traditional psychometric methods. It detected that the five-category item scoring function did not work as intended for nine items in the physical subscale and one item in the psychological subscale. This implied that these items had too many response options. In fact, when the CPCs and the threshold values were examined for the items with ordered thresholds, there was good evidence that all items would benefit from fewer response categories. This can be seen quite clearly in the threshold maps, which show the region of the continuum represented by each item response option for items with ordered thresholds (items with disordered thresholds are not shown). *Figure 40* shows the threshold maps for the two MSIS-29 subscales, and demonstrates that the area of the continuum represented by response label 2 (= 'moderately' on the questionnaire) is often very small. Because of this finding we undertook a post hoc rescoring of the items by collapsing adjacent categories (so that all items had four response categories), as suggested by each CPC. Re-analysis of the data demonstrated that all thresholds were now correctly ordered. The rest of the analyses were similar. Clearly, this needs empirical testing as collapsing categories makes assumptions about how people would respond.

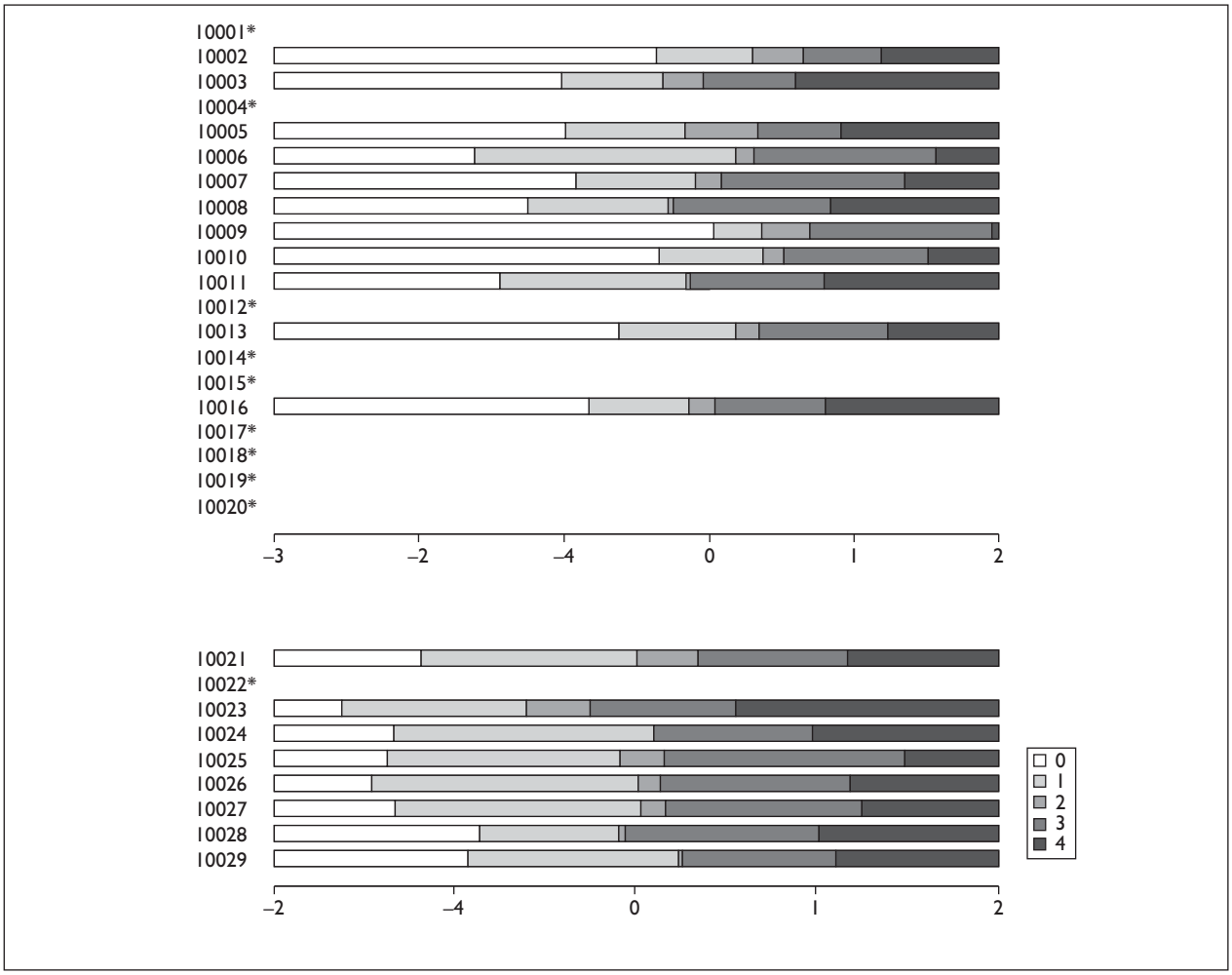


FIGURE 40 *MSIS-29 subscales – threshold maps. *Reversed thresholds. Key: 0, response category labelled 0; 1, response category labelled 1; 2, response category labelled 2; 3, response category labelled 3; 4, response category labelled 4*

A second limitation of the MSIS-29 detected by Rasch analysis, and applicable to both subscales but especially the psychological subscale, was the range and spread of the item locations and thresholds. Neither scale mapped a wide variable. Both scales demonstrated bunching of item thresholds and gaps in their continua. The implication of these findings is that measurement is suboptimal and could be improved. It is also important to note that these findings are a function of traditional psychometric methods. Correlation-based analyses (e.g. item–total correlations, item–item correlations, alphas), which are the mainstay underpinning item selection, tend to identify items in the middle of the scale range as superior. A number of difficulties arise from packing items in the central region of the scale range. One problem is that the scale is only a precise measure over a limited range. This is demonstrated by the subscale information functions (see *Figures 28* and *39*), the plots of standard errors against location (see *Figures 27* and *38*) and the fact that both scales only provide precise measurement to a proportion of the study sample. A second problem is that scales could become overprecise in the centre relative to the extreme. Thus, the relationship between raw scores and interval measures becomes increasingly S-shaped, and thus increasingly non-linear, so that the 'real' implications of raw score changes and differences becomes increasingly distorted. We have seen in this study that the change in interval-level measurement implied by a fixed change in raw score varies up to 27-fold.

A third limitation of the MSIS-29, detected by Rasch analysis, relates to the extent to which the items work together to define a single variable. Both subscales, but especially the physical subscale, demonstrated notable misfit on the two numerical indicators of fit of the data to the Rasch model (item–person fit residuals; item–trait chi-squared values). The third graphical indicator of fit implied

that this misfit was less of a concern. These findings could be interpreted in two ways. On the one hand, we could favour the graphical indicator, play down the numerical values and argue that the items in each subscale scale work well together to define a single variable. On the other hand, we could be more circumspect and try to diagnose why an instrument that passes all traditional tests of validity demonstrates notable misfit on some – but not all – tests of fit. The second approach is that favoured by leading Rasch analysts,[66,74,136] and in line with Kuhnian theory, which argues that the role of measurement is to highlight anomalies for further investigation.[137,138]

In keeping with this approach, the results of the test of fit provoked a careful examination of the items, a consideration of the scale development process and an explanation for the findings. For example, the 20 items of the physical impact subscale can be grouped into three themes: activity limitations; symptoms causing physical limitations; and limitations in social functioning. The amalgamation of these related but different subconstructs into a single set operationalises a broad physical functioning variable and would explain why there are some problems with fit. Why did this happen? Because the scale development process used a factor analysis which groups items that are related but distinct from other groups of items. Moreover, as factor analysis is based on correlations, it will tend to bring together items that measure at a similar point on the scale, irrespective of their content.

These findings have led us to revise the MSIS-29, which is now under evaluation. (MSIS-29v2; Appendix 2.2)

Implications for development of new scales

The results of this Rasch analysis have implications for scale development. The ability to formally examine item scoring functions means that this can be built into scale development so that the number of item response categories is determined empirically rather than by assumption. At the moment these are typically chosen because they seem to be sensible. Our own research demonstrates that the numbers of response options that 'work' varies from scale to scale. In addition, traditional psychometric method has tended to impose the same scoring function for each item in a scale. We think that this can affect the clinical meaningfulness of items to patients. It would be far more appropriate for the number and type of response options to be item specific and empirically determined. Rasch analysis enables this; traditional psychometric methods do not.

The fact that Rasch analysis enables investigators to visualise the relative locations of items and their thresholds, and the knowledge that these locations are independent of the distributional properties of the study sample, have implications for scale development. This means that the variable mapped out by a set of items can be constructed to cover a suitable range. Gaps in the measurement continuum can be identified early, and appropriate items can be identified and tested. Item bunching and redundancy can be minimised. Application of these methods is likely to produce more responsive scales. Traditional methods do not enable this to be done, and the fact that items are highly correlated does not mean that they are redundant.

The most important implication for scale development arose from the fit statistic findings. The examinations they invoked raise serious questions about the way many scales are developed, and this has led us to change our scale development practice.[119] Typically, scales are developed 'top down'. That is, a pool of items is generated and grouped into subscales on the basis of either statistical tests, such as factor analysis, or thematic similarity. This has two potential limitations. First, grouping items statistically does not ensure that they measure the same construct. Second, grouping items thematically does not ensure that they map out a variable in a clinically meaningful fashion.

It seems clear to us that the scale development process has to be underpinned by a clear definition, and conceptualisation, of the variable to be measured. Without this foundation it is impossible to map out a variable (referred to as 'operationalise') in a clinically meaningful and measurable fashion. It then follows that the purpose of a psychometric evaluation is to establish the extent to which a proposed quantitative conceptualisation has been successfully operationalised.[3]

Taking a 'top-down' approach to scale development may mean that scale developers might not invest fully in the process of defining, conceptualising and operationalising variables which we[111,119] now, and others,[139,140] believe is central to valid measurement. Indeed, a critical look at our development of the MSIS-29, and most other published rating scales, indicates that those processes were not formally undertaken. Many scales consist of items that are

mix-related constructs, and the variables they purport to measure may not be clear from the scale content. This is not helped by the fact that psychometric evaluations of scales can produce apparently excellent results in the face of limited content/clinical validity. This is because there is a difference, perhaps subtle, between a set of items being related and a set of items mapping out a variable. In both situations it is highly likely that responses to the items will be correlated (the basis of traditional analyses) and related probabilistically (the basis of Rasch analysis). It is also because statistical evaluations of any set of items cannot be expected to tell us directly *what* a scale measures. Conversely, the fact that a set of items appears clinically meaningful does not tell us *how* they will perform as a measurement instrument.

Implications for clinical trials

The results of this Rasch analysis have important implications for clinical trials. We have already discussed the potential for producing superior, more reliable and valid rating scales. In this section we discuss the benefits of interval-level measurement and individual patient measurement. The non-linear (S-shaped) relationship between ordinal raw scores and the interval measures is well known. However, the implications of this relationship for the measurement of change have received less attention. In this monograph we have demonstrated that a fixed 1-point change in raw scores implies a variable change in interval-level (i.e. equal-interval) measurement. The variability was scale dependent and ranged from 4.7-fold (RMI) to 27-fold (MSIS physical subscale). The direct implication is that studies using raw scores have almost certainly underestimated change in the samples; the problem is that we know neither to what extent nor the circumstances in which different inferences would have been made. However, if clinical trials strive to deliver truthful inferences then it seems hard to argue against methods that enable interval-level measurement of patient-reported outcomes.

Rasch analysis enables legitimate measurement at the individual person level. Traditional methods are not recommended for this.[107] This is because Rasch analysis generates a standard error for every person, determined by the items that person answers, their location on the continuum and the targeting of the items to their location. In contrast, traditional methods generate one standard error for all locations on the scale that is determined by the reliability of the scale and the standard deviation of the sample. Thus, even when the reliability is very high, the error is wide (unless the sample SD is low, in which case the reliability is unlikely to be high). Moreover, as we have seen (see *Tables 20* and *24*), the standard error associated with measurements of people is logical and empirically dependent on their location on the continuum. This makes the concept of a single standard error scientifically weak.

Rasch analysis, therefore, allows clinical triallists to evaluate their data at both the group and individual person levels. This is important as the treatment effect is typically variable: different people benefit to different degrees. Ideally, clinical trials should have the facility to determine responders from non-responders as well as the overall group effect. In addition, group-based statistical tests do not account for the different measurement precision of rating scales. This can influence the results from studies and the inferences made (see Chapter 8).

The standard errors demonstrated in *Figures 27* and *38*, and the information functions demonstrated in *Figures 28* and *39*, have another implication for clinical trials. They demonstrate that rating scales are really only good measures over a limited range. Typically, the samples used in clinical trials are variable in the construct of interest. This implies that we need scales with good precision over a wide range. By definition, this means scales with large numbers of items well spread over the range of the continuum. Such scales would almost certainly be unsuitable for use in clinical trials. The alternative is targeted measurement, i.e. to present each individual with a selection of items whose locations are similar to that of the person being measured. This process requires a large pool of calibrated items – known as an item bank – and a method of administering those items to individuals. These issues are discussed in Chapter 6.

Targeted measurement represents a 'sea change' from our current thinking about rating scales. Typically, trials use scales with a fixed number of items [e.g. MSIS-29, SF-36, PDQ-39 (Parkinson's Disease Questionnaire)]. These scales are inflexible in that they are good measures over a limited range that does not change. Essentially, therefore, when we use a fixed-length rating scale, we hope that the sample will fit the scale. This situation needs to be reversed: scales need to be tailored to fit samples and the individuals within them. This process is achievable with item banking and computer adaptive testing.

Are the specialised skills required for Rasch analysis justified?

Rasch analysis is a more sophisticated psychometric method. It has been regarded as a refinement of, or advance on, traditional psychometric methods.[3] Therefore, we expect it to identify limitations not identified using traditional methods. However, there is a trade-off. Investment is required to understand the underpinning concepts, read and understand an initially inaccessible literature, then use and interpret the software programs. Do the advantages outweigh the investment? We think that there are moral, theoretical and empirical arguments in their favour.

The moral argument is that there can be no compromise in the efforts made to improve the quality of measurement in clinical studies. Rating scales are increasingly the primary outcome measures in clinical trials. In this role, they are the central dependent variables on which decisions are made about the treatment of people, prescribing habits, the expenditure of public funds and future research. Moreover, vast amounts of public money are spent on clinical trials. It is hard to argue against state-of-the-art clinical trials using state-of-the-art measurement methods.

The theoretical scientific arguments in favour of Rasch analysis are many. They include the use of an explicit mathematical model that realises the requirements for measurement; the use of a mathematical model that enables construction of stable (invariant) linear measurements and sophisticated checks on the internal validity and consistency of scores; the ability to generate interval-level measurements from ordinal data; the ability to undertake legitimate individual person measurements; the ability to facilitate the development of item banks from which any subset of items can be used; and a scientific handling of missing data.

This study goes some way towards providing empirical scientific arguments for the advantages of Rasch analysis over traditional psychometric methods. We have demonstrated that a scale developed using traditional psychometric methods has important limitations that went undetected by traditional methods. In addition, we have been able to transform the ordinal scores of the MSIS-29 into interval measurements. Finally, we have demonstrated how Rasch analysis could lead to improvement in scale development and guide modification of existing scales.

It is sometimes difficult to demonstrate the advantages of one method over another. Head-to-head comparisons of scales are uncommon. Similarly, and for this reason, some people argue that there is still no 'proof' that new psychometric methods provide meaningful advances on traditional methods because there is no evidence that the results of a clinical trial are changed by using 'Rasch scoring rather than Likert scoring'. We think this argument misses the point.

Chapter 6

Test–retest reproducibility

Comparison of traditional and Rasch-based psychometric analysis

Overview

This chapter uses the examination of test–retest reproducibility to illustrate similarities and differences between traditional and Rasch-based psychometric evaluations. Test–retest reproducibility refers simply to the stability of a measuring instrument (i.e. it is the ability of the measure to produce stable scores over a given period of time in which the respondent's condition of interest is assumed to have remained the same). On the surface, the assessment of test–retest reproducibility seems straightforward. As we will see, it is not straightforward, because of an interaction between scale stability and person stability. This chapter demonstrates how Rasch analysis is able to dissect the different components of this interaction to achieve a comprehensive understanding of item and person stability and change, whereas traditional psychometric methods are unable to do so.

The test–retest reproducibility of the MSIS-29 is shown to be good. However, that is not the main point of this chapter. Its primary goal is to use the examination of test–retest reproducibility of the MSIS-29 as a vehicle for introducing, explaining, demonstrating and discussing two important aspects of a Rasch analysis. The first of these aspects is the application of 'racking' and 'stacking' data designs. These refer to different arrangements (set-ups) of rating scale data to enable different questions to be answered by different analyses. The second aspect is the concept of differential item functioning (DIF). This is the extent to which the functioning of an item differs across different circumstances. This has been mentioned, but not actively discussed, in previous chapters.

The clinical problem

On the surface, the approach to the assessment of consistency of two reports (i.e. two completions of a questionnaire) appears straightforward; simply administer a rating scale (e.g. MSIS-29) twice to a cohort of people and determine the agreement between their total scores at the two time points. Unfortunately, this oversimplifies a complex situation because it confounds two related but independent questions. First, does the scale work in the same way internally on both occasions? Second, if the scale does work in the same way on both occasions, do the people generate equivalent measurements at the two time points? Andrich has termed these as questions of equivalence in *kind* and equivalence in *degree*.[97]

It is easy to appreciate that the same rater may generate different ratings at two time points using the same rating scale. It is more difficult to appreciate that these differences may take two forms that need to be separated: first, a lack of internal stability (invariance); and, second, a lack of average measurement equivalence. The reason for this is that when rating scales are used to quantify variables, measurements are constructed from the responses of persons to a set of items. As such, there is an interaction between raters and items that must be disentangled into its component parts.

Simple indicators of agreement between total scores, even if complemented by a study of agreement between item-level scores, are unable to answer these two questions of equivalence of kind and degree independently. The reason lies in that a feature of traditional psychometric evaluations is that scale performance is dependent on the sample, which, to confound matters further, is often assumed to be normally distributed. More correctly, the performance of the scale and its items is dependent on the distribution of disability in the sample in which it was examined. Likewise, the measurement of people is dependent on the distribution of disability measured by the scale items. Thus, the performance of the scale cannot be divorced from the sample and the measurement of people cannot be divorced from the scale. Equivalence of kind and equivalence of degree cannot be separated.

Rasch analysis can overcome this problem and enable equivalence of kind and equivalence of degree to be studied independently and legitimately. This is because one mathematical property of the Rasch model is the separation of item and person parameters. To recap, the estimates of the item locations are independent of the sampling distribution of the people in whom they are studied. Likewise, estimates of person locations are independent of the sampling distribution of the items on which they were measured.

The Rasch model is able to estimate item and person locations separately because the model arises from the requirement of invariance (stability) of the operation of the items across different groups and across the quantitative trait. Therefore, checks on the fit of the data to the model are checks on the invariance of the operation of the items, i.e. a check on any difference in *kind* of the ratings such as difference among raters at two time points (i.e. test–retest reproducibility). This check is commonly known as a check on DIF. If the tests of invariance show adequate fit to the model, then as an explicit second step the measurements derived from the ratings, i.e. differences in *degree*, can be compared. More detailed explanations of Rasch analysis are provided elsewhere.[97,99,141]

It is important to distinguish these two sources of invariance in order both to understand the sources of the difference and to improve the consistency of the ratings depending on the sources of the inconsistencies.

The Multiple Sclerosis Impact Scale (MSIS-29)

The MSIS-29 has 29 items grouped into two self-report subscales aiming to measure the health impact of MS on physical (20 items) and psychological functioning (nine items). Items are summed and transformed to score from 0 (no problem) to 100 (extreme problems) for each scale. Two scores can be generated by summing items. Further details are presented in Chapter 5.

Setting, sample and procedure

We carried out a postal survey of randomly selected and geographically stratified people ($n = 150$) with MS recruited from the MS Society. People completed the MSIS-29 on two occasions separated by a 10-day interval.

Methods

We used two approaches to evaluate test–retest reproducibility: traditional psychometric methods and Rasch analyses. Analyses were undertaken separately for the physical and psychological subscales.

Traditional psychometric analyses

Reproducibility at the item level

The correlation between item-level ratings of MS patients was examined using two-way random intraclass correlation coefficients. We used the standard reproducibility criteria of > 0.80,[9,22] although some have argued that standards for adequate item-level reproducibility can be as low as 0.50.[142]

Reproducibility at the total score level

Reproducibility of MSIS-29 subscales was examined in four ways. First, time 1 (T1) and time 2 (T2) total scores were plotted to determine, visually, the extent of their agreement. Second, paired samples, Student *t*-tests and analyses of variance (ANOVAs) were computed to determine if group mean differences were significantly different.[7,8] Third, T1 to T2 change was quantified using standardised response means (mean difference/SD of difference).[143] This enabled us to apply an accepted classification to assess the size of any bias (score = 0.20 small bias, score = 0.50 moderate bias, score = 0.80 large bias).[10] Fourth, the agreement between ratings of MS patients was examined using two-way random-effects intraclass correlation coefficients. As discussed elsewhere,[15] intraclass correlation coefficients are preferable to Pearson product–moment correlation coefficients, as the former account for systematic mean differences.[144,145] We used the standard reliability criteria of > 0.80.[9,22]

Rasch analyses

The arrangement of the data for analysis and the analyses

Two distinct data arrangements, or designs, are required to enable the analyst to answer comprehensively the questions of equivalence in kind and equivalence in degree. In one design, the item response data for the two time points are arranged horizontally ('racked'). In the other

design, the item response data are arranged vertically ('stacked'). Each design enables both item-level and person-level analyses. The two designs provide complementary information. Neither design is sufficient on its own to answer completely the questions of equivalence in kind or degree.

Design 1: the horizontal-rating structure or 'racked' design

In this design, the T1 and T2 item ratings for each person are replicated horizontally (side by side). *Figure 41* is a schema showing how the data for the MSIS-29 physical impact subscale were set up for analysis, and the description below applies to this subscale. An identical approach was used for the psychological impact subscale.

The consequence of this design is that the two sets of 20 item responses for each person can be concatenated (combined end-to-end) to produce a 40-item structure overall. In this way each patient can be assessed from his or her ratings on 40 items. However, because the two sets of items arise from assessments at different time points, there is an interaction between item and time. Thus, while item 1 for the T1 rating and item 1 for the T2 rating relate to the same MSIS item, the outcome on each occasion can be different due to this interaction between time and item.

With the data arranged in this manner, we then perform a Rasch analysis of the items as if they were a 40-item scale. This examines the 40 items *relative to each other*, and computes an item location for each of the 40 items, with associated standard errors. This, of course, gives two locations (T1, T2) and two standard errors (T1 SE, T2 SE) for each item.

Item location analysis

With the data racked, the question of equivalence of kind is addressed by determining the agreement between item locations at T1 and T2. These can be visualised by means of a scatterplot. Also, we can determine the items for which the T2 location falls outside the 95% confidence intervals of the T1 location (T1 ± 1.96×T1 SE). Confidence intervals can be added to the plot to aid the visualisation.

However, here we are interested to know if the change between T1 and T2 is outside the error associated with the T1 and T2 locations. Thus, we need to compute the standard error of the difference (SED) = $\sqrt{[(\text{T1 SE})^2+(\text{T2 SE})^2]}$, and evaluate the change (change = T1 location – T2 location) with respect to this for each item. By dividing each item's change by its own standard error of the difference (change/SED), we obtain the significance of that item's change in standard error of difference units. The significance of each item's change (SigChange) is interpreted as:

SigChange ≥ +1.96 = significant improvement

0 < SigChange < +1.95 = non-significant improvement

SigChange = 0 = no change

−1.95 < SigChange < 0 = non-significant worsening

Person	Time 1	Time 2
	MSIS-29 physical subscale Items 1–20*	MSIS-29 physical subscale Items 1(21)–20 (40)
1	43154544415332334453	43554542314443534 51
2	21111122111111111112	21111121121111111112
.	...	...
.	...	...
*n***	33411134221213241223	52221445142334142443

FIGURE 41 *Horizontal 'racked' data design. *Raw scores as entered. **Final patient's worth of data entered in data set.*

$$\text{SigChange} \leq -1.96 = \text{significant worsening}$$

We can now identify items that have undergone significant change.

Person location analysis
With data in the racked design, the question of equivalence of degree is answered by determining the agreement between predicted person locations generated by the MSIS-29 physical subscale at T1 and T2. Essentially, we determine whether the same total score at the two time points produces the same person location.

As the 40 items have been examined as a single set, the items have been located on the continuum relative to each other. That is, the 40 items have been located on a common metric. Therefore, we can examine the relationships between measurements (person locations) produced by using different subsets of the 40 items. By taking items 1–20 as one subset and items 21–40 as another subset, we can examine the extent to which the predicted person locations generated by the MSIS-29 physical subscale at T1 and T2 are equivalent. It is important to note that these are *predicted* locations based on the item calibrations, not the locations for the people in the study sample. Essentially, this analysis equates the two sets of values and examines the agreement between the total scores at the two time points.

Differences can also be examined at the group level using paired-sample *t*-tests and at the individual level by determining total scores where the T2 value is outside the 95% confidence intervals of the T1 value.

Design 2: the vertical-rating structure or 'stacked' design
In this design, the two separate sets of patient ratings are replicated vertically (one on top of the other) for each patient. *Figure 42* is a schema showing how the data for the MSIS-29 physical impact subscale were arranged for this analysis, and the description below applies to this subscale. The same approach applies to the analysis of the psychological impact subscale. The consequence of this design is that we have twice as many people ($2n$) rated on the same 20-item scale.

Item location analysis
With data in the vertical design, equivalence of kind can be determined by examining if the responses to the items are stable (invariant) across each of the two occasions within the context of this design. This is achieved by examining the ICCs and the extent to which there is differential item functioning (DIF) by time point. The concept and examination of DIF is now explained in some detail.

To recap, whenever we use a school ruler or tape-measure to measure the length of objects, we have confidence that the measurement instrument is stable whatever is being measured. We cannot have that confidence with rating scales because, although we can see and handle them, the measurement rulers they subtend do not exist as visible concrete entities and measurements are inferred from people's responses to them. So, we need methods of determining their stability in different groups of people (e.g. men/women, older/younger, more disabled/less disabled, different types of MS, etc.). The list of different groups is of course potentially endless, so we need to be able to study the clinically important ones. Tests of item stability are called tests of DIF.

The basic premise for the stable (invariant) performance of an item is that for any level of patient disability, as best can be estimated from the data, the expected value on the item is the same irrespective of the group the person might belong to. In this study, each MSIS-29 item was rated twice by each patient: T1 and T2. Thus, patients can be classified into two groups (T1 and T2), and our aim is to determine the extent to which expected scores for each item, at the same levels of patient disability, were similar across these two ratings.

The basis for the study of DIF is the item characteristic curve (ICC), and *Figure 43* shows this for MSIS-29 item 1. To recap, the ICC shows the expected or mean value for that item (y-axis) for every person location (x-axis). This is to be compared with the observed rating. Therefore, also shown in *Figure 43* are the means (shown as black dots) of the ratings for four class intervals. Clearly, this should be close to the ICC curve.

If the observed means *follow* the ICC (as they do in *Figure 43*), this implies that the item is functioning invariantly across the level of disability, i.e. across the trait, and that all the variation across the trait is accounted for by the estimated parameters of the model. If the observed means *do not follow* the curve adequately, this implies that there is residual variation not explained by the model and therefore unaccounted for. However, in addition to lack of invariance across the trait, it is possible for there to be variation across rating groups for the same level of the trait. Accordingly, the data shown in *Figure 43* can be further subdivided into the relevant

Person	Time point	MSIS-29 physical subscale Items 1–20*
1	T1	43154544415332334453
2	T1	21111122111111111112
.	..	...
.	..	...
n	T1	33411134222123241223
1	T2	43554542314444353451
2	T2	21111121121111111112
.	..	...
.	..	...
*n***	T2	52221445142334142443

FIGURE 42 *Vertical 'stacked' data design. *Raw scores as entered. **Final patient's worth of data entered in data set = 2n.*

groups. *Figure 44* shows the ICC for item 1, but with the observed means separated according to the group: T1 or T2. There is no difference. *Figure 45* shows the same results for an item of the psychological subscale (item 24).

A unified way of assessing DIF across the rating groups, and across the levels of disability, is to consider the standardised residual of each person to each item and classify them by group (here T1 and T2) and by class interval. A two-way ANOVA of standardised residuals, in which one level of the ANOVA is the class intervals across the disability continuum and the other level gives the groups, provides a unified way of quantifying invariance of the operation of the items. In this case, the number of groups was two (T1 and T2), and the number of class intervals chosen, because of the small sample size, was four. A significant time effect, irrespective of class interval, indicates that there is a uniform DIF between T1 and T2. A significant interaction between class intervals and time indicates a non-uniform DIF. A significant difference across class intervals irrespective of time indicates a misfit to the model across the continuum. (This is analogous to the chi-squared tests of fit except that the chi-squared test of fit also enables an examination at the level of each class interval.)

Graphically, DIF across the disability variable measured by the MSIS-29 is present when the means of persons in the class intervals are not close

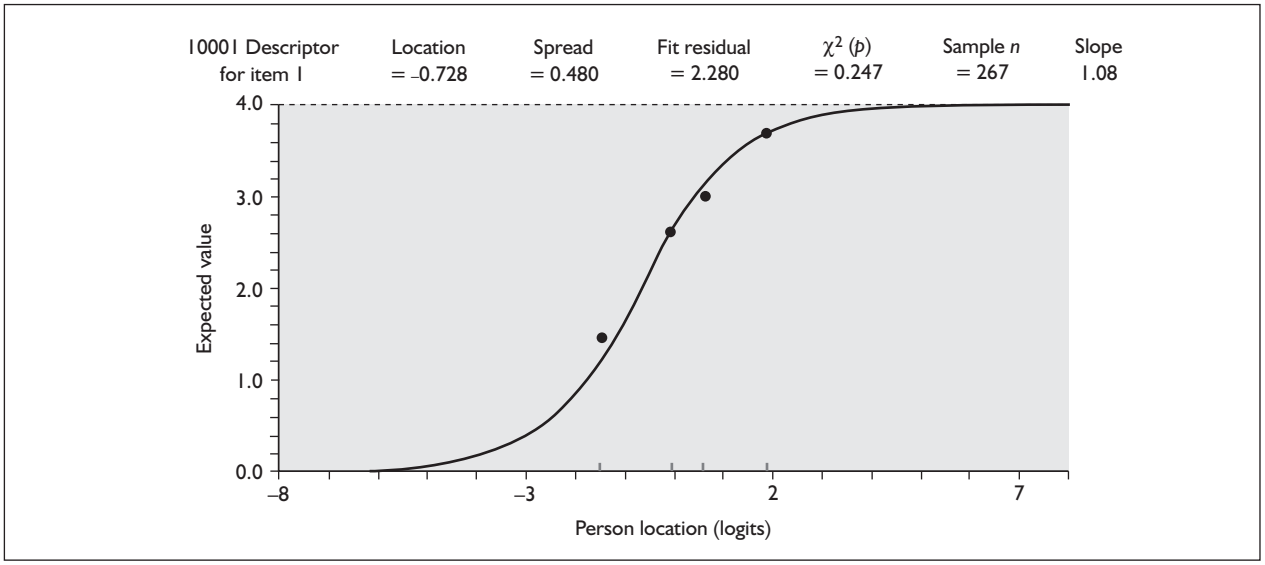


FIGURE 43 *MSIS-29 physical subscale – item characteristic curve for item 1.*

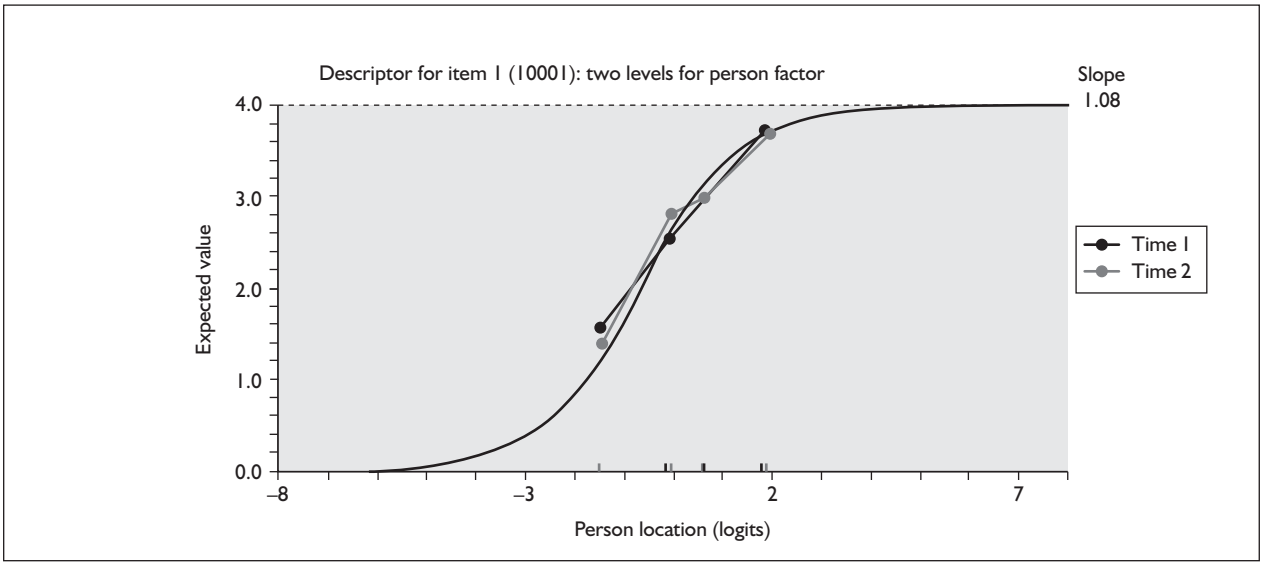


FIGURE 44 *MSIS-29 physical subscale – item characteristic curve for item 1 examining differential item functioning by time.*

to the expected value curve (ICC). DIF between time points is evident when, for a given estimate of the disability variable, the mean scores of the persons in the rating groups (T1 and T2) are significantly different from each other.

Multiple tests of fit inflate the type 1 error, for which the Bonferroni correction for a type 1 error level of 0.01 was applied. For the physical subscale, with 20 items, and for each item two probabilities (rater effect, rater-by-class interval interaction), we followed recommended guidelines[141] and set the criterion level for misfit for each statistic as $0.05/(20\times2) = 0.001250$. For the psychological subscale, with nine items, the criterion level for misfit for each statistic was set as $0.05/(9\times2) = 0.002778$.

Person location analysis

With data in the vertical, stacked design equivalence of degree is determined by examining the agreement between T1 and T2 locations for each patient in the sample to assess if any change has occurred. A number of methods can be used. The relationship can be examined visually using a scatterplot. The degree of difference can be examined at the group level using product–moment correlations, paired sample *t*-tests, and ANOVAs.

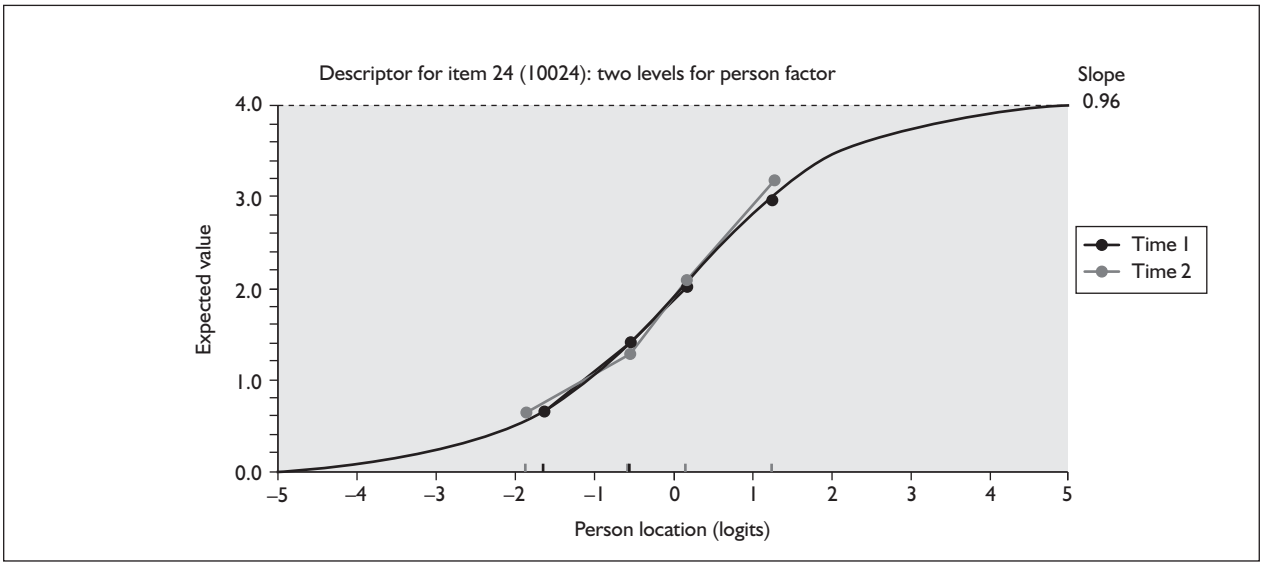


FIGURE 45 *MSIS-29 psychological subscale – typical example of differential item functioning over time 1 and time 2.*

More importantly, differences can also be examined at the individual person level by determining how many people in the sample have a T2 location that falls outside the 95% confidence intervals. The 95% confidence intervals around the T1 locations are computed as (T1 location ± 1.96×T1 SE). However, as stated before, we are interested to know if the change between T1 and T2 is outside the error associated with the T1 and T2 locations. Thus, we need to compute the standard error of the difference (SED) = $\sqrt{[(\text{T1 SE})^2+(\text{T2 SE})^2]}$, and evaluate the change (change = T1 location – T2 location) with respect to this for each person. By dividing each person's change by his or her own standard error of the difference (change/SED), the significance of that person's change is given in standard error of difference units. The significance of each person's change (SigChange) is interpreted as:

SigChange ≥ +1.96 = significant improvement

0 < SigChange < +1.95 = non-significant improvement

SigChange = 0 = no change

−1.95 < SigChange < 0 = non-significant worsening

SigChange ≤ −1.96 = significant worsening

Note that, for the MSIS-29, higher scores indicate worse disability. Thus, when change (T1 location – T2 location) is negative this implies that people have increased their scores at T2 and thus their condition has worsened.

We can now simply count the numbers of people achieving each level of significance of change and perform a chi-squared test on the results. More importantly, we can identify individuals who appear to have undergone significant improvements or deteriorations to clarify if this is the case, and if so to determine why.

Results

Participants, recruitment and data collection

Test–retest reproducibility was studied in 150 people, of whom 136 returned completed questionnaires (91% response rate). Seventy-four per cent were female and the mean age was 51 years (SD = 11; range 21–78).

Traditional psychometric analyses

Reproducibility at the item level

Table 25 reports the findings from the item-level analysis. Item-level intraclass correlation coefficients for all 29 items of the MSIS-29 ranged from 0.79 to 0.95. Thus, all but two items (Q01, Doing physically demanding things; Q02, Grip things tightly) fulfilled the minimum criterion of 0.80 that we set. None failed the more lenient criterion of > 0.50.

Reproducibility at the total score level

Table 26 reports the findings from the scale score-level analysis. Scale scores generated by the MSIS-29 subscales at T1 and T2 were not significantly different using paired-sample *t*-tests or ANOVAs. Standardised response means (SRMs) were 0.03 (physical subscale) and 0.02 (psychological subscale), indicating small change and therefore low potential for bias.

Intraclass correlation coefficients between MSIS-29 total scores were 0.97 and 0.93 for the physical and the psychological subscales, respectively. Thus, both scales fulfilled the recommended minimum criterion of 0.80.

Rasch analyses

Design 1: the horizontal-rating structure or 'racked' design

Item location analyses

Table 27 (physical subscale) and *Table 28* (psychological subscale) show the locations with standard errors at T1 and T2. *Figures 46–49* show the scatterplots for the physical and psychological subscales with item labels (*Figures 46* and *48*) and with 95% confidence intervals (*Figures 47* and *49*). The differences between the values are small and no T2 item locations for either subscale lie outside the 95% confidence intervals of the T1 locations. These findings indicate that item locations were stable across the two time points within the context of this data design. *Table 29* shows the results of the individual-level analysis of item locations. In the total sample, all items of both scales underwent no or non-significant change in their locations between the time points on each scale.

Person location analyses

Table 30 (physical subscale) and *Table 31* (psychological subscale) show the person locations, predicted from the Rasch model, implied by each and every total score on the two subscales.

Figure 50 (physical subscale) and *Figure 51* (psychological subscale) plot the T1 and T2 person

TABLE 25 *Traditional psychometric analysis: intraclass correlations at item score level*

		Intraclass correlation	
Item		**MSIS-29**	**MSIS-29**
Q01	Do physically demanding things	0.89	–
Q02	Grip things tightly	0.79	–
Q03	Carry things	0.81	–
Q04	Problems with balance	0.87	–
Q05	Difficulty moving about indoors	0.91	–
Q06	Being clumsy	0.89	–
Q07	Stiffness	0.90	–
Q08	Heavy arms and/or legs	0.91	–
Q09	Tremor of your arms or legs	0.91	–
Q10	Spasms in your limbs	0.95	–
Q11	Your body not doing what you want it to do	0.90	–
Q12	Having to depend on others to do things for you	0.94	–
Q13	Limitations in your social and leisure activities at home	0.87	–
Q14	Being stuck at home more than you would like to be	0.93	–
Q15	Difficulties using your hands in everyday tasks	0.95	–
Q16	Having to cut down time spent on work or other daily activities	0.90	–
Q17	Problems using transport	0.93	–
Q18	Taking longer to do things	0.93	–
Q19	Difficulty doing things spontaneously	0.90	–
Q20	Needing to go to the toilet urgently	0.89	–
Q21	Feeling unwell	–	0.88
Q22	Problems sleeping	–	0.88
Q23	Feeling mentally fatigued	–	0.90
Q24	Worries related to your MS	–	0.83
Q25	Feeling anxious or tense	–	0.87
Q26	Feeling irritable, impatient, or short-tempered	–	0.88
Q27	Problems concentrating	–	0.86
Q28	Lack of confidence	–	0.90
Q29	Feeling depressed	–	0.88

TABLE 26 *Traditional psychometric analysis: test–retest reproducibility at the total score level*

	MSIS-29 physical subscale		**MSIS-29 psychological subscale**	
	Time 1	**Time 2**	**Time 1**	**Time 2**
Mean (SD)	55.28 (26.81)	54.92 (27.34)	45.61 (25.65)	45.27 (25.93)
Change [mean (SD)]	0.364 (9.381)		0.335 (13.001)	
t-Value (sig)	0.443 (0.659)		0.293 (0.770)	
ANOVA *F*-value (sig)	0.062 (0.803)		0.005 (0.944)	
SRM	0.03		0.02	

ANOVA, analysis of variance; SD, standard deviation; SRM, standardised response mean.

TABLE 27 *Rasch analysis: item locations and standard errors for 20-item MSIS-29 physical subscale*

	Time 1		Time 2	
Item	**Location**	**Standard error**	**Location**	**Standard error**
1	–0.977	0.131	–0.885	0.128
2	0.746	0.122	0.572	0.122
3	–0.110	0.118	–0.334	0.120
4	–0.304	0.122	–0.318	0.121
5	0.104	0.114	0.153	0.119
6	0.162	0.125	–0.063	0.124
7	0.035	0.119	0.167	0.114
8	–0.521	0.123	–0.373	0.119
9	0.805	0.115	1.005	0.121
10	0.664	0.113	0.680	0.111
11	0.028	0.122	–0.186	0.119
12	0.014	0.116	–0.068	0.120
13	0.078	0.121	0.077	0.125
14	0.140	0.113	0.042	0.116
15	0.511	0.119	0.559	0.118
16	–0.168	0.115	–0.233	0.119
17	0.432	0.112	0.383	0.111
18	–0.661	0.127	–0.585	0.124
19	–0.421	0.115	–0.463	0.116
20	–0.302	0.115	–0.385	0.115

TABLE 28 *Rasch analysis: item locations and standard errors for nine-item MSIS-29 psychological subscale*

	Time 1		Time 2	
Item	**Location**	**Standard error**	**Location**	**Standard error**
21	0.039	0.118	0.084	0.115
22	0.183	0.112	0.227	0.112
23	–0.525	0.113	–0.440	0.113
24	0.105	0.122	–0.105	0.123
25	–0.062	0.113	–0.191	0.115
26	–0.068	0.119	–0.240	0.113
27	0.125	0.122	0.119	0.123
28	0.044	0.110	0.106	0.116
29	0.333	0.116	0.266	0.115

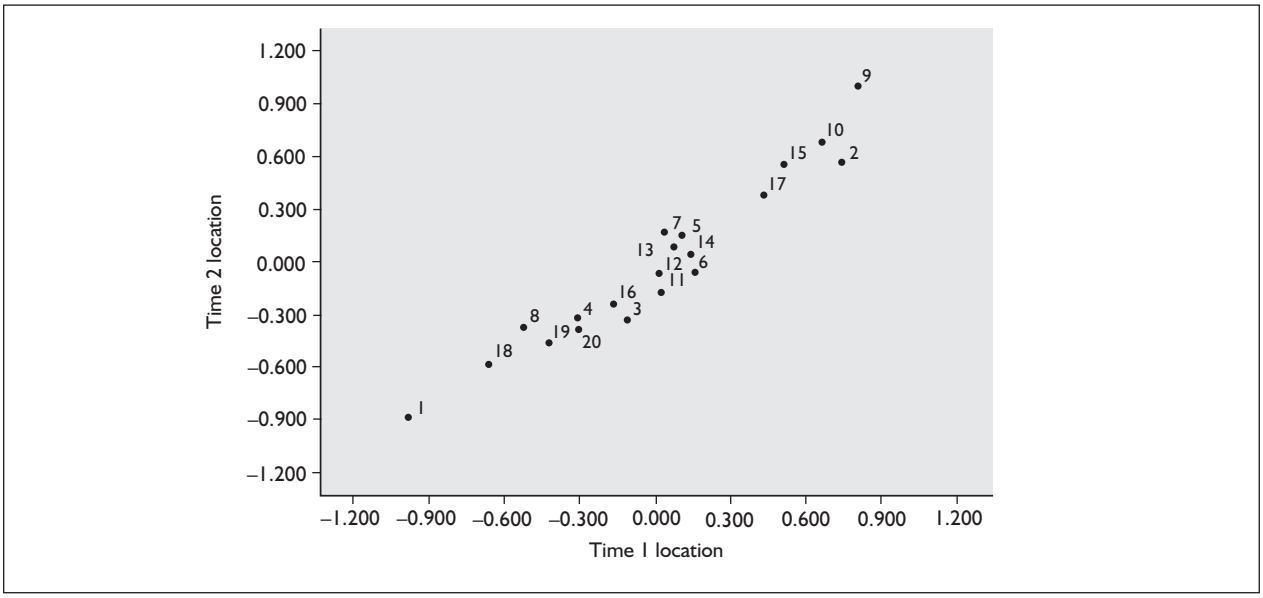


FIGURE 46 *MSIS-29 physical subscale – plot of item locations from time 1 and time 2.*

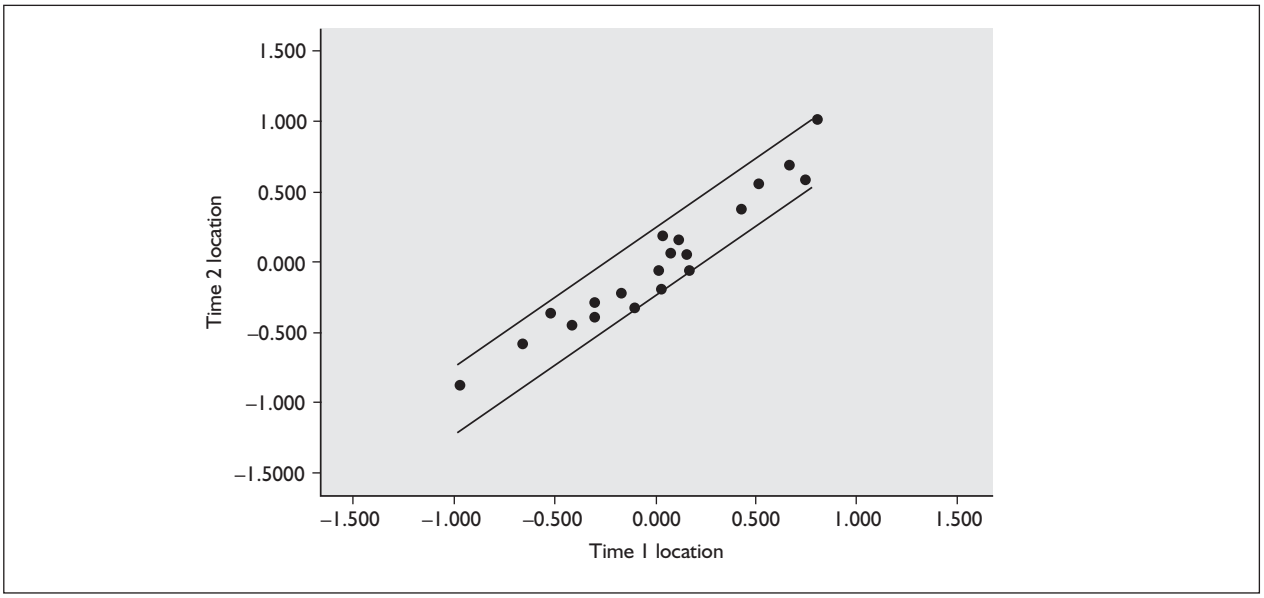


FIGURE 47 *MSIS-29 physical subscale – plot of item locations from time 1 and time 2. Black parallel lines indicate 95% confidence intervals.*

locations from the same total score. They are effectively superimposed.

These findings indicate that the person locations predicted by the Rasch model on the basis of the item locations at two time points are essentially equal when the total scores are the same.

Design 2: the vertical-rating structure or 'stacked' design

Item location analyses

Table 32 (physical subscale) and *Table 33* (psychological subscale) show the results from tests of DIF over time. No items in either subscale demonstrate statistically significant differences over time (uniform DIF), or statistically significant differences in the interaction between class interval and time (non-uniform DIF). These findings imply that the items of both subscales are adequately invariant over time across the range of the disability spectrum. *Figures 44* and *45*, as shown earlier, are representative examples of these results from the physical and psychological subscales.

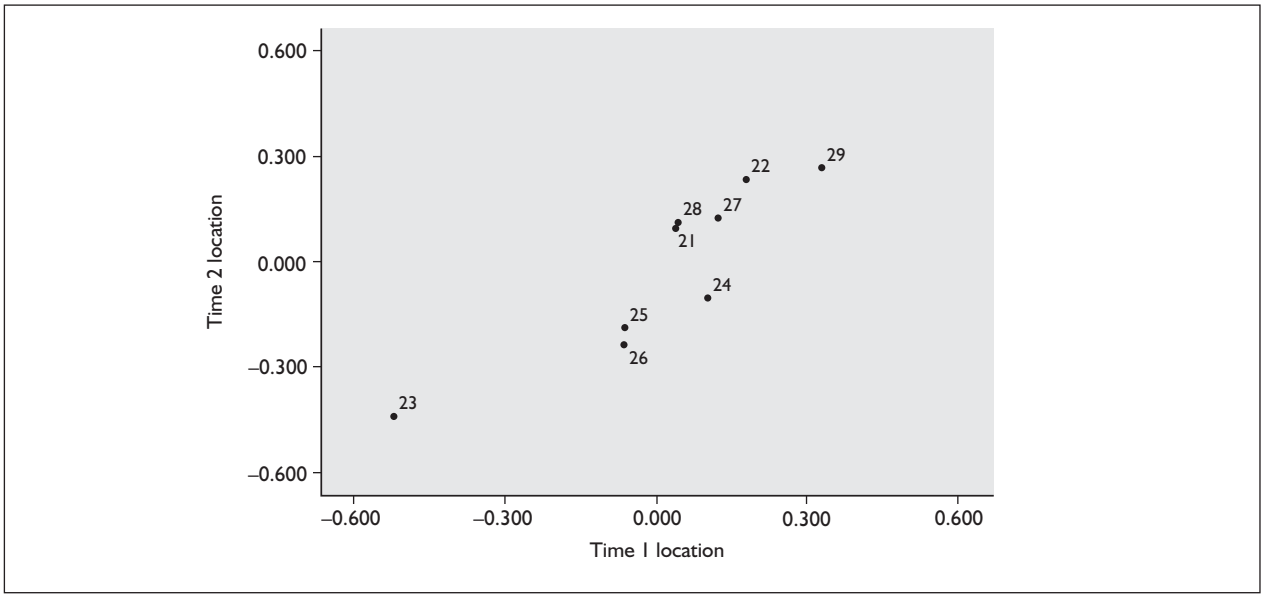


FIGURE 48 *MSIS-29 physical subscale – plot of item locations from time 1 and time 2.*

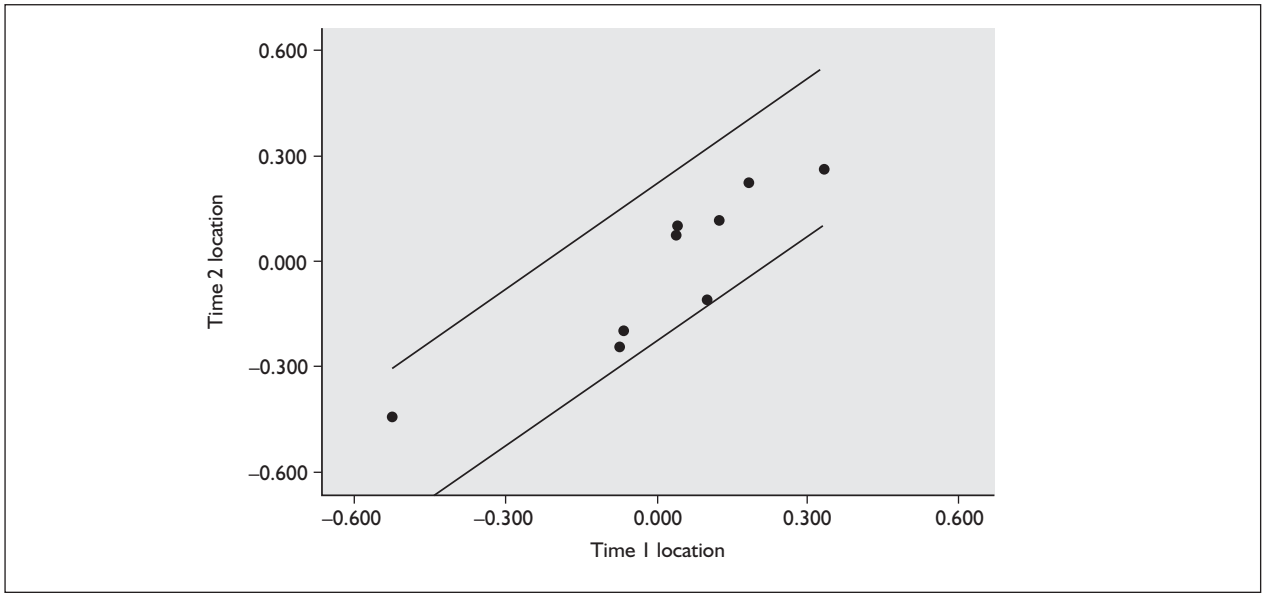


FIGURE 49 *MSIS-29 physical subscale – plot of item locations from time 1 and time 2. Black parallel lines indicate 95% confidence intervals.*

Person location analyses

Table 34 reports the findings from the *group-level analysis* of person locations. Person locations generated by the MSIS-29 physical and psychological subscales at T1 and T2 were not significantly different when examined using paired-sample *t*-tests or ANOVAs. Standardised SRM values of 0.04 (physical subscale) and 0.07 (psychological subscale) indicated little change and therefore low potential for bias.

Person locations generated by the MSIS-29 physical and psychological subscales at T1 and T2 had high intraclass correlation coefficients (physical = 0.97; psychological = 0.94). Thus, both scales satisfied the recommended minimum criterion of 0.80.

Table 35 shows the results of the *individual-level analysis* of person locations. In the total sample, the majority of people underwent no or non-significant change between the time points on each scale (78% physical subscale; 90% psychological subscale).

Figure 52 (physical subscale) and *Figure 53* (psychological subscale) show the relationships between the T1 and T2 person locations in this

TABLE 29 *Rasch analysis: equivalence of kind at the individual item level*

Change in disability (SED)	MSIS-29 physical subscale ($n = 99$)	MSIS-29 psychological subscale ($n = 121$)
Significant improvement	0%	0%
Non-significant improvement	55.6%	50%
No change	11.1%	0%
Non-significant deterioration	33.3%	50%
Significant deterioration	0%	0%

Significant improvement = SED $\geq +1.96$.
Non-significant improvement = $0 <$ SED $< +1.95$.
No change = SED = 0.
Non-significant worsening = $-1.95 <$ SED < 0.
Significant worsening = SED ≤ -1.96.

TABLE 30 *Rasch analysis: person locations implied by each MSIS-29 physical subscale total score*

	Person location	
Total score	Time 1	Time 2
0	–5.120	–4.920
1	–4.270	–4.100
2	–3.670	–3.540
3	–3.260	–3.160
4	–2.950	–2.870
5	–2.690	–2.640
6	–2.470	–2.440
7	–2.290	–2.270
8	–2.120	–2.120
9	–1.980	–1.980
10	–1.850	–1.860
11	–1.730	–1.750
12	–1.620	–1.650
13	–1.520	–1.560
14	–1.430	–1.470
15	–1.350	–1.390
16	–1.270	–1.310
17	–1.190	–1.240
18	–1.120	–1.170
19	–1.050	–1.100
20	–0.990	–1.040
21	–0.930	–0.980
22	–0.870	–0.920
23	–0.810	–0.860
24	–0.750	–0.810
25	–0.700	–0.760

TABLE 30 *Rasch analysis: person locations implied by each MSIS-29 physical subscale total score*

	Person location	
Total score	**Time 1**	**Time 2**
26	–0.640	–0.700
27	–0.590	–0.650
28	–0.540	–0.600
29	–0.490	–0.550
30	–0.440	–0.500
31	–0.400	–0.460
32	–0.350	–0.410
33	–0.300	–0.360
34	–0.260	–0.320
35	–0.210	–0.270
36	–0.160	–0.220
37	–0.120	–0.180
38	–0.070	–0.130
39	–0.030	–0.090
40	0.020	–0.040
41	0.060	0.010
42	0.110	0.050
43	0.160	0.100
44	0.200	0.150
45	0.250	0.190
46	0.300	0.240
47	0.340	0.290
48	0.390	0.340
49	0.440	0.390
50	0.490	0.440
51	0.540	0.490
52	0.590	0.540
53	0.640	0.590
54	0.690	0.650
55	0.750	0.700
56	0.800	0.760
57	0.860	0.820
58	0.920	0.880
59	0.980	0.940
60	1.040	1.010
61	1.110	1.080
62	1.180	1.150
63	1.250	1.220
64	1.320	1.300
65	1.400	1.380

continued

TABLE 30 *Rasch analysis: person locations implied by each MSIS-29 physical subscale total score (continued)*

	Person location	
Total score	**Time 1**	**Time 2**
66	1.480	1.470
67	1.570	1.560
68	1.670	1.660
69	1.770	1.760
70	1.880	1.880
71	2.010	2.000
72	2.140	2.140
73	2.290	2.300
74	2.470	2.470
75	2.670	2.670
76	2.900	2.910
77	3.190	3.200
78	3.570	3.580
79	4.140	4.130
80	4.960	4.940

sample. The vast majority of people have a T1 location with the 95% confidence intervals of the T1 location, although there are a few notable exceptions. Note, however, that the confidence intervals for these plots are computed as T1 location ± 1.96 T1 SE, hence they will differ slightly from the determination of significance of change when the standard error of the difference is used.

Summary

This chapter had two aims: first, to compare and contrast the evaluation of test–retest reproducibility using traditional and Rasch-based psychometric methods; second, to introduce and explain two methodological issues within the framework of Rasch-based evaluations – racking and stacking data for analysis and the evaluation of DIF. Both psychometric approaches came to similar conclusions, i.e. that the MSIS-29 has good test–retest reproducibility. It is therefore important to consider carefully the added value of using Rasch analysis.

Traditional methods base their conclusions on correlations between person scores that are scale dependent. Thus, they confound the performance of the scale with the measurement of the people. In contrast, Rasch analysis enabled us to study these two potentially confounding variables separately, in the knowledge that the evaluation of items and persons was independent of the sampling distribution of the other. This ability arises out of a property of the Rasch model – that the estimation of item parameters and person parameters is separated. In addition, the Rasch-based evaluation enabled a legitimate study at the individual person level because the analysis generates individualised standard errors.

The Rasch-based evaluation informed us that the scale was stable. Stability of the scale was studied with the data racked and stacked, as these data designs enable different but complementary evaluations. The item locations were stable, and items did not demonstrate differential functioning over time. Thus, the MSIS-29 was a stable ruler for measuring these people at these two points.

Ruler stability is a prerequisite for assessing and interpreting person stability. We were able to examine the equivalence in degree of disability of this sample across two time points, at both the group and individual levels. The group-level data implied no significant differences. The individual-level analyses implied that 22% of the sample had significantly different physical disability and 10% had significantly different psychological disability at the two time points. This is unexpected and out of keeping with the group-based inference of no difference between time points. These individuals

TABLE 31 *Rasch analysis: person locations implied by each MSIS-29 psychological subscale total score*

	Person location	
Total score	**Time 1**	**Time 2**
0	–4.010	–4.250
1	–3.160	–3.380
2	–2.590	–2.770
3	–2.200	–2.340
4	–1.910	–2.010
5	–1.670	–1.740
6	–1.460	–1.520
7	–1.290	–1.320
8	–1.130	–1.150
9	–0.990	–1.000
10	–0.860	–0.870
11	–0.740	–0.740
12	–0.620	–0.620
13	–0.510	–0.510
14	–0.410	–0.400
15	–0.300	–0.290
16	–0.200	–0.190
17	–0.100	–0.090
18	0.000	0.010
19	0.110	0.110
20	0.210	0.210
21	0.310	0.310
22	0.420	0.420
23	0.530	0.520
24	0.640	0.640
25	0.760	0.750
26	0.890	0.880
27	1.030	1.010
28	1.180	1.150
29	1.340	1.310
30	1.520	1.490
31	1.730	1.690
32	1.970	1.920
33	2.270	2.210
34	2.660	2.590
35	3.240	3.150
36	4.080	3.980

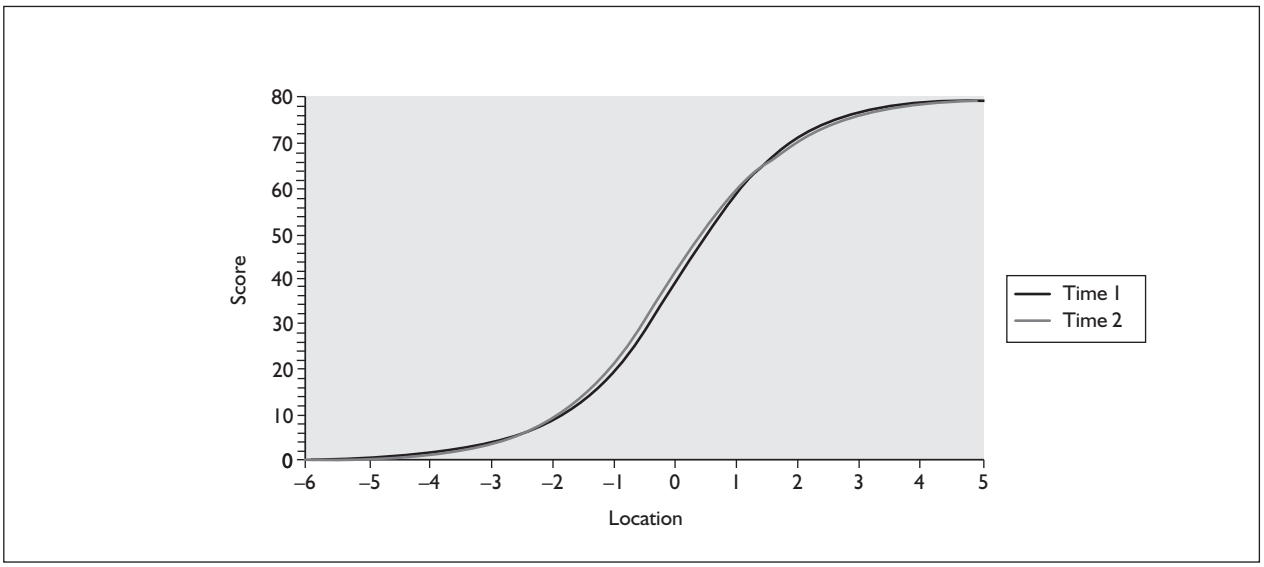


FIGURE 50 *MSIS-29 physical subscale – plot of predicted person locations from time 1 and time 2.*

can be identified and their results explored. For example, they could be evaluated qualitatively by interview to determine if there was a suitable explanation for the results. Also, the response patterns of these individuals could be examined. For example, we might hypothesise that the people who appear to change are those who respond inconsistently to items. To study this we examined whether people who changed between T1 and T2 were those people whose item response patterns were the most misfitting. *Figure 54* is a plot of significance of change against person–item fit residual for the physical subscale at T1 and T2, and suggests that this hypothesis does not explain these findings.

Another related point worthy of recapitulation is that variable standard errors are associated with different person location estimates. This is important for another reason. It demonstrates that significant change for any one individual is not simply a function of the magnitude of the change. It also depends on the location on the continuum at the measurement time points. This important fact is not accounted for in group-based analyses of change (see Chapter 8).

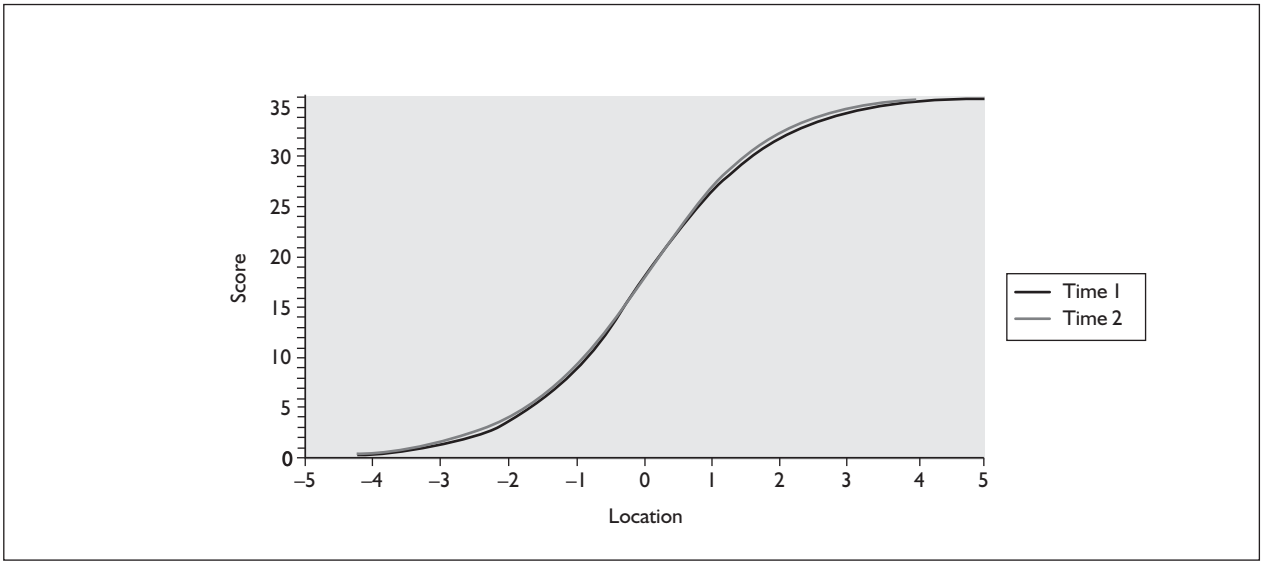


FIGURE 51 *MSIS-29 physical subscale – plot of predicted person locations from time 1 and time 2.*

TABLE 32 *Rasch analysis: MSIS-29 physical subscale – summary of differential functioning analysis*

	Time				Time by class interval			
Item	**MS**	***F***	**df**	***p***	**MS**	***F***	**df**	***p***
1	0.06	0.046	1	0.831037	0.4	0.328	3	0.80547
2	0	0.002	1	0.961278	0.52	0.473	3	0.701233
3	1.33	1.214	1	0.271541	0.96	0.88	3	0.452063
4	0.6	0.515	1	0.473765	0.49	0.419	3	0.73929
5	0.24	0.38	1	0.537942	0.11	0.177	3	0.91203
6	3.6	4.094	1	0.04409	0.81	0.927	3	0.428346
7	1.93	1.6	1	0.207059	1.18	0.974	3	0.405679
8	0.09	0.075	1	0.784432	1.5	1.312	3	0.270896
9	0.06	0.042	1	0.837056	0.1	0.063	3	0.979145
10	0.49	0.371	1	0.54288	1.21	0.909	3	0.437092
11	0.67	0.914	1	0.339995	0.61	0.822	3	0.483019
12	0.01	0.011	1	0.91557	0.55	1.021	3	0.383982
13	0.24	0.214	1	0.644198	0.11	0.1	3	0.960123
14	0.38	0.335	1	0.563494	2.85	2.52	3	0.058428
15	0.12	0.163	1	0.686378	0.22	0.302	3	0.823926
16	0.19	0.194	1	0.660105	0.37	0.387	3	0.762495
17	0.01	0.013	1	0.909383	0.19	0.237	3	0.870582
18	0.01	0.013	1	0.909857	0.14	0.256	3	0.856881
19	0.05	0.047	1	0.827792	0.45	0.438	3	0.726183
20	0.32	0.179	1	0.672953	0.27	0.148	3	0.931112

df, degrees of freedom; *F*, *F*-value; MS, mean square.

TABLE 33 *Rasch analysis: MSIS-29 psychological subscale – summary of differential functioning analysis*

	Time				Time by class interval			
Item	**MS**	***F***	**df**	***p***	**MS**	***F***	**df**	***p***
21	0.18	0.197	1	0.657797	2.02	2.173	3	0.091665
22	0.02	0.011	1	0.918075	0.84	0.588	3	0.623498
23	1.17	1.279	1	0.259108	0.26	0.281	3	0.838798
24	0.15	0.157	1	0.691999	0.89	0.952	3	0.415992
25	1.14	1.953	1	0.163484	0.37	0.638	3	0.591391
26	0.13	0.173	1	0.677653	2.71	3.548	3	0.015102
27	0.37	0.394	1	0.530592	0.08	0.083	3	0.969362
28	0.74	0.808	1	0.369414	0.26	0.278	3	0.841192
29	0.02	0.027	1	0.869227	0.32	0.463	3	0.70839

df, degrees of freedom; *F*, *F*-value; MS, mean square.

TABLE 34 *Rasch analysis: equivalence of degree at the sample level*

	MSIS-29 physical subscale		MSIS-29 psychological subscale	
	Time 1	**Time 2**	**Time 1**	**Time 2**
Mean (SD)	0.161 (1.813)	0.186 (1.742)	–0.369 (1.517)	–0.315 (1.492)
Change [mean (SD)]	–0.252 (0.583)		–0.054 (0.722)	
t–value (sig)	–0.428 ($p = 0.669$)		–0.739 ($p = 0.462$)	
ANOVA *F*-value (sig)	0.010 (0.921)		0.063 (0.802)	
Standardised response mean	0.04		0.07	
Intraclass correlation	0.97		0.94	

ANOVA, analysis of variance; SD, standard deviation.

TABLE 35 *Rasch analysis: equivalence of degree at the individual person level*

Change in disability (SED)	**MSIS-29 physical subscale ($n = 99$)**	**MSIS-29 psychological subscale ($n = 121$)**
Significant improvement	14%	4%
Non-significant improvement	29%	50%
No change	1%	1%
Non-significant deterioration	48%	39%
Significant deterioration	8%	6%

Significant improvement = SED ≥ +1.96.
Non-significant improvement = 0 < SED < +1.95.
No change = SED = 0.
Non-significant worsening = –1.95 < SED < 0.
Significant worsening = SED ≤ –1.96.

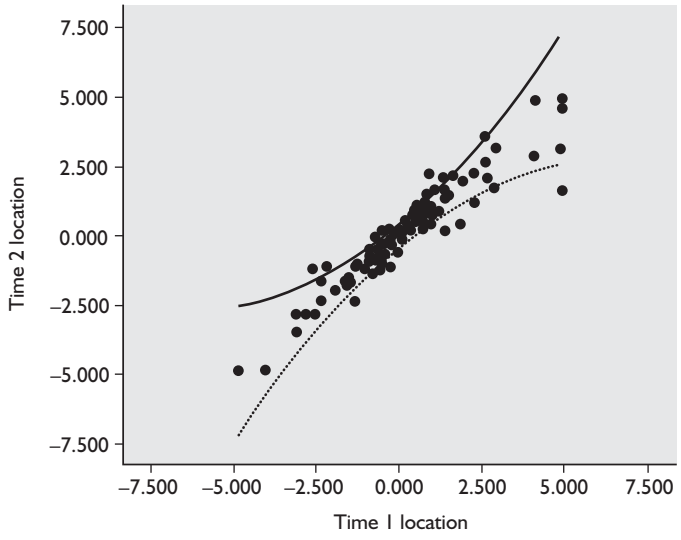


FIGURE 52 *MSIS-29 physical subscale – plot of person locations at time 1 and time 2. Curved lines indicate 95% confidence intervals.*

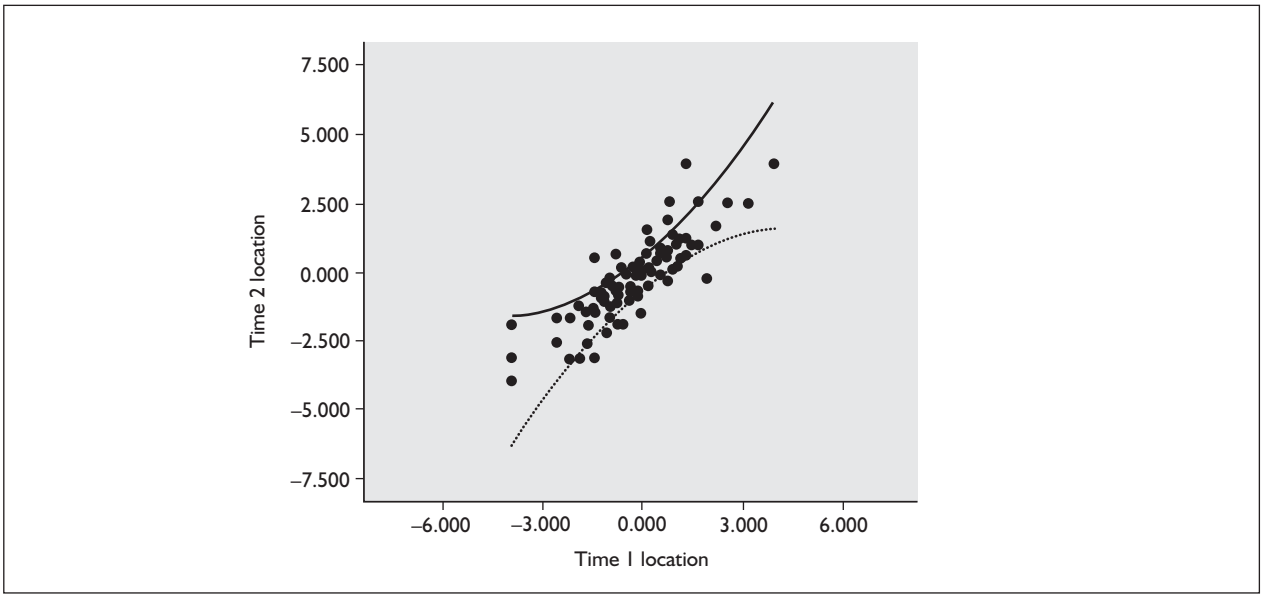


FIGURE 53 *MSIS-29 psychological subscale – plot of person locations at time 1 and time 2. Curved lines indicate 95% confidence intervals.*

In essence, Rasch analysis enables a very detailed and sophisticated examination of data. It brings notable added value to the evaluation and understanding of measurement problems.

The second aim of this chapter was to use the evaluation of test–retest reproducibility as a vehicle for introducing, explaining and demonstrating racking and stacking data designs and DIF. Both have uses way beyond the examination of test–retest reproducibility. Racking data, that is lining items up side by side, can be used to examine item stability under any circumstance in which the same person responds to the same item on more than one occasion (e.g. pre or post intervention). In addition, racking data is the basis for examining the extent to which items measure the same construct, i.e. testing dimensionality. This is demonstrated and discussed in Chapter 7. Stacking data, that is lining responses to the same item on top of each other, is the basis for examining DIF across different administrations (e.g. pre or post treatment, admission, discharge, etc.).

Racked and stacked data designs enable the generation of complementary information.

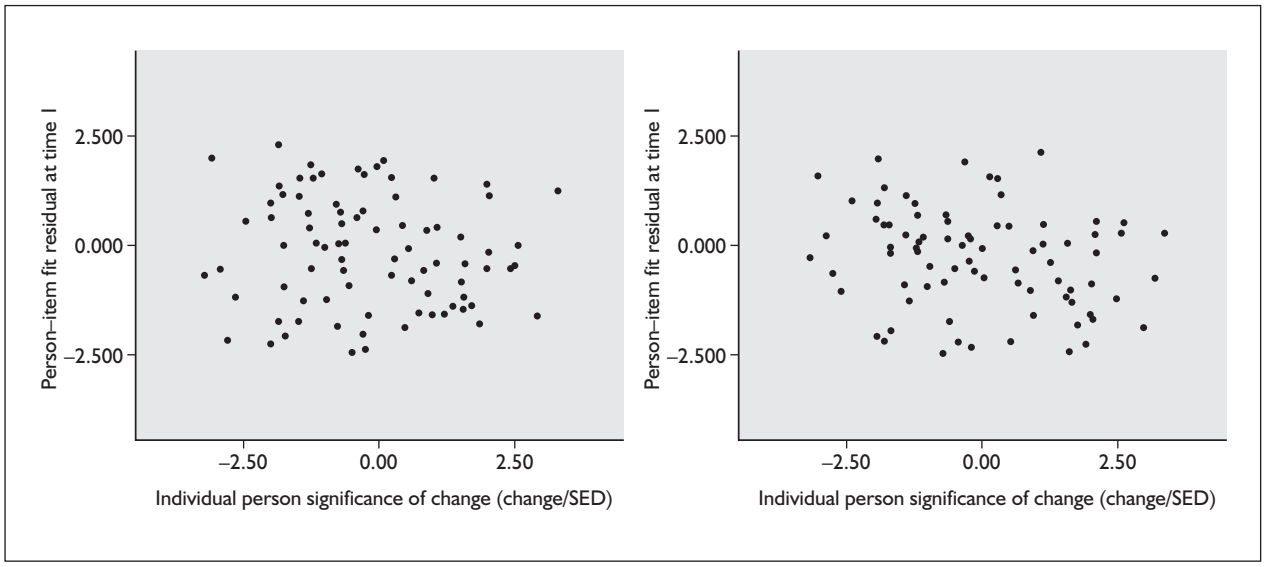


FIGURE 54 *MSIS-29 physical subscale – plot of significance of change against person fit residual.*

For example, in the examination of test–retest reproducibility, racking data give two locations for each item (T1 and T2) within the frame of reference determined by both sets of items (i.e. 40 physical or 18 psychological items). In contrast, stacking data gives one location for each item based on analysis of the data across two time points.

Differential item functioning (DIF) is a central concept in all item response models. The Rasch model was deduced on the *requirement* of invariance (stability). This does not mean, as has sometimes been believed, that examining data with the Rasch model automatically produces invariant results. It means that the Rasch model provides us with the facilities for testing if the items of our scales are invariant across the trait they measure and the clinically different groups in which they are used, and for testing the invariance of person locations measured by different sets of items.

Finally, a few words on Rasch analysis versus Item Response Theory (IRT). There is no doubt that both Rasch analysis and IRT offer the ability to undertake vastly more sophisticated analyses than traditional psychometric methods. However, in the context of analysing item and person locations, their stability and their changes, there is one facet in which Rasch analysis is superior to IRT. Rasch analysis, as we have stressed throughout this chapter, enables the item and person locations to be estimated separately – independently of the sampling distribution of each other. Thus, these parameters are not confounded. This is a mathematical property of the Rasch model. Unfortunately, the addition of other item parameters (e.g. discrimination, guessing) or person parameters (e.g. guessing) destroys the ability to estimate the parameters separately.[66] Massof[33] provides an excellent demonstration of this using the axioms of measurement.The inability to separate the parameters makes it much more difficult in practice, and impossible mathematically, to formally study invariance. This is one of the reasons underpinning Massof's statement that IRT models are not measurement models.[33]

Chapter 7

Equating rating scales using Rasch analysis

Overview

The aim of this chapter is to demonstrate one of the advantages of Rasch analysis over traditional psychometric methods: the ability to determine the extent to which different scales measure the same construct and can be equated, and if so to equate them on the same metric. We start by discussing the clinical problem and one of the fundamental reasons why equating scales is difficult with traditional psychometric methods. Next we provide an example of using Rasch analysis to equate scales on common metrics in physical functioning and psychological functioning.

Data were generated by a study to validate the MSIS-29 in which people with MS were sent a selection of scales to complete at the same point in time. This resulted in data from 563 people who completed four rating scales purporting to measure aspects of physical functioning and four scales purporting to measure psychological functioning.

For each dimension (physical and psychological functioning), scale response data were combined and analysed as if they were a single set of items. We examined four criteria indicating the extent to which each set of four measured a common construct: threshold estimates; item–person fit residuals; item–trait chi-squared values; and item characteristic curves. From the results of these analyses, we determined what modifications might improve the accuracy of equating. Equating tables were produced for both unmodified and modified versions.

Results supported the clinical impression that the four scales in each set measured common variables. There were some misfitting items whose removal improved the accuracy of equating. We present the scales equated on a common metric for each domain.

This study highlights some of the clinical advantages of Rasch analysis over more traditional psychometric methods in terms of assessing dimensionality and scale equating.

The clinical problem

Patient-completed rating scales are increasingly used as outcome measures in clinical trials of MS. Two domains are particularly important for measurement in MS and other disabling diseases: physical and psychological functioning. Consequently, many self-report rating scales now exist for these two dimensions. If these scales could be equated in a rigorous way, so that the scores on one scale could be interpreted in terms of the scores on another scale, comparisons of different studies and meta-analyses would be facilitated.

In the process of validating the MSIS-29[2,126] we collected data from additional rating scales purporting to measure physical and psychological functioning. These scales were the Medical Outcomes Study 36-item Short Form Health Survey (SF-36),[36] the Functional Assessment of MS (FAMS),[146] the self-report version of the 10-item Barthel Index (BI)[147] and the 12-item version of the General Health Questionnaire (GHQ).[148] This validation study gave us the opportunity to examine the potential to equate four scales for measuring physical functioning (the 20-item MSIS-29 physical impact scale; the 10-item SF-36 physical functioning dimension; the seven-item FAMS mobility scale; and the 10-item self-report version of the BI), and four scales for measuring psychological functioning (the nine-item MSIS-29 psychological impact scale; the five-item SF-36 mental health dimension; the seven-item FAMS emotional well-being scale; and the 12-item GHQ). Clinically, equating could be feasible as the scales purport to measure common variables. Statistically, however, correlations between the scale scores measuring the same dimensions ranged from 0.50 to 0.70 (mean 0.60), which raises concerns about the extent to which they measure a common construct.

Unfortunately, correlations between scale scores are not reliable indicators of the extent to which they measure a common construct. This is because the size of correlation coefficients is confounded by the relative distribution of the people on the variable

being measured.[122,149] For example, consider two physical functioning scales, the BI and SF-36 physical functioning (SF-36PF) scales. The BI is a measure of physical function for people with moderate to severe disability, and the SF-36PF is a measure for people with mild to moderate disability. The fact that these two scales measure at different distributions of physical functioning means that some people will obtain high scores on the BI and low scores on the SF-36PF. This will attenuate the correlation between these two scales. This inherent limitation of correlations can influence all correlation-based analyses including factor analysis and regression, which are used frequently in rating scale development and evaluation.

Rasch analysis [65,69,70,97] is a modern psychometric method for constructing and evaluating rating scales that can overcome the limitations of correlation coefficients and aid the equating of scales. There are four main reasons for this.

1. Rasch analysis is underpinned by testing the goodness-of-fit of observed data to a mathematical (Rasch) model, so it does not rely on correlations.
2. Rasch analysis determines the relationships between individual items, in terms of their relative locations on the hypothesised variable rather than the relationships between people's scale scores.
3. A mathematical property of the Rasch model is that the item location estimates are independent of the distribution of disability locations in the study sample.
4. By testing goodness-of-fit of observed data to a mathematical model, Rasch analysis determines formally the extent to which any group of items measure a common variable. Thus, it provides a formal test of dimensionality.

The aim of this study was to determine, using Rasch analysis, if and under what circumstances we could equate four physical functioning rating scales and four psychological functioning rating scales with each other, so that the results from different studies can be compared on the same metric.

Sample

The data were generated during the validation stage of the development of the MSIS-29.[2] Data were analysed from a postal survey of 1000 people randomly selected from the MS Society membership database. One random half-sample were sent a booklet (B1) containing the MSIS-29,[2] the SF-36,[36] the EuroQol Five Dimensions (EQ-5D)[150] and the self-report version of the BI.[147] The other random half-sample were sent a booklet (B2) containing the MSIS-29, the FAMS,[146] the EQ-5D and 12-item version of the GHQ (GHQ-12).[151] These scales are all described elsewhere.[2,126] Reminders were sent to non-responders at 3 and 5 weeks after the initial mail-out. Data collection was closed at 8 weeks. Ethics committees from the Institute of Neurology and National Hospital for Neurology and Neurosurgery approved the study.

Methods

Rating scales

In this analysis we focused only on selected scales. The four physical functioning scales were: the 20-item MSIS-29 physical impact scale (MSISphys); the 10-item SF-36 physical functioning dimension (SF-36PF); the seven-item FAMS mobility scale (FAMSmob) and the 10-item self-report version of the BI. The four psychological function scales were: the nine-item MSIS-29 psychological impact scale (MSISpsych); the five-item SF-36 mental health dimension (SF-36MH); the seven-item FAMS emotional well-being scale (FEW) and the 12-item GHQ.

Analysis plans

Data analysis for each dimension had three stages. The first stage was to determine, in broad terms, the extent to which the pooled items measured common dimensions, and thus determine whether it would be possible to equate the different scales on common metrics of physical and psychological functioning. The second stage was to determine what modifications, if any, to the two pools of items (e.g. removal of specific items) might improve the accuracy (reliability and validity) of equating. The third stage was to equate the scales, with and without modifications, so that clinicians could compare results generated using each instrument on a common metric. Data were analysed using the software program Rasch Unidimensional Measurement Model (RUMM2020).[100]

Stage 1: analysis of all items

In order to determine the extent to which the relevant four scales measured the same domain, we used the horizontal rating or 'racking' data design described in Chapter 6. Thus, items were

pooled and analysed as if they were a single rating scale. This gave two pools of items: 47 for physical functioning and 33 for psychological functioning. Equating scales is feasible when the majority of their items constitute a conformable set that map out a clinically and statistically meaningful variable. The set of items is clinically conformable if it appears to define, from more to less, a variable (in this case either physical or psychological functioning) in which the ordering of the items is meaningful. A set of items maps out in a statistically conformable way: if the observed data satisfy ('fit') the requirements of the Rasch measurement model; if the items spread out to produce a continuum from more to less; and if these items reliably separate people within the sample in terms of their level of physical functioning.

No one indicator of observed data-to-model fit is necessary and sufficient to summarise fit. Rather, decisions are informed by a combination of information. In this study, we examined four complementary indicators: ordering of item threshold estimates; item–person fit residuals; item–trait chi-squared statistics; and item–trait characteristic curves. These indicators have been described in detail previously, so only the key features are summarised below.

Threshold estimates
This issue was discussed in Chapter 5. To recap, the ordering of item threshold statistics indicates the extent to which the item response categories are working as intended, to define a progression from 'less' to 'more' functioning. Thresholds are transition points for adjacent categories. They mark the points on the continuum at which a person is equally likely to respond to one or other of two adjacent categories. There is empirical support for the response categories when the threshold estimates are appropriately ordered.[99] For example, on the MSIS-29, if the item response options are working as intended, the order of threshold estimates for each item should be 'not at all'/'a little'; 'a little'/'moderately'; 'moderately'/'quite a lot'; 'quite a lot'/'extremely'.

Item–person fit residuals
The item–person fit residual for each item summarises the fit of the observed data to the statistical model from the perspective of the items. A residual is the difference between the observed response (score) of a person to an item and the expected value of that person to that item as predicted by the model. For each item, residuals are generated for every person in the sample who responds to that item. These residuals are then combined across persons to give a summary value which is then standardised and transformed so that perfect fit has a mean of 0 and standard deviation of 1. Larger fit residuals mean worse fit of observed data to the measurement model. Values in the range −2.5 to +2.5 are considered within statistically acceptable limits.[118] Thus, the size of the fit residual indicates the degree to which observed responses to items are inconsistent with predictions based on the mathematical model. The accompanying positive or negative sign gives information as to the nature of this misfit.

Item–trait chi-squared values
The item–trait chi-squared value for each item summarises the fit of the data to the model from the perspective of the variable measured by the items (in this case, physical or psychological functioning). In essence, this chi-squared statistic compares, for groups of people with a similar level of disability (known as 'class intervals'), the observed score and expected value on each item. The larger the difference between these two values, the greater the difference between observed score and expected value, the greater the chi-squared value and the worse the fit of observed data to model expectations.

Item characteristic curves (ICC)
The ICC for each item is a graphic indicator of fit which provides complementary qualitative information about the fit of the observed data to the model, from the perspective of the trait measured by the items. The ICC is the plot of the expected value for an item (y-axis) against the latent variable measured by the set of items (in this case, physical or psychological functioning). That plot includes the observed mean scores for the people in each class interval defined by their level of functioning. The better the fit of the data to the model, the closer the proximity of the observed scores to expected values.

Fit statistics provide complementary information. Therefore, each indicator of fit should be interpreted in the context of the others and, most importantly, within the clinical context of the variable we are intending to define and measure. Statistical test of fit provides stringent tests of the extent to which observed data satisfy model requirements. Consequently, misfit is to be expected. A key focus of the analysis is to go beyond the identification of misfit and seek to explain why items initially hypothesised to

belong to a common variable do not support that prediction.[66]

Item locations and person separation indices

The range and spread of the item locations, for each individual scale and all four scales combined, was examined to determine the extent to which they mapped out a variable. The person separation indices were examined to determine the extent to which the combined item pools were useful for discriminating among people.

Stage 2: modifications and reanalysis

The aim of stage 2 was to determine what modifications (if any) are required for, or might improve, accuracy of equating. All items misfitting on any of the four criteria described above were identified and examined in detail in terms of the nature of the misfit, the nature of the item and the nature of the functioning variables. Consequences of item removal were examined in terms of the four fit indicators described above, reliability indices and targeting of the items to the distribution of functioning in the sample.

Stage 3: equating

The four rating scales were equated for each variable, with (if possible) and without the modifications suggested by the analysis. Equating without modifications is valuable as it enables people who only have total scores available, or who do not have the facilities to undertake further analyses to benefit from best estimate equating on an interval-level common metric (logit scale). From the tables we provide in the Results section below, investigators can determine the best estimate equivalent raw score on each of the other three scales if they wish.

Results

Sample characteristics

Responses were analysed from a total of 563 people (B1 = 288; B2 = 275). The response rate

TABLE 36 *Sample characteristics (n = 563)*

Characteristic	Numerical value
Age [mean (SD); range]	52.5 (12.0); 18–82
Sex (percentage female)	72.1%
Duration of MS	
≤7 years	14.9%
8–12 years	14.7%
13–20 years	26.7%
≥21 years	43.7%
Indoor mobility	
Unaided	29.9%
Uses a walking aid	40.7%
Uses a wheelchair	29.4%
Marital status	
Married	68.6%
Lives alone	18.4%
Lives with others	81.6%
Employment	
Retired as a result of MS	57.6%
Employed	20%
Educational level	
Degree/professional qualification	26.7%

MS, multiple sclerosis; SD, standard deviation.

was 63% (for further details see references 2 and 126). *Table 36* shows the sample characteristics. Although a wide range of age and disease duration was represented, this was an older sample of people with MS with relatively long disease duration. There were no differences between the characteristics of the random half-samples. It should be noted that, documenting the development and validation of the MSIS-29,[126] in the original article this was defined by chi-squared analyses of the categorical data (e.g. gender and marital status) and *t*-test for continuous data.

Physical functioning

Stage 1 – analysis of all 47 physical functioning items as a single set

Table 37 shows, for each of the 47 items, the item locations relative to each other on an interval-level continuum; the associated standard errors for these estimates; which items had reversed threshold estimates; and the item–person fit residuals and the item–trait chi-squared values and their associated probabilities. The power for detecting misfitting items was considered excellent.

Threshold estimates

A total of 16 items had reversed thresholds (six MSIS, one SF-36, five FAMS and four BI). This indicates that the ordering of response options did not work as intended for these items. Closer examination of the reversed thresholds, and the category probability curves for the items, showed that for 13 of 16 items (all except SF-36 Q08, BI Q06 and BI Q09) the values of the reversed thresholds were very similar. This finding implies that people with MS had difficulty discriminating reliably between the five response options, and that these items would probably operate better with fewer response categories.

Item–person fit residuals

Table 37 shows the numerical values of the fit residuals. *Figure 55* represents these values graphically. Most of the item–person fit residuals (34 of 47) lie within recommended confidence intervals (−2.5 to +2.5). A further 10 items lie outside, but near to, these limits. Three items, FAMS Q05 ('my legs are strong'), MSIS Q20 ('needing to go to the toilet urgently') and FAMS Q02 ('I am able to work') lie considerably outside the confidence intervals and notably away from the other items.

Item–trait chi-squared values

Table 37 shows the numerical values of the chi-squared statistics. *Figure 56* represents these values graphically. Most items (39 of 47) have similar chi-squared values ($\chi^2 < 20$), a further five items have larger values ($\chi^2 = 23$–38), then there are notable gaps until items FAMS Q05 ($\chi^2 = 48$) and MSIS Q20 ($\chi^2 = 68$), and a huge gap until item FAMS Q02 ($\chi^2 = 218$).

Item characteristic curves (ICCs)

Figures 57–59 show the ICCs for the three items with particularly large item–person fit residuals and item–trait chi-squared statistics (FAMS Q02, MSIS Q20, FAMS Q05). These demonstrate that item FAMS Q02 (*Figure 57*) had notable discrepancies between the observed and predicted item scores for most class intervals. The other two items had less severe graphical, but still notable numerical, discrepancies between observed and predicted values. Examination of the ICCs for the items with smaller degrees of misfit showed good coherence between the observed and predicted item scores for each class interval.

Item locations and person separation index (PSI)

Figure 60 shows the relative item locations for the four scales and for all 47 items combined. Each black dot represents the item location for a single item, which is the mean of the threshold locations. Item locations ranged from −4.08 to +3.14 units (logits). This indicates that the items are spread out and therefore map out a continuum from less to more.

Figure 60 shows that the BI measures towards the more disabled end of the physical functioning spectrum, the SF-36PF measures towards the less disabled end and the FAMS and MSIS-29 measure in the middle of the continuum. As a 47-item pool there is a reasonable spread of coverage but there is bunching of items centrally, and there are notable gaps in the continuum.

Figure 61 shows the distribution of the sample (person locations) relative to the distribution of the item locations. The sample is spread over a wide range. The person distribution covers the items well, but the items do not provide good coverage of the persons. The PSI was high at 0.969. This indicates that the 47-item scale is a useful measure for discriminating among people in this variable.

TABLE 37 *Physical functioning (all 47 items)*

Scale/item	Location	Standard error	Thresholds	Fit residual	χ^2	χ^2 probability
MSISphys						
01	–0.753	0.055		1.068	8.995	0.061
02	0.597	0.048		1.655	12.663	0.013
03	–0.166	0.051		–0.461	18.570	0.001
04	–0.378	0.051		2.550	11.079	0.026
05	0.252	0.048		–5.947	37.693	0.000
06	0.048	0.051		–0.225	6.578	0.160
07	0.213	0.049		4.859	21.196	0.000
08	–0.119	0.049		3.869	13.035	0.011
09	0.937	0.047		3.719	29.449	0.000
10	0.702	0.047		5.104	35.831	0.000
11	–0.020	0.049		–1.376	9.899	0.042
12	0.033	0.047		–5.482	34.076	0.000
13	0.230	0.048		0.744	5.126	0.275
14	0.134	0.045	R	0.788	2.880	0.578
15	0.504	0.048	R	–0.705	15.065	0.005
16	–0.180	0.049		0.344	11.084	0.026
17	0.286	0.044	R	–2.022	13.724	0.008
18	–0.473	0.052	R	–3.542	22.161	0.000
19	–0.222	0.046	R	–2.136	8.483	0.075
20	–0.110	0.047	R	9.956	68.299	0.000
SF-36PF						
01	–4.081	0.246		–1.217	19.793	0.001
02	–1.485	0.124		–1.755	10.831	0.029
03	–1.438	0.123		–1.591	17.263	0.002
04	–2.773	0.163		–1.098	13.183	0.010
05	–0.811	0.114		–2.092	10.338	0.035
06	–1.014	0.114		–1.192	4.700	0.319
07	–2.671	0.175		–1.195	17.384	0.002
08	–1.655	0.132	R	–1.381	8.032	0.090
09	–0.626	0.106		–1.738	9.379	0.052
10	0.470	0.107		–2.617	17.868	0.001
FAMSmob						
01	0.174	0.063	R	4.653	1.670	0.796
02	–0.114	0.064	R	10.679	206.640	0.000
03	–0.622	0.071	R	–0.947	4.624	0.328
04	–0.419	0.069		–0.837	5.996	0.199
05	–0.497	0.069	R	6.805	44.192	0.000
06	–0.347	0.068		–1.294	4.511	0.341
07	–0.729	0.070	R	–0.927	1.581	0.812

TABLE 37 *Physical functioning (all 47 items)*

Scale/item	Location	Standard error	Thresholds	Fit residual	χ^2	χ^2 probability
BI						
01	0.213	0.140		–1.354	11.309	0.023
02	2.600	0.110		1.017	9.350	0.053
03	1.110	0.100		–2.196	11.689	0.020
04	2.962	0.141		–1.285	10.551	0.032
05	1.489	0.085	R	3.626	5.265	0.261
06	0.108	0.088	R	1.569	20.053	0.000
07	2.191	0.123	R	–0.367	8.875	0.064
08	3.143	0.215		–0.887	14.356	0.006
09	0.818	0.090	R	1.476	15.550	0.004
10	2.487	0.127		0.199	6.441	0.169

BI, Barthel Index; FAMSmob, functional assessment of multiple sclerosis mobility scale; MSISphys, Multiple Sclerosis Impact Scale physical subscale; R, reversed thresholds; SF-36PF, Medical Outcomes Study 36-item Short Form Health Survey physical functioning dimension.

That is, it can provide reliable measures of different levels of physical functioning.

Interpretation of stage 1 analyses

The findings, when considered together, indicate that most items work well together to form a set that could be used to measure patients reliably on a physical functioning continuum. This indicates that all four scales measured a common dimension and that equating of the four scales was feasible. However, the analyses suggested that deleting the three items that failed all fit criteria would improve the accuracy of equating, and that this accuracy might be improved further if the response options were working as intended. Although 10 items demonstrated lesser degrees of misfit on some indicators, the ICCs for all these items showed good coherence between observed and predicted responses supporting the retention of these items.

Stage 2: modification and re-analysis

All misfitting items were examined in an attempt to explain the reason for their misfit. The three notably statistically misfitting items were: 'I am

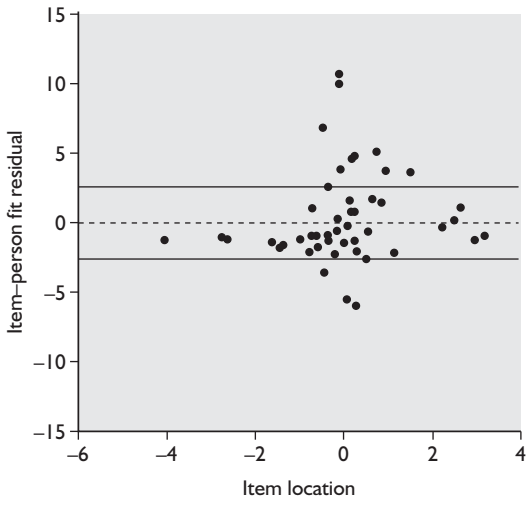


FIGURE 55 *Physical function scale (47 items) – plot of fit residuals.*

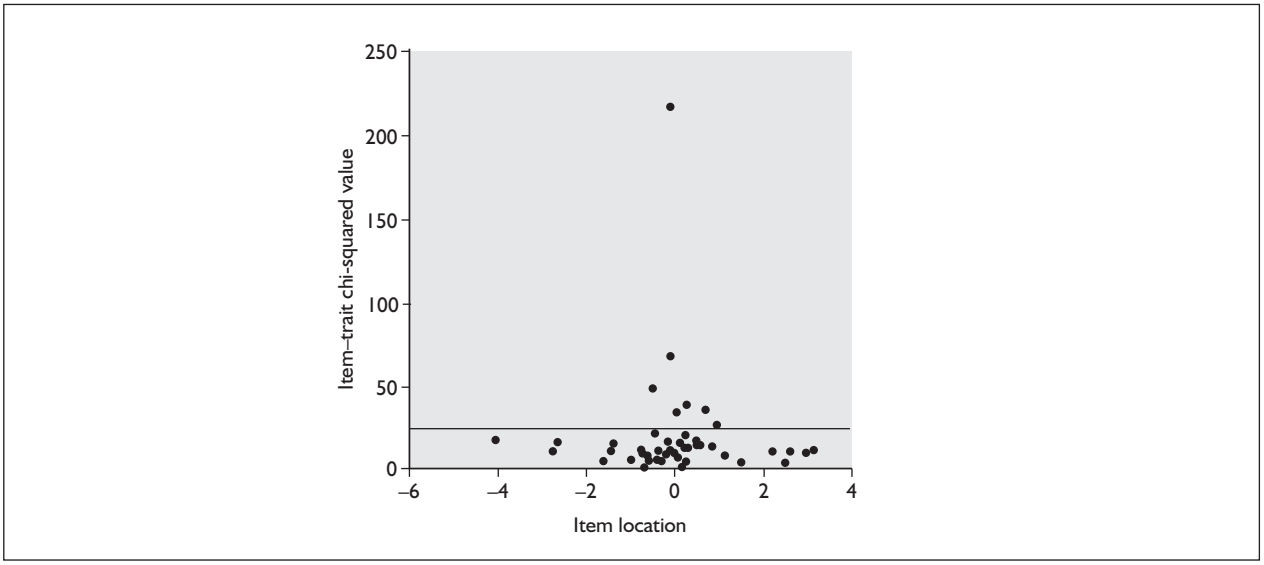


FIGURE 56 *Physical function scale (47 items) – plot of chi-squared values.*

able to work' (FAMS Q02); 'needing to go to the toilet urgently' (MSIS Q20); and 'my legs are strong' (FAMS Q05). Clinically, all three items are non-specific indicators of physical functioning. This was supported by the ICCs for these items, which demonstrated that the curve representing the observed scores was flatter than the curves of the expected values. Thus, misfit was due to limited discrimination across the range of the scale. *Figure 57* shows this for the worst item (FAMS Q02). These findings suggested that equating would be more accurate if these three items were removed. Examination of the content of items with lesser degrees of misfit indicated that it was more appropriate that they were retained in a measure of physical functioning. This was supported statistically by the ICCs, which demonstrated good coherence between observed and predicted responses.

Items FAMS Q05, MSIS Q20 and FAMS Q02 were removed and a Rasch analysis was performed on the remaining 44 items. The purpose of this analysis was to determine if deleting three items significantly altered the fit indicators. No fit indicators were altered significantly. A total of 12 items still had reversed thresholds. The targeting of the 44 items to patients was good. The reliability of the 44 items was high (PSI = 0.97).

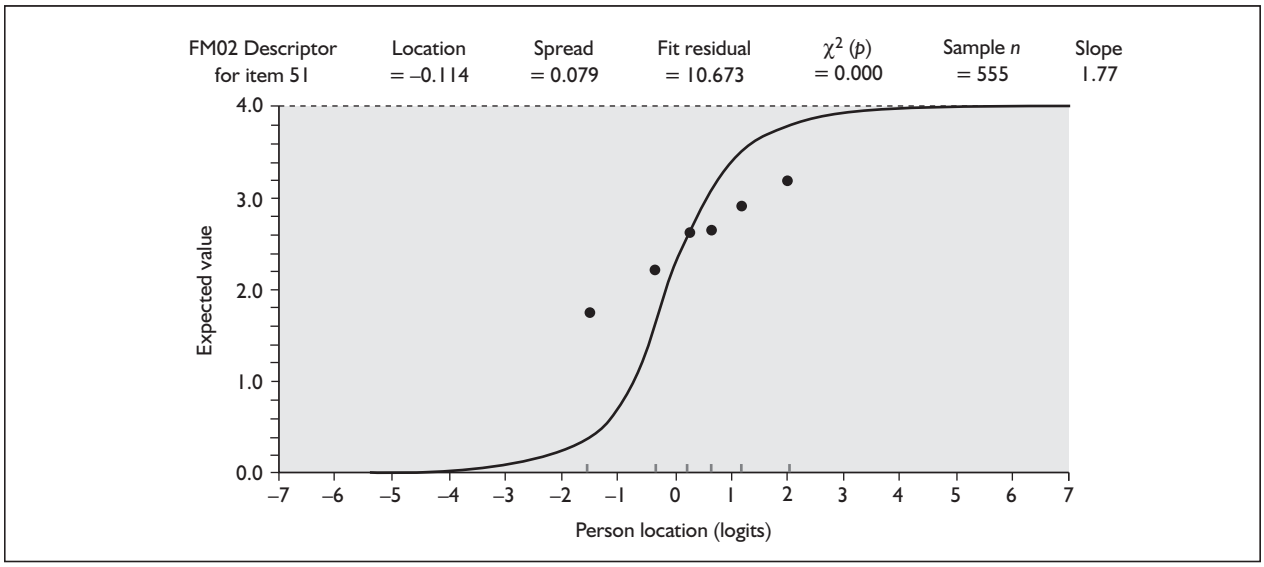


FIGURE 57 *Item characteristic curve for item FAMS 02.*

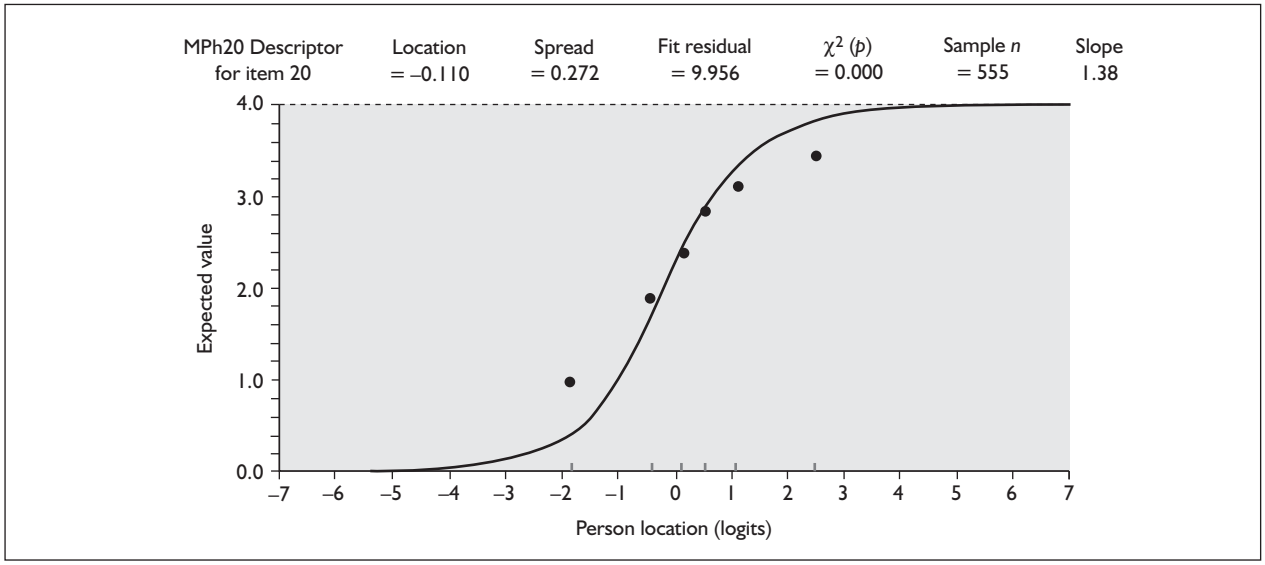


FIGURE 58 *Item characteristic curve for item MSIS 20.*

Stage 3: equating

Table 38 shows the values on a common physical functioning interval-level metric associated with any raw score on any of the scales. This table was computed for the scales without any items removed. Note that all raw scores for all scales have been set to start at 0. Thus, values for the MSIS-29 physical scale, which on the questionnaire range from 20 to 100, now range from 0 to 80. Similarly, values for the SF-36PF, which on the questionnaire range from 10 to 30, now range from 0 to 20. To use *Table 38* simply find the raw score for any scale in the first column, and then read across to the column of the scale of interest to find the value on an interval scale, shared with all the other scales. For example, a raw score of 10 on the SF-36 implies a measure of −1.58 logits. To find the corresponding raw score for the MSIS-29 physical scale, find the nearest value to −1.58 in the second column (−1.57), and read off the associated raw score (11). *Figure 62* is a graphical representation of the relationships between the four scales and the common variable. The *y*-axis is the raw score. The *x*-axis is the common physical functioning variable defined by the four scales.

Table 39 shows the same equating scales when three items were removed from the analysis. Note

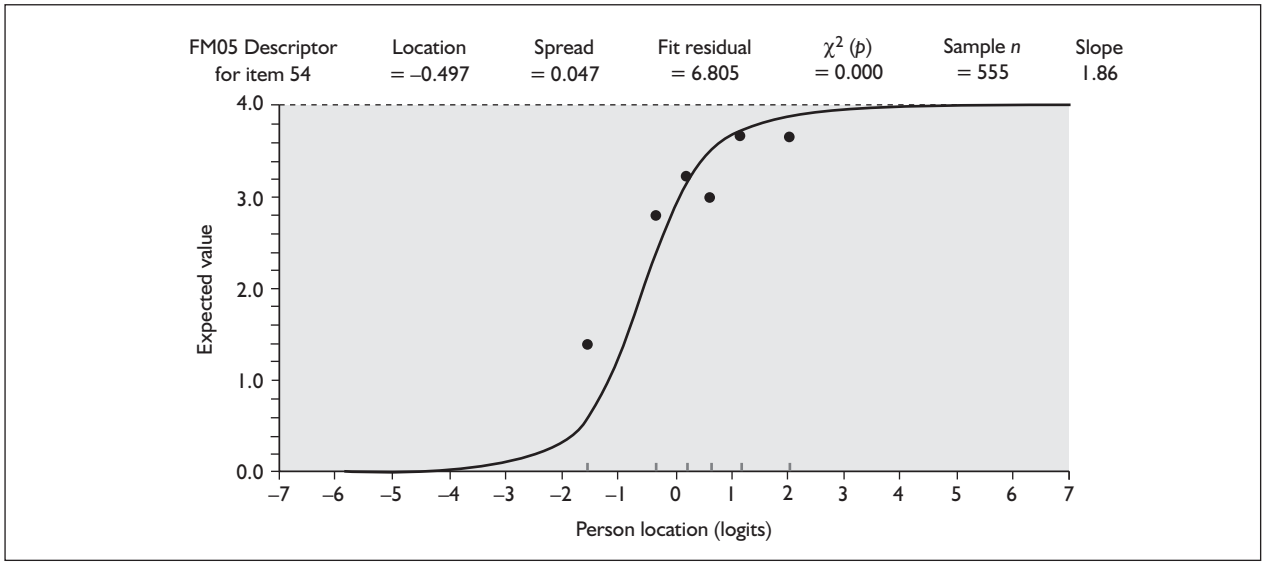


FIGURE 59 *Item characteristic curve for item FAMS 0.5.*

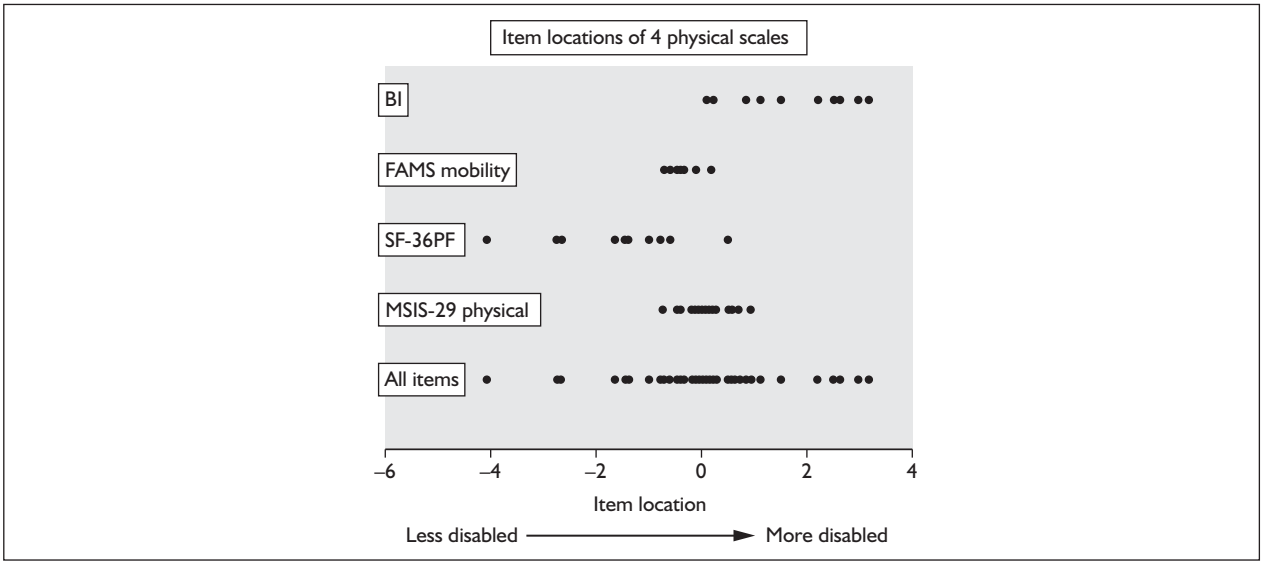


FIGURE 60 *Plot of item location.*

that the values in *Tables 38* and *39* differ. This is because the removal of three items changes the frame of reference, albeit slightly, and the mean item location which is always centred on 0.

Psychological functioning

Stage 1 – Analysis of all 33 psychological functioning items as a single set

Table 40 shows, for each of the 33 items, the item locations relative to each other on an interval-level continuum, the associated standard errors for these estimates, which items had reversed threshold estimates, the item–person fit residuals, and the item–trait chi-squared values and their associated probabilities. The power for detecting misfitting items was considered excellent.

Threshold estimates

A total of eight items had reversed thresholds, which indicates that the ordering of response options did not work as intended for these items. Closer examination of the reversed thresholds and the category probability curves for the items showed that for four of eight items (all except SF-36MH Q01, FEW Q04, FEW Q05 and FEW Q06)

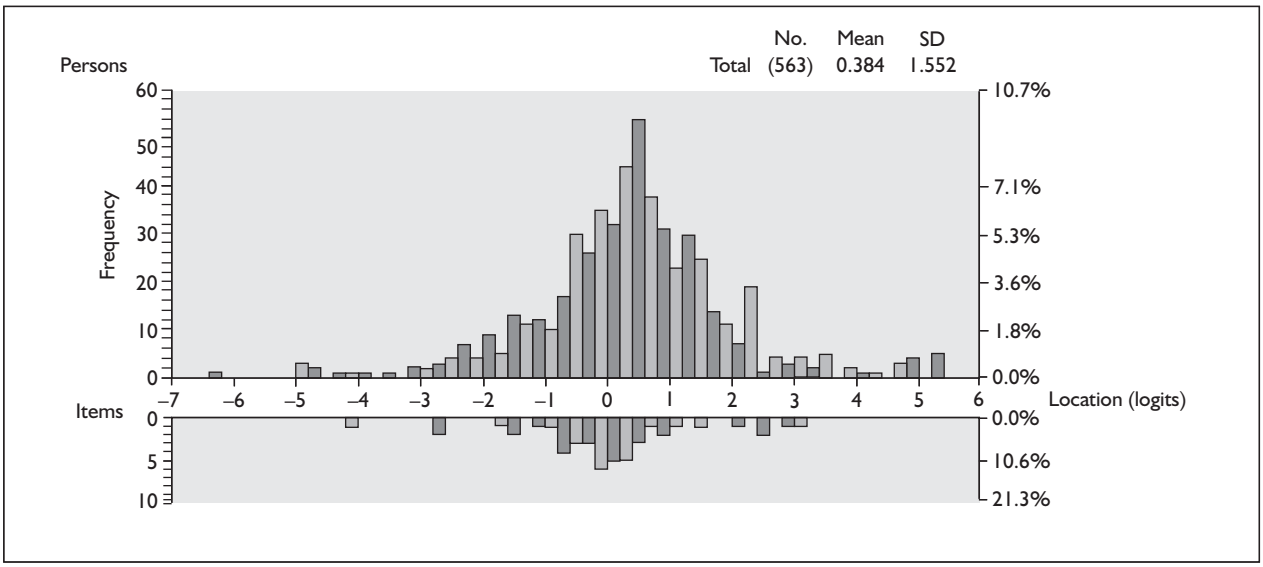


FIGURE 61 *Targeting of sample to 47 physical items. Person–item location distribution (grouping set to interval length of 0.20, making 65 groups).*

TABLE 38 *Equating physical functioning scales for MS (all 47 items)*

	Values on a common metric			
Raw score	**MSISphys**	**SF-36PF**	**FAMSmob**	**BI**
0	–4.62	–6.07	–3.28	–1.46
1	–3.82	–4.99	–2.50	–0.82
2	–3.28	–4.22	–2.01	–0.35
3	–2.91	–3.69	–1.70	–0.01
4	–2.63	–3.27	–1.48	0.26
5	–2.41	–2.93	–1.30	0.51
6	–2.22	–2.62	–1.16	0.74
7	–2.06	–2.34	–1.03	0.96
8	–1.92	–2.08	–0.92	1.17
9	–1.79	–1.83	–0.82	1.38
10	–1.68	–1.58	–0.72	1.59
11	–1.57	–1.34	–0.63	1.81
12	–1.47	–1.09	–0.54	2.04
13	–1.38	–0.83	–0.45	2.28
14	–1.30	–0.55	–0.37	2.55
15	–1.22	–0.25	–0.28	2.84
16	–1.15	0.08	–0.19	3.19
17	–1.08	0.47	–0.10	3.61
18	–1.01	0.96	–0.01	4.13
19	–0.95	1.65	0.09	4.85
20	–0.89	2.62	0.20	5.82
21	–0.83		0.32	
22	–0.77		0.45	
23	–0.72		0.60	
24	–0.66		0.77	
25	–0.61		0.99	
26	–0.56		1.28	
27	–0.51		1.74	
28	–0.46		2.44	
29	–0.42		–3.28	
30	–0.37			
31	–0.32			
32	–0.28			
33	–0.23			
34	–0.19			
35	–0.15			
36	–0.10			
37	–0.06			
38	–0.02			
39	0.03			

continued

TABLE 38 *Equating physical functioning scales for MS (all 47 items)*

	Values on a common metric			
Raw score	**MSISphys**	**SF-36PF**	**FAMSmob**	**BI**
40	0.07			
41	0.11			
42	0.16			
43	0.20			
44	0.24			
45	0.29			
46	0.33			
47	0.38			
48	0.42			
49	0.47			
50	0.51			
51	0.56			
52	0.61			
53	0.66			
54	0.71			
55	0.76			
56	0.81			
57	0.87			
58	0.92			
59	0.98			
60	1.04			
61	1.10			
62	1.17			
63	1.24			
64	1.31			
65	1.39			
66	1.46			
67	1.55			
68	1.64			
69	1.74			
70	1.84			
71	1.96			
72	2.08			
73	2.23			
74	2.39			
75	2.57			
76	2.79			
77	3.06			
78	3.42			
79	3.95			
80	4.74			

BI, Barthel Index; FAMSmob, functional assessment of multiple sclerosis mobility scale; MSISphys, Multiple Sclerosis Impact Scale physical subscale; SF-36PF, Medical Outcomes Study 36-item Short Form Health Survey physical functioning dimension.

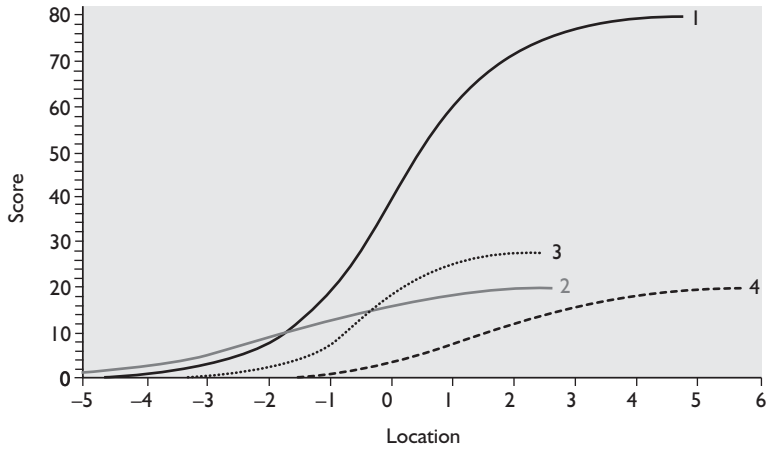


FIGURE 62 *Equating of the four physical functioning scales. 1, MSIS-29 physical; 2, SF-36PF; 3, FAMS mobility; 4, Barthel Index.*

TABLE 39 *Equating physical functioning scales for MS (excluding three items)*

	Values on a common metric			
Raw score	**MSISphys**	**SF36PF**	**FAMSmob**	**BI**
0	–4.69	–6.29	–3.30	–1.46
1	–3.88	–5.13	–2.47	–0.82
2	–3.33	–4.33	–1.94	–0.35
3	–2.96	–3.78	–1.60	–0.01
4	–2.68	–3.36	–1.35	0.27
5	–2.45	–3.01	–1.15	0.52
6	–2.26	–2.69	–0.98	0.75
7	–2.09	–2.41	–0.83	0.97
8	–1.94	–2.14	–0.69	1.19
9	–1.81	–1.89	–0.55	1.40
10	–1.69	–1.64	–0.42	1.62
11	–1.58	–1.38	–0.28	1.84
12	–1.48	–1.13	–0.14	2.08
13	–1.39	–0.86	0.00	2.33
14	–1.30	–0.58	0.15	2.59
15	–1.22	–0.28	0.33	2.90
16	–1.14	0.06	0.52	3.24
17	–1.06	0.46	0.76	3.67
18	–0.99	0.95	1.06	4.20
19	–0.92	1.65	1.52	4.92
20	–0.86	2.64	2.19	5.90
21	–0.80			

continued

TABLE 39 *Equating physical functioning scales for MS (excluding three items)*

	Values on a common metric			
Raw score	**MSISphys**	**SF36PF**	**FAMSmob**	**BI**
22	–0.74			
23	–0.68			
24	–0.62			
25	–0.57			
26	–0.51			
27	–0.46			
28	–0.41			
29	–0.36			
30	–0.31			
31	–0.26			
32	–0.21			
33	–0.16			
34	–0.11			
35	–0.06			
36	–0.02			
37	0.03			
38	0.08			
39	0.12			
40	0.17			
41	0.22			
42	0.27			
43	0.32			
44	0.36			
45	0.41			
46	0.46			
47	0.51			
48	0.56			
49	0.62			
50	0.67			
51	0.72			
52	0.78			
53	0.84			
54	0.90			
55	0.96			
56	1.02			
57	1.08			
58	1.15			
59	1.22			
60	1.30			
61	1.37			
62	1.46			

TABLE 39 *Equating physical functioning scales for MS (excluding three items)*

	Values on a common metric			
Raw score	**MSISphys**	**SF36PF**	**FAMSmob**	**BI**
63	1.54			
64	1.64			
65	1.74			
66	1.84			
67	1.96			
68	2.09			
69	2.23			
70	2.40			
71	2.58			
72	2.80			
73	3.08			
74	3.44			
75	3.97			
76	4.77			

BI, Barthel Index; FAMSmob, functional assessment of multiple sclerosis mobility scale; MSISphys, Multiple Sclerosis Impact Scale physical subscale; SF-36PF, Medical Outcomes Study 36-item Short Form Health Survey physical functioning dimension.

the values of the reversed thresholds were very similar. This finding implies that people with MS had difficulty in discriminating reliably between the multiple response options of these items, and that these items would probably operate better with fewer response categories.

Item–person fit residuals

Table 40 shows the numerical values of the fit residuals. *Figure 63* represents these values graphically. Most of the item–person fit residuals (23 of 33) lie within the recommended confidence intervals (−2.5 to +2.5), with 26 of 33 items lying in the range −3.0 to +3.0. Only three items lie notably away from this boundary: MSIS Q22 ('problems sleeping') = +9.6; FEW Q03 ('I am able to enjoy life') = +5.9; MSIS Q29 ('feeling depressed') = −4.7.

Item–trait chi-squared values

Table 40 shows the numerical values of the chi-squared statistics. *Figure 64* represents these values graphically. One item (MSIS Q22, 'problems sleeping = 165.4) is vastly different to the others, with two additional items (MSIS Q29, 'bothered by feeling depressed' = 36.3; FEW Q03, 'I am able to enjoy life' = 34.4) being relatively distant from the main group.

Item characteristic curves (ICCs)

Figures 65–67 show the ICCs for the three most misfitting items (MSIS Q22, MSIS Q29, FEW Q03). These demonstrate that item MSIS Q22 (see *Figure 64*) had notable discrepancies between the observed and predicted item scores for most class intervals. However, the other two items had less disturbing graphical appearances of misfit. Examination of the ICCs for the items with smaller degrees of misfit showed good coherence between the observed and predicted item scores for each class interval.

Item locations and person separation index

Figure 68 shows the relative item locations for the four scales and for all 33 items combined. Relative item locations for all 33 ranged from −0.80 to +0.76 logits. This indicates that the items do map out a continuum from less to more, but have a narrow spread.

Figure 67 shows that the GHQ-12 has the widest coverage across the continuum, but with notable gaps. The 33-item pool provided reasonably consistent coverage over the continuum, albeit over a narrow range.

TABLE 40 *Psychological functioning (all 33 items).*

Scale/item	Location	SE	Thresholds	Fit residual	χ^2	χ^2 probability
MSISpsych						
21	–0.023	0.048		0.167	3.476	0.627
22	0.055	0.044		9.634	165.360	0.000
23	–0.627	0.046		3.490	11.821	0.037
24	–0.067	0.047		–1.175	15.640	0.008
25	–0.072	0.048		–3.187	21.407	0.001
26	–0.182	0.047		–0.069	2.797	0.731
27	–0.275	0.046		2.322	9.577	0.088
28	–0.113	0.045		–1.308	9.947	0.077
29	0.132	0.045	R	–4.745	36.285	0.000
SF-36MH						
01	0.691	0.062	R	2.841	28.606	0.000
02	0.731	0.063		–0.685	17.026	0.004
03	–0.724	0.063		1.103	3.312	0.652
04	0.384	0.066		–0.820	4.656	0.459
05	0.171	0.063		2.639	8.438	0.134
FEW						
01	0.329	0.064	R	–1.753	11.635	0.040
02	0.314	0.060	R	0.796	13.237	0.021
03	–0.040	0.066		5.865	34.375	0.000
04	–0.675	0.060	R	1.891	9.949	0.077
05	0.053	0.062	R	–3.336	12.401	0.030
06	–0.322	0.058	R	–0.668	7.910	0.161
07	0.272	0.062	R	–1.606	6.917	0.227
GHQ						
01	–0.583	0.125		0.316	6.302	0.278
02	0.764	0.096		3.545	24.460	0.000
03	–0.653	0.098		–0.989	3.636	0.603
04	–0.283	0.113		–0.399	9.667	0.085
05	0.199	0.102		–0.269	11.932	0.036
06	0.246	0.099		–1.119	15.561	0.008
07	–0.800	0.098		0.117	1.902	0.863
08	–0.213	0.117		–1.065	10.649	0.059
09	0.412	0.089		–2.266	22.479	0.000
10	0.489	0.089		–1.899	14.694	0.012
11	0.555	0.084		–2.972	19.327	0.002
12	–0.146	0.118		–1.519	15.502	0.008

FEW, functional assessment of multiple sclerosis emotional well-being scale; GHQ, General Health Questionnaire; MSISpsych, Multiple Sclerosis Impact Scale psychological subscale; R, reversed thresholds; SF-36MH, Medical Outcomes Study 36-item Short Form Health Survey mental health dimension.

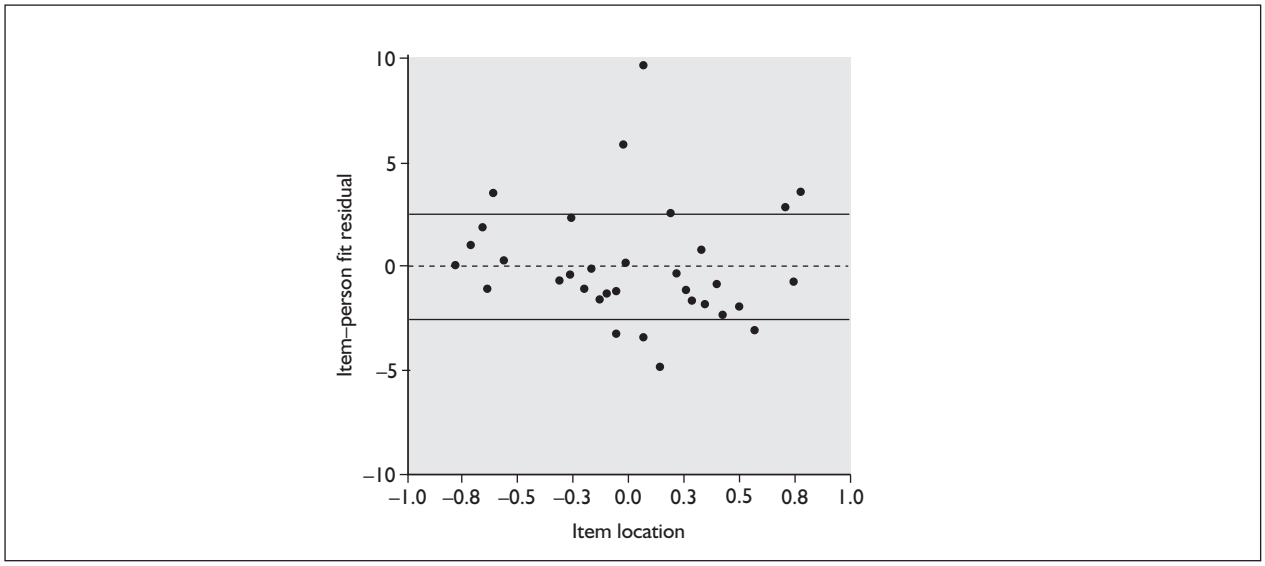


FIGURE 63 *Plot of fit residuals. Psychological impact 33 items.*

Figure 69 shows the distribution of the sample (person locations) relative to the distribution of the item locations. The sample is spread over a reasonably wide range (−4.5 to +4.5 logits). The PSI was high at 0.934, indicating that the 33-item 'scale' is a useful measure for discriminating among people in terms of their psychological functioning. That is, it can provide reliable measures of different levels of psychological functioning. However, most notable is the fact that the persons cover the items, but that the items have very limited coverage of the persons. It is, however, important to note that *Figure 68* shows only the single location for each item. This is the mean of the threshold estimates, which spread over a wider range.

Interpretation of stage 1 analyses

Most items work together to form a set that could be used to measure patients reliably on a psychological functioning continuum. Three items performed poorly, implying that their removal would improve the reliability and validity

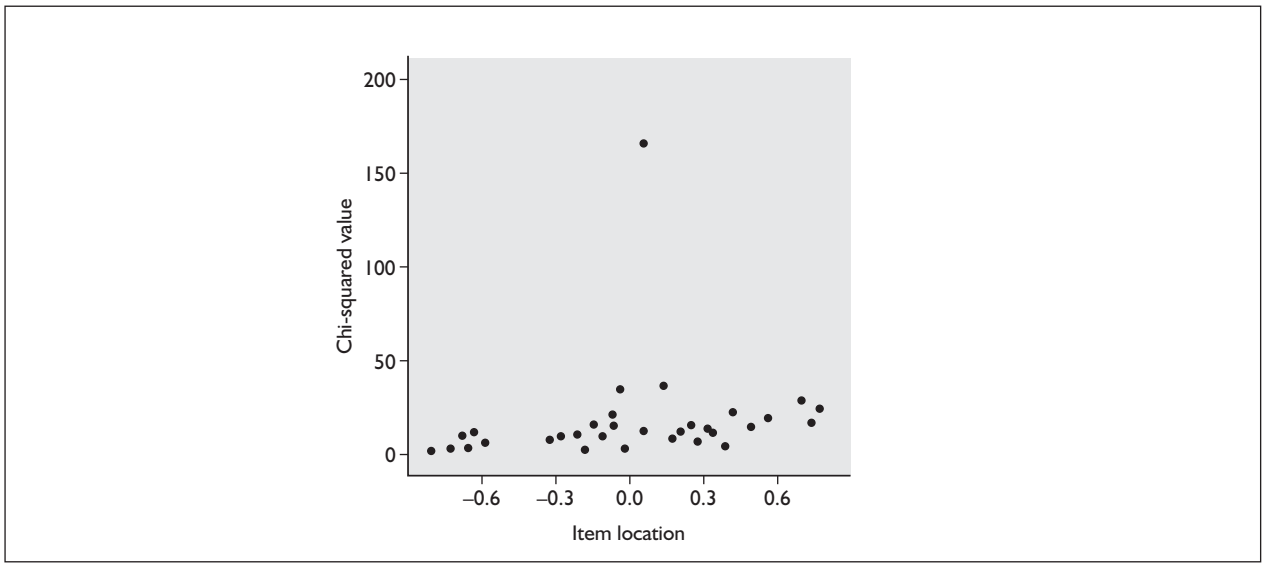


FIGURE 64 *Plot of chi-squared values. Psychological functioning 33 items.*

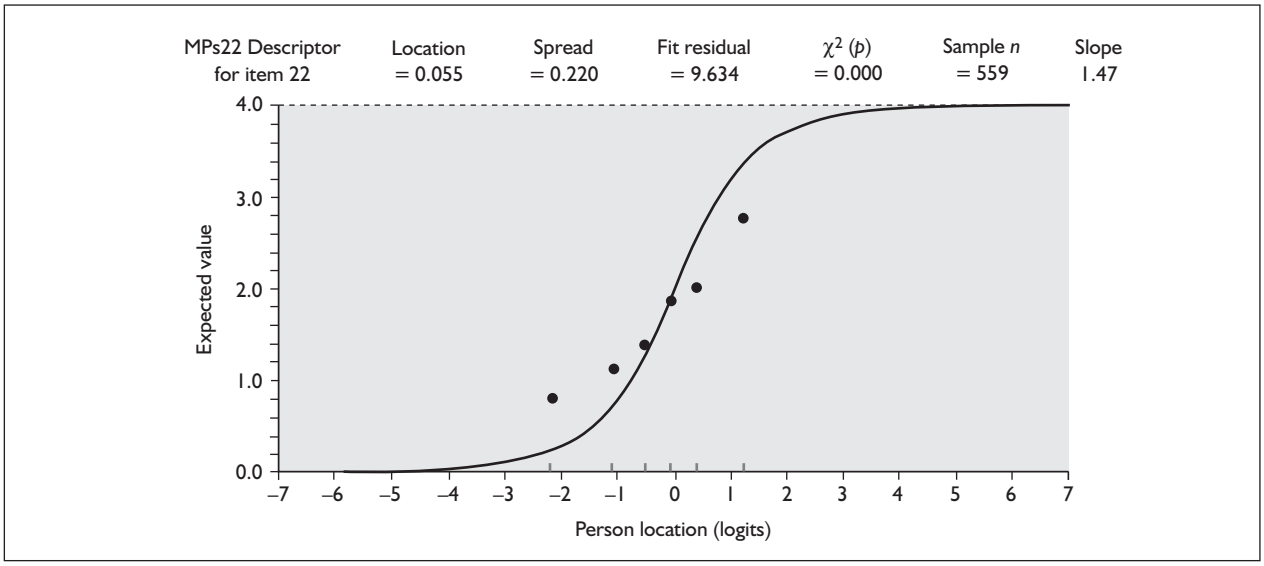


FIGURE 65 *Item characteristic curve item MSIS-22.*

of measurement. A number of items had lesser degrees of misfit, but all showed good coherence between observed and predicted responses, supporting the retention of these items.

Stage 2: modification and re-analysis

All misfitting items were examined to try and explain the reason for their misfit. The three notably statistically misfitting items were MSIS Q22 ('problems sleeping'), FEW Q03 ('I am able to enjoy life'), and MSIS Q29 ('feeling depressed').

From a clinical perspective, problems with sleeping (MSIS Q22) in people with MS can be caused by physical as well as psychological disturbance. Similarly, the ability to enjoy life (FEW Q03) is influenced by many factors. This may explain why these two items are poor discriminators relative to the frame of reference of the item group.

Item MSIS Q29 ('feeling depressed') has a different ICC pattern to the other items. The slope of the observed scores is steeper than the slope of the expected values. Thus, this item is overdiscriminating relative to the frame of reference of the other items. However, this item differs from the other two notable misfitters in

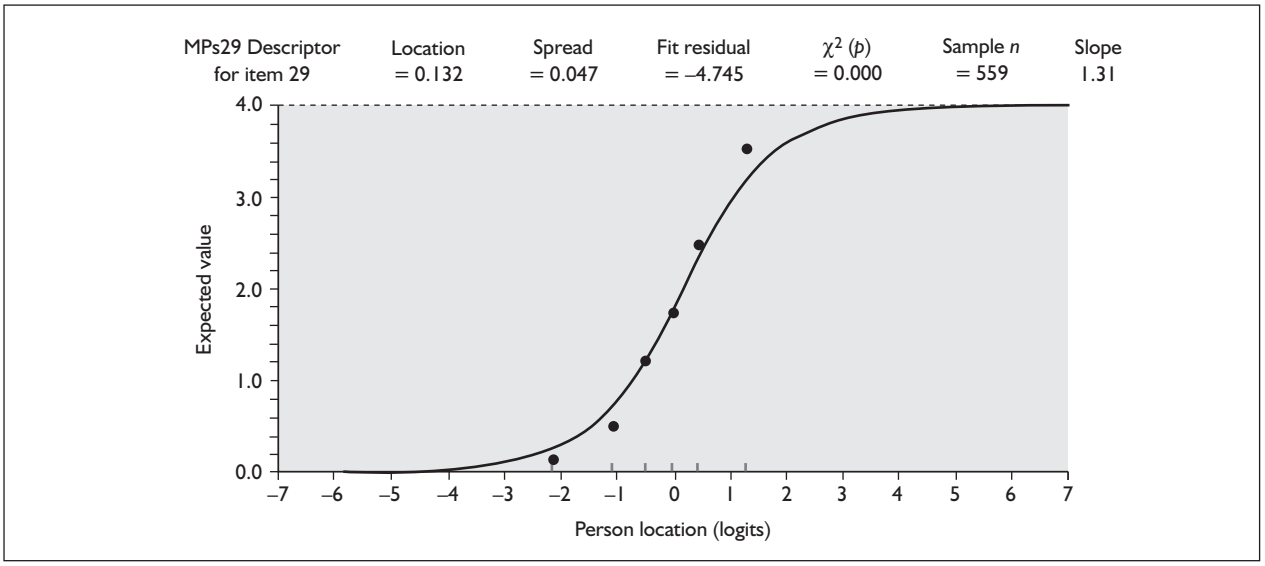


FIGURE 66 *Item characteristic curves item MSIS-29.*

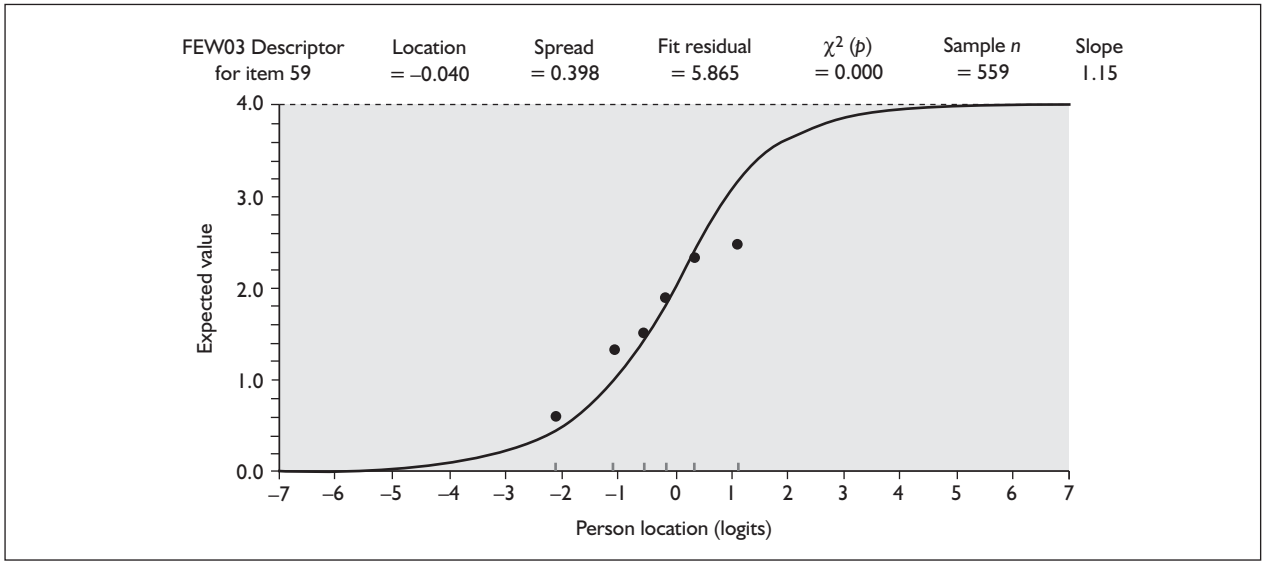


FIGURE 67 *Item characteristic curve for item FAMS-03.*

terms of the clinical relevance of the item with respect to the construct purportedly measured.

Thus, in the first instance we reanalysed the data with item MSIS Q22 removed (i.e. a total of 32 items). The fit statistics for item FEW Q03 (fit residual = 6.76; $\chi^2 = 41.0$) worsened and the fit statistics for item MSIS Q29 improved (fit residual = 4.57; $\chi^2 = 30.6$). Thus, we reanalysed the data with item FEW Q03 removed (i.e. a total of 31 items). The fit statistics for item MSIS Q29 continued to improve (fit residual = 4.39; $\chi^2 = 29.4$) and other items became relatively more misfitting. But no one item consistently misfit all criteria relative to the others. Examination of the content of these items indicated that the three most misfitting were: 'have you recently lost much sleep over worry?' (GHQ Q02), 'have you been bothered by mental fatigue?' (MSIS Q23), and 'have you been a nervous person?' (SF-36MH Q01). Of these three, the least specific with respect to a psychological functioning variable was item MSIS Q23. This was removed, and the data were reanalysed (i.e. a total of 30 items). Although a few items failed single criteria for misfit, no one item misfit relatively more than the others. Thus, it was considered sensible to stop removing items at this stage. The fit 30-item scale had a PSI of

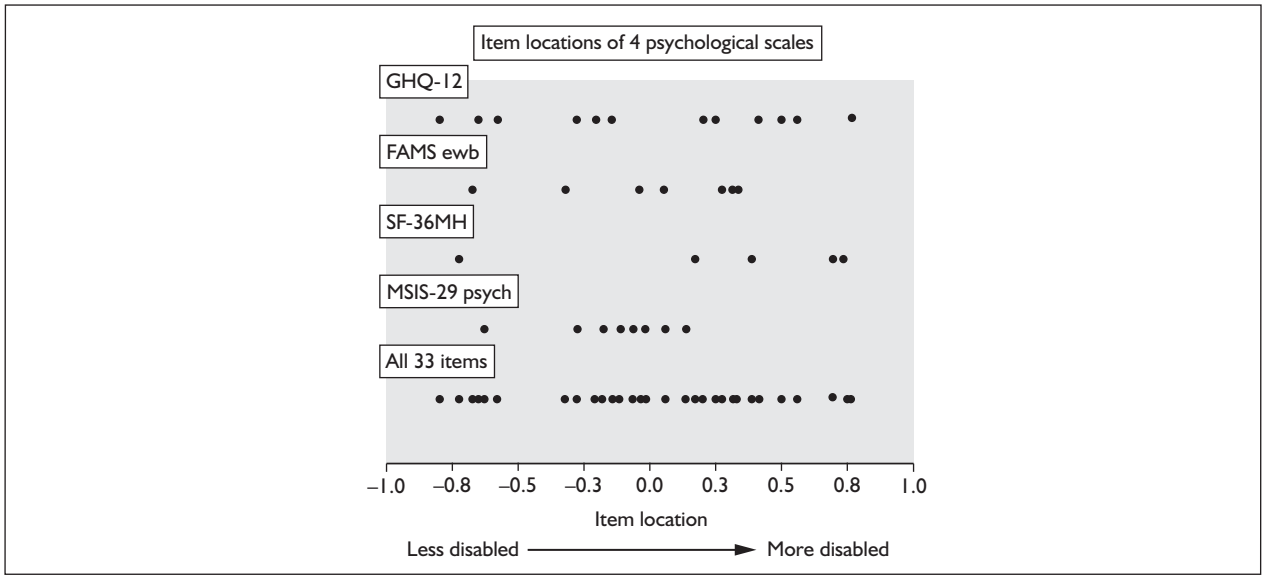


FIGURE 68 *Plot of item locations of the four psychological scales.*

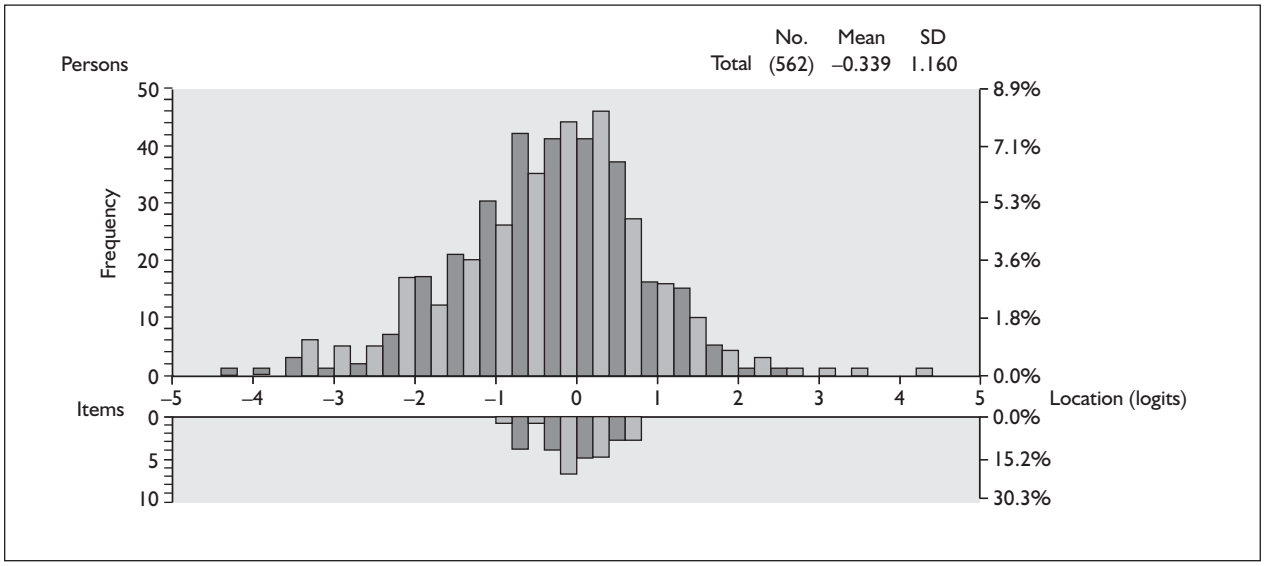


FIGURE 69 *Targeting of sample to 33 psychological items. Person–item location distribution (grouping set to interval length of 0.20, making 50 groups).*

0.936, indicating that the removal of three items (essentially 10% of the items) did not compromise the reliability of the scale; in fact, it improved it marginally.

Stage 3: equating

Table 41 is an equating table for the four psychological functioning scales. It shows the values, on a common psychological functioning interval-level metric, associated with any raw score on any of the four scales. As for the physical functioning variable, all raw scores range from 0 upwards and thus total scores for the MSIS-29 psychological scale and SF-36 mental health scale have been adjusted for this.

Table 41 shows a scale equating for the four scales without any items removed. To use this table, simply find the raw score for any scale in the first column, and then read across to the column of the scale of interest to find the value on an interval scale, shared with all the other scales. For example, a raw score of 10 on the FEW scale implies a measure of −0.32 logits. To find the corresponding raw score for the MSIS-29 psychological scale, find the nearest value to −0.32 in the second column (−0.29), and read off the associated raw score, which is 16. *Figure 70* is a graphical representation of the relationships between the four scales and the common variable. The *y*-axis is the raw score. The *x*-axis is the common psychological variable defined by the four scales.

Table 42 shows the same equating scales when three items were removed from the analysis. Note that the values in *Tables 41* and *42* differ. This is because the removal of three items changes the frame of reference, albeit slightly, and the mean item location which is always centred on 0.

Summary

General issues

The aim of this chapter was to demonstrate how Rasch analysis addresses the problem of determining whether different scales measure the same variable and the possibility of equating different scales on common metrics. We used data from the MSIS-29 validation study, have equated four physical functioning scales and four psychological scales, and have produced equating tables for clinical use.

The ability to equate scales is clinically important. There are now literally hundreds of health rating scales available. Many of these measure a small number of variables, particularly physical and psychological disability. Clinical trials often use different scales, for a variety of reasons, including the distribution of disability in the sample, the user-friendliness of the scale and investigator preference. The ability to compare results from these studies relies on the ability to determine the extent to which they measure the same variable and, where appropriate, equate results from the different instruments used.

Equating scales using traditional methods is difficult. It relies on very large samples of people

TABLE 41 *Equating psychological functioning scales for MS (all 33 items)*

	Values on a common metric			
Raw score	**MSISpsych**	**SF-36MH**	**FEW**	**GHQ**
0	–4.02	–3.91	–3.36	–6.52
1	–3.20	–2.93	–2.46	–5.58
2	–2.64	–2.24	–1.86	–4.86
3	–2.26	–1.77	–1.48	–4.31
4	–1.96	–1.40	–1.20	–3.82
5	–1.73	–1.10	–0.99	–3.36
6	–1.52	–0.84	–0.82	–2.91
7	–1.35	–0.61	–0.67	–2.47
8	–1.19	–0.40	–0.54	–2.04
9	–1.05	–0.20	–0.43	–1.65
10	–0.92	–0.01	–0.32	–1.29
11	–0.80	0.17	–0.22	–0.96
12	–0.69	0.35	–0.12	–0.67
13	–0.58	0.52	–0.03	–0.42
14	–0.48	0.68	0.06	–0.19
15	–0.38	0.84	0.15	0.02
16	–0.29	0.99	0.24	0.21
17	–0.19	1.13	0.33	0.39
18	–0.10	1.28	0.42	0.55
19	–0.01	1.43	0.52	0.71
20	0.09	1.59	0.62	0.86
21	0.18	1.78	0.73	1.00
22	0.27	2.01	0.86	1.15
23	0.37	2.32	0.99	1.29
24	0.47	2.83	1.15	1.43
25	0.58	3.65	1.36	1.57
26	0.69		1.64	1.72
27	0.81		2.08	1.87
28	0.94		2.79	2.03
29	1.08			2.20
30	1.25			2.38
31	1.43			2.59
32	1.65			2.83
33	1.93			3.11
34	2.29			3.48
35	2.83			4.02
36	3.63			4.82

FEW, functional assessment of multiple sclerosis emotional well-being scale; GHQ, General Health Questionnaire; MSISpsych, Multiple Sclerosis Impact Scale psychological subscale; SF-36MH, Medical Outcomes Study 36-item Short Form Health Survey mental health dimension.

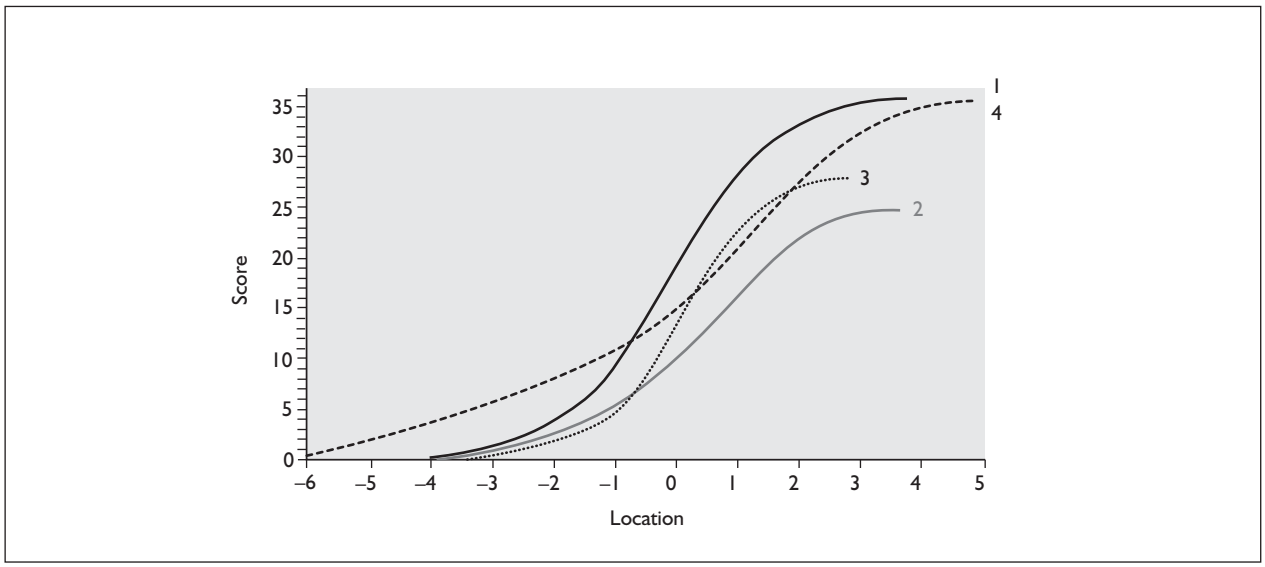


FIGURE 70 *Equating of the four psychological functioning scales. 1, MSIS-29 psychological; 2, SF-36MH; 3, FAMS emotional well-being (FEW); 4, GHQ*

to ensure generalisability, and every person needs to complete all the scales. To complicate matters, equating can be achieved only for the area of the continuum where measurement overlaps. Equating cannot be achieved at places where scales do not overlap, and results will be in raw score units, which are ordinal.

In contrast, as we have seen in this chapter, equating using Rasch analysis is relatively straightforward. Particularly large samples are not required, although they are preferable. Every person does not have to complete all the scales. As long as one scale, or in fact a few items, are common to everyone then different scales from different subsamples can be equated on the same common metric (provided, of course that there is evidence that they measure a common variable). Here, equating of different scales from the two samples was achieved by linking through the MSIS-29, which was completed by all subjects. Scales measuring different disability distributions can be equated, and equating is achieved on a common interval-level metric.

For each of physical and psychological functioning scales we have provided two equating tables. *Table 38* (physical) and *Table 41* (psychological) give the equating values for the scales without any modifications. These values can be used by clinicians who only have total scores available for these scales. *Table 39* (physical) and *Table 42* (psychological) give the equating values for the scales after a few of the most misfitting, less clinically relevant items have been removed. Removing these items provides more accurate equating, and these values can be used by clinicians who have item scores available for the scales. For the physical scales, simply omit items FAMS Q02 and Q05 and MSIS Q20 (where appropriate) and sum the remainder for the appropriate scale. For the psychological scales, simply omit MSIS-29 items Q22 and Q23 and FAMS emotional well-being item Q03 (FAMS scale item Q17), and sum the remaining items for the appropriate scale. With the use of these tables, previous studies using any of the different physical and psychological scales can be compared. In addition, future studies can choose the most appropriate scale for their study, in the knowledge that they can be referred to a metric shared by all four.

Different scales can be equated in two different ways. One option is to transform the raw score from one scale into the best-estimate equivalent raw score on the others. Alternatively, raw scores can be transformed into estimates (locations) on the shared metric. We recommend the latter approach, because the common metric is an interval-level measurement metric and, as such, a difference or change of one scale unit has the same meaning (implication) across the whole range of the scale. In contrast, raw scores are ordinal-level measurement metrics and the implication of a 1-point difference or change varies up to 13-fold across the physical scale and nine fold across the psychological scale.

A second reason to recommend using interval-level locations, rather than ordinal-level raw scores, is that the best-estimate raw score comparisons will

TABLE 42 *Equating psychological functioning scales for MS (30 items)*

	Values on a common metric			
Raw score	**MSISpsych**	**SF-36MH**	**FEW**	**GHQ**
0	–3.94	–4.13	–3.26	–6.66
1	–3.09	–3.12	–2.32	–5.73
2	–2.51	–2.40	–1.71	–5.01
3	–2.10	–1.90	–1.33	–4.46
4	–1.78	–1.51	–1.06	–3.97
5	–1.51	–1.19	–0.85	–3.50
6	–1.29	–0.91	–0.68	–3.04
7	–1.09	–0.67	–0.53	–2.59
8	–0.91	–0.44	–0.40	–2.15
9	–0.75	–0.23	–0.28	–1.74
10	–0.60	–0.03	–0.16	–1.37
11	–0.46	0.17	–0.05	–1.03
12	–0.33	0.36	0.06	–0.72
13	–0.20	0.54	0.16	–0.45
14	–0.07	0.71	0.27	–0.21
15	0.05	0.88	0.38	0.01
16	0.18	1.04	0.49	0.21
17	0.31	1.19	0.61	0.39
18	0.44	1.34	0.74	0.56
19	0.58	1.49	0.88	0.72
20	0.73	1.66	1.04	0.87
21	0.88	1.85	1.23	1.02
22	1.06	2.09	1.50	1.17
23	1.26	2.42	1.91	1.32
24	1.49	2.93	2.56	1.46
25	1.78	3.77		1.61
26	2.15			1.76
27	2.70			1.91
28	3.52			2.07
29				2.25
30				2.44
31				2.65
32				2.89
33				3.18
34				3.55
35				4.10
36				4.90

FEW, functional assessment of multiple sclerosis emotional well-being scale; GHQ, General Health Questionnaire; MSISpsych, Multiple Sclerosis Impact Scale psychological subscale; SF-36MH, Medical Outcomes Study 36-item Short Form Health Survey mental health dimension.

vary in how close they are. This is particularly noticeable at the extremes of the scale ranges because a 1-point change in raw score implies a larger change in intervalised measurement.

A third reason to recommend transforming raw scores into interval-level measurements is that every estimate on this metric has an associated standard error. A careful look at *Table 37* (physical) and *Table 40* (psychological) shows that these standard errors vary between scales, and across the range of each scale. This is because measurement precision, the confidence limits around measurements, are related to the number of items and item response options of a rating scale, and the location of the score in that scale's range. Logically, we have the least confidence (greatest expected error) about the precision of measurements associated with people who score at the extremes of the score range (floor or ceiling effect). Similarly, we have the most confidence about the measurements made on people in the centre of the scale range. Consequently, standard errors for all four scales are greatest at the extremes. These facts are reflected in the fact that the standard errors corresponding with the measures for each scale are greatest at the extremes and least in the centre. Similarly, greater precision is achieved with more items and more response options. In contrast, the standard error associated with raw scores, computed as $SD \times \sqrt{(1 - \text{reliability})}$, is wide and constant across the scale range.

The equating tables assume complete data. That is, raw scores can be equated according to the tables only when a person has answered all the items in a scale. What should investigators do when patients have missed one or more items of a scale? It is widely accepted practice, provided an individual has answered at least half of the items in a scale, to compute the mean score of the answered items and use this value as the estimated score for each of the missing items.[36] Summed scores are then generated in the usual way by adding up the actual and imputed item scores. Such an approach could be extrapolated so our equating tables can be used with missing data, and perhaps be justified on the basis of widespread practice. However, if this is done, the summed scores, their derived measures on the common metric, should be considered less reliable, to an unknown extent, than if data were complete. This is because the process of imputing for missing data makes assumptions about how an individual would respond to an item.

Another method of handling missing data is for individual investigators to use our Rasch-derived item locations (available from JH on request) in the analysis of their own data. This process, called anchoring, would be the best way of analysing their data. In addition, because our item location estimates have been used, any measurements of people will be made on the same metric as the one we have used.

To what extent are the results from this study generalisable in MS? Generalisability is always a concern with traditional psychometric analyses because their results are dependent on the distribution of the variable being measured (in this case, physical and psychological functioning) in the sample. Rasch analysis causes less concern because the results are independent of that distribution. This is a property of the mathematical model. However, it does not mean that the results are 'sample independent', as has sometimes been implied or expected.[72] The clinical importance of this mathematical fact is that Rasch analysis enables clinicians to examine, empirically, the stability (invariance) of results from different samples.

As we have discussed in previous chapters, a distinctive feature of Rasch analysis is the ability to examine how the item response options work. In this study, 12 of 44 items in the physical functioning pool of items, and 8 of 33 items in the psychological pool of items, have reversed threshold estimates, indicating that the item response functions for these items are not functioning as intended. This can occur for many reasons, including too many response options, ambiguous response options or multidimensionality within a set of items. When there are too many response options, responder uncertainty is created; this leads to response inconsistency and reduces item reliability and validity.

Nine of the 12 items with reversed thresholds had five response options. The CPCs for these items indicate that the middle category was not functioning as intended, and imply inconsistency distinguishing 'a little' from 'moderately', and 'moderately' from 'quite a bit'. To examine if these items worked better with four response categories, we collapsed adjacent categories, guided by the CPCs, and undertook a post hoc analysis. Although thresholds became ordered appropriately, this post hoc analysis makes assumptions about how people may respond to revised item response options. Consequently, the impact of revisions to response options in measurement terms should be examined prospectively. We have not reported the equating results post rescoring. This is because the scales

we have studied and equated are currently used in the forms that we have used them. It did not seem valuable to clinicians to complicate the process further at this stage.

This example of equating scales helps explain the concept of item banking.[152] An item bank is effectively a huge rating scale, a large collection of organised and catalogued items (often 100–200 or more) proven to be measuring the same underlying construct. It is simply an extrapolation of the equating process that we have demonstrated here. An item bank does not necessarily have to be created by equating existing scales: any items from any sources (e.g. existing items from a range of scales, modified versions of existing items or newly written items) can be used. The aim is to have good coverage across the range of the variable of interest. The idea of an item bank is that investigators can select any items from the bank, as dictated by the measurement problem, to make up a scale.[68] Similarly, from this study we could administer any selection of items from the physical and psychological pools (there are really item 'banklets') and use them as a scale.

The reason that it is possible to select any combination of items (within reason) from an item pool to form a 'scale' is that each item has a defined location relative to the others. Thus, we use the locations of the items we select to form our scale in the analysis of the response data to achieve measurement on the common variable defined by the bank. Clearly, the fact that the Rasch model generates estimates of item locations that are independent of the sample distribution facilitates greatly the process of item banking.[153] Some argue that this mathematical property underpins equating and is essential for it. The essential outcome of item banking is that it is possible to calculate, for any set of items measuring a common trait and drawn from an item bank, and for any set of responses to these items, a scaled score that is interpretable with respect to the entire bank and not just those questions included in the test. Since all other sets of items measuring the same trait will lead back to the same scale, however short or long, hard or easy the particular scale, we may think of this scale as being a common standardised scale for that trait.[68]

The main purpose of item banking is to achieve targeted and precise individual person measurement. It is clear that the best items for measuring any one individual accurately are those that are located at a very similar same place on the continuum. Items that are far away from a person's location, in either direction (too 'easy' or too 'difficult'), are quite unhelpful as they simply tell us that a person is above or below that location. The challenge then is to give the right items to a person we are trying to measure (locate on the continuum) but whose measurement we have yet to determine.

This can be achieved in a number of ways. One method is by computer.[154,155] Here, a computer algorithm uses the response to the first item (usually a broad item) to determine the next item (computer-adaptive measurement) presented to the respondent, and the response to the second item determines the response to the third. This process continues until an accurate measurement is achieved. It works because, once the items of an item bank have been calibrated (i.e. their relative locations determined), every response to every item has a specific meaning in terms of the most likely location of any person who gives that response. Thus, from the response to the first item the computer has, immediately, a best-guess estimate for that person so it can choose an item located near that estimated location. As more items are presented the estimate of the person's location becomes refined (unless the responses are poorly misfitting) and the error around the estimate narrows (because the standard error is related to the number of items answered). The investigator predetermines the standard error required, and when that is reached the job is finished, a person's location has been determined, and no further items are administered. Surprisingly few items are needed in computer-adaptive measurement, often as few as five, to achieve very precise measurement.

A second method of targeting the items of an item bank to an individual person is a simplified version of computer-adaptive testing. Instead of administering a single item in response to the first items, the computer delivers a set of items covering an appropriate area of the continuum. This approach can also be used with postal questionnaires (traditionally called the paper and pencil format). For example, the first question for a mobility questionnaire might be 'Do you walk unaided, with an aid, use a wheelchair?' Each response to this question will direct the person to certain other questions in that questionnaire. For example: unaided = answer items 1–20; with aid = answer items 10–30; wheelchair = answer items 20–40. All in all, adaptive measurement offers the potential to achieve rapid, efficient, user-friendly, and precise individual person measurement with substantial relevance to individual patient care.

Equating of physical functioning scales

One aspect of this study was to equate the scores of four patient-report physical functioning rating scales. Before this could be done we had to test the hypothesis that the four scales measure a common variable. From a clinical perspective, while all four scales measure physical functioning, and thus it seems clinically reasonable that they be equatable, they are somewhat different. A closer look at the 47 items shows three main categories of items within each scale: symptoms that impact on physical functioning; the ability to undertake specific physical tasks; and physical functioning in a wider social perspective. This is reasonable because physical functioning is a broad, loosely defined variable. However, this helps to explain in part why some of the items were misfitting. Nevertheless, statistically, most items satisfied all tests of fit, some items had small degrees of misfit on only some tests, and only three items failed all tests of fit. These findings indicated that 44 of the 47 items worked together to define a conformable set that can be used to measure physical functioning in MS, and supported, empirically, the equating of the four scales.

Figure 60 shows how the four different physical functioning scales relate to each other in terms of their coverage of the variable. This figure also shows the coverage provided by the 47-item banklet that arises from pooling the items of the scales. *Figure 60* identifies the obvious 'gaps' in the continuum and serves to act as the evidence base for developing the banklet further.

It is important to note that *Figure 60* and the corresponding figure for psychological scale (*Figure 67*) are simplifications, as one black dot represents one item location. This would be the case if the items were dichotomous. However, when items have more than two response options, i.e. are polytomous items, each item has multiple thresholds (number of response categories – 1), and the resulting item location is that of the thresholds. Nevertheless, *Figures 60* and *67* are helpful in illustrating some of the ideas behind further development of item banks.

Equating of psychological functioning scales

The combined analysis of four psychological scales supported their measurement of a common underlying variable. As with the physical scales there were some misfitting items. This should provoke a critical examination of the items, and their response categories, in relation to the underlying variable of interest.

When considered critically, there were clear reasons why these items might not 'work with' the others. Such concerns may not necessarily be apparent in advance of the analysis. For example, sleep problems are a common symptom of psychological disturbance; so, it is not surprising that they appear in psychological disability scales. In MS, sleep disturbance may be associated with physical problems such as nocturnal spasms and sphincter disturbance. It is easy to see why such an item appears to misfit with the construct measured by the others. This non-specificity of the MSIS-29 sleep item is indicated further by the finding that a sleep item worded to be more specific to psychological problems – 'Have you recently lost much sleep over worry?' (GHQ-12 Q02) – has much better fit statistics and adherence to the ICC.

This sleep item example illustrates one of the ways in which Rasch analysis helps the investigator to understand, clarify and specify the construct being measured. It would seem logical that such evaluations of items are less critical, and less developed, when the aim of the item analysis is to find the mathematical model that best fits the data.

Finally, but importantly…

This chapter has examined the equating of existing scales to facilitate meta-analyses and to illustrate the concept of item banking. In the course of creating item banks de novo, we would advocate very strongly a somewhat different approach: first, define, explicitly, the variable(s) for measurement; second, select and/or develop items that articulate this explicit definition; third, test using Rasch analysis the extent to which the pool of items so generated functions as a measurement instrument.

Chapter 8

Rating scale responsiveness

Traditional psychometric methods versus Rasch measurement

Overview

It is essential that rating scales are able to detect statistically and clinically important change, when it occurs, if they are to be useful in clinical trials. As there are no guidelines as to what constitutes adequate responsiveness, it is important to evaluate the relative responsiveness of competing measures. This chapter uses traditional psychometric methods and Rasch analysis to compare the relative responsiveness of widely used neurological clinician-rated scales in a rehabilitation setting, in order to examine the extent to which the newer psychometric technique is able to provide information that may add to our understanding of clinical change. We do this by presenting three studies that compare two closely related scales. These scales are the Barthel Index (BI) and the Functional Independence Measure (FIM).

In study 1, we examine item and total score distributions on admission and discharge, and item and total score effect sizes (mean change/SD time 1 score) for both measures in a sample of patients undergoing neurorehabilitation.

In study 2, we explore the data further using subsample analyses and alternative group-level indices of responsiveness, including examining the number of patients whose scores changed; the impact of floor and ceiling effects; the number of people who scored towards the BI ceiling on admission; the eight items common to both scales; and alternative responsiveness statistics (i.e. standardised response mean and analysis of variance).

In study 3, we analyse the same data set using Rasch analysis to compare the relative responsiveness of the two measures at the group and individual levels, including examinations of effect sizes; standardised response mean; relative efficiency (pair-wise squared *t*-values from paired *t*-tests); relative precision (ratio of pair-wise *F*-values from the ANOVA); paired-sample *t*-tests; ANOVA; and individual significance of change (SigChange). In each instance, responsiveness was measured against the expected clinical improvement brought about by neurorehabilitation.

Background

The ability of rating scales to detect clinical change is known as responsiveness.[156] For rating scales with multiple items, responsiveness is usually determined by computing an effect-size statistic (standardised change score) from pre- and post-treatment total scores. Adding up the item scores generates a total score. Therefore, the extent to which total score changes accurately reflect clinical change is determined by the extent to which item scores *can*, and *do*, change. Consequently, analysing total scores alone could be misleading if there are problems at the item level, for example notable ceiling or floor effects when a substantial proportion of the sample endorses the maximum (ceiling) or minimum (floor) item scores. People at the ceiling *cannot* change their item score regardless of clinical improvement. People at the floor *may not* change their item score despite clinical improvement. Despite these facts, item-level responsiveness is rarely examined.

In a previous study,[157] we determined whether total score changes accurately reflect item score changes, using the BI in a group of patients undergoing neurological rehabilitation. We found that the responsiveness of the BI for the whole sample was moderate to large when computed as an effect size (0.77) from pre- and post-rehabilitation total scores. In addition, the distribution of admission BI total scores demonstrated minimal floor and ceiling effects, indicating that the *potential* for the BI to detect change associated with rehabilitation in this sample appears to be good. However, item-level analyses revealed that effect sizes varied widely, indicating that some items detected more change than others. Nine of the 10 BI items had notable admission ceiling effects (> 22%), with

five items having particularly large effects (46.5–80.6%). These patients could not improve their item score irrespective of any clinically significant improvement in that task. These item-level ceiling effects increased on discharge, indicating that an additional proportion of patients (up to 51.7% for some items) might have improved more than these items have detected. In addition, seven items had notable floor effects (20.4–78.0%). These represented patients who might have undergone clinically important improvements but did not record changes in their item score, or whose extent of change was undermeasured by the items.

The study findings indicated that, in contrast to the total score analyses, item-level analyses raised important questions about the suitability of the BI as a rating scale for measuring the impact of neurological rehabilitation. We concluded that total score analyses can be a limited indicator of a rating scale's potential to detect clinical change, and therefore it is vital to examine item-level responsiveness to get a clearer picture. The three studies described in this chapter build on this conclusion.

Setting

The three studies described in this chapter were based on a cohort of patients with neurological disability, who were admitted to a single neurorehabilitation unit (National Hospital of Neurology and Neurosurgery, London, UK) between May 1993 and March 2003. The neurorehabilitation unit specialises in intensive, individually tailored, goal-orientated rehabilitation. Patients received input from at least two disciplines other than medical and nursing staff , which included physiotherapy, occupational therapy, speech and language therapy, social work and neuropsychology.

Procedure and sample

Information on patients was prospectively included on a database within the unit. Of the patients admitted (N=1495), only those with complete admission and discharge data with a length of stay exceeding 10 days were included in the studies. As part of a larger battery of measures, the BI and FIM were scored within 3 days of admission and 2 days of discharge. Complete BI and FIM data were available for 1396 people (93% of the total sample). The mean age and length of rehabilitation were 48 years (SD 15) and 34 days (SD 24). The main diagnostic groups were MS (42%), stroke (20%) and cord syndromes (17%) (*Table 43*).

Measures

Barthel Index (BI)

The BI (see Appendix 3) is a clinician-scored 10-item measure of personal activities of daily living (pADLs). It includes items with a choice of two response categories (two items), three response categories (six items) or four response categories (two items). BI total scores (which range from 0 to 20) are generated by summing scores for the 10 items. Higher values indicate better functioning.[158]

Functional Independence Measure (FIM)

The FIM (see Appendix 4) also measures pADLs, and was developed because of dissatisfaction with existing scales, in particular the BI, which was considered crude, insensitive to change and unable to account for cognitive impairment. The FIM comprises 18 items grouped into two domains – the motor scale (FIMm; 13 items) and the cognitive scale (five items). Two BI items ('dressing' and 'transferring') are represented in the FIMm by a total of five items ('dressing upper body', 'dressing lower body', 'bed transfer', 'toilet transfer' and 'bath transfer').[159] However, the two scales share eight items that are effectively identical apart from the number of response options.

Each item on the FIM is scored on a seven-point scale, and each domain gives a subtotal. The FIMm, the subscale we are interested in because of its similarity to the BI, gives total scores ranging from 13 to 91 that are generated by summing scores for the 13 items. Higher values indicate better functioning. Not surprisingly, the FIMm and the BI are highly correlated (r = 0.95).

The FIMm and BI thus offer excellent potential for comparing rating scales. The increased number of item response options in the FIM should improve its potential (i.e. smaller floor/ceiling effects) and therefore also affect its ability (i.e. increased effect sizes) to measure change when compared with the BI. Interestingly, we and others have shown that the BI and FIMm have similar responsiveness in small samples.[160–164]

TABLE 43 *Sample characteristics*

	Total group	MS	Stroke	Spinal cord injury
Total number of patients	1495	622 (42%)	291 (20%)	250 (17%)
Patients' data available	1418	596 (42%)	282 (20%)	237 (17%)
Age [mean (SD); range]	48 (15); 16–88	44 (12); 16–75	53 (15); 16–87	52 (16); 16–85
Gender (percentage male)	46%	33%	60%	57%
Length of stay (days) [mean (SD); range]	34 (24); 10–184	23 (11); 10–102	51 (30); 10–149	43 (27); 10–184

MS, multiple sclerosis; SD, standard deviation.

TABLE 44 *Sample characteristics and comparison of FIMm and BI total scores on admission and discharge and effect sizes*

	Total group	MS	Stroke	Spinal cord injury
FIM motor scale score				
Admission [mean (SD); range]	58.2 (19.5); 13–91	59.7 (19.4); 13–90	57.6 (18.4); 13–91	56.6 (19.6); 13–88
Admission floor/ceiling	0.8/0.3	0.5/0.0	0.4/1.4	1.7/0.0
Discharge [mean (SD); range]	72.7 (17.5); 13–91	68.3 (19.0); 13–91	77.2 (14.7); 13–91	74.4 (15.3); 21–91
Discharge floor/ceiling	0.2/1.7	0.4/0.2	0.4/3.2	0.0/0.4
Effect size	0.74	0.44	1.04	0.90
Number with score changed (%)	1333 (96%)	572 (96%)	271 (96%)	230 (97%)
Improved	1267 (91%)	524 (88%)	265 (94%)	220 (93%)
Same	61 (4%)	24 (4%)	11 (4%)	7 (3%)
Worsened	68 (5%)	48 (8%)	6 (2%)	10 (4%)
Barthel Index score				
Admission [mean (SD); range]	11.8 (5.3); 0–20	12.2 (5.4); 0–20	11.7 (5.0); 0–20	11.2 (5.3); 0–20
Admission floor/ceiling	1.1/5.3	1.0/5.7	0.7/5.3	2.5/5.5
Discharge [mean (SD); range]	15.9 (4.8); 0–20	14.8 (5.4); 0–20	17.2 (4.0); 2–20	16.3 (4.2); 3–20
Discharge floor/ceiling	0.1/27.9	0.2/19.3	0.0/40.1	0.0/24.1
Effect size	0.77	0.47	1.09	0.98
Number with score changed (%)	1233 (88%)	489 (82%)	259 (92%)	211 (89%)
Improved	1176 (84%)	453 (76%)	257 (91%)	204 (86%)
Same	192 (14%)	107 (18%)	22 (8%)	26 (11%)
Worsened	28 (2%)	36 (6%)	3 (1%)	7 (3%)

BI, Barthel Index; FIM, Functional Independence Measure; MS, multiple sclerosis; SD, standard deviation.

Study 1: comparison of BI and FIMm using traditional psychometric methods

Hypothesis

Study 1 builds upon the study described above, in which the BI exhibited high item floor and ceiling effects as eight of its 10 items have a limited number of response alternatives [i.e. either 'dependent/independent' (two items), or 'fully dependent/partly dependent/independent' (six items)].[157] We hypothesised that increasing the number of item response options should improve the potential [i.e. smaller item floor (per cent of sample scoring the minimum possible value)/ ceiling (per cent of sample scoring the maximum value) effects] and therefore improve ability (i.e. increased scale effect sizes) of the BI to detect change.

Analysis

For both scales, we examined the item and total score distributions (mean, SD and per cent floor and ceiling effects) on admission and discharge and the item and total score effect sizes (mean change/ SD admission[165]) and compared the findings.

Results

FIMm and BI total scores

At admission and discharge, total score ceiling effects were lower for the FIMm than the BI (0.4/1.7 and 5.4/27.8), implying that the FIMm had better potential to detect change (*Table 44*). However, total score effect sizes were nearly identical (FIM = 0.74; BI = 0.77), implying that the FIMm was no better at detecting change in this sample.

FIMm and BI item-level scores

Floor and ceiling effects for all FIMm items were less than for the comparable BI items, implying greater potential to detect change (*Tables 45* and *46*). However, this potential was not reflected in item effect sizes. These were better for two BI items ('feeding' and 'bathing') and two FIMm items ('bowels' and 'walk/wheelchair use'), and were equivalent for four items ('grooming', 'toileting', 'bladder' and 'stairs'). Thus, at the item level, the better potential of the FIMm to detect change was not reflected in the effect sizes. These findings were consistent across the major disease groups (*Table 46*).

TABLE 45 *FIMm and BI effect sizes[a] and item response option frequency distributions at floor/ceiling (all patients)*

	FIMm			BI		
	Effect size	Floor/ceiling		Effect size	Floor/ceiling	
		Admission	Discharge		Admission	Discharge
Feeding	0.42	6.4/36.1	2.2/54.5	0.55	8.9/48.5	4.3/78.9
Grooming	0.43	4.7/41.1	2.4/64.1	0.44	31.1/68.9	10.9/89.1
Bathing	0.60	9.1/14.7	2.8/38.1	0.80	78.1/21.9	45.3/54.6
Dressing upper body	0.49	6.7/30.4	3.0/53.4	–	–	–
Dressing lower body	0.67	22.2/8.2	11.1/28.8	–	–	–
Toileting	0.52	17.5/21.6	9.1/46.6	0.51	22.3/46.4	10.6/74.6
Bladder	0.31	13.0/35.7	6.3/45.4	0.33	20.7/60.5	10.4/76.9
Bowels	0.24	6.8/37.3	4.3/51.2	0.20	9.3/80.9	5.0/88.9
Bed transfer	0.67	12.3/13.8	4.3/44.7	0.59	8.5/39.3	3.4/72.6
Toilet transfer	0.63	12.8/9.6	5.8/33.2	–	–	–
Bath transfer	0.72	22.2/2.6	8.3/10.3	–	–	–
Walk/wheelchair use	0.82	30.7/3.9	5.3/16.8	0.68	22.1/27.8	4.2/59.5
Stairs	0.76	60.4/1.1	31.5/5.5	0.78	63.2/14.6	32.6/41.4

BI, Barthel Index; FIMm, Functional Independence Measure motor scale.
a Computed as mean change/SD admission.

DOI: 10.3310/hta13120

TABLE 46 *FIMm item floor/ceiling effects and effect sizes for MS, stroke and SCI disease groups*

	MS			Stroke			SCI		
	Effect size	Floor/ceiling		Effect size	Floor/ceiling		Effect size	Floor/ceiling	
		Admission	Discharge		Admission	Discharge		Admission	Discharge
Feeding	0.32	6.3/39.7	2.8/53.8	0.72	5.7/13.4	1.4/36.7	0.34	4.2/61.0	0.0/73.7
Grooming	0.25	4.0/46.4	3.5/62.2	0.76	3.2/21.9	0.7/58.7	0.41	5.5/58.1	1.7/74.2
Bathing	0.31	8.4/17.0	6.0/31.2	0.94	7.8/9.5	1.4/41.3	0.77	10.2/13.1	2.1/43.2
Dressing upper body	0.26	6.2/33.2	4.7/48.3	0.89	5.3/15.9	0.7/47.3	0.49	11.0/42.4	2.5/64.4
Dressing lower body	0.34	22.3/8.4	15.9/19.8	1.05	16.6/8.5	4.6/35.5	0.92	23.3/5.1	8.1/31.8
Toileting	0.24	15.8/21.6	13.0/35.0	0.80	9.9/20.8	3.9/61.8	0.67	22.5/20.3	6.8/46.2
Bladder	0.27	15.1/19.9	9.6/23.3	0.30	8.1/58.0	2.8/73.9	0.46	16.5/22.0	4.2/34.7
Bowels	0.14	7.7/34.1	6.3/41.1	0.26	2.8/51.9	1.4/72.8	0.44	13.6/17.4	5.5/37.3
Bed transfer	0.42	15.1/11.1	9.3/29.6	0.93	7.1/18.0	1.8/62.5	0.80	13.1/14.0	1.3/51.3
Toilet transfer	0.38	15.8/6.3	10.0/19.6	0.88	7.1/15.9	2.1/52.3	0.75	13.6/7.6	3.0/34.3
Bath transfer	0.45	22.3/1.4	12.3/5.1	0.92	19.4/4.9	4.2/14.1	0.86	24.2/1.7	6.4/8.9
Walk/wheelchair use	0.58	22.3/1.8	5.6/6.1	1.03	37.1/8.1	3.9/30.7	0.94	40.3/3.0	5.9/13.6
Stairs	0.40	59.4/0.2	44.4/1.9	1.05	53.4/3.9	13.4/11.7	1.11	71.6/0.8	29.2/3.8

MS, multiple sclerosis; SCI, spinal cord injury.

Conclusions

The FIMm had smaller floor and ceiling effects than the BI, at both the item score and the total score levels. This supported our hypothesis, and is intuitively sound as the FIMm has more item response options and items than the BI. These findings imply that the FIMm has the greater potential to detect change. However, the finding that the FIMm total score had a similar effect size to that of the BI implies that the ability of the FIMm to detect clinical change was no better than that of the BI in our sample, thus refuting our hypothesis. Therefore, study 1 demonstrates that the FIMm had a greater potential to detect change, and yet appears to be no better at detecting change than the BI in this sample.

Study 2: can we reconcile the potential-ability discrepancy using traditional psychometric methods?

Hypothesis

Study 1 uncovered a *potential-ability* discrepancy in the relative responsiveness of the BI and FIMm. Although the FIMm had the greater potential to detect change, it had an almost identical effect size to the BI, which suggests that its ability to detect clinical change was no better than that of the BI. This seems counter-intuitive clinically.

Analysis

In order to try to address this counter-intuitive finding, we explored our data in five different ways using traditional group-level statistics.

Examination of people whose scores changed between admission and discharge

We focused on the people whose BI score did not change between admission and discharge, but whose BI scores did change. We hypothesised that if the FIMm and the BI were truly equally responsive then two things may be found. Either (1) the total number of people whose FIMm total score changed but whose BI score did not was too small to impact on the findings; or (2) the number of people whose FIMm score improved was so similar to the number of people whose FIMm score worsened that they cancelled each other out.

Examination of the impact of floor and ceiling effects

Our second hypothesis was that the ceiling effect of the BI total score at discharge may give a false impression of responsiveness. This would occur if the variance of BI total scores at admission was small relative to their change scores. The effect of this would be to artificially elevate the effect size (computed as mean change/SD T1 score). In order to explore this idea, we compared the effect sizes of the BI and FIMm in the subsample of people whose admission *and* discharge scores were within the floor and ceiling of both scales ($n = 976$), i.e. those people whose scores, on both admission and discharge, and therefore also their score changes, were not limited by the finite range of the scale.

Examination of people who scored towards the BI ceiling on admission (> 15)

We hypothesised that the subgroup of patients for whom the FIMm should show greater responsiveness than the BI are those towards the BI ceiling on admission. Thus, we examined the effect sizes in the subsample of people with a BI admission score of > 15 and compared these with effect sizes in the total sample.

Examination of the eight items common to both scales

We wondered if the non-shared items were having a deleterious effect on the effect size of the FIMm. Thus, we re-examined total score effect size statistics by focusing on the eight items common to both scales. Essentially, we generated an eight-item FIMm and an eight-item BI. This had the effect of making the two scales even more equivalent.

Examination of other responsiveness statistics

Another possible reason why the FIMm and BI have similar effect sizes, which is counter-intuitive, is that this may be a quirk of the specific effect size statistic we used. The term 'effect size' refers to a family of calculations that standardise change scores.[156] Thus, we examined whether our findings were consistent across different calculations. First, we computed the standardised response mean (mean change/SD of change score).[143] This differs from the Kazis effect size statistic[165] by using the variance in change scores as the denominator, as opposed to the variance in admission (T1) scores. Second, we computed an ANOVA on the admission and discharge scores and used the resulting *F*-statistic (ratio of between-group to within-group variance) as the indicator of responsiveness. This differs from the effect size statistic by instead being an indicator of the extent to which the group means differ, taking into account between-group to within-group variance.

Results

Examination of actual number of patients whose scores changed

A total of 192 people did not have a change in their BI score (14% of total sample). Of these people, 79% ($n = 151$; 10% of total sample) did have a change in their FIM score (120 improved, 31 worsened). Thus, the FIMm detected change in most people who would be considered stable by the BI, and the ratio of those improved to those worsened was not equal.

Examination of impact of floor and ceiling effects

A total of 976 people had admission and discharge scores within the floor and ceiling of the FIMm and BI, and thus their scores and score changes were not limited by the finite range of the scale. There were no notable differences in effect sizes of the FIMm and BI.

Examination of people who scored towards the BI ceiling on admission (> 15)

A total of 538 people scored > 15 on the BI at admission. Almost half of these people (48%) had a discharge BI score at the ceiling (score = 20). In this group of people the values of the effect sizes were elevated in general, but there was no difference in the relative effect sizes of the two scales (FIMm = 0.95; BI = 0.99). We did not witness the expected finding of a higher effect size in the FIMm.

Conversely, when we examined the subsample who scored towards the FIMm ceiling on admission (FIMm > 85; $n = 67$), we found that the effect sizes were smaller and favoured the FIMm (FIMm = 0.62; BI = 0.47). In this subsample of people, the ceiling effect for the BI on admission was 60% of these people ($n = 40$); thus, these findings were to be expected.

Examination of the eight items common to both scales

The eight-item versions of the FIMm and BI had near-equivalent effect sizes (FIMm = 0.71; BI = 0.77).

Examination of other responsiveness statistics

Table 47 shows that the use of alternative responsiveness statistics supported the findings from the effect sizes. Again, the BI appeared marginally more responsive than the FIMm.

Conclusions

In order to address the potential-ability discrepancy in the relative responsiveness of the BI and FIMm revealed in study 1, we explored the data set traditional group-level statistics in five ways: examining the number of patients whose scores changed; accounting for the impact of floor and ceiling effects; focusing on the subsample who scored towards the BI ceiling on admission (> 15); focusing on the eight items shared by both the BI and FIMm; and calculating standardised response means and analysis of variance. Each of these analyses supported the findings from study 1 (i.e. that the BI is marginally more responsive than the FIMm in our data set), and, importantly, failed to identify the cause of the potential-ability discrepancy uncovered in study 1.

Studies 1 and 2 demonstrate that the FIMm had a greater potential to detect change and identified more individuals who actually changed, and yet appears no better at detecting change than the BI. Although we used a wide range of traditional group-level total score statistics in a variety of subsamples, we were unable to uncover the reason for this counter-intuitive finding. We therefore conclude that rather than being an artefact of

TABLE 47 *Examination of standardised response means and analysis of variance for the BI and FIMm (n = 1396)*

	BI	FIMm
SRM	1.07	1.05
F-value (*p*)	461.8 (0.000)	416.6 (0.000)
RMP	100%	90%

BI, Barthel Index; FIMm, Functional Independence Measure motor scale.
SRM, standardised response mean = mean change/SD change score.
F-value from ANOVA admission and discharge mean scores.
RMP, relative measurement precision of BI compared with FIMm = (*F*-value BI)/(*F*-value FIMm).

our data set, our findings may raise concerns about traditional group-level indicators of scale responsiveness.[166] We tested this in study 3.

Study 3: can the potential-ability discrepancy be reconciled using Rasch measurement?

Hypothesis

In light of the findings from studies 1 and 2, and in particular our concerns about the ability of traditional group-level statistics, in study 3 we examined the relative responsiveness of the BI and FIMm using Rasch analysis. The potential advantages of Rasch analysis over traditional analytic methods have been introduced in other chapters. We hypothesised that Rasch analysis may provide some solutions for the potential-ability discrepancy over and above traditional approaches for three reasons.

First, the transformation of ordinal-level BI and FIMm total scores into interval-level measurements (linearisation) helps to account for the fact that fixed changes in ordinal scores (e.g. 10 points) imply variable changes in interval-level measurements.[167] As such, analysing ordinal scores may hide true differences between scales. Traditional psychometric methods use total scores to locate people on an ordinal scale, while Rasch analysis generates locations of people on an interval scale. Total scores, which increase in successive integer counts, are non-linear because they have a non-linear relationship to the underlying trait they seek to measure (in the case of the BI and FIMm, personal activities of daily living). In contrast, Rasch-derived person locations (and item locations) are linear measures because they have a linear relationship to the underlying trait they seek to measure. The implication of this is that a change in measurement implied by a 1-point change in BI or FIMm total score may vary across the range of the scale (see Chapters 4 and 5).

Second, Rasch analysis allows for legitimate examination of change at the individual person level, in addition to group comparison. The availability of individualised standard errors makes this a legitimate process. This means that BI and FIM data can be compared for each individual person as well as for the group. This enables a more sophisticated and detailed analysis. In contrast, traditional psychometric analyses are not recommended for individual person decision making.[107]

Third, Rasch analysis can be used to compare the BI and FIMm on a common metric. There are four main reasons for this:

1. Rasch analysis is underpinned by testing the goodness-of-fit of observed data to a mathematical (Rasch) model. As such, it does not rely on correlations.
2. Rasch analysis determines the relationships between individual items, in terms of their relative locations on the hypothesised variable, rather than the relationships between people's scale scores.
3. A mathematical property of the Rasch model, repeatedly mentioned in this monograph, is that the item location estimates are independent of the distribution of disability in the study sample.
4. By testing goodness-of-fit of observed data to a mathematical model, Rasch analysis determines formally the extent to which any group of items measure a common variable. As such, it provides a formal test of dimensionality. This is described further in Chapter 7.

Analysis

Arrangement of the data for analysis

In Chapter 6 we introduced the idea of horizontal (racking) and vertical (stacking) data arrangements. In this example, we needed to ensure that the FIMm and BI were measured on the same metric. In order to achieve this the 13 items of the FIMm and the 10 items of the BI were *racked* side by side to produce a 23-item structure. However, we also needed to ensure that the admission and discharge measures were on the same metric in a way that enabled people to obtain different locations. In order to achieve this we *stacked* admission data onto discharge data. Thus, this analysis was undertaken on one data set with the data both racked and stacked. *Figure 71* shows the layout of the data.All Rasch analyses were performed using RUMM2020.[100]

All Rasch analyses were performed using RUMM2020.[100]

Group-level comparison

In order to evaluate the impact of the transformation of BI and FIM ordinal-level

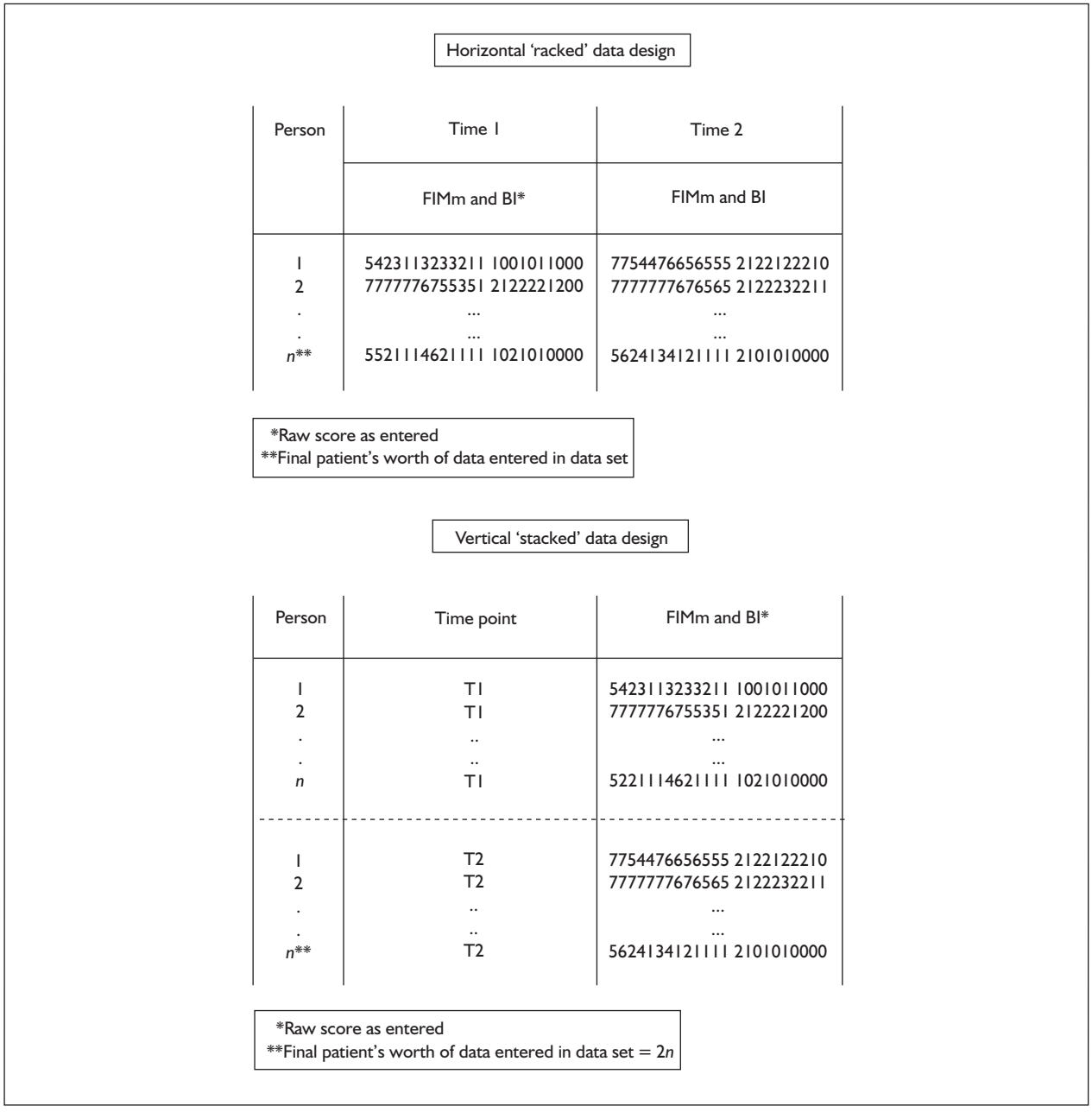


FIGURE 71 *Horizontal 'racked' data design (top); vertical 'stacked' data design (bottom).*

total scores into interval-level person locations, the relative responsiveness of the BI and FIMm was examined at the group level using five methods: comparing admission and discharge linear measurements using effect sizes (Kazis[165] standardised response mean),[143] relative efficiency (pair-wise squared *t*-values from paired *t*-tests),[168] relative precision (ratio of pair-wise *F*-values from the ANOVA)[169] and paired sample *t*-tests and ANOVA.

Individual person-level comparison

Differences were examined at the individual person level by determining how many people in the sample had a discharge person location that fell outside the 95% confidence intervals. The 95% confidence intervals around the T1 locations are computed as (T1 location ± 1.96×SE T1). However, as before, we are interested to know if the change between T1 and T2 is outside the error associated with the T1 and T2 locations. Thus, we need to compute the standard error of the difference

(SED) = $\sqrt{[(SE\ T1)^2 + (SE\ T2)^2]}$, and evaluate the change (change = T1 location – T2 location) with respect to this for each person. By dividing each person's change by his or her own standard error of the difference (change/SED), the significance of that person's change in given standard error of difference units. The significance of each person's change (SigChange) is interpreted as:

SigChange ≥ +1.96 = significant improvement

0 < SigChange < +1.95 = non-significant improvement

SigChange = 0 = no change

−1.95 < SigChange < 0 = non-significant worsening

SigChange ≤ −1.96 = significant worsening

We can now simply count the numbers of people achieving each level of significance of change. We also conducted a chi-squared test of the proportions of people in each SigChange classification to test for statistical significance.

In order to further justify the legitimacy of comparing the BI and FIMm on the same metric, we also determined the extent to which BI and FIMm estimates on admission and discharge differed at the individual person level by computing the significance of the difference (SigDiff) using the following formula:

$$\text{SigDiff for BI and FIM on admission} = \frac{\text{admission BI} - \text{admission FIM}}{\sqrt{\text{SE admission BI}^2 + \text{SE adm FIM}^2}}$$

This tested the extent to which locations generated by the BI and locations generated by the FIMm at the same time point were equivalent.

Results

Group-level comparison

The BI and FIMm measured significant changes between admission and discharge at the group level (*Table 48*). These differences represented large effect sizes (based on Cohen's criterion > 0.80). These analyses implied that the two scales had similar responsiveness.

Disability estimates generated by the BI and FIMm at *admission* were significantly different using both paired-sample *t*-tests [mean difference (SD) = −0.071 (0.559); $t = -4.73$; $p < 0.000$)] and ANOVA ($p < 0.000$). Similarly, disability estimates generated by the BI and FIMm at *discharge* were significantly different using both paired *t*-tests [mean difference (SD) = 0.054 (0.668); $t = 2.97$; $p = 0.003$)] and ANOVA ($p < 0.000$).

Individual person comparison

The FIMm detected significant improvement in 721 people and the BI detected significant improvement in 366 people (51.6% vs 26.2% of sample, respectively). The FIMm recorded 24 people as unchanged on discharge, while the BI recorded 138 people as unchanged (1.7% vs 9.9%, respectively; *Table 49*). A chi-squared test of the proportion of people in each SigChange classification revealed significant differences (BI $\chi^2 = 1403.484$, FIMm $\chi^2 = 1617.172$, $p < 0.000$).

On admission, the locations for each individual person measured by the BI and FIMm were significantly different for nobody. On discharge, the locations for each individual person measured by the BI and FIMm were significantly different for only four people (> 0.01%). This indicated that the BI and FIMm generate the same measurement for individual people at admission and discharge.

Figure 72 shows the plot of admission versus discharge scores for the BI within upper limit 95% confidence intervals around the BI admission location. *Figure 73* shows the same plot for the FIMm locations.

Conclusions

In study 3, we hypothesised that the potential advantages of Rasch analysis might help us to explore, above and beyond traditional psychometric methods, the counter-intuitive findings that the FIM and BI have similar ability to detect change. With the data arranged in a combined racked and stacked design we were able to ensure that the two scales measured on the same metric on both admission and discharge. We were able to make legitimate individual person-level comparisons as well as group comparisons.

Our findings support the clinical opinion, and logical expectation, that the FIMm is more responsive than the BI. However, this fact only became manifest when data were examined at the individual person level. Essentially, the FIMm detected significant change in twice as many people as the BI, and detected change in 133 people with unchanged BI scores. Our group-level analyses,

TABLE 48 *Rasch analysis: group-level analyses (n = 1396)*

	BI		FIMm	
	Admission	**Discharge**	**Admission**	**Discharge**
PSI (Cronbach's alpha)	0.85 (0.88)	0.87 (0.91)	0.94 (0.94)	0.92 (0.95)
Floor/ceiling effect (%)	1.1/5.3	0.1/27.9	0.9/0.3	0.3/1.6
Mean (SD)	0.265 (1.413)	1.536 (1.567)	0.336 (1.134)	1.481 (1.470)
Change [mean (SD)]		–1.270 (1.170)	–1.146 (1.023)	
F-value (*p*)		506.05 (0.000)	531.65 (0.000)	
RMP		95%	100%	
t-Value (*p*)		–40.56 (0.000)	–41.84 (0.000)	
RME		94%	100%	
SRM		1.08	1.11	
ES		0.90	1.01	

BI, Barthel Index; FIMm, Functional Independence Measure motor scale; SD, standard deviation.
PSI, person separation index. This is equivalent to Cronbach's alpha but is computed from linear measures rather than total scores.
F-value from ANOVA admission and discharge mean scores.
RMP, relative measurement precision of BI compared with FIMm = (*F*-value BI)/(*F*-value FIMm).
t-value from admission–discharge paired sample *t*-tests.
RME, relative measurement efficiency of BI compared with FIM = $(t\text{-value BI})^2/(t\text{-value FIM})^2$.
SRM, standardised response mean = mean change/SD change score.
ES, Kazis effect size = mean change/SD admission score.

using interval-level locations rather than ordinal-level scores detect no differences.

This study raises the question as to why these differences were not detected by group-level analyses. The most likely reason is that none of the group-level analyses take into account the greater precision of the FIMm, as group-level calculations use the variance of scores, in the form of either the SD (√variance) or standard error (SD adjusted for n), as the denominator. As such, these statistics are not able to detect such differences. In contrast, the improved precision of the FIMm created by the additional item response categories is manifest in the standard error of scores for individuals (confidence intervals around individual person scores). *Figure 72* shows the different measurement precisions of the BI and FIMm by plotting the standard error of measurement (y-axis) for each location on the disability continuum mapped out

TABLE 49 *Rasch analysis: individual person analyses (*n = *1396)*

Change in disability (SigChange)	**BI**	**FIM**
Significant improvement	26.2%	51.6%
Non-significant improvement	56.3%	40.1%
No change	9.9%	1.7%
Non-significant deterioration	7.4%	6.1%
Significant deterioration	0.1%	0.4%

BI, Barthel Index; FIMm, Functional Independence Measure motor scale.
Significant improvement = SigDiff ≥ +1.96.
Non-significant improvement = 0 < SigDiff < +1.95.
No change = SigDiff = 0.
Non-significant worsening = –1.95 < SigDiff < 0.
Significant worsening = SigDiff ≤ –1.96.
SigDiff, significance of difference = (admission measure – discharge measure)/√(SE admission^2 + SE discharge^2).

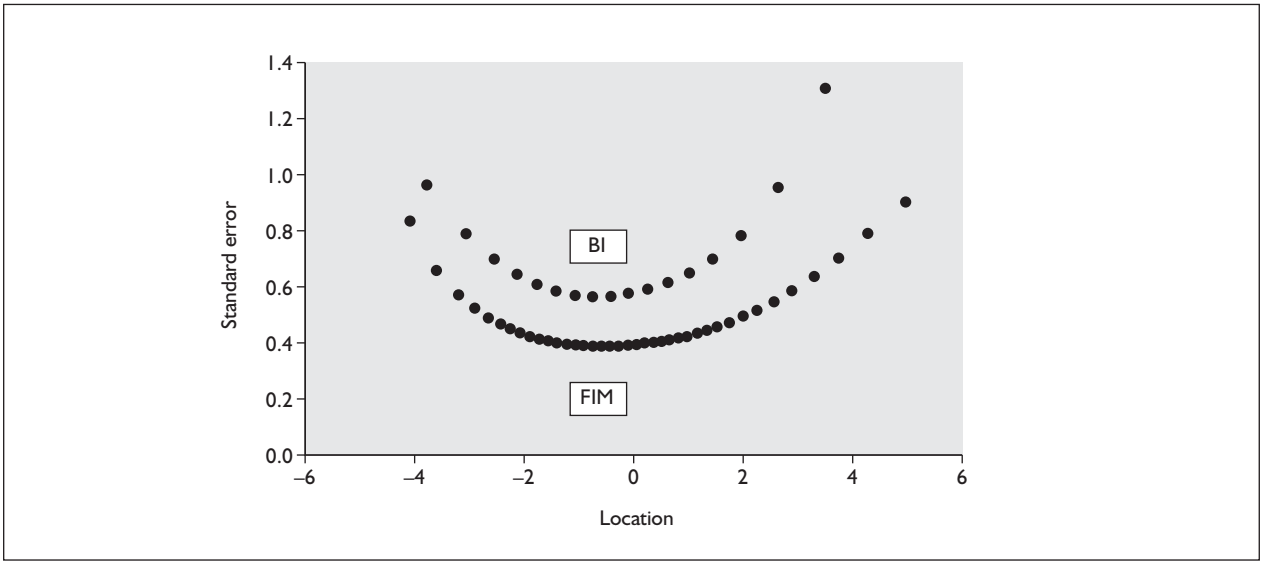


FIGURE 72 *Relative standard errors of the Barthel Index and Functional Independence Measure motor scale. Comparison of standard errors across the location.*

by the scales. The graph shows that for any point on the disability continuum the standard error associated with an FIMm measurement is smaller than that generated by the BI. That is, the FIMm is more precise than the BI.

This difference between the BI and FIMm is also shown in *Figures 73* and *74*. These plots show admission versus discharge scores for each scale and include the confidence intervals around the admission scores. It is clear that more people change significantly on the FIMm than on the BI.

Our findings have two important implications for clinicians, clinical practice and clinical trials. First, they demonstrate that group-based statistics can be misleading, not of their own volition, when representing the ability, and relative ability, of scales to detect change. As such, they demonstrate the added value of using Rasch

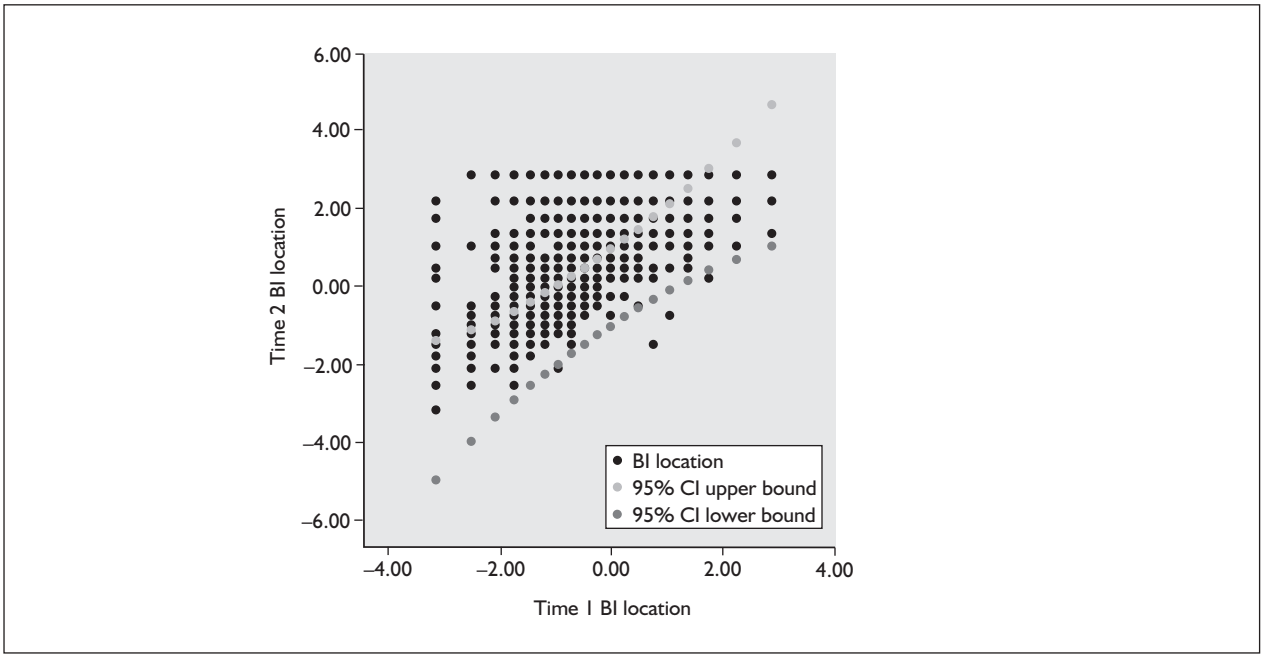


FIGURE 73 *Barthel Index – plot of item locations from time 1 and time 2 with 95% CI.*

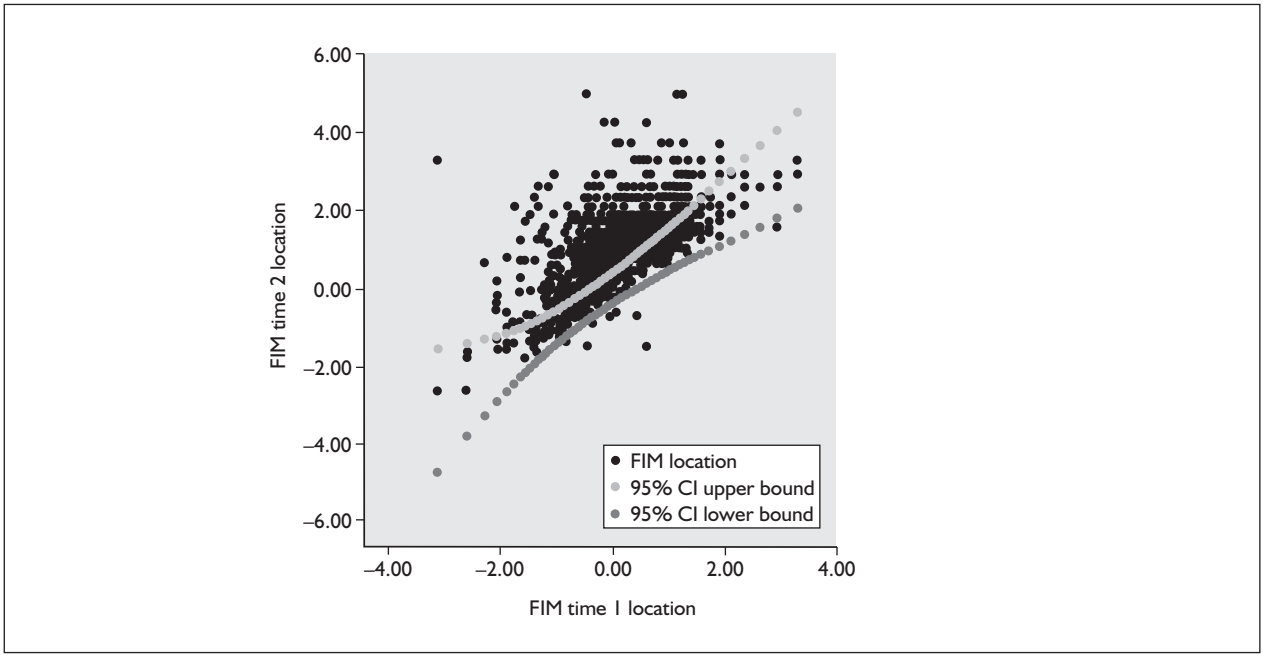


FIGURE 74 *Functional Independence Measure motor scale – plot of item locations from time 1 and time 2 with 95% CI.*

analysis and indicate that group-based analyses should be complemented by legitimate analyses at the individual person level. The second implication, a consequence of the first, is that clinical investigators need to become familiar with, and apply, new psychometric methods that enable legitimate comparisons at the individual person level. Traditional psychometric analyses, using raw scores, are not suitable for that purpose. When considered together, the findings of study 3 demonstrate the added value that Rasch analysis brings to examining and understanding rating scale responsiveness.

Summary

In this chapter we carried out three studies to test hypotheses based around the relative responsiveness of the BI and the FIMm. In the first study, we tested the hypothesis that the FIMm would be more responsive than the BI as it has more response options and therefore, theoretically, greater precision. We found that the FIM had greater potential for responsiveness, and identified more people whose scores changed, but its actual responsiveness, measured by effect sizes, equalled that of the BI. We tested this counter-intuitive finding in the second study by re-analysing the data using different group-level statistics and subsamples. However, our findings supported those of study 1. In study 3, we suggested that the reason for our findings may be lie in the limitations of traditional group-level statistics, and therefore we re-examined our data using Rasch analysis. Individual patient-level analyses showed that the FIMm detected significant change in twice as many people as the BI, supporting our hypothesis.

Our findings demonstrate that group-based statistical tests can be misleading when assessing responsiveness of outcome measures. However, more importantly, our findings demonstrate that Rasch analysis was able to detect difference between scales that traditional methods could not find.[18]

Chapter 9

Concluding remarks

Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods

Overview

The purpose of this body of work was to determine the role, if any, of new psychometric methods in health measurement. It followed on from a study supported by the NHS HTA to develop a patient-reported rating scale for MS – the MSIS-29.[126] That scale was developed using traditional psychometric methods which were at that stage, had been for a long time, and remain the dominant paradigm for rating scale development, evaluation and data analysis.

Although the new psychometric methods had been around for 40 years by then, there was felt to be limited evidence to demonstrate their superiority empirically and uncertainty of their role. A piece of work comparing and contrasting the two approaches was therefore timely. Perhaps more importantly, there were increasing calls from health-care researchers for an accessible account of the new psychometric methods. There were two main reasons for this. First, the field is based largely in educational and psychological measurement. As such the literature is alien. Second, we think that even the accounts that claim to be non-technical are at best abstruse. Essentially, most health-care professionals can barely get to first base, and have neither the time nor the facilities to obtain tuition in some of the basic language required to set a foundation for their knowledge.

One of us (JH) was fortunate to be able to do that. Initially, supported by the HTA, I was able to spend time at the University of Chicago with Professor Ben Wright and Dr Mike Linacre. Latterly, by means of a sabbatical supported by the Peninsula Medical School and the Royal Society of Medicine in the form of an Ellison-Cliffe Travelling Fellowship, I was able to work at Murdoch University with Professor David Andrich, the acknowledged leader in the field of Rasch measurement.

There is no doubt that the field of the new psychometric methods is complex. It combines mathematics with social science, philosophy and the history of science. We have found that there are many misconceptions and misunderstandings. It is no wonder that the field has been slow to take off. Hence we have tried to explain basic issues and repeat principles and interpretations. We do not apologise for the length, repetition and explanation in the text. We wish we had had this document a few years ago.

Content

This monograph tries to bridge the gap that we believe exists in front of the so-called 'basic' texts. We have tried to address the role of the new psychometric methods from a theoretical and a practical perspective, and to provide explanations and answers to the questions we had when entering the field.

Consequently, Chapter 1 presents a discussion of what rating scales are trying to achieve – to map out a meaningful continuum on which people can be located. Next we discussed the dominant traditional paradigm for the construction, evaluation and analysis of scales. The conclusions of that section are quite alarming and a cause for concern. This is because the theory underpinning most rating scale work is so very weak, in fact so a-theoretical that it cannot be tested. As Massof concludes, Classical Test Theory is a tautology.[33] We describe the conclusion of the first section about traditional psychometric methods as alarming. The reason for this is that the rating scales are used increasingly as the primary and secondary outcomes of 'state-of-the-art' clinical trials. Thus, they are the central dependent variables on which decisions about people's treatments and the spending of public funds are made.

In Chapter 2 we looked at the new psychometric methods, examining the impetus behind their development. Both Rasch measurement and Item Response Theory represent a concerted attempt to bring theory and structure to an inherently weak field. There is little doubt from a theoretical perspective (mathematical models to test theories) and a moral perspective (there can be no compromise in the efforts made to improve patient care) that they offer a major advance on traditional psychometric methods and should take over as the dominant paradigms.

Within the field of new psychometric methods there are two approaches: Rasch measurement and IRT. We have gone to some lengths to explain the difference because those explanations are rare. In fact, they are notable by their absence, which begs the question 'why?'.

The difference between IRT and Rasch measurement is striking. At its simplest, IRT seeks to find the model that best explains the observed data, whereas Rasch measurement seeks to find data that fit the model. This means that proponents of the two approaches, while often having seemingly compatible goals, have different research agendas. We have specifically not argued that one of the perspectives is correct and the other flawed. We have tried to give the background, facts and perspectives to enable researchers to decide for themselves which approach best meets their scientific need.

The different research agendas of Rasch measurement and IRT make a head-to-head comparison of their techniques somewhat meaningless. Whether they come to similar or dissimilar conclusions about a data set is not the issue. This is the main reason why this monograph has focused on one approach.

Chapter 3 demonstrates that proponents of Rasch measurement give primacy to the model, rather than to the data, because of its inherent properties. It is a method that offers the potential to achieve measurement of health variables of the nature taken for granted in the physical sciences. This makes it a potentially powerful tool for health measurement and patient care.

In the second half of the monograph we focused on worked examples comparing Rasch measurement with traditional psychometric methods. In Chapters 4 and 5, we undertook comprehensive evaluations of two existing scales: the Rivermead Mobility Index (RMI) and the Multiple Sclerosis Impact Scale (MSIS-29). In both cases, we demonstrated the limitations of traditional psychometric methods and the advantages of Rasch measurement. In addition, the amount of information gleaned about a scale, the constructs measured and the sample was profound. This highlights explicit ways of improving scales and the development of scales in general.

In Chapter 6 we used the examination of test–retest reliability as a vehicle for demonstrating how a Rasch analysis can systematically dissect measurement problems. Specifically, we demonstrated the use of different data designs to answer the various components of complex problems, and examined the examination of concept item functioning.

In Chapter 7 we demonstrated the use of Rasch analysis to equate scales. We produced best-estimate equating tables that enable users of different scales to compare their results. These would be of considerable assistance in performing meta-analyses. We also showed how equating forms the basis for item banking and computer-adaptive testing, and that these can offer quick and precise measurement.

In Chapter 8 we looked at the evaluation of responsiveness. We used Rasch analysis in an attempt to find a solution for the paradox of why two rating scales, clinically known to be different in their ability to detect change, appear to be equally responsive. We found that group-based statistics may mislead, and highlighted the value and importance of being able to examine change data at the individual person level.

Suggestions for further research

There are a multitude of future research issues open to exploration. The following recommended directions represent some of those that we believe to be important next steps. First, it is vital that other researchers and clinicians reproduce our findings in a range of clinical populations and scenarios in order to demonstrate the utility of methods used. For example, one potential avenue would be to examine the performance of the scales we used in our studies across a range of clinical populations. This is because the various scales that we used were developed for different purposes. For example, the MSIS-29 was created

for people with MS, the FAMS was generated as a modification of a cancer scale for people with MS, the BI was originally developed for people with musculoskeletal disorders, the SF-36 was created for the general population, the RMI was generated for stroke and the GHQ was developed for psychiatric conditions. As the use of a scale in a sample for which it was not necessarily developed can lead to differences in performance, and as Rasch analysis allows this potential difference to be studied empirically through an examination of DIF by condition, it is important for these examinations to be undertaken in other patient groups.

Second, detailed head-to-head comparisons of Rasch analysis and IRT are required, along with clarification of the relative roles in outcomes research. We believe that it may be possible for the two approaches to co-exist in harmony despite their differing research agendas: to use Rasch analysis to develop measurement instruments and dynamic latent trait models (essentially IRT models) to examine changes over time. Third, work is needed to determine further sample size requirements for adequate person and item estimations. As such, although evidence exists that small sample sizes are adequate, it is important to test this across the range of health measurement.

The application of Rasch analysis to clinical practice needs to be fully examined and tested. It may be naïve to expect that the use of Rasch analysis will rapidly gain acceptance for clinical practice. Therefore, its applicability may be questioned, and multiple presentations of its utility in a range of media will be needed to confirm its appropriateness. We believe that the use of clinically meaningful and psychometrically sound clinician- and patient-reported rating scales has many potential benefits in clinical practice. These include prioritising problems; facilitation of communication; screening potential problems; identifying preferences; monitoring changes or responses to treatment; training new staff; and clinical audit. The key additional benefit of Rasch analysis over traditional rating scale techniques is that we are able to move towards justifying the assessment and management of individual patients based on the scores generated by rating scales. To date, as traditional psychometric methods are the dominant paradigm, this has not been possible.

One area we have not addressed in this monograph is the extent to which a scale measures the variable it purports to measure: validity testing. While psychometric methods give sophisticated information about scales and samples they cannot tell us what is being measured. This is a related, but separate, area. We have addressed this issue in a recent article in *Lancet Neurology*,[170] which also provides further information on the new psychometric methods and can be viewed as a companion to this monograph. In brief, that article explains that the current methods of testing validity, from which conclusions are made about the extent to which a set of items measures the variable they purport to measure, are very weak. At best they provide weak circumstantial evidence for validity. Over the last 25 years one group, led by Dr Jack Stenner in the US, has developed a method for explicitly determining what is being measured.[171–174] Their work on theory-referenced measurement of reading ability is illuminating and a 'must-read'. Sadly, there is a lack of awareness of Stenner's work, and it is disappointing to find that other leaders in the field of measurement who are proponents of theory-driven approaches to rating scale development (but do not articulate how this might be achieved)[94,175,176] seem to have completely missed Stenner's ground-breaking contribution that has solved the validity problem. The potential benefits to medicine of theory-referenced measurement-derived scales, which combine Rasch's mathematical model with Stenner's construct specification equations, are enormous. Health measurement finally has the vehicles to develop rating scales that generate measurement (not just numbers) of explicit clinically meaningful variables that we are able to understand.

Conclusion

We think the arguments and demonstrations in this monograph, both theoretical and empirical, illustrate that Rasch analysis is vastly superior to traditional psychometric methods. Although we have only highlighted the value of Rasch analysis in the context of a few scales for people with MS, we feel that it has much to offer all health measurement, state-of-the-art clinical trials and, most importantly, the individual patients that clinicians treat. We think that it is time that psychometricians took the step, albeit a cerebrally painful one, of seeking to become knowledgeable about the applications of Rasch measurement to health care.

Acknowledgements

We would like to thank Professor David Andrich for his enormous contribution to our understanding of rating scale science and its practical application to health measurement.

We would also like to thank Dr Barry Sheridan for his major contribution to the practical application of the RUMM2020 Rasch measurement software program to the analysis of our data.

The following people have made significant contributions to our academic development over the years, which underpin this monograph: Professors Alan Thompson, John Zajicek, Ben Wright, Ray Fitzpatrick, Irene Styles, Ian McDonald and Donna Lamping; and Drs Jack Stenner, Mike Linacre and William Fisher.

Contribution of authors

Jeremy Hobart (Senior Lecturer and Consultant Neurologist) conceptualised the study, undertook the data analysis and wrote the report. Stefan Cano (Lecturer) contributed to the data analysis, the writing of the report, the preparation of tables and figures, the search for references and formatting of the report.

References

1. Collen FM, Wade DT, Robb GF, Bradshaw CM. The Rivermead Mobility Index: a further development of the Rivermead Motor Assessment. *Int Disabil Stud* 1991;**13**:50–4.

2. Hobart JC, Lamping DL, Fitzpatrick R, Riazi A, Thompson AJ. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. *Brain* 2001;**124**:962–73.

3. Andrich D, Styles IM. Report on the psychometric analysis of the early development instrument (EDI) using the Rasch model. Perth, WA: Murdoch University; 2004.

4. Thurstone LL. The method of paired comparisons for social values. *J Abnorm Soc Psychol* 1927;**21**:384–400.

5. Thurstone LL, Chave EJ. *The measurement of attitude: a psychophysical method and some experiments with a scale for measuring attitude toward the church*. Chicago, IL: University of Chicago Press; 1929.

6. Likert RA. A technique for the measurement of attitudes. *Arch Psychol* 1932;**140**:5–55.

7. Guttman L. A basis for scaling qualitative data. *Am Sociol Rev* 1944;**9**:139–50.

8. Edwards AL. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts; 1957.

9. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality of life instruments: attributes and review criteria. *Qual Life Res* 2002;**11**:193–205.

10. Stewart AL, Ware JE Jr, editors. *Measuring functioning and well-being: the Medical Outcomes Study approach*. Durham, NC: Duke University Press; 1992.

11. McHorney CA, Ware JE Jr, Lu JFR, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. *Med Care* 1994;**32**(1):40–66.

12. Hobart JC, Freeman JA, Lamping DL, Fitzpatrick R, Thompson AJ. The SF-36 in multiple sclerosis (MS): why basic assumptions must be tested. *J Neurol Neurosurg Psychiatry* 2001;**71**:363–70.

13. Nunnally JC Jr. *Introduction to psychological measurement*. New York: McGraw-Hill; 1970.

14. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. How responsive is the MSIS-29? A comparison with other self report scales. *J Neurol Neurosurg Psychiatry* 2005;**76**(11):1539–43.

15. Hobart JC, Lamping DL, Thompson AJ. Evaluating neurological outcome measures: the bare essentials. *J Neurol Neurosurg Psychiatry* 1996;**60**:127–30.

16. Hobart JC, Freeman JA, Thompson AJ. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. *Brain* 2000;**123**:1027–40.

17. Lord FM, Novick MR (with contributions by Birnbaum A). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley; 1968.

18. Hambleton RK, Swaminathan H. *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff; 1985.

19. Allen MJ, Yen WM. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole; 1979 (re-issued 2002 by Waveland Press Inc).

20. Lord FM. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.

21. Gulliksen H. *Theory of mental tests*. New York: Wiley; 1950.

22. Nunnally JC, Bernstein IH. *Psychometric theory*. 3rd edn. New York: McGraw-Hill; 1994.

23. Novick MR. The axioms and principal results of classical test theory. *J Math Psychol* 1966;**3**:1–18.

24. Spearman CE. Correlations of sums and differences. *Br J Psychol* 1913;**5**:417–26.

25. Spearman CE. The proof and measurement of association between two things. *Am J Psychol* 1904;**15**:72–101.

26. Spearman CE. 'General intelligence' objectively determined and measured. *Am J Psychol* 1904;**15**:201–92.

27. Spearman CE. Demonstration for true formulae of true measurement of correlation. *Am J Psychol* 1907;**18**:161–9.

28. Spearman CE. Correlation calculated from faulty data. *Br J Psychol* 1910;**3**:271–95.

29. Likert RA, Roslow S, Murphy G. A simple and reliable method of scoring the Thurstone attitude scales. *J Soc Psychol* 1934;**5**:228–38.

30. Murphy G, Likert R. *Public opinion and the individual*. New York: Harper Brothers; 1938.

31. Thurstone LL. Attitudes can be measured. *Am J Sociol* 1928;**33**(4):529–54.

32. Spector PE. *Summated rating scale construction: an introduction*. Newbury Park, CA: Sage; 1992.

33. Massof R. The measurement of vision disability. *Optom Vis Sci* 2002;**79**:516–52.

34. Zajicek J, Fox P, Sanders H, Wright D, Vickery J, Nunn A, *et al*.; UK MS Research Group. Cannabinoids for treatment of spasticity and other symptoms related to multiple sclerosis (CAMS study): multi-centre randomised placebo-controlled trial. *Lancet* 2003;**362**:1517–26.

35. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 2nd edn. Oxford: Oxford University Press; 1995.

36. Ware JE Jr, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey manual and interpretation guide*. Boston, MA: Nimrod Press; 1993.

37. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Md Med J* 1965;**14**:61–5.

38. Wright, BD. Sample-free test calibration and person measurement. *Proceedings of the 1967 invitational conference on testing problems*. Princeton, NJ: Educational Testing Service; 1968.

39. Wright BD, Masters G. *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA; 1982.

40. Thurstone LL. Theory of attitude measurement. *Psychol Rev* 1929;**36**:222–41.

41. Spector PE. Choosing response categories for summated rating scales. *J Appl Psychol* 1976;**61**(3):374–5.

42. Thurstone LL. *The reliability and validity of tests*. Ann Arbour, MI: Edwards Brothers; 1931.

43. Thurstone LL. *The vectors of mind: multiple-factor analysis for the isolation of primary traits*. Chicago, IL: University of Chicago Press; 1935.

44. Richardson M. The relationship between item difficulty and the differential validity of a test. *Psychometrika* 1936;**1**:33–49.

45. Ferguson GA. Item selection by the constant process. *Psychometrika* 1942;**7**:19–29.

46. Ferguson GA. On the theory of test development. *Psychometrika* 1949;**14**:61–8.

47. Ferguson L. A study of the Likert technique of attitude scale construct. *J Soc Psychol* 1941;**13**:51–7.

48. Lawley DN. The factorial analysis of multiple item tests. *Proc Royal Soc Edin* 1944;**62**-A:74–82.

49. Lawley DN. On problems connected with item selection and test construction. *Proc Royal Soc Edin* 1943;**6**:273–87.

50. Tucker LR. Maximum validity of a test with equivalent items. *Psychometrika* 1946;**11**:1–13.

51. Brogden H. Variation in test validity with variation in the distribution of item difficulties, number of items and degree of their intercorrelations. *Psychometrika* 1946;**11**:197–214.

52. Lazarsfeld PF. The logical and mathematical foundation of latent structure analysis. In Stouffer SA, Guttman, L, Suchman EA, Lazarsfeld PF, Star SA, Clausen JA, editors. *Measurement and Prediction*. Princeton: Princeton University Press; 1950.

53. Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika* 1937;**2**(3):151–60.

54. Cronbach LJ. Test 'reliability': its meaning and determination. *Psychometrika* 1947;**12**(1):1–16.

55. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**(3):297–334.

56. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**(4):281–302.

57. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychol Bull* 1959;**56**(2):81–105.

58. Guilford JP. New standards for test evaluation. *Educ Psychol Measurement* 1946;**6**:427–38.

59. Lord FM. A theory of test scores. *Psychometric monographs* 1952;No. 7.

60. Lord FM. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika* 1952;**17**(2):181–94.

61. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In Lord FM, editor. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley; 1968.

62. Waller M. Estimating parameters in the Rasch model: removing the effects of random guessing. Princeton, NJ: Educational Testing Service; 1976.

63. Lumsden J. Person reliability. *Appl Psychol Measurement* 1977;**1**:477–82.

64. Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika* 1986;**51**(4):567–77.

65. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Chicago: Danish Institute for Education Research; 1960.

66. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med Care* 2004;**42**(1):I7–16.

67. Andrich D. Georg Rasch. Personal communication with Hobart JC; 2006.

68. Wright BD. Solving measurement problems with the Rasch model. *J Educ Measurement* 1977;**14**(2):97–116.

69. Wright BD, Stone MH. *Best test design: Rasch measurement*. Chicago, IL: MESA; 1979.

70. Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978;**43**:561–73.

71. Wright BD. IRT in the 1990s: which models work best? *Rasch Measurement Transactions* 1992;**6**(1):196–200.

72. Divgi D. Does the Rasch model really work for multiple choice items? Not if you look closely. *J Educ Measurement* 1986;**23**(4):283–98.

73. Suen HK. *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.

74. Andrich D, de Jong JHAL, Sheridan BE. Diagnostic opportunities with the Rasch model for ordered response categories. In Rost J, Langeheine R, editors. *Applications of latent trait and latent class models in the social sciences*. Munster: Waxmann Verlag GmbH; 1997. pp. 59–70.

75. Andrich D. Distinctions between assumptions and requirements in measurement in the social sciences. In Keats JA, Taft R, Heath RA, Lovibond SH, editors. *Proceedings of the XXIVth International Congress of Psychology*. North Holland: Elsevier Science Publications BV; 1989. pp. 7–16.

76. Cook K, Monahan P, McHorney C. Delicate balance between theory and practice. *Med Care* 2003;**41**(5):571–4.

77. Stevens SS. On the theory of scales of measurement. *Science* 1946;**103**(2684):677–80.

78. Guilford JP. *Psychometric methods*. 2nd edn. New York: McGraw-Hill; 1954.

79. Torgerson WS. *Theory and methods of scaling*. New York: John Wiley and Sons; 1958.

80. Campbell NR. Symposium: measurement and its importance for philosophy. *Proceedings of the Aristotelian Society Supplement* 1938;**17**(Suppl):121–42.

81. Nunnally JC. *Psychometric theory*. 1st edn. New York: McGraw-Hill; 1967.

82. Cronbach LJ. *Essentials of psychological testing*. 5th edn. New York: HarperCollins; 1990.

83. Anastasi A, Urbina S. *Psychological testing*. 7th edn. Upper Saddle River, NJ: Prentice-Hall; 1997.

84. Ghiselli E. *Theory of psychological measurement*. New York: McGraw-Hill; 1964.

85. Helmstadter G. *Principles of psychological measurement*. New York: Appleton-Century-Crofts; 1964.

86. Horst P. *Psychological measurement and prediction*. Belmont, CA: Wadsworth; 1966.

87. Kaplan RM, Saccuzzo DP. *Psychological testing: principles, applications, and issues*. 3rd edn. Pacific Grove, CA: Brooks/Cole; 1993.

88. Brown FG. *Principles of educational and psychological testing*. Hinsdale, IL: Dryden Press; 1970.

89. Bohrnstedt GW. Measurement. In Rossi PH, Wright JD, Anderson AB, editors. *Handbook of survey research*. New York: Academic Press; 1983. pp. 69–121.

90. Helmholtz H. All science is measurement. Unreferenced citation in Bradford Hill A, editor. *Principles of medical statistics*. 7th edn. New York: Oxford University Press;1961. p. 6.

91. Campbell N. *Physics: the elements*. London: Cambridge University Press; 1920.

92. Perline R, Wright BD, Wainer H. The Rasch model as additive conjoint measurement. *Appl Psychol Measurement* 1979;**3**(2):237–55.

93. Luce RD, Tukey JW. Simultaneous conjoint measurement: a new type of fundamental measurement. *J Math Psychol* 1964;**1**:1–27.

94. Michell J. *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.

95. Wright BD, Stone MH. *Measurement essentials*. Wilmington, DE: Wide Range Inc; 1999.

96. Wright BD, Stone MH. *Making measures*. Chicago: Phaneron Press; 2004.

97. Andrich D. *Rasch models for measurement*. Beverley Hills, CA: Sage Publications; 1988.

98. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd; 1925.

99. Andrich D. The Rasch model explained. In Alagumalai S, Curtis DD, Hungi N, editors. *Applied Rasch measurement: a book of exemplars*. *Papers in honour of John P. Keeves*. Dordrecht: Springer-Kluwer; 2005, pp. 308–28.

100. RUMM 2020 [program]. 4.0 for windows (upgrade 4600.0109) version. Perth, WA: RUMM Laboratory Pty Ltd; 1997–2004.

101. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;**2**(14).

102. McDowell I, Jenkinson C. Development standards for health measures. *J Health Serv Res Policy* 1996;**1**(4):238–46.

103. Ware JE Jr, Harris WJ, Gandek B, Rogers BW, Reese PR. *MAP-R for windows: multitrait/multi-item analysis program – revised user's guide*. Boston, MA: Health Assessment Lab; 1997.

104. Howard KI, Forehand GC. A method for correcting item–total correlations for the effect of relevant item inclusion. *Educ Psychol Measurement* 1962;**22**:731–5.

105. Nunnally JC. *Psychometric theory*. 2nd edn. New York: McGraw-Hill; 1978.

106. Hays RD, Hayashi T. Beyond internal consistency reliability: rationale and user's guide for Multi-Trait Analysis Program on the microcomputer. *Behav Res Methods, Instruments Computers* 1990;**22**:167–75.

107. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res*1995;**4**:293–307.

108. Holmes WC, Shea JA. Performance of a new, HIV/AIDS-targeted quality of life (HAT-QoL) instrument in asymptomatic seropositive individuals. *Qual Life Res* 1997;**6**(6):561–71.

109. Cortina JM. What is coefficient alpha? An examination of theory and applications. *J Appl Psychol* 1993;**78**(1):98–104.

110. Eisen M, Ware JE Jr, Donald CA, Brook RH. Measuring components of children's health status. *Med Care* 1979;**17**(9):902–21.

111. Nicholl L, Hobart JC, Cramp AFL, Lowe-Strong AS. Measuring quality of life in multiple sclerosis: not as simple as it sounds. *Mult Scler* 2005;**11**:708–12.

112. Kerlinger FN. *Foundations of behavioural research*. 2nd edn. New York: Holt, Rinehart and Winston; 1973.

113. Ware JE Jr, Brook RH, Davies-Avery A, Williams KN, Stewart AL, Rogers WH, *et al*. *Conceptualization and measurement of health for adults in the health insurance study*. Vol. I: Model of health and methodology. Santa Monica, CA: Rand Corporation; 1980.

114. Guttman LA. Some necessary conditions for common-factor analysis. *Psychometrika* 1954;**19**(2):149–61.

115. Guertin WH, Bailey JP Jr. *Introduction to modern factor analysis*. Ann Arbor, MI: Edwards Brothers; 1970.

116. Cattell RB. The scree test for the number of factors. *Multivariate Behavioural Research* 1966;**1**(2):245–76.

117. Smith EV Jr. Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *J Appl Measurement* 2001;**2**:281–311.

118. Andrich D, Luo G, Sheridan BE. Interpreting RUMM2020. Perth, WA: RUMM Laboratory; 2004.

119. Hobart JC, Riazi A, Thompson AJ, Styles IM, Ingram W, Vickery PJ, *et al*. Getting the measure of spasticity in multiple sclerosis: the Multiple Sclerosis Spasticity Scale (MSSS-88). *Brain* 2006;**129**(1):224–34.

120. Andrich D. An index of person separation in latent trait theory, the traditional KR20 index and the Guttman scale response pattern. *Educ Psychol Res* 1982;**9**(1):95–104.

121. Sheridan BE. Traditional psychometric analyses. Personal communication with Hobart JC; 2006.

122. Duncan OD. Probability, disposition and the inconsistency of attitudes and behaviours. *Syntheses* 1985;**42**:21–34.

123. McHorney CA, Haley SM, Ware JE Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol* 1997;**50**(4):451–61.

124. Andrich D. Measuring criteria for choosing among models with graded responses. In von Eye A, Clogg

CC, editors. *Categorical variables in developmental research: methods and analysis*. San Diego, CA: Academic Press; 1996. pp. 3–35.

125. Thurstone LL. A method for scaling psychological and educational tests. *J Educ Psychol* 1925;**16**(7):433–51.

126. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome. *Health Technol Assess* 2004;**8**(9):1–48.

127. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. Measuring the impact of MS on walking ability: the 12-item MS Walking Scale (MSWS-12). *Neurology* 2003;**60**:31–6.

128. Holland A, O'Connor RJ, Thompson AJ, Playford ED, Hobart JC. Talking the talk on walking the walk: a 12-item generic walking scale for neurological conditions. *J Neurol* 2006; 253(12):1594–602.

129. Graham RC, Hughes RA. Clinimetric properties of a walking scale in peripheral neuropathy. *J Neurol Neurosurg Psychiatry* 2006;**77**(8):977–9.

130. Riazi A, Hobart J, Lamping D, Fitzpatrick R, Thompson A. Multiple Sclerosis Impact Scale (MSIS-29): reliability and validity in hospital based samples. *J Neurol Neurosurg Psychiatry* 2002;**73**:701–4.

131. Riazi A, Hobart J, Lamping D, Fitzpatrick R, Thompson A. Evidence-based measurement in multiple sclerosis: the psychometric properties of the physical and psychological dimensions of three quality of life rating scales. *Mult Scler* 2003;**9**(4):411–19.

132. Hoogervorst EL, Zwemmer JN, Jelles B, Polman CH, Uitdehaag BM. Multiple Sclerosis Impact Scale (MSIS-29): relation to established measures of impairment and disability. *Mult Scler* 2004;**10**(5):569–74.

133. McGuigan C, Hutchinson M. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. *J Neurol Neurosurg Psychiatry* 2004;**75**(275):266–9.

134. Sheridan BE. Chi-square. Personal communication with Hobart JC; 2006.

135. Herndon R. *Handbook of neurological rating scales*. New York: Demos Medical Publishing; 2006.

136. Wright BD. Misunderstanding the Rasch model. *J Educ Measurement* 1977;**14**:219–25.

137. Kuhn TS. *The structure of scientific revolutions*. Chicago: University of Chicago Press; 1962.

138. Kuhn TS. *The essential tension*. Chicago: University of Chicago Press; 1977.

139. Andrich D. A framework relating outcomes based education and the taxonomy of educational objectives. *Stud Educ Eval* 2002;**28**:35–59.

140. Andrich D. Implication and applications of modern test theory in the context of outcomes based research. *Stud Educ Eval* 2002;**28**:103–21.

141. Hagquist C, Andrich D. Is the Sense of Coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling. *Personality Individ Diff* 2004;**36**:955–68.

142. Duruoz MT, Poiraudeau S, Fermanian J, Menkes C-J, Amor B, Dougados M, *et al*. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. *J Rheumatol* 1996;**23**:1167–72.

143. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;**28**(7):632–8.

144. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**(2):420–8.

145. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods of measuring a quantitative variable. *Computers Biol Med* 1989;**19**:61–70.

146. Cella DF, Dineen K, Arnason B, Reder A, Webster KA, Karabatsos G, *et al*. Validation of the Functional Assessment of Multiple Sclerosis quality of life instrument. *Neurology* 1996;**47**:129–39.

147. Gompertz P, Pound P, Ebrahim S. A postal version of the Barthel Index. *Clin Rehab* 1994;**8**:233–9.

148. Goldberg DP. *Manual of the General Health Questionnaire*. Windsor: NFER-Nelson; 1978.

149. Allerup P, Bech P, Loldrup D, Alvarez P, Baneil T, Styles I, *et al*. Psychiatric, business and psychological applications of fundamental measurement models. *Int J Educ Res*1994;**21**(6):611–22.

150. EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. *Health Policy* 1990;**16**:199–208.

151. Goldberg DP, Hillier VF. A scaled version of the General Health Questionnaire. *Psychol Med* 1979;**9**:139–45.

152. Revicki D, Cella D. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res* 1997;**6**:595–600.

153. Choppin B. An item bank using sample free calibration. *Nature* 1968;**219**:870–2.

154. Ware JE Jr, Bjorner JB, Kosinski M. Practical implications of item response theory and computer adaptive testing. A brief summary of ongoing studies of widely used headache impact scales. *Med Care* 2000;**38**(Suppl 11):73–82.

155. Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ, Steinberg L, Thissen D, editors. *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.

156. Beaton D, Bombardier C, Katz J, Wright J. A taxonomy for responsiveness. *J Clin Epidemiol* 2001;**54**:1204–17.

157. O'Connor R, Cano S, Thompson A, Hobart J. Exploring rating scale responsiveness: does the total score reflect the sum of its parts? *Neurology* 2004;**62**:1842–4.

158. Mahoney F, Barthel D. Functional evaluation: the Barthel index. *Md Med J* 1965;**16**:61–5.

159. Granger CV, Hamilton B, Keith R, Zielezny M, Sherwin F. Advances in functional assessment for medical rehabilitation. *Top Geriatr Rehab* 1986;**1**:59–74.

160. van der Putten JJ, Hobart J, Freeman J, Thompson A. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. *J Neurol Neurosurg Psychiatry* 1999;**66**:480–4.

161. Hobart JC, Lamping DL, Freeman JA, Langdon DW, McLellan DL, Greenwood RJ, *et al*. Evidence-based measurement: which disability scale for neurological rehabilitation? *Neurology* 2001;**57**:639–44.

162. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. *J Clin Epidemiol* 2002;**55**:922–8.

163. Hsueh IP, Lin JH, Jeng JS, Hsieh CL. Comparison of the psychometric characteristics of the Functional Independence Measure, 5-item Barthel Index and 10-item Barthel Index in patients with stroke. *Neurol Practice* 2002;**73**:188–90.

164. Houlden H, Edwards M, McNeil J, Greenwood R. Use of the Barthel Index and the Functional Independence Measure during inpatient rehabilitation after single brain injury. *Clin Rehab* 2006;**20**:153–9.

165. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**(Suppl 3):178–89.

166. Cano S, O'Connor R, Thompson A, Hobart J. Exploring rating scale responsiveness II: Is bigger better? *Neurology* 2006;**67**:2056–9.

167. Hobart JC. Rating scales for neurologists. *J Neurol Neurosurg Psychiatry* 2003;**74**(Suppl IV):22–6.

168. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;**28**(5):542–7.

169. McHorney CA, Ware JE Jr, Rogers W, Raczek AE, Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. *Med Care* 1992;**30**(5):MS253–65.

170. Hobart J, Cano S, Zajicek J, Thompson A. Rating scales as outcome measures for clinical trials in neurology: problems, solutions and recommendations. *Lancet Neurol* 2007;**6**:1094–105.

171. Stenner AJ, Burdick H, Sandford EE, Burdick DS. How accurate are lexile text measures? *J Appl Measurement* 2006;**7**(3):307–22.

172. Stenner AJ, Smith M. Testing construct theories. *Percept Mot Skills* 1982;**55**:415–26.

173. Stenner AJ, Smith M, Burdick D. Towards a theory of construct definition. *J Educ Measurement* 1983;**20**(4):305–16.

174. Stone MH, Wright BD, Stenner AJ. Mapping variables. *J Outcomes Measurement* 1999;**3**(4):308–22.

175. Borsboom D. *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press; 2005.

176. Michell J. *Measurement in psychology: critical history of a methodological concept*. Cambridge: Cambridge University Press; 1999.

Appendix 1

Rivermead Mobility Index

Please tick 'no' or 'yes' for each question	No	Yes
1. Turning over in bed Do you turn over from your back to your side without help?	❑	❑
2. Lying to sitting From lying in bed, do you get up to sit on the edge of the bed on your own?	❑	❑
3. Sitting balance Do you sit on the edge of the bed without holding on for 10 seconds?	❑	❑
4. Sitting to standing Do you stand up (from any chair) in less than 15 seconds (using hands, and with an aid if necessary)?	❑	❑
5. Standing unsupported Observe standing for 10 seconds without any aid	❑	❑
6. Transfer Do you manage to move from bed to chair and back without any help?	❑	❑
7. Stairs[a] Do you manage a flight of stairs without help?	❑	❑
8. Walking inside, with an aid if needed[a] Do you walk 10 metres with an aid if necessary but with no standby help?	❑	❑
9. Walking outside (even ground) Do you walk around outside on pavements without help?	❑	❑
10. Walking inside with no aid Do you walk 10 metres inside with no caliper, splint or aid, and no standby help?	❑	❑
11. Picking off the floor If you drop something on the floor, do you manage to walk 5 metres, pick it up and then walk back?	❑	❑
12. Walking outside (uneven ground) Do you walk over uneven ground (grass, gravel, dirt, snow, ice, etc) without help?	❑	❑
13. Bathing Do you get in/out of bath or shower unsupervised and wash yourself?	❑	❑
14. Up and down four steps Do you manage to go up and down four steps with no rail, but using an aid if necessary?	❑	❑
15. Running Do you run 10 metres without limping in 4 seconds (fast walk is acceptable)?	❑	❑
Score (total number of 'yes' responses) =		

a In other versions of the Rivermead Mobility Index, item 7 is 'Walking inside, with an aid if needed' and item 8 is 'Stairs'. Here, they are the other way around.

Appendix 2

Multiple Sclerosis Impact Scale (MSIS-29)

Appendix 2.1: Multiple Sclerosis Impact Scale version 1 (MSIS-29v1)

- The following questions ask for your views on the impact of MS on your day-to-day life *during the past 2 weeks*.
- For each statement, please circle the *one* number that *best* describes your situation.
- Please answer *all* questions.

In the *past 2 weeks*, how much has your MS limited your ability to …	Not at all	A little	Moderately	Quite a bit	Extremely
1. Do physically demanding tasks?	1	2	3	4	5
2. Grip things tightly (e.g. turning on taps)?	1	2	3	4	5
3. Carry things?	1	2	3	4	5

In the *past 2 weeks*, how much have you been bothered by …	Not at all	A little	Moderately	Quite a bit	Extremely
4. Problems with your balance?	1	2	3	4	5
5. Difficulties moving about indoors?	1	2	3	4	5
6. Being clumsy?	1	2	3	4	5
7. Stiffness?	1	2	3	4	5
8. Heavy arms and/or legs?	1	2	3	4	5
9. Tremor of your arms or legs?	1	2	3	4	5
10. Spasms in your limbs?	1	2	3	4	5
11. Your body not doing what you want it to do?	1	2	3	4	5
12. Having to depend on others to do things for you?	1	2	3	4	5

In the *past 2 weeks*, how much have you been bothered by …	Not at all	A little	Moderately	Quite a bit	Extremely
13. Limitations in your social and leisure activities at home?	1	2	3	4	5
14. Being stuck at home more than you would like to be?	1	2	3	4	5
15. Difficulties using your hands in everyday tasks?	1	2	3	4	5
16. Having to cut down the amount of time you spent on work or other daily activities?	1	2	3	4	5
17. Problems using transport (e.g. car, bus, train, taxi, etc.)?	1	2	3	4	5
18. Taking longer to do things?	1	2	3	4	5
19. Difficulty doing things spontaneously (e.g. going out on the spur of the moment)?	1	2	3	4	5
20. Needing to go to the toilet urgently?	1	2	3	4	5
21. Feeling unwell?	1	2	3	4	5
22. Problems sleeping?	1	2	3	4	5
23. Feeling mentally fatigued?	1	2	3	4	5
24. Worries related to your MS?	1	2	3	4	5
25. Feeling anxious or tense?	1	2	3	4	5
26. Feeling irritable, impatient or short-tempered?	1	2	3	4	5
27. Problems concentrating?	1	2	3	4	5
28. Lack of confidence?	1	2	3	4	5
29. Feeling depressed?	1	2	3	4	5

Appendix 2.2: Multiple Sclerosis Impact Scale version 2 (MSIS-29v2)

- The following questions ask for your views on the impact of MS on your day-to-day life *during the past 2 weeks*.
- For each statement, please circle the *one* number that *best* describes your situation.
- Please answer *all* questions.

In the *past 2 weeks*, how much has your MS limited your ability to …	Not at all	A little	Moderately	Extremely
1. Do physically demanding tasks?	1	2	3	4
2. Grip things tightly (e.g. turning on taps)?	1	2	3	4
3. Carry things?	1	2	3	4

In the *past 2 weeks*, how much have you been bothered by …	Not at all	A little	Moderately	Extremely
4. Problems with your balance?	1	2	3	4
5. Difficulties moving about indoors?	1	2	3	4
6. Being clumsy?	1	2	3	4
7. Stiffness?	1	2	3	4
8. Heavy arms and/or legs?	1	2	3	4
9. Tremor of your arms or legs?	1	2	3	4
10. Spasms in your limbs?	1	2	3	4
11. Your body not doing what you want it to do?	1	2	3	4
12. Having to depend on others to do things for you?	1	2	3	4

In the *past 2 weeks*, how much have you been bothered by …	Not at all	A little	Moderately	Extremely
13. Limitations in your social and leisure activities at home?	1	2	3	4
14. Being stuck at home more than you would like to be?	1	2	3	4
15. Difficulties using your hands in everyday tasks?	1	2	3	4
16. Having to cut down the amount of time you spent on work or other daily activities?	1	2	3	4
17. Problems using transport (e.g. car, bus, train, taxi, etc.)?	1	2	3	4
18. Taking longer to do things?	1	2	3	4
19. Difficulty doing things spontaneously (e.g. going out on the spur of the moment)?	1	2	3	4
20. Needing to go to the toilet urgently?	1	2	3	4
21. Feeling unwell?	1	2	3	4
22. Problems sleeping?	1	2	3	4
23. Feeling mentally fatigued?	1	2	3	4
24. Worries related to your MS?	1	2	3	4
25. Feeling anxious or tense?	1	2	3	4
26. Feeling irritable, impatient or short-tempered?	1	2	3	4
27. Problems concentrating?	1	2	3	4
28. Lack of confidence?	1	2	3	4
29. Feeling depressed?	1	2	3	4

Appendix 3

Barthel Index

Activity
Feeding 0 = unable 1 = needs help cutting, spreading butter, etc., or requires modified diet 2 = independent
Bathing 0 = dependent 1 = independent (or in shower)
Grooming 0 = needs help with personal care 1 = independent face/hair/teeth/shaving (implements provided)
Dressing 0 = dependent 1 = needs help but can do about half unaided 2 = independent (including buttons, zips, laces, etc.)
Bowels 0 = incontinent (or needs to be given enemas) 1 = occasional accident 2 = continent
Bladder 0 = incontinent or catheterised and unable to manage alone 1 = occasional accident 2 = continent
Toilet use 0 = dependent 1 = needs some help, but can do something alone 2 = independent (on and off, dressing, wiping)
Transfers (bed to chair and back) 0 = unable, no sitting balance 1 = major help (one or two people, physical), can sit 2 = minor help (verbal or physical) 3 = independent
Mobility (on level surfaces) 0 = immobile or < 2 yards 1 = wheelchair independent, including corners, > 2 yards 2 = walks with help of one person (verbal or physical) > 2 yards 3 = independent (but may use any aid; for example, stick) > 2 yards
Stairs 0 = unable 1 = needs help (verbal, physical, carrying aid) 2 = independent

Appendix 4

Functional Independence Measure motor scale

Self-care
A. Eating
B. Grooming
C. Bathing
D. Dressing – upper body
E. Dressing – lower body
F. Toileting
Sphincter control
G. Bladder management
H. Bowel management
Transfers
I. Bed, chair, wheelchair
J. Toilet
K. Tub, shower
Locomotion
L. Walk/wheelchair
M. Stairs

LEVELS	**Independent** 7 Complete independence (timely, safely) 6 Modified independence (device)	**NO HELPER**
	Modified dependence 5 Supervision (subject = 100%+) 4 Minimal assistance (subject = 75%+) 3 Moderate assistance (subject = 50%+) **Complete dependence** 2 Maximal assistance (subject = 25%+) 1 Total assistance (subject = less than 25%)	**HELPER**

Health Technology Assessment reports published to date

Volume 1, 1997

No. 1
Home parenteral nutrition: a systematic review.
By Richards DM, Deeks JJ, Sheldon TA, Shaffer JL.

No. 2
Diagnosis, management and screening of early localised prostate cancer.
A review by Selley S, Donovan J, Faulkner A, Coast J, Gillatt D.

No. 3
The diagnosis, management, treatment and costs of prostate cancer in England and Wales.
A review by Chamberlain J, Melia J, Moss S, Brown J.

No. 4
Screening for fragile X syndrome.
A review by Murray J, Cuckle H, Taylor G, Hewison J.

No. 5
A review of near patient testing in primary care.
By Hobbs FDR, Delaney BC, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH, *et al*.

No. 6
Systematic review of outpatient services for chronic pain control.
By McQuay HJ, Moore RA, Eccleston C, Morley S, de C Williams AC.

No. 7
Neonatal screening for inborn errors of metabolism: cost, yield and outcome.
A review by Pollitt RJ, Green A, McCabe CJ, Booth A, Cooper NJ, Leonard JV, *et al*.

No. 8
Preschool vision screening.
A review by Snowdon SK, Stewart-Brown SL.

No. 9
Implications of socio-cultural contexts for the ethics of clinical trials.
A review by Ashcroft RE, Chadwick DW, Clark SRL, Edwards RHT, Frith L, Hutton JL.

No. 10
A critical review of the role of neonatal hearing screening in the detection of congenital hearing impairment.
By Davis A, Bamford J, Wilson I, Ramkalawan T, Forshaw M, Wright S.

No. 11
Newborn screening for inborn errors of metabolism: a systematic review.
By Seymour CA, Thomason MJ, Chalmers RA, Addison GM, Bain MD, Cockburn F, *et al*.

No. 12
Routine preoperative testing: a systematic review of the evidence.
By Munro J, Booth A, Nicholl J.

No. 13
Systematic review of the effectiveness of laxatives in the elderly.
By Petticrew M, Watt I, Sheldon T.

No. 14
When and how to assess fast-changing technologies: a comparative study of medical applications of four generic technologies.
A review by Mowatt G, Bower DJ, Brebner JA, Cairns JA, Grant AM, McKee L.

Volume 2, 1998

No. 1
Antenatal screening for Down's syndrome.
A review by Wald NJ, Kennard A, Hackshaw A, McGuire A.

No. 2
Screening for ovarian cancer: a systematic review.
By Bell R, Petticrew M, Luengo S, Sheldon TA.

No. 3
Consensus development methods, and their use in clinical guideline development.
A review by Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, *et al*.

No. 4
A cost–utility analysis of interferon beta for multiple sclerosis.
By Parkin D, McNamee P, Jacoby A, Miller P, Thomas S, Bates D.

No. 5
Effectiveness and efficiency of methods of dialysis therapy for end-stage renal disease: systematic reviews.
By MacLeod A, Grant A, Donaldson C, Khan I, Campbell M, Daly C, *et al*.

No. 6
Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model.
By Faulkner A, Kennedy LG, Baxter K, Donovan J, Wilkinson M, Bevan G.

No. 7
Antimicrobial prophylaxis in colorectal surgery: a systematic review of randomised controlled trials.
By Song F, Glenny AM.

No. 8
Bone marrow and peripheral blood stem cell transplantation for malignancy.
A review by Johnson PWM, Simnett SJ, Sweetenham JW, Morgan GJ, Stewart LA.

No. 9
Screening for speech and language delay: a systematic review of the literature.
By Law J, Boyle J, Harris F, Harkness A, Nye C.

No. 10
Resource allocation for chronic stable angina: a systematic review of effectiveness, costs and cost-effectiveness of alternative interventions.
By Sculpher MJ, Petticrew M, Kelland JL, Elliott RA, Holdright DR, Buxton MJ.

No. 11
Detection, adherence and control of hypertension for the prevention of stroke: a systematic review.
By Ebrahim S.

No. 12
Postoperative analgesia and vomiting, with special reference to day-case surgery: a systematic review.
By McQuay HJ, Moore RA.

No. 13
Choosing between randomised and nonrandomised studies: a systematic review.
By Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C.

No. 14
Evaluating patient-based outcome measures for use in clinical trials.
A review by Fitzpatrick R, Davey C, Buxton MJ, Jones DR.

No. 15
Ethical issues in the design and conduct of randomised controlled trials.
A review by Edwards SJL, Lilford RJ, Braunholtz DA, Jackson JC, Hewison J, Thornton J.

No. 16
Qualitative research methods in health technology assessment: a review of the literature.
By Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P.

No. 17
The costs and benefits of paramedic skills in pre-hospital trauma care.
By Nicholl J, Hughes S, Dixon S, Turner J, Yates D.

No. 18
Systematic review of endoscopic ultrasound in gastro-oesophageal cancer.
By Harris KM, Kelly S, Berry E, Hutton J, Roderick P, Cullingworth J, *et al*.

No. 19
Systematic reviews of trials and other studies.
By Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F.

No. 20
Primary total hip replacement surgery: a systematic review of outcomes and modelling of cost-effectiveness associated with different prostheses.
A review by Fitzpatrick R, Shortall E, Sculpher M, Murray D, Morris R, Lodge M, *et al*.

Volume 3, 1999

No. 1
Informed decision making: an annotated bibliography and systematic review.
By Bekker H, Thornton JG, Airey CM, Connelly JB, Hewison J, Robinson MB, *et al*.

No. 2
Handling uncertainty when performing economic evaluation of healthcare interventions.
A review by Briggs AH, Gray AM.

No. 3
The role of expectancies in the placebo effect and their use in the delivery of health care: a systematic review.
By Crow R, Gage H, Hampson S, Hart J, Kimber A, Thomas H.

No. 4
A randomised controlled trial of different approaches to universal antenatal HIV testing: uptake and acceptability. Annex: Antenatal HIV testing – assessment of a routine voluntary approach.
By Simpson WM, Johnstone FD, Boyd FM, Goldberg DJ, Hart GJ, Gormley SM, *et al*.

No. 5
Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review.
By Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ.

No. 6
Assessing the costs of healthcare technologies in clinical trials.
A review by Johnston K, Buxton MJ, Jones DR, Fitzpatrick R.

No. 7
Cooperatives and their primary care emergency centres: organisation and impact.
By Hallam L, Henthorne K.

No. 8
Screening for cystic fibrosis.
A review by Murray J, Cuckle H, Taylor G, Littlewood J, Hewison J.

No. 9
A review of the use of health status measures in economic evaluation.
By Brazier J, Deverill M, Green C, Harper R, Booth A.

No. 10
Methods for the analysis of quality-of-life and survival data in health technology assessment.
A review by Billingham LJ, Abrams KR, Jones DR.

No. 11
Antenatal and neonatal haemoglobinopathy screening in the UK: review and economic analysis.
By Zeuner D, Ades AE, Karnon J, Brown J, Dezateux C, Anionwu EN.

No. 12
Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses.
A review by Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, *et al*.

No. 13
'Early warning systems' for identifying new healthcare technologies.
By Robert G, Stevens A, Gabbay J.

No. 14
A systematic review of the role of human papillomavirus testing within a cervical screening programme.
By Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A, *et al*.

No. 15
Near patient testing in diabetes clinics: appraising the costs and outcomes.
By Grieve R, Beech R, Vincent J, Mazurkiewicz J.

No. 16
Positron emission tomography: establishing priorities for health technology assessment.
A review by Robert G, Milne R.

No. 17 (Pt 1)
The debridement of chronic wounds: a systematic review.
By Bradley M, Cullum N, Sheldon T.

No. 17 (Pt 2)
Systematic reviews of wound care management: (2) Dressings and topical agents used in the healing of chronic wounds.
By Bradley M, Cullum N, Nelson EA, Petticrew M, Sheldon T, Torgerson D.

No. 18
A systematic literature review of spiral and electron beam computed tomography: with particular reference to clinical applications in hepatic lesions, pulmonary embolus and coronary artery disease.
By Berry E, Kelly S, Hutton J, Harris KM, Roderick P, Boyce JC, *et al*.

No. 19
What role for statins? A review and economic model.
By Ebrahim S, Davey Smith G, McCabe C, Payne N, Pickin M, Sheldon TA, *et al*.

No. 20
Factors that limit the quality, number and progress of randomised controlled trials.
A review by Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, *et al*.

No. 21
Antimicrobial prophylaxis in total hip replacement: a systematic review.
By Glenny AM, Song F.

No. 22
Health promoting schools and health promotion in schools: two systematic reviews.
By Lister-Sharp D, Chapman S, Stewart-Brown S, Sowden A.

No. 23
Economic evaluation of a primary care-based education programme for patients with osteoarthritis of the knee.
A review by Lord J, Victor C, Littlejohns P, Ross FM, Axford JS.

Volume 4, 2000

No. 1
The estimation of marginal time preference in a UK-wide sample (TEMPUS) project.
A review by Cairns JA, van der Pol MM.

No. 2
Geriatric rehabilitation following fractures in older people: a systematic review.
By Cameron I, Crotty M, Currie C, Finnegan T, Gillespie L, Gillespie W, *et al*.

No. 3
Screening for sickle cell disease and thalassaemia: a systematic review with supplementary research.
By Davies SC, Cronin E, Gill M, Greengross P, Hickman M, Normand C.

No. 4
Community provision of hearing aids and related audiology services.
A review by Reeves DJ, Alborz A, Hickson FS, Bamford JM.

No. 5
False-negative results in screening programmes: systematic review of impact and implications.
By Petticrew MP, Sowden AJ, Lister-Sharp D, Wright K.

No. 6
Costs and benefits of community postnatal support workers: a randomised controlled trial.
By Morrell CJ, Spiby H, Stewart P, Walters S, Morgan A.

No. 7
Implantable contraceptives (subdermal implants and hormonally impregnated intrauterine systems) versus other forms of reversible contraceptives: two systematic reviews to assess relative effectiveness, acceptability, tolerability and cost-effectiveness.
By French RS, Cowan FM, Mansour DJA, Morris S, Procter T, Hughes D, *et al*.

No. 8
An introduction to statistical methods for health technology assessment.
A review by White SJ, Ashby D, Brown PJ.

No. 9
Disease-modifying drugs for multiple sclerosis: a rapid and systematic review.
By Clegg A, Bryant J, Milne R.

No. 10
Publication and related biases.
A review by Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ.

No. 11
Cost and outcome implications of the organisation of vascular services.
By Michaels J, Brazier J, Palfreyman S, Shackley P, Slack R.

No. 12
Monitoring blood glucose control in diabetes mellitus: a systematic review.
By Coster S, Gulliford MC, Seed PT, Powrie JK, Swaminathan R.

No. 13
The effectiveness of domiciliary health visiting: a systematic review of international studies and a selective review of the British literature.
By Elkan R, Kendrick D, Hewitt M, Robinson JJA, Tolley K, Blair M, *et al*.

No. 14
The determinants of screening uptake and interventions for increasing uptake: a systematic review.
By Jepson R, Clegg A, Forbes C, Lewis R, Sowden A, Kleijnen J.

No. 15
The effectiveness and cost-effectiveness of prophylactic removal of wisdom teeth.
A rapid review by Song F, O'Meara S, Wilson P, Golder S, Kleijnen J.

No. 16
Ultrasound screening in pregnancy: a systematic review of the clinical effectiveness, cost-effectiveness and women's views.
By Bricker L, Garcia J, Henderson J, Mugford M, Neilson J, Roberts T, *et al*.

No. 17
A rapid and systematic review of the effectiveness and cost-effectiveness of the taxanes used in the treatment of advanced breast and ovarian cancer.
By Lister-Sharp D, McDonagh MS, Khan KS, Kleijnen J.

No. 18
Liquid-based cytology in cervical screening: a rapid and systematic review.
By Payne N, Chilcott J, McGoogan E.

No. 19
Randomised controlled trial of non-directive counselling, cognitive–behaviour therapy and usual general practitioner care in the management of depression as well as mixed anxiety and depression in primary care.
By King M, Sibbald B, Ward E, Bower P, Lloyd M, Gabbay M, *et al*.

No. 20
Routine referral for radiography of patients presenting with low back pain: is patients' outcome influenced by GPs' referral for plain radiography?
By Kerry S, Hilton S, Patel S, Dundas D, Rink E, Lord J.

No. 21
Systematic reviews of wound care management: (3) antimicrobial agents for chronic wounds; (4) diabetic foot ulceration.
By O'Meara S, Cullum N, Majid M, Sheldon T.

No. 22
Using routine data to complement and enhance the results of randomised controlled trials.
By Lewsey JD, Leyland AH, Murray GD, Boddy FA.

No. 23
Coronary artery stents in the treatment of ischaemic heart disease: a rapid and systematic review.
By Meads C, Cummins C, Jolly K, Stevens A, Burls A, Hyde C.

No. 24
Outcome measures for adult critical care: a systematic review.
By Hayes JA, Black NA, Jenkinson C, Young JD, Rowan KM, Daly K, *et al*.

No. 25
A systematic review to evaluate the effectiveness of interventions to promote the initiation of breastfeeding.
By Fairbank L, O'Meara S, Renfrew MJ, Woolridge M, Sowden AJ, Lister-Sharp D.

No. 26
Implantable cardioverter defibrillators: arrhythmias. A rapid and systematic review.
By Parkes J, Bryant J, Milne R.

No. 27
Treatments for fatigue in multiple sclerosis: a rapid and systematic review.
By Brañas P, Jordan R, Fry-Smith A, Burls A, Hyde C.

No. 28
Early asthma prophylaxis, natural history, skeletal development and economy (EASE): a pilot randomised controlled trial.
By Baxter-Jones ADG, Helms PJ, Russell G, Grant A, Ross S, Cairns JA, *et al*.

No. 29
Screening for hypercholesterolaemia versus case finding for familial hypercholesterolaemia: a systematic review and cost-effectiveness analysis.
By Marks D, Wonderling D, Thorogood M, Lambert H, Humphries SE, Neil HAW.

No. 30
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of glycoprotein IIb/IIIa antagonists in the medical management of unstable angina.
By McDonagh MS, Bachmann LM, Golder S, Kleijnen J, ter Riet G.

No. 31
A randomised controlled trial of prehospital intravenous fluid replacement therapy in serious trauma.
By Turner J, Nicholl J, Webber L, Cox H, Dixon S, Yates D.

No. 32
Intrathecal pumps for giving opioids in chronic pain: a systematic review.
By Williams JE, Louw G, Towlerton G.

No. 33
Combination therapy (interferon alfa and ribavirin) in the treatment of chronic hepatitis C: a rapid and systematic review.
By Shepherd J, Waugh N, Hewitson P.

No. 34
A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies.
By MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS.

No. 35
Intravascular ultrasound-guided interventions in coronary artery disease: a systematic literature review, with decision-analytic modelling, of outcomes and cost-effectiveness.
By Berry E, Kelly S, Hutton J, Lindsay HSJ, Blaxill JM, Evans JA, *et al*.

No. 36
A randomised controlled trial to evaluate the effectiveness and cost-effectiveness of counselling patients with chronic depression.
By Simpson S, Corney R, Fitzgerald P, Beecham J.

No. 37
Systematic review of treatments for atopic eczema.
By Hoare C, Li Wan Po A, Williams H.

No. 38
Bayesian methods in health technology assessment: a review.
By Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR.

No. 39
The management of dyspepsia: a systematic review.
By Delaney B, Moayyedi P, Deeks J, Innes M, Soo S, Barton P, *et al*.

No. 40
A systematic review of treatments for severe psoriasis.
By Griffiths CEM, Clark CM, Chalmers RJG, Li Wan Po A, Williams HC.

Volume 5, 2001

No. 1
Clinical and cost-effectiveness of donepezil, rivastigmine and galantamine for Alzheimer's disease: a rapid and systematic review.
By Clegg A, Bryant J, Nicholson T, McIntyre L, De Broe S, Gerard K, *et al*.

No. 2
The clinical effectiveness and cost-effectiveness of riluzole for motor neurone disease: a rapid and systematic review.
By Stewart A, Sandercock J, Bryan S, Hyde C, Barton PM, Fry-Smith A, *et al*.

No. 3
Equity and the economic evaluation of healthcare.
By Sassi F, Archard L, Le Grand J.

No. 4
Quality-of-life measures in chronic diseases of childhood.
By Eiser C, Morse R.

No. 5
Eliciting public preferences for healthcare: a systematic review of techniques.
By Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, *et al*.

No. 6
General health status measures for people with cognitive impairment: learning disability and acquired brain injury.
By Riemsma RP, Forbes CA, Glanville JM, Eastwood AJ, Kleijnen J.

No. 7
An assessment of screening strategies for fragile X syndrome in the UK.
By Pembrey ME, Barnicoat AJ, Carmichael B, Bobrow M, Turner G.

No. 8
Issues in methodological research: perspectives from researchers and commissioners.
By Lilford RJ, Richardson A, Stevens A, Fitzpatrick R, Edwards S, Rock F, *et al*.

No. 9
Systematic reviews of wound care management: (5) beds; (6) compression; (7) laser therapy, therapeutic ultrasound, electrotherapy and electromagnetic therapy.
By Cullum N, Nelson EA, Flemming K, Sheldon T.

No. 10
Effects of educational and psychosocial interventions for adolescents with diabetes mellitus: a systematic review.
By Hampson SE, Skinner TC, Hart J, Storey L, Gage H, Foxcroft D, *et al*.

No. 11
Effectiveness of autologous chondrocyte transplantation for hyaline cartilage defects in knees: a rapid and systematic review.
By Jobanputra P, Parry D, Fry-Smith A, Burls A.

No. 12
Statistical assessment of the learning curves of health technologies.
By Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT.

No. 13
The effectiveness and cost-effectiveness of temozolomide for the treatment of recurrent malignant glioma: a rapid and systematic review.
By Dinnes J, Cave C, Huang S, Major K, Milne R.

No. 14
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of debriding agents in treating surgical wounds healing by secondary intention.
By Lewis R, Whiting P, ter Riet G, O'Meara S, Glanville J.

No. 15
Home treatment for mental health problems: a systematic review.
By Burns T, Knapp M, Catty J, Healey A, Henderson J, Watt H, *et al*.

No. 16
How to develop cost-conscious guidelines.
By Eccles M, Mason J.

No. 17
The role of specialist nurses in multiple sclerosis: a rapid and systematic review.
By De Broe S, Christopher F, Waugh N.

No. 18
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of orlistat in the management of obesity.
By O'Meara S, Riemsma R, Shirran L, Mather L, ter Riet G.

No. 19
The clinical effectiveness and cost-effectiveness of pioglitazone for type 2 diabetes mellitus: a rapid and systematic review.
By Chilcott J, Wight J, Lloyd Jones M, Tappenden P.

No. 20
Extended scope of nursing practice: a multicentre randomised controlled trial of appropriately trained nurses and preregistration house officers in preoperative assessment in elective general surgery.
By Kinley H, Czoski-Murray C, George S, McCabe C, Primrose J, Reilly C, *et al*.

No. 21
Systematic reviews of the effectiveness of day care for people with severe mental disorders: (1) Acute day hospital versus admission; (2) Vocational rehabilitation; (3) Day hospital versus outpatient care.
By Marshall M, Crowther R, Almaraz- Serrano A, Creed F, Sledge W, Kluiter H, *et al*.

No. 22
The measurement and monitoring of surgical adverse events.
By Bruce J, Russell EM, Mollison J, Krukowski ZH.

No. 23
Action research: a systematic review and guidance for assessment.
By Waterman H, Tillen D, Dickson R, de Koning K.

No. 24
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of gemcitabine for the treatment of pancreatic cancer.
By Ward S, Morris E, Bansback N, Calvert N, Crellin A, Forman D, *et al*.

No. 25
A rapid and systematic review of the evidence for the clinical effectiveness and cost-effectiveness of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer.
By Lloyd Jones M, Hummel S, Bansback N, Orr B, Seymour M.

No. 26
Comparison of the effectiveness of inhaler devices in asthma and chronic obstructive airways disease: a systematic review of the literature.
By Brocklebank D, Ram F, Wright J, Barry P, Cates C, Davies L, *et al*.

No. 27
The cost-effectiveness of magnetic resonance imaging for investigation of the knee joint.
By Bryan S, Weatherburn G, Bungay H, Hatrick C, Salas C, Parry D, *et al*.

No. 28
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of topotecan for ovarian cancer.
By Forbes C, Shirran L, Bagnall A-M, Duffy S, ter Riet G.

No. 29
Superseded by a report published in a later volume.

No. 30
The role of radiography in primary care patients with low back pain of at least 6 weeks duration: a randomised (unblinded) controlled trial.
By Kendrick D, Fielding K, Bentley E, Miller P, Kerslake R, Pringle M.

No. 31
Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients.
By McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, *et al*.

No. 32
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of paclitaxel, docetaxel, gemcitabine and vinorelbine in non-small-cell lung cancer.
By Clegg A, Scott DA, Sidhu M, Hewitson P, Waugh N.

No. 33
Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives.
By Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G.

No. 34
Depot antipsychotic medication in the treatment of patients with schizophrenia: (1) Meta-review; (2) Patient and nurse attitudes.
By David AS, Adams C.

No. 35
A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression.
By Churchill R, Hunot V, Corney R, Knapp M, McGuire H, Tylee A, *et al*.

No. 36
Cost analysis of child health surveillance.
By Sanderson D, Wright D, Acton C, Duree D.

Volume 6, 2002

No. 1
A study of the methods used to select review criteria for clinical audit.
By Hearnshaw H, Harker R, Cheater F, Baker R, Grimshaw G.

No. 2
Fludarabine as second-line therapy for B cell chronic lymphocytic leukaemia: a technology assessment.
By Hyde C, Wake B, Bryan S, Barton P, Fry-Smith A, Davenport C, *et al*.

No. 3
Rituximab as third-line treatment for refractory or recurrent Stage III or IV follicular non-Hodgkin's lymphoma: a systematic review and economic evaluation.
By Wake B, Hyde C, Bryan S, Barton P, Song F, Fry-Smith A, *et al*.

No. 4
A systematic review of discharge arrangements for older people.
By Parker SG, Peet SM, McPherson A, Cannaby AM, Baker R, Wilson A, *et al*.

No. 5
The clinical effectiveness and cost-effectiveness of inhaler devices used in the routine management of chronic asthma in older children: a systematic review and economic evaluation.
By Peters J, Stevenson M, Beverley C, Lim J, Smith S.

No. 6
The clinical effectiveness and cost-effectiveness of sibutramine in the management of obesity: a technology assessment.
By O'Meara S, Riemsma R, Shirran L, Mather L, ter Riet G.

No. 7
The cost-effectiveness of magnetic resonance angiography for carotid artery stenosis and peripheral vascular disease: a systematic review.
By Berry E, Kelly S, Westwood ME, Davies LM, Gough MJ, Bamford JM, *et al*.

No. 8
Promoting physical activity in South Asian Muslim women through 'exercise on prescription'.
By Carroll B, Ali N, Azam N.

No. 9
Zanamivir for the treatment of influenza in adults: a systematic review and economic evaluation.
By Burls A, Clark W, Stewart T, Preston C, Bryan S, Jefferson T, *et al*.

No. 10
A review of the natural history and epidemiology of multiple sclerosis: implications for resource allocation and health economic models.
By Richards RG, Sampson FC, Beard SM, Tappenden P.

No. 11
Screening for gestational diabetes: a systematic review and economic evaluation.
By Scott DA, Loveman E, McIntyre L, Waugh N.

No. 12
The clinical effectiveness and cost-effectiveness of surgery for people with morbid obesity: a systematic review and economic evaluation.
By Clegg AJ, Colquitt J, Sidhu MK, Royle P, Loveman E, Walker A.

No. 13
The clinical effectiveness of trastuzumab for breast cancer: a systematic review.
By Lewis R, Bagnall A-M, Forbes C, Shirran E, Duffy S, Kleijnen J, *et al*.

No. 14
The clinical effectiveness and cost-effectiveness of vinorelbine for breast cancer: a systematic review and economic evaluation.
By Lewis R, Bagnall A-M, King S, Woolacott N, Forbes C, Shirran L, *et al*.

No. 15
A systematic review of the effectiveness and cost-effectiveness of metal-on-metal hip resurfacing arthroplasty for treatment of hip disease.
By Vale L, Wyness L, McCormack K, McKenzie L, Brazzelli M, Stearns SC.

No. 16
The clinical effectiveness and cost-effectiveness of bupropion and nicotine replacement therapy for smoking cessation: a systematic review and economic evaluation.
By Woolacott NF, Jones L, Forbes CA, Mather LC, Sowden AJ, Song FJ, *et al*.

No. 17
A systematic review of effectiveness and economic evaluation of new drug treatments for juvenile idiopathic arthritis: etanercept.
By Cummins C, Connock M, Fry-Smith A, Burls A.

No. 18
Clinical effectiveness and cost-effectiveness of growth hormone in children: a systematic review and economic evaluation.
By Bryant J, Cave C, Mihaylova B, Chase D, McIntyre L, Gerard K, *et al*.

No. 19
Clinical effectiveness and cost-effectiveness of growth hormone in adults in relation to impact on quality of life: a systematic review and economic evaluation.
By Bryant J, Loveman E, Chase D, Mihaylova B, Cave C, Gerard K, *et al*.

No. 20
Clinical medication review by a pharmacist of patients on repeat prescriptions in general practice: a randomised controlled trial.
By Zermansky AG, Petty DR, Raynor DK, Lowe CJ, Freementle N, Vail A.

No. 21
The effectiveness of infliximab and etanercept for the treatment of rheumatoid arthritis: a systematic review and economic evaluation.
By Jobanputra P, Barton P, Bryan S, Burls A.

No. 22
A systematic review and economic evaluation of computerised cognitive behaviour therapy for depression and anxiety.
By Kaltenthaler E, Shackley P, Stevens K, Beverley C, Parry G, Chilcott J.

No. 23
A systematic review and economic evaluation of pegylated liposomal doxorubicin hydrochloride for ovarian cancer.
By Forbes C, Wilby J, Richardson G, Sculpher M, Mather L, Reimsma R.

No. 24
A systematic review of the effectiveness of interventions based on a stages-of-change approach to promote individual behaviour change.
By Riemsma RP, Pattenden J, Bridle C, Sowden AJ, Mather L, Watt IS, *et al*.

No. 25
A systematic review update of the clinical effectiveness and cost-effectiveness of glycoprotein IIb/IIIa antagonists.
By Robinson M, Ginnelly L, Sculpher M, Jones L, Riemsma R, Palmer S, *et al*.

No. 26
A systematic review of the effectiveness, cost-effectiveness and barriers to implementation of thrombolytic and neuroprotective therapy for acute ischaemic stroke in the NHS.
By Sandercock P, Berge E, Dennis M, Forbes J, Hand P, Kwan J, *et al*.

No. 27
A randomised controlled crossover trial of nurse practitioner versus doctor-led outpatient care in a bronchiectasis clinic.
By Caine N, Sharples LD, Hollingworth W, French J, Keogan M, Exley A, *et al*.

No. 28
Clinical effectiveness and cost – consequences of selective serotonin reuptake inhibitors in the treatment of sex offenders.
By Adi Y, Ashcroft D, Browne K, Beech A, Fry-Smith A, Hyde C.

No. 29
Treatment of established osteoporosis: a systematic review and cost–utility analysis.
By Kanis JA, Brazier JE, Stevenson M, Calvert NW, Lloyd Jones M.

No. 30
Which anaesthetic agents are cost-effective in day surgery? Literature review, national survey of practice and randomised controlled trial.
By Elliott RA Payne K, Moore JK, Davies LM, Harper NJN, St Leger AS, *et al*.

No. 31
Screening for hepatitis C among injecting drug users and in genitourinary medicine clinics: systematic reviews of effectiveness, modelling study and national survey of current practice.
By Stein K, Dalziel K, Walker A, McIntyre L, Jenkins B, Horne J, *et al*.

No. 32
The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature.
By Crow R, Gage H, Hampson S, Hart J, Kimber A, Storey L, *et al*.

No. 33
The effectiveness and cost-effectiveness of imatinib in chronic myeloid leukaemia: a systematic review.
By Garside R, Round A, Dalziel K, Stein K, Royle R.

No. 34
A comparative study of hypertonic saline, daily and alternate-day rhDNase in children with cystic fibrosis.
By Suri R, Wallis C, Bush A, Thompson S, Normand C, Flather M, *et al*.

No. 35
A systematic review of the costs and effectiveness of different models of paediatric home care.
By Parker G, Bhakta P, Lovett CA, Paisley S, Olsen R, Turner D, *et al*.

Volume 7, 2003

No. 1
How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study.
By Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J.

No. 2
Systematic review of the effectiveness and cost-effectiveness, and economic evaluation, of home versus hospital or satellite unit haemodialysis for people with end-stage renal failure.
By Mowatt G, Vale L, Perez J, Wyness L, Fraser C, MacLeod A, *et al*.

No. 3
Systematic review and economic evaluation of the effectiveness of infliximab for the treatment of Crohn's disease.
By Clark W, Raftery J, Barton P, Song F, Fry-Smith A, Burls A.

No. 4
A review of the clinical effectiveness and cost-effectiveness of routine anti-D prophylaxis for pregnant women who are rhesus negative.
By Chilcott J, Lloyd Jones M, Wight J, Forman K, Wray J, Beverley C, *et al*.

No. 5
Systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing's sarcoma and neuroblastoma.
By Riley RD, Burchill SA, Abrams KR, Heney D, Lambert PC, Jones DR, *et al*.

No. 6
The cost-effectiveness of screening for *Helicobacter pylori* to reduce mortality and morbidity from gastric cancer and peptic ulcer disease: a discrete-event simulation model.
By Roderick P, Davies R, Raftery J, Crabbe D, Pearce R, Bhandari P, *et al*.

No. 7
The clinical effectiveness and cost-effectiveness of routine dental checks: a systematic review and economic evaluation.
By Davenport C, Elley K, Salas C, Taylor-Weetman CL, Fry-Smith A, Bryan S, *et al*.

No. 8
A multicentre randomised controlled trial assessing the costs and benefits of using structured information and analysis of women's preferences in the management of menorrhagia.
By Kennedy ADM, Sculpher MJ, Coulter A, Dwyer N, Rees M, Horsley S, *et al*.

No. 9
Clinical effectiveness and cost–utility of photodynamic therapy for wet age-related macular degeneration: a systematic review and economic evaluation.
By Meads C, Salas C, Roberts T, Moore D, Fry-Smith A, Hyde C.

No. 10
Evaluation of molecular tests for prenatal diagnosis of chromosome abnormalities.
By Grimshaw GM, Szczepura A, Hultén M, MacDonald F, Nevin NC, Sutton F, *et al*.

No. 11
First and second trimester antenatal screening for Down's syndrome: the results of the Serum, Urine and Ultrasound Screening Study (SURUSS).
By Wald NJ, Rodeck C, Hackshaw AK, Walters J, Chitty L, Mackinson AM.

No. 12
The effectiveness and cost-effectiveness of ultrasound locating devices for central venous access: a systematic review and economic evaluation.
By Calvert N, Hind D, McWilliams RG, Thomas SM, Beverley C, Davidson A.

No. 13
A systematic review of atypical antipsychotics in schizophrenia.
By Bagnall A-M, Jones L, Lewis R, Ginnelly L, Glanville J, Torgerson D, *et al.*

No. 14
Prostate Testing for Cancer and Treatment (ProtecT) feasibility study.
By Donovan J, Hamdy F, Neal D, Peters T, Oliver S, Brindle L, *et al*.

No. 15
Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation.
By Boland A, Dundar Y, Bagust A, Haycox A, Hill R, Mujica Mota R, *et al*.

No. 16
Screening for fragile X syndrome: a literature review and modelling.
By Song FJ, Barton P, Sleightholme V, Yao GL, Fry-Smith A.

No. 17
Systematic review of endoscopic sinus surgery for nasal polyps.
By Dalziel K, Stein K, Round A, Garside R, Royle P.

No. 18
Towards efficient guidelines: how to monitor guideline use in primary care.
By Hutchinson A, McIntosh A, Cox S, Gilbert C.

No. 19
Effectiveness and cost-effectiveness of acute hospital-based spinal cord injuries services: systematic review.
By Bagnall A-M, Jones L, Richardson G, Duffy S, Riemsma R.

No. 20
Prioritisation of health technology assessment. The PATHS model: methods and case studies.
By Townsend J, Buxton M, Harper G.

No. 21
Systematic review of the clinical effectiveness and cost-effectiveness of tension-free vaginal tape for treatment of urinary stress incontinence.
By Cody J, Wyness L, Wallace S, Glazener C, Kilonzo M, Stearns S, *et al*.

No. 22
The clinical and cost-effectiveness of patient education models for diabetes: a systematic review and economic evaluation.
By Loveman E, Cave C, Green C, Royle P, Dunn N, Waugh N.

No. 23
The role of modelling in prioritising and planning clinical trials.
By Chilcott J, Brennan A, Booth A, Karnon J, Tappenden P.

No. 24
Cost–benefit evaluation of routine influenza immunisation in people 65–74 years of age.
By Allsup S, Gosney M, Haycox A, Regan M.

No. 25
The clinical and cost-effectiveness of pulsatile machine perfusion versus cold storage of kidneys for transplantation retrieved from heart-beating and non-heart-beating donors.
By Wight J, Chilcott J, Holmes M, Brewer N.

No. 26
Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment.
By Williams JG, Cheung WY, Cohen DR, Hutchings HA, Longo MF, Russell IT.

No. 27
Evaluating non-randomised intervention studies.
By Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, *et al*.

No. 28
A randomised controlled trial to assess the impact of a package comprising a patient-orientated, evidence-based self-help guidebook and patient-centred consultations on disease management and satisfaction in inflammatory bowel disease.
By Kennedy A, Nelson E, Reeves D, Richardson G, Roberts C, Robinson A, *et al*.

No. 29
The effectiveness of diagnostic tests for the assessment of shoulder pain due to soft tissue disorders: a systematic review.
By Dinnes J, Loveman E, McIntyre L, Waugh N.

No. 30
The value of digital imaging in diabetic retinopathy.
By Sharp PF, Olson J, Strachan F, Hipwell J, Ludbrook A, O'Donnell M, *et al*.

No. 31
Lowering blood pressure to prevent myocardial infarction and stroke: a new preventive strategy.
By Law M, Wald N, Morris J.

No. 32
Clinical and cost-effectiveness of capecitabine and tegafur with uracil for the treatment of metastatic colorectal cancer: systematic review and economic evaluation.
By Ward S, Kaltenthaler E, Cowan J, Brewer N.

No. 33
Clinical and cost-effectiveness of new and emerging technologies for early localised prostate cancer: a systematic review.
By Hummel S, Paisley S, Morgan A, Currie E, Brewer N.

No. 34
Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system.
By Royle P, Waugh N.

No. 35
Systematic review and economic decision modelling for the prevention and treatment of influenza A and B.
By Turner D, Wailoo A, Nicholson K, Cooper N, Sutton A, Abrams K.

No. 36
A randomised controlled trial to evaluate the clinical and cost-effectiveness of Hickman line insertions in adult cancer patients by nurses.
By Boland A, Haycox A, Bagust A, Fitzsimmons L.

No. 37
Redesigning postnatal care: a randomised controlled trial of protocol-based midwifery-led care focused on individual women's physical and psychological health needs.
By MacArthur C, Winter HR, Bick DE, Lilford RJ, Lancashire RJ, Knowles H, *et al*.

No. 38
Estimating implied rates of discount in healthcare decision-making.
By West RR, McNabb R, Thompson AGH, Sheldon TA, Grimley Evans J.

No. 39
Systematic review of isolation policies in the hospital management of methicillin-resistant *Staphylococcus aureus*: a review of the literature with epidemiological and economic modelling.
By Cooper BS, Stone SP, Kibbler CC, Cookson BD, Roberts JA, Medley GF, *et al*.

No. 40
Treatments for spasticity and pain in multiple sclerosis: a systematic review.
By Beard S, Hunn A, Wight J.

No. 41
The inclusion of reports of randomised trials published in languages other than English in systematic reviews.
By Moher D, Pham B, Lawson ML, Klassen TP.

No. 42
The impact of screening on future health-promoting behaviours and health beliefs: a systematic review.
By Bankhead CR, Brett J, Bukach C, Webster P, Stewart-Brown S, Munafo M, *et al*.

Volume 8, 2004

No. 1
What is the best imaging strategy for acute stroke?
By Wardlaw JM, Keir SL, Seymour J, Lewis S, Sandercock PAG, Dennis MS, *et al*.

No. 2
Systematic review and modelling of the investigation of acute and chronic chest pain presenting in primary care.
By Mant J, McManus RJ, Oakes RAL, Delaney BC, Barton PM, Deeks JJ, *et al*.

No. 3
The effectiveness and cost-effectiveness of microwave and thermal balloon endometrial ablation for heavy menstrual bleeding: a systematic review and economic modelling.
By Garside R, Stein K, Wyatt K, Round A, Price A.

No. 4
A systematic review of the role of bisphosphonates in metastatic disease.
By Ross JR, Saunders Y, Edmonds PM, Patel S, Wonderling D, Normand C, *et al*.

No. 5
Systematic review of the clinical effectiveness and cost-effectiveness of capecitabine (Xeloda®) for locally advanced and/or metastatic breast cancer.
By Jones L, Hawkins N, Westwood M, Wright K, Richardson G, Riemsma R.

No. 6
Effectiveness and efficiency of guideline dissemination and implementation strategies.
By Grimshaw JM, Thomas RE, MacLennan G, Fraser C, Ramsay CR, Vale L, *et al*.

No. 7
Clinical effectiveness and costs of the Sugarbaker procedure for the treatment of pseudomyxoma peritonei.
By Bryant J, Clegg AJ, Sidhu MK, Brodin H, Royle P, Davidson P.

No. 8
Psychological treatment for insomnia in the regulation of long-term hypnotic drug use.
By Morgan K, Dixon S, Mathers N, Thompson J, Tomeny M.

No. 9
Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome.
By Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ.

No. 10
A systematic review and economic evaluation of magnetic resonance cholangiopancreatography compared with diagnostic endoscopic retrograde cholangiopancreatography.
By Kaltenthaler E, Bravo Vergel Y, Chilcott J, Thomas S, Blakeborough T, Walters SJ, *et al*.

No. 11
The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis.
By Barton P, Jobanputra P, Wilson J, Bryan S, Burls A.

No. 12
Clinical effectiveness and cost-effectiveness of neonatal screening for inborn errors of metabolism using tandem mass spectrometry: a systematic review.
By Pandor A, Eastham J, Beverley C, Chilcott J, Paisley S.

No. 13
Clinical effectiveness and cost-effectiveness of pioglitazone and rosiglitazone in the treatment of type 2 diabetes: a systematic review and economic evaluation.
By Czoski-Murray C, Warren E, Chilcott J, Beverley C, Psyllaki MA, Cowan J.

No. 14
Routine examination of the newborn: the EMREN study. Evaluation of an extension of the midwife role including a randomised controlled trial of appropriately trained midwives and paediatric senior house officers.
By Townsend J, Wolke D, Hayes J, Davé S, Rogers C, Bloomfield L, *et al*.

No. 15
Involving consumers in research and development agenda setting for the NHS: developing an evidence-based approach.
By Oliver S, Clarke-Jones L, Rees R, Milne R, Buchanan P, Gabbay J, *et al*.

No. 16
A multi-centre randomised controlled trial of minimally invasive direct coronary bypass grafting versus percutaneous transluminal coronary angioplasty with stenting for proximal stenosis of the left anterior descending coronary artery.
By Reeves BC, Angelini GD, Bryan AJ, Taylor FC, Cripps T, Spyt TJ, *et al*.

No. 17
Does early magnetic resonance imaging influence management or improve outcome in patients referred to secondary care with low back pain? A pragmatic randomised controlled trial.
By Gilbert FJ, Grant AM, Gillan MGC, Vale L, Scott NW, Campbell MK, *et al*.

No. 18
The clinical and cost-effectiveness of anakinra for the treatment of rheumatoid arthritis in adults: a systematic review and economic analysis.
By Clark W, Jobanputra P, Barton P, Burls A.

No. 19
A rapid and systematic review and economic evaluation of the clinical and cost-effectiveness of newer drugs for treatment of mania associated with bipolar affective disorder.
By Bridle C, Palmer S, Bagnall A-M, Darba J, Duffy S, Sculpher M, *et al*.

No. 20
Liquid-based cytology in cervical screening: an updated rapid and systematic review and economic analysis.
By Karnon J, Peters J, Platt J, Chilcott J, McGoogan E, Brewer N.

No. 21
Systematic review of the long-term effects and economic consequences of treatments for obesity and implications for health improvement.
By Avenell A, Broom J, Brown TJ, Poobalan A, Aucott L, Stearns SC, *et al*.

No. 22
Autoantibody testing in children with newly diagnosed type 1 diabetes mellitus.
By Dretzke J, Cummins C, Sandercock J, Fry-Smith A, Barrett T, Burls A.

No. 23
Clinical effectiveness and cost-effectiveness of prehospital intravenous fluids in trauma patients.
By Dretzke J, Sandercock J, Bayliss S, Burls A.

No. 24
Newer hypnotic drugs for the short-term management of insomnia: a systematic review and economic evaluation.
By Dündar Y, Boland A, Strobl J, Dodd S, Haycox A, Bagust A, *et al*.

No. 25
Development and validation of methods for assessing the quality of diagnostic accuracy studies.
By Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J.

No. 26
EVALUATE hysterectomy trial: a multicentre randomised trial comparing abdominal, vaginal and laparoscopic methods of hysterectomy.
By Garry R, Fountain J, Brown J, Manca A, Mason S, Sculpher M, *et al*.

No. 27
Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-β and glatiramer acetate for multiple sclerosis.
By Tappenden P, Chilcott JB, Eggington S, Oakley J, McCabe C.

No. 28
Effectiveness and cost-effectiveness of imatinib for first-line treatment of chronic myeloid leukaemia in chronic phase: a systematic review and economic analysis.
By Dalziel K, Round A, Stein K, Garside R, Price A.

No. 29
VenUS I: a randomised controlled trial of two types of bandage for treating venous leg ulcers.
By Iglesias C, Nelson EA, Cullum NA, Torgerson DJ, on behalf of the VenUS Team.

No. 30
Systematic review of the effectiveness and cost-effectiveness, and economic evaluation, of myocardial perfusion scintigraphy for the diagnosis and management of angina and myocardial infarction.
By Mowatt G, Vale L, Brazzelli M, Hernandez R, Murray A, Scott N, *et al*.

No. 31
A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme.
By Claxton K, Ginnelly L, Sculpher M, Philips Z, Palmer S.

No. 32
The Social Support and Family Health Study: a randomised controlled trial and economic evaluation of two alternative forms of postnatal support for mothers living in disadvantaged inner-city areas.
By Wiggins M, Oakley A, Roberts I, Turner H, Rajan L, Austerberry H, *et al*.

No. 33
Psychosocial aspects of genetic screening of pregnant women and newborns: a systematic review.
By Green JM, Hewison J, Bekker HL, Bryant, Cuckle HS.

No. 34
Evaluation of abnormal uterine bleeding: comparison of three outpatient procedures within cohorts defined by age and menopausal status.
By Critchley HOD, Warner P, Lee AJ, Brechin S, Guise J, Graham B.

No. 35
Coronary artery stents: a rapid systematic review and economic evaluation.
By Hill R, Bagust A, Bakhai A, Dickson R, Dündar Y, Haycox A, *et al*.

No. 36
Review of guidelines for good practice in decision-analytic modelling in health technology assessment.
By Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, *et al*.

No. 37
Rituximab (MabThera®) for aggressive non-Hodgkin's lymphoma: systematic review and economic evaluation.
By Knight C, Hind D, Brewer N, Abbott V.

No. 38
Clinical effectiveness and cost-effectiveness of clopidogrel and modified-release dipyridamole in the secondary prevention of occlusive vascular events: a systematic review and economic evaluation.
By Jones L, Griffin S, Palmer S, Main C, Orton V, Sculpher M, *et al.*

No. 39
Pegylated interferon α-2a and -2b in combination with ribavirin in the treatment of chronic hepatitis C: a systematic review and economic evaluation.
By Shepherd J, Brodin H, Cave C, Waugh N, Price A, Gabbay J.

No. 40
Clopidogrel used in combination with aspirin compared with aspirin alone in the treatment of non-ST-segment-elevation acute coronary syndromes: a systematic review and economic evaluation.
By Main C, Palmer S, Griffin S, Jones L, Orton V, Sculpher M, *et al.*

No. 41
Provision, uptake and cost of cardiac rehabilitation programmes: improving services to under-represented groups.
By Beswick AD, Rees K, Griebsch I, Taylor FC, Burke M, West RR, *et al.*

No. 42
Involving South Asian patients in clinical trials.
By Hussain-Gambles M, Leese B, Atkin K, Brown J, Mason S, Tovey P.

No. 43
Clinical and cost-effectiveness of continuous subcutaneous insulin infusion for diabetes.
By Colquitt JL, Green C, Sidhu MK, Hartwell D, Waugh N.

No. 44
Identification and assessment of ongoing trials in health technology assessment reviews.
By Song FJ, Fry-Smith A, Davenport C, Bayliss S, Adi Y, Wilson JS, *et al*.

No. 45
Systematic review and economic evaluation of a long-acting insulin analogue, insulin glargine
By Warren E, Weatherley-Jones E, Chilcott J, Beverley C.

No. 46
Supplementation of a home-based exercise programme with a class-based programme for people with osteoarthritis of the knees: a randomised controlled trial and health economic analysis.
By McCarthy CJ, Mills PM, Pullen R, Richardson G, Hawkins N, Roberts CR, *et al*.

No. 47
Clinical and cost-effectiveness of once-daily versus more frequent use of same potency topical corticosteroids for atopic eczema: a systematic review and economic evaluation.
By Green C, Colquitt JL, Kirby J, Davidson P, Payne E.

No. 48
Acupuncture of chronic headache disorders in primary care: randomised controlled trial and economic analysis.
By Vickers AJ, Rees RW, Zollman CE, McCarney R, Smith CM, Ellis N, *et al*.

No. 49
Generalisability in economic evaluation studies in healthcare: a review and case studies.
By Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H, *et al*.

No. 50
Virtual outreach: a randomised controlled trial and economic evaluation of joint teleconferenced medical consultations.
By Wallace P, Barber J, Clayton W, Currell R, Fleming K, Garner P, *et al*.

Volume 9, 2005

No. 1
Randomised controlled multiple treatment comparison to provide a cost-effectiveness rationale for the selection of antimicrobial therapy in acne.
By Ozolins M, Eady EA, Avery A, Cunliffe WJ, O'Neill C, Simpson NB, *et al*.

No. 2
Do the findings of case series studies vary significantly according to methodological characteristics?
By Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L.

No. 3
Improving the referral process for familial breast cancer genetic counselling: findings of three randomised controlled trials of two interventions.
By Wilson BJ, Torrance N, Mollison J, Wordsworth S, Gray JR, Haites NE, *et al*.

No. 4
Randomised evaluation of alternative electrosurgical modalities to treat bladder outflow obstruction in men with benign prostatic hyperplasia.
By Fowler C, McAllister W, Plail R, Karim O, Yang Q.

No. 5
A pragmatic randomised controlled trial of the cost-effectiveness of palliative therapies for patients with inoperable oesophageal cancer.
By Shenfine J, McNamee P, Steen N, Bond J, Griffin SM.

No. 6
Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography.
By Taylor P, Champness J, Given-Wilson R, Johnston K, Potts H.

No. 7
Issues in data monitoring and interim analysis of trials.
By Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, Darbyshire JH, *et al*.

No. 8
Lay public's understanding of equipoise and randomisation in randomised controlled trials.
By Robinson EJ, Kerr CEP, Stevens AJ, Lilford RJ, Braunholtz DA, Edwards SJ, *et al*.

No. 9
Clinical and cost-effectiveness of electroconvulsive therapy for depressive illness, schizophrenia, catatonia and mania: systematic reviews and economic modelling studies.
By Greenhalgh J, Knight C, Hind D, Beverley C, Walters S.

No. 10
Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology.
By Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, *et al*.

No. 11
Clinical effectiveness and cost-effectiveness of drotrecogin alfa (activated) (Xigris®) for the treatment of severe sepsis in adults: a systematic review and economic evaluation.
By Green C, Dinnes J, Takeda A, Shepherd J, Hartwell D, Cave C, *et al*.

No. 12
A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy.
By Dinnes J, Deeks J, Kirby J, Roderick P.

No. 13
Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK.
By Willis BH, Barton P, Pearmain P, Bryan S, Hyde C.

No. 14
Laparoscopic surgery for inguinal hernia repair: systematic review of effectiveness and economic evaluation.
By McCormack K, Wake B, Perez J, Fraser C, Cook J, McIntosh E, *et al*.

No. 15
Clinical effectiveness, tolerability and cost-effectiveness of newer drugs for epilepsy in adults: a systematic review and economic evaluation.
By Wilby J, Kainth A, Hawkins N, Epstein D, McIntosh H, McDaid C, *et al*.

No. 16
A randomised controlled trial to compare the cost-effectiveness of tricyclic antidepressants, selective serotonin reuptake inhibitors and lofepramine.
By Peveler R, Kendrick T, Buxton M, Longworth L, Baldwin D, Moore M, *et al*.

No. 17
Clinical effectiveness and cost-effectiveness of immediate angioplasty for acute myocardial infarction: systematic review and economic evaluation.
By Hartwell D, Colquitt J, Loveman E, Clegg AJ, Brodin H, Waugh N, *et al*.

No. 18
A randomised controlled comparison of alternative strategies in stroke care.
By Kalra L, Evans A, Perez I, Knapp M, Swift C, Donaldson N.

No. 19
The investigation and analysis of critical incidents and adverse events in healthcare.
By Woloshynowych M, Rogers S, Taylor-Adams S, Vincent C.

No. 20
Potential use of routine databases in health technology assessment.
By Raftery J, Roderick P, Stevens A.

No. 21
Clinical and cost-effectiveness of newer immunosuppressive regimens in renal transplantation: a systematic review and modelling study.
By Woodroffe R, Yao GL, Meads C, Bayliss S, Ready A, Raftery J, *et al*.

No. 22
A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis.
By Stevenson M, Lloyd Jones M, De Nigris E, Brewer N, Davis S, Oakley J.

No. 23
A systematic review to examine the impact of psycho-educational interventions on health outcomes and costs in adults and children with difficult asthma.
By Smith JR, Mugford M, Holland R, Candy B, Noble MJ, Harrison BDW, *et al*.

No. 24
An evaluation of the costs, effectiveness and quality of renal replacement therapy provision in renal satellite units in England and Wales.
By Roderick P, Nicholson T, Armitage A, Mehta R, Mullee M, Gerard K, *et al*.

No. 25
Imatinib for the treatment of patients with unresectable and/or metastatic gastrointestinal stromal tumours: systematic review and economic evaluation.
By Wilson J, Connock M, Song F, Yao G, Fry-Smith A, Raftery J, *et al*.

No. 26
Indirect comparisons of competing interventions.
By Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, *et al*.

No. 27
Cost-effectiveness of alternative strategies for the initial medical management of non-ST elevation acute coronary syndrome: systematic review and decision-analytical modelling.
By Robinson M, Palmer S, Sculpher M, Philips Z, Ginnelly L, Bowens A, *et al*.

No. 28
Outcomes of electrically stimulated gracilis neosphincter surgery.
By Tillin T, Chambers M, Feldman R.

No. 29
The effectiveness and cost-effectiveness of pimecrolimus and tacrolimus for atopic eczema: a systematic review and economic evaluation.
By Garside R, Stein K, Castelnuovo E, Pitt M, Ashcroft D, Dimmock P, *et al*.

No. 30
Systematic review on urine albumin testing for early detection of diabetic complications.
By Newman DJ, Mattock MB, Dawnay ABS, Kerry S, McGuire A, Yaqoob M, *et al*.

No. 31
Randomised controlled trial of the cost-effectiveness of water-based therapy for lower limb osteoarthritis.
By Cochrane T, Davey RC, Matthes Edwards SM.

No. 32
Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain.
By Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M, *et al*.

No. 33
Cost-effectiveness and safety of epidural steroids in the management of sciatica.
By Price C, Arden N, Coglan L, Rogers P.

No. 34
The British Rheumatoid Outcome Study Group (BROSG) randomised controlled trial to compare the effectiveness and cost-effectiveness of aggressive versus symptomatic therapy in established rheumatoid arthritis.
By Symmons D, Tricker K, Roberts C, Davies L, Dawes P, Scott DL.

No. 35
Conceptual framework and systematic review of the effects of participants' and professionals' preferences in randomised controlled trials.
By King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, *et al*.

No. 36
The clinical and cost-effectiveness of implantable cardioverter defibrillators: a systematic review.
By Bryant J, Brodin H, Loveman E, Payne E, Clegg A.

No. 37
A trial of problem-solving by community mental health nurses for anxiety, depression and life difficulties among general practice patients. The CPN-GP study.
By Kendrick T, Simons L, Mynors-Wallis L, Gray A, Lathlean J, Pickering R, *et al*.

No. 38
The causes and effects of socio-demographic exclusions from clinical trials.
By Bartlett C, Doyal L, Ebrahim S, Davey P, Bachmann M, Egger M, *et al*.

No. 39
Is hydrotherapy cost-effective? A randomised controlled trial of combined hydrotherapy programmes compared with physiotherapy land techniques in children with juvenile idiopathic arthritis.
By Epps H, Ginnelly L, Utley M, Southwood T, Gallivan S, Sculpher M, *et al*.

No. 40
A randomised controlled trial and cost-effectiveness study of systematic screening (targeted and total population screening) versus routine practice for the detection of atrial fibrillation in people aged 65 and over. The SAFE study.
By Hobbs FDR, Fitzmaurice DA, Mant J, Murray E, Jowett S, Bryan S, *et al*.

No. 41
Displaced intracapsular hip fractures in fit, older people: a randomised comparison of reduction and fixation, bipolar hemiarthroplasty and total hip arthroplasty.
By Keating JF, Grant A, Masson M, Scott NW, Forbes JF.

No. 42
Long-term outcome of cognitive behaviour therapy clinical trials in central Scotland.
By Durham RC, Chambers JA, Power KG, Sharp DM, Macdonald RR, Major KA, *et al*.

No. 43
The effectiveness and cost-effectiveness of dual-chamber pacemakers compared with single-chamber pacemakers for bradycardia due to atrioventricular block or sick sinus syndrome: systematic review and economic evaluation.
By Castelnuovo E, Stein K, Pitt M, Garside R, Payne E.

No. 44
Newborn screening for congenital heart defects: a systematic review and cost-effectiveness analysis.
By Knowles R, Griebsch I, Dezateux C, Brown J, Bull C, Wren C.

No. 45
The clinical and cost-effectiveness of left ventricular assist devices for end-stage heart failure: a systematic review and economic evaluation.
By Clegg AJ, Scott DA, Loveman E, Colquitt J, Hutchinson J, Royle P, *et al*.

No. 46
The effectiveness of the Heidelberg Retina Tomograph and laser diagnostic glaucoma scanning system (GDx) in detecting and monitoring glaucoma.
By Kwartz AJ, Henson DB, Harper RA, Spencer AF, McLeod D.

No. 47
Clinical and cost-effectiveness of autologous chondrocyte implantation for cartilage defects in knee joints: systematic review and economic evaluation.
By Clar C, Cummins E, McIntyre L, Thomas S, Lamb J, Bain L, *et al*.

No. 48
Systematic review of effectiveness of different treatments for childhood retinoblastoma.
By McDaid C, Hartley S, Bagnall A-M, Ritchie G, Light K, Riemsma R.

No. 49
Towards evidence-based guidelines for the prevention of venous thromboembolism: systematic reviews of mechanical methods, oral anticoagulation, dextran and regional anaesthesia as thromboprophylaxis.
By Roderick P, Ferris G, Wilson K, Halls H, Jackson D, Collins R, *et al*.

No. 50
The effectiveness and cost-effectiveness of parent training/education programmes for the treatment of conduct disorder, including oppositional defiant disorder, in children.
By Dretzke J, Frew E, Davenport C, Barlow J, Stewart-Brown S, Sandercock J, *et al*.

Volume 10, 2006

No. 1
The clinical and cost-effectiveness of donepezil, rivastigmine, galantamine and memantine for Alzheimer's disease.
By Loveman E, Green C, Kirby J, Takeda A, Picot J, Payne E, *et al*.

No. 2
FOOD: a multicentre randomised trial evaluating feeding policies in patients admitted to hospital with a recent stroke.
By Dennis M, Lewis S, Cranswick G, Forbes J.

No. 3
The clinical effectiveness and cost-effectiveness of computed tomography screening for lung cancer: systematic reviews.
By Black C, Bagust A, Boland A, Walker S, McLeod C, De Verteuil R, *et al*.

No. 4
A systematic review of the effectiveness and cost-effectiveness of neuroimaging assessments used to visualise the seizure focus in people with refractory epilepsy being considered for surgery.
By Whiting P, Gupta R, Burch J, Mujica Mota RE, Wright K, Marson A, *et al*.

No. 5
Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies.
By Dundar Y, Dodd S, Dickson R, Walley T, Haycox A, Williamson PR.

No. 6
Systematic review and evaluation of methods of assessing urinary incontinence.
By Martin JL, Williams KS, Abrams KR, Turner DA, Sutton AJ, Chapple C, *et al*.

No. 7
The clinical effectiveness and cost-effectiveness of newer drugs for children with epilepsy. A systematic review.
By Connock M, Frew E, Evans B-W, Bryan S, Cummins C, Fry-Smith A, *et al*.

No. 8
Surveillance of Barrett's oesophagus: exploring the uncertainty through systematic review, expert workshop and economic modelling.
By Garside R, Pitt M, Somerville M, Stein K, Price A, Gilbert N.

No. 9
Topotecan, pegylated liposomal doxorubicin hydrochloride and paclitaxel for second-line or subsequent treatment of advanced ovarian cancer: a systematic review and economic evaluation.
By Main C, Bojke L, Griffin S, Norman G, Barbieri M, Mather L, *et al*.

No. 10
Evaluation of molecular techniques in prediction and diagnosis of cytomegalovirus disease in immunocompromised patients.
By Szczepura A, Westmoreland D, Vinogradova Y, Fox J, Clark M.

No. 11
Screening for thrombophilia in high-risk situations: systematic review and cost-effectiveness analysis. The Thrombosis: Risk and Economic Assessment of Thrombophilia Screening (TREATS) study.
By Wu O, Robertson L, Twaddle S, Lowe GDO, Clark P, Greaves M, *et al*.

No. 12
A series of systematic reviews to inform a decision analysis for sampling and treating infected diabetic foot ulcers.
By Nelson EA, O'Meara S, Craig D, Iglesias C, Golder S, Dalton J, *et al*.

No. 13
Randomised clinical trial, observational study and assessment of cost-effectiveness of the treatment of varicose veins (REACTIV trial).
By Michaels JA, Campbell WB, Brazier JE, MacIntyre JB, Palfreyman SJ, Ratcliffe J, *et al*.

No. 14
The cost-effectiveness of screening for oral cancer in primary care.
By Speight PM, Palmer S, Moles DR, Downer MC, Smith DH, Henriksson M, *et al*.

No. 15
Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis.
By Goodacre S, Sampson F, Stevenson M, Wailoo A, Sutton A, Thomas S, *et al*.

No. 16
Systematic review of the effectiveness and cost-effectiveness of HealOzone® for the treatment of occlusal pit/fissure caries and root caries.
By Brazzelli M, McKenzie L, Fielding S, Fraser C, Clarkson J, Kilonzo M, *et al*.

No. 17
Randomised controlled trials of conventional antipsychotic versus new atypical drugs, and new atypical drugs versus clozapine, in people with schizophrenia responding poorly to, or intolerant of, current drug treatment.
By Lewis SW, Davies L, Jones PB, Barnes TRE, Murray RM, Kerwin R, *et al*.

No. 18
Diagnostic tests and algorithms used in the investigation of haematuria: systematic reviews and economic evaluation.
By Rodgers M, Nixon J, Hempel S, Aho T, Kelly J, Neal D, *et al*.

No. 19
Cognitive behavioural therapy in addition to antispasmodic therapy for irritable bowel syndrome in primary care: randomised controlled trial.
By Kennedy TM, Chalder T, McCrone P, Darnley S, Knapp M, Jones RH, *et al*.

No. 20
A systematic review of the clinical effectiveness and cost-effectiveness of enzyme replacement therapies for Fabry's disease and mucopolysaccharidosis type 1.
By Connock M, Juarez-Garcia A, Frew E, Mans A, Dretzke J, Fry-Smith A, *et al*.

No. 21
Health benefits of antiviral therapy for mild chronic hepatitis C: randomised controlled trial and economic evaluation.
By Wright M, Grieve R, Roberts J, Main J, Thomas HC, on behalf of the UK Mild Hepatitis C Trial Investigators.

No. 22
Pressure relieving support surfaces: a randomised evaluation.
By Nixon J, Nelson EA, Cranny G, Iglesias CP, Hawkins K, Cullum NA, *et al*.

No. 23
A systematic review and economic model of the effectiveness and cost-effectiveness of methylphenidate, dexamfetamine and atomoxetine for the treatment of attention deficit hyperactivity disorder in children and adolescents.
By King S, Griffin S, Hodges Z, Weatherly H, Asseburg C, Richardson G, *et al.*

No. 24
The clinical effectiveness and cost-effectiveness of enzyme replacement therapy for Gaucher's disease: a systematic review.
By Connock M, Burls A, Frew E, Fry-Smith A, Juarez-Garcia A, McCabe C, *et al*.

No. 25
Effectiveness and cost-effectiveness of salicylic acid and cryotherapy for cutaneous warts. An economic decision model.
By Thomas KS, Keogh-Brown MR, Chalmers JR, Fordham RJ, Holland RC, Armstrong SJ, *et al*.

No. 26
A systematic literature review of the effectiveness of non-pharmacological interventions to prevent wandering in dementia and evaluation of the ethical implications and acceptability of their use.
By Robinson L, Hutchings D, Corner L, Beyer F, Dickinson H, Vanoli A, *et al*.

No. 27
A review of the evidence on the effects and costs of implantable cardioverter defibrillator therapy in different patient groups, and modelling of cost-effectiveness and cost–utility for these groups in a UK context.
By Buxton M, Caine N, Chase D, Connelly D, Grace A, Jackson C, *et al*.

No. 28
Adefovir dipivoxil and pegylated interferon alfa-2a for the treatment of chronic hepatitis B: a systematic review and economic evaluation.
By Shepherd J, Jones J, Takeda A, Davidson P, Price A.

No. 29
An evaluation of the clinical and cost-effectiveness of pulmonary artery catheters in patient management in intensive care: a systematic review and a randomised controlled trial.
By Harvey S, Stevens K, Harrison D, Young D, Brampton W, McCabe C, *et al*.

No. 30
Accurate, practical and cost-effective assessment of carotid stenosis in the UK.
By Wardlaw JM, Chappell FM, Stevenson M, De Nigris E, Thomas S, Gillard J, *et al*.

No. 31
Etanercept and infliximab for the treatment of psoriatic arthritis: a systematic review and economic evaluation.
By Woolacott N, Bravo Vergel Y, Hawkins N, Kainth A, Khadjesari Z, Misso K, *et al*.

No. 32
The cost-effectiveness of testing for hepatitis C in former injecting drug users.
By Castelnuovo E, Thompson-Coon J, Pitt M, Cramp M, Siebert U, Price A, *et al*.

No. 33
Computerised cognitive behaviour therapy for depression and anxiety update: a systematic review and economic evaluation.
By Kaltenthaler E, Brazier J, De Nigris E, Tumur I, Ferriter M, Beverley C, *et al*.

No. 34
Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy.
By Williams C, Brunskill S, Altman D, Briggs A, Campbell H, Clarke M, *et al*.

No. 35
Psychological therapies including dialectical behaviour therapy for borderline personality disorder: a systematic review and preliminary economic evaluation.
By Brazier J, Tumur I, Holmes M, Ferriter M, Parry G, Dent-Brown K, *et al*.

No. 36
Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model.
By Whiting P, Westwood M, Bojke L, Palmer S, Richardson G, Cooper J, *et al*.

No. 37
Cognitive behavioural therapy in chronic fatigue syndrome: a randomised controlled trial of an outpatient group programme.
By O'Dowd H, Gladwell P, Rogers CA, Hollinghurst S, Gregory A.

No. 38
A comparison of the cost-effectiveness of five strategies for the prevention of nonsteroidal anti-inflammatory drug-induced gastrointestinal toxicity: a systematic review with economic modelling.
By Brown TJ, Hooper L, Elliott RA, Payne K, Webb R, Roberts C, et al.

No. 39
The effectiveness and cost-effectiveness of computed tomography screening for coronary artery disease: systematic review.
By Waugh N, Black C, Walker S, McIntyre L, Cummins E, Hillis G.

No. 40
What are the clinical outcome and cost-effectiveness of endoscopy undertaken by nurses when compared with doctors? A Multi-Institution Nurse Endoscopy Trial (MINuET).
By Williams J, Russell I, Durai D, Cheung W-Y, Farrin A, Bloor K, *et al*.

No. 41
The clinical and cost-effectiveness of oxaliplatin and capecitabine for the adjuvant treatment of colon cancer: systematic review and economic evaluation.
By Pandor A, Eggington S, Paisley S, Tappenden P, Sutcliffe P.

No. 42
A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness.
By Chen Y-F, Jobanputra P, Barton P, Jowett S, Bryan S, Clark W, *et al*.

No. 43
Telemedicine in dermatology: a randomised controlled trial.
By Bowns IR, Collins K, Walters SJ, McDonagh AJG.

No. 44
Cost-effectiveness of cell salvage and alternative methods of minimising perioperative allogeneic blood transfusion: a systematic review and economic model.
By Davies L, Brown TJ, Haynes S, Payne K, Elliott RA, McCollum C.

No. 45
Clinical effectiveness and cost-effectiveness of laparoscopic surgery for colorectal cancer: systematic reviews and economic evaluation.
By Murray A, Lourenco T, de Verteuil R, Hernandez R, Fraser C, McKinley A, *et al*.

No. 46
Etanercept and efalizumab for the treatment of psoriasis: a systematic review.
By Woolacott N, Hawkins N, Mason A, Kainth A, Khadjesari Z, Bravo Vergel Y, *et al*.

No. 47
Systematic reviews of clinical decision tools for acute abdominal pain.
By Liu JLY, Wyatt JC, Deeks JJ, Clamp S, Keen J, Verde P, *et al*.

No. 48
Evaluation of the ventricular assist device programme in the UK.
By Sharples L, Buxton M, Caine N, Cafferty F, Demiris N, Dyer M, *et al*.

No. 49
A systematic review and economic model of the clinical and cost-effectiveness of immunosuppressive therapy for renal transplantation in children.
By Yao G, Albon E, Adi Y, Milford D, Bayliss S, Ready A, *et al*.

No. 50
Amniocentesis results: investigation of anxiety. The ARIA trial.
By Hewison J, Nixon J, Fountain J, Cocks K, Jones C, Mason G, *et al*.

Volume 11, 2007

No. 1
Pemetrexed disodium for the treatment of malignant pleural mesothelioma: a systematic review and economic evaluation.
By Dundar Y, Bagust A, Dickson R, Dodd S, Green J, Haycox A, *et al*.

No. 2
A systematic review and economic model of the clinical effectiveness and cost-effectiveness of docetaxel in combination with prednisone or prednisolone for the treatment of hormone-refractory metastatic prostate cancer.
By Collins R, Fenwick E, Trowman R, Perard R, Norman G, Light K, *et al*.

No. 3
A systematic review of rapid diagnostic tests for the detection of tuberculosis infection.
By Dinnes J, Deeks J, Kunst H, Gibson A, Cummins E, Waugh N, *et al*.

No. 4
The clinical effectiveness and cost-effectiveness of strontium ranelate for the prevention of osteoporotic fragility fractures in postmenopausal women.
By Stevenson M, Davis S, Lloyd-Jones M, Beverley C.

No. 5
A systematic review of quantitative and qualitative research on the role and effectiveness of written information available to patients about individual medicines.
By Raynor DK, Blenkinsopp A, Knapp P, Grime J, Nicolson DJ, Pollock K, *et al*.

No. 6
Oral naltrexone as a treatment for relapse prevention in formerly opioid-dependent drug users: a systematic review and economic evaluation.
By Adi Y, Juarez-Garcia A, Wang D, Jowett S, Frew E, Day E, *et al*.

No. 7
Glucocorticoid-induced osteoporosis: a systematic review and cost–utility analysis.
By Kanis JA, Stevenson M, McCloskey EV, Davis S, Lloyd-Jones M.

No. 8
Epidemiological, social, diagnostic and economic evaluation of population screening for genital chlamydial infection.
By Low N, McCarthy A, Macleod J, Salisbury C, Campbell R, Roberts TE, *et al*.

No. 9
Methadone and buprenorphine for the management of opioid dependence: a systematic review and economic evaluation.
By Connock M, Juarez-Garcia A, Jowett S, Frew E, Liu Z, Taylor RJ, *et al*.

No. 10
Exercise Evaluation Randomised Trial (EXERT): a randomised trial comparing GP referral for leisure centre-based exercise, community-based walking and advice only.
By Isaacs AJ, Critchley JA, See Tai S, Buckingham K, Westley D, Harridge SDR, *et al*.

No. 11
Interferon alfa (pegylated and non-pegylated) and ribavirin for the treatment of mild chronic hepatitis C: a systematic review and economic evaluation.
By Shepherd J, Jones J, Hartwell D, Davidson P, Price A, Waugh N.

No. 12
Systematic review and economic evaluation of bevacizumab and cetuximab for the treatment of metastatic colorectal cancer.
By Tappenden P, Jones R, Paisley S, Carroll C.

No. 13
A systematic review and economic evaluation of epoetin alfa, epoetin beta and darbepoetin alfa in anaemia associated with cancer, especially that attributable to cancer treatment.
By Wilson J, Yao GL, Raftery J, Bohlius J, Brunskill S, Sandercock J, *et al*.

No. 14
A systematic review and economic evaluation of statins for the prevention of coronary events.
By Ward S, Lloyd Jones M, Pandor A, Holmes M, Ara R, Ryan A, *et al*.

No. 15
A systematic review of the effectiveness and cost-effectiveness of different models of community-based respite care for frail older people and their carers.
By Mason A, Weatherly H, Spilsbury K, Arksey H, Golder S, Adamson J, *et al*.

No. 16
Additional therapy for young children with spastic cerebral palsy: a randomised controlled trial.
By Weindling AM, Cunningham CC, Glenn SM, Edwards RT, Reeves DJ.

No. 17
Screening for type 2 diabetes: literature review and economic modelling.
By Waugh N, Scotland G, McNamee P, Gillett M, Brennan A, Goyder E, *et al*.

No. 18
The effectiveness and cost-effectiveness of cinacalcet for secondary hyperparathyroidism in end-stage renal disease patients on dialysis: a systematic review and economic evaluation.
By Garside R, Pitt M, Anderson R, Mealing S, Roome C, Snaith A, *et al*.

No. 19
The clinical effectiveness and cost-effectiveness of gemcitabine for metastatic breast cancer: a systematic review and economic evaluation.
By Takeda AL, Jones J, Loveman E, Tan SC, Clegg AJ.

No. 20
A systematic review of duplex ultrasound, magnetic resonance angiography and computed tomography angiography for the diagnosis and assessment of symptomatic, lower limb peripheral arterial disease.
By Collins R, Cranny G, Burch J, Aguiar-Ibáñez R, Craig D, Wright K, *et al*.

No. 21
The clinical effectiveness and cost-effectiveness of treatments for children with idiopathic steroid-resistant nephrotic syndrome: a systematic review.
By Colquitt JL, Kirby J, Green C, Cooper K, Trompeter RS.

No. 22
A systematic review of the routine monitoring of growth in children of primary school age to identify growth-related conditions.
By Fayter D, Nixon J, Hartley S, Rithalia A, Butler G, Rudolf M, *et al*.

No. 23
Systematic review of the effectiveness of preventing and treating *Staphylococcus aureus* carriage in reducing peritoneal catheter-related infections.
By McCormack K, Rabindranath K, Kilonzo M, Vale L, Fraser C, McIntyre L, *et al*.

No. 24
The clinical effectiveness and cost of repetitive transcranial magnetic stimulation versus electroconvulsive therapy in severe depression: a multicentre pragmatic randomised controlled trial and economic analysis.
By McLoughlin DM, Mogg A, Eranti S, Pluck G, Purvis R, Edwards D, *et al*.

No. 25
A randomised controlled trial and economic evaluation of direct versus indirect and individual versus group modes of speech and language therapy for children with primary language impairment.
By Boyle J, McCartney E, Forbes J, O'Hare A.

No. 26
Hormonal therapies for early breast cancer: systematic review and economic evaluation.
By Hind D, Ward S, De Nigris E, Simpson E, Carroll C, Wyld L.

No. 27
Cardioprotection against the toxic effects of anthracyclines given to children with cancer: a systematic review.
By Bryant J, Picot J, Levitt G, Sullivan I, Baxter L, Clegg A.

No. 28
Adalimumab, etanercept and infliximab for the treatment of ankylosing spondylitis: a systematic review and economic evaluation.
By McLeod C, Bagust A, Boland A, Dagenais P, Dickson R, Dundar Y, *et al*.

No. 29
Prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and expected value of information analyses.
By Colbourn T, Asseburg C, Bojke L, Philips Z, Claxton K, Ades AE, *et al*.

No. 30
Clinical effectiveness and cost-effectiveness of bone morphogenetic proteins in the non-healing of fractures and spinal fusion: a systematic review.
By Garrison KR, Donell S, Ryder J, Shemilt I, Mugford M, Harvey I, *et al*.

No. 31
A randomised controlled trial of postoperative radiotherapy following breast-conserving surgery in a minimum-risk older population. The PRIME trial.
By Prescott RJ, Kunkler IH, Williams LJ, King CC, Jack W, van der Pol M, *et al*.

No. 32
Current practice, accuracy, effectiveness and cost-effectiveness of the school entry hearing screen.
By Bamford J, Fortnum H, Bristow K, Smith J, Vamvakas G, Davies L, *et al*.

No. 33
The clinical effectiveness and cost-effectiveness of inhaled insulin in diabetes mellitus: a systematic review and economic evaluation.
By Black C, Cummins E, Royle P, Philip S, Waugh N.

No. 34
Surveillance of cirrhosis for hepatocellular carcinoma: systematic review and economic analysis.
By Thompson Coon J, Rogers G, Hewson P, Wright D, Anderson R, Cramp M, *et al*.

No. 35
The Birmingham Rehabilitation Uptake Maximisation Study (BRUM). Homebased compared with hospital-based cardiac rehabilitation in a multi-ethnic population: cost-effectiveness and patient adherence.
By Jolly K, Taylor R, Lip GYH, Greenfield S, Raftery J, Mant J, *et al*.

No. 36
A systematic review of the clinical, public health and cost-effectiveness of rapid diagnostic tests for the detection and identification of bacterial intestinal pathogens in faeces and food.
By Abubakar I, Irvine L, Aldus CF, Wyatt GM, Fordham R, Schelenz S, *et al*.

No. 37
A randomised controlled trial examining the longer-term outcomes of standard versus new antiepileptic drugs. The SANAD trial.
By Marson AG, Appleton R, Baker GA, Chadwick DW, Doughty J, Eaton B, *et al*.

No. 38
Clinical effectiveness and cost-effectiveness of different models of managing long-term oral anti-coagulation therapy: a systematic review and economic modelling.
By Connock M, Stevens C, Fry-Smith A, Jowett S, Fitzmaurice D, Moore D, *et al*.

No. 39
A systematic review and economic model of the clinical effectiveness and cost-effectiveness of interventions for preventing relapse in people with bipolar disorder.
By Soares-Weiser K, Bravo Vergel Y, Beynon S, Dunn G, Barbieri M, Duffy S, *et al*.

No. 40
Taxanes for the adjuvant treatment of early breast cancer: systematic review and economic evaluation.
By Ward S, Simpson E, Davis S, Hind D, Rees A, Wilkinson A.

No. 41
The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation.
By Burr JM, Mowatt G, Hernández R, Siddiqui MAR, Cook J, Lourenco T, *et al*.

No. 42
Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models.
By Davis A, Smith P, Ferguson M, Stephens D, Gianopoulos I.

No. 43
Contamination in trials of educational interventions.
By Keogh-Brown MR, Bachmann MO, Shepstone L, Hewitt C, Howe A, Ramsay CR, *et al*.

No. 44
Overview of the clinical effectiveness of positron emission tomography imaging in selected cancers.
By Facey K, Bradbury I, Laking G, Payne E.

No. 45
The effectiveness and cost-effectiveness of carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma: a systematic review and economic evaluation.
By Garside R, Pitt M, Anderson R, Rogers G, Dyer M, Mealing S, *et al*.

No. 46
Drug-eluting stents: a systematic review and economic evaluation.
By Hill RA, Boland A, Dickson R, Dündar Y, Haycox A, McLeod C, *et al*.

No. 47
The clinical effectiveness and cost-effectiveness of cardiac resynchronisation (biventricular pacing) for heart failure: systematic review and economic model.
By Fox M, Mealing S, Anderson R, Dean J, Stein K, Price A, *et al*.

No. 48
Recruitment to randomised trials: strategies for trial enrolment and participation study. The STEPS study.
By Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, *et al*.

No. 49
Cost-effectiveness of functional cardiac testing in the diagnosis and management of coronary artery disease: a randomised controlled trial. The CECaT trial.
By Sharples L, Hughes V, Crean A, Dyer M, Buxton M, Goldsmith K, *et al*.

No. 50
Evaluation of diagnostic tests when there is no gold standard. A review of methods.
By Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM.

No. 51
Systematic reviews of the clinical effectiveness and cost-effectiveness of proton pump inhibitors in acute upper gastrointestinal bleeding.
By Leontiadis GI, Sreedharan A, Dorward S, Barton P, Delaney B, Howden CW, *et al*.

No. 52
A review and critique of modelling in prioritising and designing screening programmes.
By Karnon J, Goyder E, Tappenden P, McPhie S, Towers I, Brazier J, *et al*.

No. 53
An assessment of the impact of the NHS Health Technology Assessment Programme.
By Hanney S, Buxton M, Green C, Coulson D, Raftery J.

Volume 12, 2008

No. 1
A systematic review and economic model of switching from nonglycopeptide to glycopeptide antibiotic prophylaxis for surgery.
By Cranny G, Elliott R, Weatherly H, Chambers D, Hawkins N, Myers L, *et al*.

No. 2
'Cut down to quit' with nicotine replacement therapies in smoking cessation: a systematic review of effectiveness and economic analysis.
By Wang D, Connock M, Barton P, Fry-Smith A, Aveyard P, Moore D.

No. 3
A systematic review of the effectiveness of strategies for reducing fracture risk in children with juvenile idiopathic arthritis with additional data on long-term risk of fracture and cost of disease management.
By Thornton J, Ashcroft D, O'Neill T, Elliott R, Adams J, Roberts C, *et al*.

No. 4
Does befriending by trained lay workers improve psychological well-being and quality of life for carers of people with dementia, and at what cost? A randomised controlled trial.
By Charlesworth G, Shepstone L, Wilson E, Thalanany M, Mugford M, Poland F.

No. 5
A multi-centre retrospective cohort study comparing the efficacy, safety and cost-effectiveness of hysterectomy and uterine artery embolisation for the treatment of symptomatic uterine fibroids. The HOPEFUL study.
By Hirst A, Dutton S, Wu O, Briggs A, Edwards C, Waldenmaier L, *et al*.

No. 6
Methods of prediction and prevention of pre-eclampsia: systematic reviews of accuracy and effectiveness literature with economic modelling.
By Meads CA, Cnossen JS, Meher S, Juarez-Garcia A, ter Riet G, Duley L, *et al*.

No. 7
The use of economic evaluations in NHS decision-making: a review and empirical investigation.
By Williams I, McIver S, Moore D, Bryan S.

No. 8
Stapled haemorrhoidectomy (haemorrhoidopexy) for the treatment of haemorrhoids: a systematic review and economic evaluation.
By Burch J, Epstein D, Baba-Akbari A, Weatherly H, Fox D, Golder S, *et al*.

No. 9
The clinical effectiveness of diabetes education models for Type 2 diabetes: a systematic review.
By Loveman E, Frampton GK, Clegg AJ.

No. 10
Payment to healthcare professionals for patient recruitment to trials: systematic review and qualitative study.
By Raftery J, Bryant J, Powell J, Kerr C, Hawker S.

No. 11
Cyclooxygenase-2 selective non-steroidal anti-inflammatory drugs (etodolac, meloxicam, celecoxib, rofecoxib, etoricoxib, valdecoxib and lumiracoxib) for osteoarthritis and rheumatoid arthritis: a systematic review and economic evaluation.
By Chen Y-F, Jobanputra P, Barton P, Bryan S, Fry-Smith A, Harris G, *et al*.

No. 12
The clinical effectiveness and cost-effectiveness of central venous catheters treated with anti-infective agents in preventing bloodstream infections: a systematic review and economic evaluation.
By Hockenhull JC, Dwan K, Boland A, Smith G, Bagust A, Dundar Y, *et al*.

No. 13
Stepped treatment of older adults on laxatives. The STOOL trial.
By Mihaylov S, Stark C, McColl E, Steen N, Vanoli A, Rubin G, *et al*.

No. 14
A randomised controlled trial of cognitive behaviour therapy in adolescents with major depression treated by selective serotonin reuptake inhibitors. The ADAPT trial.
By Goodyer IM, Dubicka B, Wilkinson P, Kelvin R, Roberts C, Byford S, *et al*.

No. 15
The use of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: systematic review and economic evaluation.
By Hind D, Tappenden P, Tumur I, Eggington E, Sutcliffe P, Ryan A.

No. 16
Ranibizumab and pegaptanib for the treatment of age-related macular degeneration: a systematic review and economic evaluation.
By Colquitt JL, Jones J, Tan SC, Takeda A, Clegg AJ, Price A.

No. 17
Systematic review of the clinical effectiveness and cost-effectiveness of 64-slice or higher computed tomography angiography as an alternative to invasive coronary angiography in the investigation of coronary artery disease.
By Mowatt G, Cummins E, Waugh N, Walker S, Cook J, Jia X, *et al*.

No. 18
Structural neuroimaging in psychosis: a systematic review and economic evaluation.
By Albon E, Tsourapas A, Frew E, Davenport C, Oyebode F, Bayliss S, *et al*.

No. 19
Systematic review and economic analysis of the comparative effectiveness of different inhaled corticosteroids and their usage with long-acting beta$_2$ agonists for the treatment of chronic asthma in adults and children aged 12 years and over.
By Shepherd J, Rogers G, Anderson R, Main C, Thompson-Coon J, Hartwell D, *et al*.

No. 20
Systematic review and economic analysis of the comparative effectiveness of different inhaled corticosteroids and their usage with long-acting beta$_2$ agonists for the treatment of chronic asthma in children under the age of 12 years.
By Main C, Shepherd J, Anderson R, Rogers G, Thompson-Coon J, Liu Z, *et al*.

No. 21
Ezetimibe for the treatment of hypercholesterolaemia: a systematic review and economic evaluation.
By Ara R, Tumur I, Pandor A, Duenas A, Williams R, Wilkinson A, *et al*.

No. 22
Topical or oral ibuprofen for chronic knee pain in older people. The TOIB study.
By Underwood M, Ashby D, Carnes D, Castelnuovo E, Cross P, Harding G, *et al*.

No. 23
A prospective randomised comparison of minor surgery in primary and secondary care. The MiSTIC trial.
By George S, Pockney P, Primrose J, Smith H, Little P, Kinley H, *et al*.

No. 24
A review and critical appraisal of measures of therapist–patient interactions in mental health settings.
By Cahill J, Barkham M, Hardy G, Gilbody S, Richards D, Bower P, *et al*.

No. 25
The clinical effectiveness and cost-effectiveness of screening programmes for amblyopia and strabismus in children up to the age of 4–5 years: a systematic review and economic evaluation.
By Carlton J, Karnon J, Czoski-Murray C, Smith KJ, Marr J.

No. 26
A systematic review of the clinical effectiveness and cost-effectiveness and economic modelling of minimal incision total hip replacement approaches in the management of arthritic disease of the hip.
By de Verteuil R, Imamura M, Zhu S, Glazener C, Fraser C, Munro N, *et al*.

No. 27
A preliminary model-based assessment of the cost–utility of a screening programme for early age-related macular degeneration.
By Karnon J, Czoski-Murray C, Smith K, Brand C, Chakravarthy U, Davis S, *et al*.

No. 28
Intravenous magnesium sulphate and sotalol for prevention of atrial fibrillation after coronary artery bypass surgery: a systematic review and economic evaluation.
By Shepherd J, Jones J, Frampton GK, Tanajewski L, Turner D, Price A.

No. 29
Absorbent products for urinary/faecal incontinence: a comparative evaluation of key product categories.
By Fader M, Cottenden A, Getliffe K, Gage H, Clarke-O'Neill S, Jamieson K, *et al*.

No. 30
A systematic review of repetitive functional task practice with modelling of resource use, costs and effectiveness.
By French B, Leathley M, Sutton C, McAdam J, Thomas L, Forster A, *et al*.

No. 31
The effectiveness and cost-effectivness of minimal access surgery amongst people with gastro-oesophageal reflux disease – a UK collaborative study. The REFLUX trial.
By Grant A, Wileman S, Ramsay C, Bojke L, Epstein D, Sculpher M, *et al*.

No. 32
Time to full publication of studies of anti-cancer medicines for breast cancer and the potential for publication bias: a short systematic review.
By Takeda A, Loveman E, Harris P, Hartwell D, Welch K.

No. 33
Performance of screening tests for child physical abuse in accident and emergency departments.
By Woodman J, Pitt M, Wentz R, Taylor B, Hodes D, Gilbert RE.

No. 34
Curative catheter ablation in atrial fibrillation and typical atrial flutter: systematic review and economic evaluation.
By Rodgers M, McKenna C, Palmer S, Chambers D, Van Hout S, Golder S, *et al*.

No. 35
Systematic review and economic modelling of effectiveness and cost utility of surgical treatments for men with benign prostatic enlargement.
By Lourenco T, Armstrong N, N'Dow J, Nabi G, Deverill M, Pickard R, *et al*.

No. 36
Immunoprophylaxis against respiratory syncytial virus (RSV) with palivizumab in children: a systematic review and economic evaluation.
By Wang D, Cummins C, Bayliss S, Sandercock J, Burls A.

Volume 13, 2009

No. 1
Deferasirox for the treatment of iron overload associated with regular blood transfusions (transfusional haemosiderosis) in patients suffering with chronic anaemia: a systematic review and economic evaluation.
By McLeod C, Fleeman N, Kirkham J, Bagust A, Boland A, Chu P, *et al*.

No. 2
Thrombophilia testing in people with venous thromboembolism: systematic review and cost-effectiveness analysis.
By Simpson EL, Stevenson MD, Rawdin A, Papaioannou D.

No. 3
Surgical procedures and non-surgical devices for the management of non-apnoeic snoring: a systematic review of clinical effects and associated treatment costs.
By Main C, Liu Z, Welch K, Weiner G, Quentin Jones S, Stein K.

No. 4
Continuous positive airway pressure devices for the treatment of obstructive sleep apnoea–hypopnoea syndrome: a systematic review and economic analysis.
By McDaid C, Griffin S, Weatherly H, Durée K, van der Burgt M, van Hout S, Akers J, *et al*.

No. 5
Use of classical and novel biomarkers as prognostic risk factors for localised prostate cancer: a systematic review.
By Sutcliffe P, Hummel S, Simpson E, Young T, Rees A, Wilkinson A, *et al*.

No. 6
The harmful health effects of recreational ecstasy: a systematic review of observational evidence.
By Rogers G, Elston J, Garside R, Roome C, Taylor R, Younger P, *et al*.

No. 7
Systematic review of the clinical effectiveness and cost-effectiveness of oesophageal Doppler monitoring in critically ill and high-risk surgical patients.
By Mowatt G, Houston G, Hernández R, de Verteuil R, Fraser C, Cuthbertson B, *et al*.

No. 8
The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports.
By Taylor RS, Elston J.

No. 9
Controlling Hypertension and Hypotension Immediately Post Stroke (CHHIPS) – a randomised controlled trial.
By Potter J, Mistri A, Brodie F, Chernova J, Wilson E, Jagger C, *et al*.

No. 10
Routine antenatal anti-D prophylaxis for RhD-negative women: a systematic review and economic evaluation.
By Pilgrim H, Lloyd-Jones M, Rees A.

No. 11
Amantadine, oseltamivir and zanamivir for the prophylaxis of influenza (including a review of existing guidance no. 67): a systematic review and economic evaluation.
By Tappenden P, Jackson R, Cooper K, Rees A, Simpson E, Read R, *et al*.

Health Technology Assessment Programme

Prioritisation Strategy Group

Members

HTA Commissioning Board

Members

Observers

Diagnostic Technologies & Screening Panel

Members

Chair,
Professor Paul Glasziou,
Professor of Evidence-Based Medicine, University of Oxford

Deputy Chair,
Dr David Elliman,
Consultant Paediatrician and Honorary Senior Lecturer, Great Ormond Street Hospital, London

Professor Judith E Adams, Consultant Radiologist, Manchester Royal Infirmary, Central Manchester & Manchester Children's University Hospitals NHS Trust, and Professor of Diagnostic Radiology, Imaging Science and Biomedical Engineering, Cancer & Imaging Sciences, University of Manchester

Ms Jane Bates, Consultant Ultrasound Practitioner, Ultrasound Department, Leeds Teaching Hospital NHS Trust

Dr Stephanie Dancer, Consultant Microbiologist, Hairmyres Hospital, East Kilbride

Professor Glyn Elwyn, Primary Medical Care Research Group, Swansea Clinical School, University of Wales

Dr Ron Gray, Consultant Clinical Epidemiologist, Department of Public Health, University of Oxford

Professor Paul D Griffiths, Professor of Radiology, University of Sheffield

Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford

Dr Susanne M Ludgate, Medical Director, Medicines & Healthcare Products Regulatory Agency, London

Dr Anne Mackie, Director of Programmes, UK National Screening Committee

Dr Michael Millar, Consultant Senior Lecturer in Microbiology, Barts and The London NHS Trust, Royal London Hospital

Mr Stephen Pilling, Director, Centre for Outcomes, Research & Effectiveness, Joint Director, National Collaborating Centre for Mental Health, University College London

Mrs Una Rennard, Service User Representative

Dr Phil Shackley, Senior Lecturer in Health Economics, School of Population and Health Sciences, University of Newcastle upon Tyne

Dr W Stuart A Smellie, Consultant in Chemical Pathology, Bishop Auckland General Hospital

Dr Nicholas Summerton, Consultant Clinical and Public Health Advisor, NICE

Ms Dawn Talbot, Service User Representative

Dr Graham Taylor, Scientific Advisor, Regional DNA Laboratory, St James's University Hospital, Leeds

Professor Lindsay Wilson Turnbull, Scientific Director of the Centre for Magnetic Resonance Investigations and YCR Professor of Radiology, Hull Royal Infirmary

Observers

Dr Tim Elliott, Team Leader, Cancer Screening, Department of Health

Dr Catherine Moody, Programme Manager, Neuroscience and Mental Health Board

Dr Ursula Wells, Principal Research Officer, Department of Health

Pharmaceuticals Panel

Members

Chair,
Professor Robin Ferner,
Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Deputy Chair,
Professor Imti Choonara,
Professor in Child Health, University of Nottingham

Mrs Nicola Carey, Senior Research Fellow, School of Health and Social Care, The University of Reading

Mr John Chapman, Service User Representative

Dr Peter Elton, Director of Public Health, Bury Primary Care Trust

Dr Ben Goldacre, Research Fellow, Division of Psychological Medicine and Psychiatry, King's College London

Mrs Barbara Greggains, Service User Representative

Dr Bill Gutteridge, Medical Adviser, London Strategic Health Authority

Dr Dyfrig Hughes, Reader in Pharmacoeconomics and Deputy Director, Centre for Economics and Policy in Health, IMSCaR, Bangor University

Professor Jonathan Ledermann, Professor of Medical Oncology and Director of the Cancer Research UK and University College London Cancer Trials Centre

Dr Yoon K Loke, Senior Lecturer in Clinical Pharmacology, University of East Anglia

Professor Femi Oyebode, Consultant Psychiatrist and Head of Department, University of Birmingham

Dr Andrew Prentice, Senior Lecturer and Consultant Obstetrician and Gynaecologist, The Rosie Hospital, University of Cambridge

Dr Martin Shelly, General Practitioner, Leeds, and Associate Director, NHS Clinical Governance Support Team, Leicester

Dr Gillian Shepherd, Director, Health and Clinical Excellence, Merck Serono Ltd

Mrs Katrina Simister, Assistant Director New Medicines, National Prescribing Centre, Liverpool

Mr David Symes, Service User Representative

Dr Lesley Wise, Unit Manager, Pharmacoepidemiology Research Unit, VRMM, Medicines & Healthcare Products Regulatory Agency

Observers

Ms Kay Pattison, Section Head, NHS R&D Programme, Department of Health

Mr Simon Reeve, Head of Clinical and Cost-Effectiveness, Medicines, Pharmacy and Industry Group, Department of Health

Dr Heike Weber, Programme Manager, Medical Research Council

Dr Ursula Wells, Principal Research Officer, Department of Health

Therapeutic Procedures Panel

Members

Chair,
Dr John C Pounsford,
Consultant Physician, North Bristol NHS Trust

Deputy Chair,
Professor Scott Weich,
Professor of Psychiatry, Division of Health in the Community, University of Warwick, Coventry

Professor Jane Barlow,
Professor of Public Health in the Early Years, Health Sciences Research Institute, Warwick Medical School, Coventry

Ms Maree Barnett,
Acting Branch Head of Vascular Programme, Department of Health

Mrs Val Carlill,
Service User Representative

Mrs Anthea De Barton-Watson,
Service User Representative

Mr Mark Emberton,
Senior Lecturer in Oncological Urology, Institute of Urology, University College Hospital, London

Professor Steve Goodacre,
Professor of Emergency Medicine, University of Sheffield

Professor Christopher Griffiths,
Professor of Primary Care, Barts and The London School of Medicine and Dentistry

Mr Paul Hilton,
Consultant Gynaecologist and Urogynaecologist, Royal Victoria Infirmary, Newcastle upon Tyne

Professor Nicholas James,
Professor of Clinical Oncology, University of Birmingham, and Consultant in Clinical Oncology, Queen Elizabeth Hospital

Dr Peter Martin,
Consultant Neurologist, Addenbrooke's Hospital, Cambridge

Dr Kate Radford,
Senior Lecturer (Research), Clinical Practice Research Unit, University of Central Lancashire, Preston

Mr Jim Reece
Service User Representative

Dr Karen Roberts,
Nurse Consultant, Dunston Hill Hospital Cottages

Observers

Dr Phillip Leech,
Principal Medical Officer for Primary Care, Department of Health

Ms Kay Pattison,
Section Head, NHS R&D Programme, Department of Health

Dr Morven Roberts,
Clinical Trials Manager, Medical Research Council

Professor Tom Walley,
Director, NIHR HTA Programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Department of Health

Disease Prevention Panel

Members

Chair,
Dr Edmund Jessop,
Medical Adviser, National Specialist, National Commissioning Group (NCG), London

Deputy Chair,
Dr David Pencheon,
Director, NHS Sustainable Development Unit, Cambridge

Dr Elizabeth Fellow-Smith,
Medical Director, West London Mental Health Trust, Middlesex

Dr John Jackson,
General Practitioner, Parkway Medical Centre, Newcastle upon Tyne

Professor Mike Kelly,
Director, Centre for Public Health Excellence, NICE, London

Dr Chris McCall,
General Practitioner, The Hadleigh Practice, Corfe Mullen, Dorset

Ms Jeanett Martin,
Director of Nursing, BarnDoc Limited, Lewisham Primary Care Trust

Dr Julie Mytton,
Locum Consultant in Public Health Medicine, Bristol Primary Care Trust

Miss Nicky Mullany,
Service User Representative

Professor Ian Roberts,
Professor of Epidemiology and Public Health, London School of Hygiene & Tropical Medicine

Professor Ken Stein,
Senior Clinical Lecturer in Public Health, University of Exeter

Dr Kieran Sweeney,
Honorary Clinical Senior Lecturer, Peninsula College of Medicine and Dentistry, Universities of Exeter and Plymouth

Professor Carol Tannahill,
Glasgow Centre for Population Health

Professor Margaret Thorogood,
Professor of Epidemiology, University of Warwick Medical School, Coventry

Observers

Ms Christine McGuire,
Research & Development, Department of Health

Dr Caroline Stone,
Programme Manager, Medical Research Council

Expert Advisory Network

Members

Professor Douglas Altman, Professor of Statistics in Medicine, Centre for Statistics in Medicine, University of Oxford

Professor John Bond, Professor of Social Gerontology & Health Services Research, University of Newcastle upon Tyne

Professor Andrew Bradbury, Professor of Vascular Surgery, Solihull Hospital, Birmingham

Mr Shaun Brogan, Chief Executive, Ridgeway Primary Care Group, Aylesbury

Mrs Stella Burnside OBE, Chief Executive, Regulation and Improvement Authority, Belfast

Ms Tracy Bury, Project Manager, World Confederation for Physical Therapy, London

Professor Iain T Cameron, Professor of Obstetrics and Gynaecology and Head of the School of Medicine, University of Southampton

Dr Christine Clark, Medical Writer and Consultant Pharmacist, Rossendale

Professor Collette Clifford, Professor of Nursing and Head of Research, The Medical School, University of Birmingham

Professor Barry Cookson, Director, Laboratory of Hospital Infection, Public Health Laboratory Service, London

Dr Carl Counsell, Clinical Senior Lecturer in Neurology, University of Aberdeen

Professor Howard Cuckle, Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Dr Katherine Darton, Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux, Professor of Paediatric Epidemiology, Institute of Child Health, London

Mr John Dunning, Consultant Cardiothoracic Surgeon, Papworth Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw, Consultant Vascular Surgeon, Gloucestershire Royal Hospital, Gloucester

Professor Martin Eccles, Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby, Dean of Faculty of Medicine, Institute of General Practice and Primary Care, University of Sheffield

Professor Gene Feder, Professor of Primary Care Research & Development, Centre for Health Sciences, Barts and The London School of Medicine and Dentistry

Mr Leonard R Fenwick, Chief Executive, Freeman Hospital, Newcastle upon Tyne

Mrs Gillian Fletcher, Antenatal Teacher and Tutor and President, National Childbirth Trust, Henfield

Professor Jayne Franklyn, Professor of Medicine, University of Birmingham

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Professor Fiona Gilbert, Consultant Radiologist and NCRN Member, University of Aberdeen

Professor Paul Gregg, Professor of Orthopaedic Surgical Science, South Tees Hospital NHS Trust

Bec Hanley, Co-director, TwoCan Associates, West Sussex

Dr Maryann L Hardy, Senior Lecturer, University of Bradford

Mrs Sharon Hart, Healthcare Management Consultant, Reading

Professor Robert E Hawkins, CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor Richard Hobbs, Head of Department of Primary Care & General Practice, University of Birmingham

Professor Alan Horwich, Dean and Section Chairman, The Institute of Cancer Research, London

Professor Allen Hutchinson, Director of Public Health and Deputy Dean of ScHARR, University of Sheffield

Professor Peter Jones, Professor of Psychiatry, University of Cambridge, Cambridge

Professor Stan Kaye, Cancer Research UK Professor of Medical Oncology, Royal Marsden Hospital and Institute of Cancer Research, Surrey

Dr Duncan Keeley, General Practitioner (Dr Burch & Ptnrs), The Health Centre, Thame

Dr Donna Lamping, Research Degrees Programme Director and Reader in Psychology, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London

Mr George Levvy, Chief Executive, Motor Neurone Disease Association, Northampton

Professor James Lindesay, Professor of Psychiatry for the Elderly, University of Leicester

Professor Julian Little, Professor of Human Genome Epidemiology, University of Ottawa

Professor Alistaire McGuire, Professor of Health Economics, London School of Economics

Professor Rajan Madhok, Medical Director and Director of Public Health, Directorate of Clinical Strategy & Public Health, North & East Yorkshire & Northern Lincolnshire Health Authority, York

Professor Alexander Markham, Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Peter Moore, Freelance Science Writer, Ashtead

Dr Andrew Mortimore, Public Health Director, Southampton City Primary Care Trust

Dr Sue Moss, Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton

Professor Miranda Mugford, Professor of Health Economics and Group Co-ordinator, University of East Anglia

Professor Jim Neilson, Head of School of Reproductive & Developmental Medicine and Professor of Obstetrics and Gynaecology, University of Liverpool

Mrs Julietta Patnick, National Co-ordinator, NHS Cancer Screening Programmes, Sheffield

Professor Robert Peveler, Professor of Liaison Psychiatry, Royal South Hants Hospital, Southampton

Professor Chris Price, Director of Clinical Research, Bayer Diagnostics Europe, Stoke Poges

Professor William Rosenberg, Professor of Hepatology and Consultant Physician, University of Southampton

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh

Dr Susan Schonfield, Consultant in Public Health, Hillingdon Primary Care Trust, Middlesex

Dr Eamonn Sheridan, Consultant in Clinical Genetics, St James's University Hospital, Leeds

Dr Margaret Somerville, Director of Public Health Learning, Peninsula Medical School, University of Plymouth

Professor Sarah Stewart-Brown, Professor of Public Health, Division of Health in the Community, University of Warwick, Coventry

Professor Ala Szczepura, Professor of Health Service Research, Centre for Health Services Studies, University of Warwick, Coventry

Mrs Joan Webster, Consumer Member, Southern Derbyshire Community Health Council

Professor Martin Whittle, Clinical Co-director, National Co-ordinating Centre for Women's and Children's Health, Lymington

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (www.hta.ac.uk) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

The NIHR Coordinating Centre for Health Technology Assessment
Alpha House, Enterprise Road
Southampton Science Park
Chilworth
Southampton SO16 7NS, UK
Email: hta@hta.ac.uk
www.hta.ac.uk

ISSN 1366-5278