JCN Journal of Clinical Nursing

Journal of Clinical Nursing

ORIGINAL ARTICLE

Item response theory: How Mokken scaling can be used in clinical practice

Roger Watson, L Andries van der Ark, Li-Chan Lin, Robert Fieo, Ian J Deary and Rob R Meijer

Aims. To demonstrate the principles and application of Mokken scaling.

Background. The history and development of Mokken scaling is described, some examples of applications are given, and some recent development of the method are summarised.

Design. Secondary analysis of data obtained by cross-sectional survey methods, including self-report and observation.

Methods. Data from the Edinburgh Feeding Evaluation in Dementia scale and the Townsend Functional Ability Scale were analysed using the Mokken scaling procedure within the 'R' statistical package. Specifically, invariant item ordering (the extent to which the order of the items in terms of difficulty was the same for all respondents whatever their total scale score) was studied.

Results. The Edinburgh Feeding Evaluation in Dementia scale and the Townsend Functional Ability Scale showed no violations of invariant item ordering, although only the Townsend Functional Ability Scale showed a medium accuracy.

Conclusion. Mokken scaling is an established method for item response theory analysis with wide application in the social sciences. It provides psychometricians with an additional tool in the development of questionnaires and in the study of individuals and their responses to latent traits. Specifically, with regard to the analyses conducted in this study, the Edinburgh Feeding Evaluation in Dementia scale requires further development and study across different levels of severity of dementia and feeding difficulty.

Relevance to clinical practice. Good scales are required for assessment in clinical practice and the present paper shows how a relatively recently developed method for analysing Mokken scales can contribute to this. The two scales used as examples for analysis are highly clinically relevant.

Key words: activities of daily living, dementia, item response theory, Mokken scaling, nurses, nursing, psychometrics, scalability

Accepted for publication: 12 July 2011

Introduction

The use of questionnaires is important and widespread in nursing research and practice. A well-designed questionnaire is a valuable instrument for the measurement of phenomena such as psychological morbidity, quality of life and clinical symptoms. Also, in nursing and other areas of social research, the measurement of attitudes,

Authors: *Roger Watson*, PhD, Professor of Nursing, School of Nursing & Midwifery, The University of Sheffield, Sheffield, UK and School of Nursing & Midwifery, University of Western Sydney, Campbelltown, Australia; *L Andries van der Ark*, PhD, Associate Professor, Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands; *Li-Chan Lin*, PhD, Professor, School of Nursing, National Yang-Ming University, Taipei, Taiwan; *Robert Fieo*, PhD, Research Scientist, Columbia University, Division of Geriatric Psychiatry, New York, NY, USA; *Ian J Deary*, PhD, Professor of Differential Psychology, Department of Psychology, Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK; *Rob R Meijer*, PhD, Professor of Psychometric and Statistical Techniques, Department of Psychometrics and Statistics, University of Groningen, Groningen, The Netherlands

Correspondence: Roger Watson, Professor of Nursing, School of Nursing & Midwifery, The University of Sheffield, Sheffield S7 5AU, UK.

E-mail: roger.watson@sheffield.ac.uk

doi: 10.1111/j.1365-2702.2011.03893.x

opinions and educational achievement is common and valuable.

The development of questionnaires requires some obligatory steps including the clarification of the concept being studied, the selection of items, validation of content, establishing the reliability of the items, and then further steps to investigate the construct validity of the questionnaire (Bannigan & Watson 2009). This field of research is known as psychometrics (Rust & Golombock 1999) and depends heavily on methods developed in psychology which are equally applicable across the different fields where questionnaires are used. The methods employed to ensure that questionnaires are psychometrically sound range from common sense (selection of items and validation of content) to some sophisticated mathematical and statistical methods for establishing reliability and validity.

Classical test theory

The methods used to establish reliability and validity rely heavily on what is referred to as classical test theory, which includes methods such as Cronbach's alpha for the estimation of reliability of a test score. Classical test theory – which will not be expounded on further here – is concerned with the estimation of measurement error and establishing, within the bounds of the methods available, an estimate of the true score.

Item response theory

A more recently developed and very powerful alternative to classical test theory is item response theory (Hulin *et al.* 1983) which seeks to solve much the same problems as classical test theory and, indeed, can often be complementary to classical test theory in terms of identifying sets of items that measure the same concept (i.e. that are unidimensional) (van Schuur 2003). This is exemplified in the case of the EdFED (Edinburgh Feeding Evaluation in Dementia) scale (Watson & Deary 1994, 1997, Watson 1996) where both exploratory factor analysis and Mokken scaling identified the same set of items to measure feeding difficulty in older people with dementia. Item response theory is less concerned with scores on sets of items (test level scores) and more concerned with the responses to individual items. Additionally, item response theory enhances the interpretive power by providing measurement precision that varies with a person's ability level (Hambleton *et al.* 1991). The discrepancies between observed and true scores indicate how much the test score differs from the true scores, and are summarised by the standard error of measurement; better reliability estimates result from high precision or relatively small measurement error (Alagumalai & Curtis 2005).This information (i.e. error that varies by person performance) can be used to identify the most sensitive part of the instrument or scale under investigation (Wilson 2005). The range of item response theories will not be covered here but the method that has most recently come to characterise and be described as an example of item response theory (although it differs from some other methods) is Rasch modelling (Meijer & Sijtsma 1990).

However, a convenient, useful and relatively easy to understand method which follows the principles of item response theory has also been developed and this is known as Mokken scaling after Robert Mokken (http://staff.science.uva.nl/~mokken/; retrieved 30 September 2010) who initially developed it. The remainder of this paper will explain the principles of Mokken scaling and provide some recent examples of its application.

Background

Mokken scaling is a non-parametric method for investigating the relationship between items and latent traits which evolved from the Guttman method of investigating hierarchies of items within scales.

Guttman scaling

Guttman scaling, developed by Louis Guttman (Stouffer *et al.* 1950, Katz 1988) is deterministic in the sense that it does not take any stochastic elements into account. This means that it does not view the relationship between an item and the latent trait that is being measured in terms of probability (as in the probability of a score on the item being related to the extent to which the latent trait is present); rather, it sees the relationship as being one of complete discrimination between the presence or absence of the latent trait based on the endorsement or lack of endorsement of an item or an item step response (i.e. the response to any one of the points on, for example, a Likert scale as opposed to a response to a dichotomous item – we will only consider dichotomous items here in relation to Guttman scales for clarity and will return to polytomous items under Mokken scaling). This relationship is shown in Fig. 1.

The way Guttman scaling works can be demonstrated by considering the questions listed in Table 1. The idea behind these questions – which, for the purposes of this example are deliberately ordered as they appear in the table – is to measure the attitude of individuals towards the supporters of an opposing football team. The questionnaire is based on the fact that most people will not openly express any particular

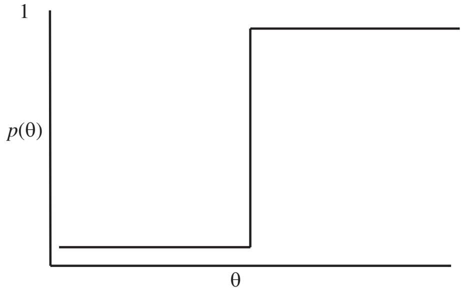

Figure 1 An example of an item behaving according to the deterministic Guttman model along a latent trait θ on the abscissa with the probability of a positive response to the items on the ordinate.

Table 1 Questions used to demonstrate hierarchical item ordering

Item	Label	Response
1	I like supporters of opposing football teams	Yes/no
2	I would sit next to an opposing supporter on a bus	Yes/no
3	I would speak to a supporter of an opposing team	Yes/no
4	I would invite an opposing supporter to my house	Yes/no
5	I would allow an opposing supporter to marry one of my children	Yes/no

prejudice towards other people but, when pushed further; for example, inviting such a person to their house or welcoming them into their family, then they will eventually 'draw the line' somewhere. A set of responses to the questions is shown in Table 2 and this demonstrates that all the respondents, except one, show no general prejudice; however, as we progress through the questions, they become increasingly hard for these respondents to endorse with only one person endorsing all of the items. Of course, the response pattern shown here is perfect: people always endorse an 'easier' or more popular item before they endorse a 'harder' or less popular item – and things are rarely like that in reality and we are more likely to obtain a response pattern like the one shown in Table 3. In reality, answer patterns are subject to more than the underlying construct or trait; answer patterns are influenced by factors like mood of the respondent and interpretation of the question (Kempen *et al.* 1995). Therefore, the relationship to item responses and the construct is better defined as probabilistic rather than deterministic. Furthermore, to be sure that a scale follows the expectation of clear-cut pass/fail point (i.e. deterministic nature) for each person, the differences between item difficulties must be large (Fisher & Fisher 1993). Thus, the sensitivity of such scales to small changes in functioning within individuals over time or to small differences between individuals can be reduced quite substantially (Finch *et al.* 1994). Without explaining the Guttman method in detail, the method is concerned with calculating acceptable levels of reproducibility and scalability (Menzel 1953, Schuessler 1971) based on the number of violations of the underlying pattern. This involves identifying the incidences when an individual does not answer in a way that is consistent with the majority of the other respondents and comparing these incidences with the number of correct responses and calculating coefficients that tell us how well the data conform to a Guttman scale.

Mokken scaling

Guttman scaling is described as deterministic in that it does not allow for randomness: a score on the scale should closely indicate the responses to the items in the scale and items are written (or discarded) to achieve this. Mokken scaling refers to a series of methods for investigating whether stochastic

Table 2 Responses to questions in Table 1 showing perfect Guttman scalability

Item	Respondent 1	Respondent 2	Respondent 3	Respondent 4	Respondent 5
1	No	Yes	Yes	Yes	Yes
2	No	Yes	Yes	Yes	Yes
3	No	No	Yes	No	Yes
4	No	No	Yes	No	No
5	No	No	Yes	No	No

Table 3 Responses to questions in Table 1 showing violations of Guttman scalability*

Item	Respondent 1	Respondent 2	Respondent 3	Respondent 4	Respondent 5
1	No	Yes	Yes	Yes	Yes
2	No	No	Yes	Yes	No
3	No	**Yes**	Yes	No	**Yes**
4	**Yes**	No	Yes	No	No
5	No	No	Yes	No	No

*Violations of Guttman scalability are shown in bold.

versions of the Guttman scale hold in the data. These stochastic versions of a Guttman scale are known as the model of monotone homogeneity (MHM) and the double monotonicity model (DMM), which fall within the broad spectrum of methods described as item response theory (IRT). The scores on a scale, representing the presence of the latent trait, are related to the scores on individual items by probability (Sijtsma & Molenaar 2002); that is, an item is more likely to be endorsed if the person scores high in the latent trait; an individual is more likely to agree to an item indicating that they are anxious if they suffer from anxiety – but the model does allow for error and individual counter-intuitive responses to some items. Thus, if we represent the response to an item on a Guttman scale as shown in Fig. 1, we can see that the response to the item is either endorsement or not; that is 'yes' or 'no' to an item where agreement with a statement is required – this is a deterministic model. However, Fig. 2 shows how responses to an item are viewed in a stochastic model such as the MMH; as the amount of the latent trait increases, the probability of endorsing the item also increases and it does so in a characteristic fashion to be described below.

Item characteristic curve

The behaviour of individual items in a scale relative to the latent trait is described by item characteristic curves (ICCs) (Hambleton & Swaminathan 1985). If the Greek symbol theta (θ) represents the latent trait then the ICC for any particular item represents $P(\theta)$ – the probability of an individual's score on an item being obtained in the presence of a particular level of the latent trait. As the latent trait increases, then the probability of a score on the item increases in a non-linear fashion described by a probability distribution function as shown in Fig. 2. The shape of and relationship between ICCs is part of Mokken scaling analysis and this will be described below.

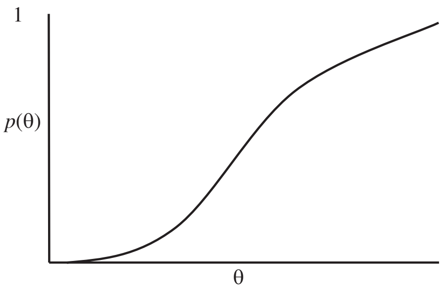

Figure 2 An item responding stochastically in the presence of latent trait θ on the abscissa with the probability of a positive response to the items on the ordinate.

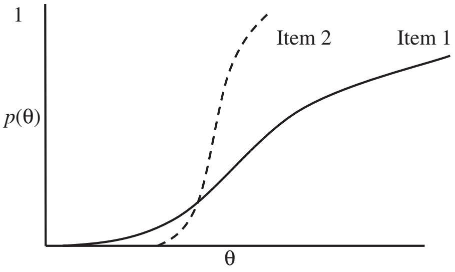

Figure 3 Two items showing different levels of difficulty and discrimination: item 1 is more difficult than item 2 and item 2 is more discriminating than item 1.

Discrimination and difficulty

ICCs, as shown in Fig. 3 can differ in steepness and those that are steeper are described as more discriminating than those that are less steep. Item 1 is more difficult than item 2 because for the majority of trait levels, the probability of endorsing item 2 is greater than the probability of endorsing item 1. In this sense 'difficulty' refers to the ease with which an item is endorsed by respondents. Ideally, in general trait measurement, what is required is highly discriminating items with varying levels of difficulty.

Assumptions of IRT

Before proceeding to describe Mokken scaling in more detail, three assumptions of IRT which are common across IRT methods should be explained. These assumptions are: unidimensionality, local independence, and monotonicity. IRT assumes that, for the items that form a scale, there is a dominant single latent trait that determines the answers to the items; this is known as unidimensionality (Hulin *et al.* 1983). This does not mean that, in a large set of items, more than one dimension may not exist; rather, that sets of items fitting an IRT model are unidimensional. This distinction is noteworthy if we emphasise that, despite meeting the assumption of unidiminsionality for this scale, minor abilities can still influence response patterns. 'It has been long argued that responses to a set of items are multiply determined, in that several minor abilities are required to respond to items' (Nandakumar 1994, p. 17). IRT also assumes local stochastic independence of items in a scale, which means that an individual's responses to items in a scale are dependent on the individual's level on the latent trait being measured: the response to one item is not influenced by the score on another (Nunnally 1978). It should be emphasised that this is, largely, an assumption as complete local stochastic independence is very hard to achieve, especially where items in a scale form an item chain where success on one item depends on success on the previous item or where items overlap (Balazs & De Boeck undated). Montonicity refers to the increasing probability of the score on the item increasing as the level if the latent trait increases; this is illustrated in Fig. 2.

The model of monotone homogeneity

The next important stage of the Mokken procedure is investigating if items fit the MMH (Mokken & Lewis 1982). The MMH is a very unrestrictive model, assuming only the three common assumptions in IRT: unidimensionality, local independence, and monotonicity. While unidimensionality and local independence cannot be visualised, Fig. 4 illustrates one item (item 1) which has monotonicity and one (item 2) which has not. MMH means that, for all items, as the score on the latent trait increases, the score on the items in the scale should also increase. The MMH allows for the ordering of persons on the latent trait by the sum of the item score – an essential requirement for a psychological test. The software for Mokken scale analysis (Molenaar & Sijtsma 2000) provides diagnostics which allow you to detect items which violate MMH and remove them from the analysis.

The double monotonicity model

The next stage of the Mokken scaling procedure is investigating whether the items fit the double monotonicity model (DMM) (Mokken & Lewis 1982). In addition to the three assumptions of the MMH, the DMM assumes that the ICCs do not intersect. This is shown in Fig. 5 where items 1 and 2 do not intersect and item 3 intersects items 1 and 2; therefore, item 3 violates the DMM. The DMM allows for the ordering of persons on the latent trait by the sum score and allows for an invariant ordering (IIO) of the items in terms of difficulty. The IIO property is crucial for establishing hierarchical scales. The order of the items in terms of difficulty should be the same for all respondents whatever their latent trait value (Sijtsma & Junker 1996). The software for Mokken scale analysis also provides diagnostics which allow you to detect items which violate the DMM and remove them from the analysis.

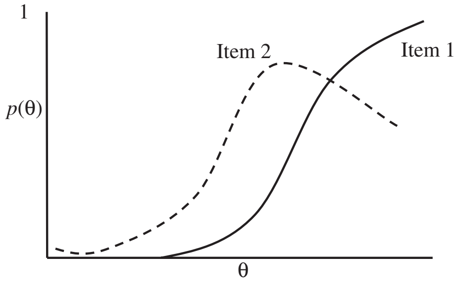

Figure 4 Two items with Item 1 showing MMH and Item 2 violating MMH.

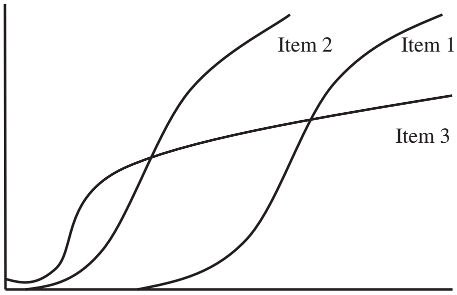

Figure 5 Three items with items 1 and 2 showing DMM and item 3 violating DMM.

Mokken scaling analysis

Mokken scaling is an improved version of Guttman scaling in the sense that it is stochastic and, thereby, less restrictive in nature: it does not assume that the responses of individuals to items in a scale are deterministic. Therefore, Mokken scaling produces a similar outcome to Guttman scaling but is likely to include more items from a pool of items into a hierarchical scale and also to provide more information about the behaviour of individual items. For scale developers, Mokken scaling is an additional tool for the identification of unidimensionsal sets of items and it provides additional information about the relationship between items in such scales.

Loevinger's coefficient (H)

The first parameter to be established in Mokken scaling is whether or not, based on their mean scores, the items in a questionnaire form a hierarchy. In Mokken scaling this is done by inspecting the item scores relative to one another to see how consistently they are ordered. This is precisely the same as Guttman scaling and it is low violations of Guttman ordering that lead to the initial inclusion of items in a Mokken scale. The principals of this can easily be demonstrated by looking at Table 4 which shows, in a 2 × 2 contingency table (based on Niemöller & van Schuur 1983), two items *i* and *j* and all possible patterns of endorsement. If item *i* has greater 'difficulty' than item *j* then responses in cells a, c and d are all acceptable, but responses in cell b are not, as these are clearly violations of the Guttman ordering and arise from respondents who have scored item *i* as being less difficult than item *j*. Cell b, therefore, is the error cell and the proportion of responses in cell b to those in the remaining cells gives an estimation of the scalability: the extent to which items form a hierarchy. This proportion is used to compute item-pair scalability coefficient H_{ij}: if there are no observations in error cell b, then H_{ij} equals 1 and items *i* and *j* form a perfect Guttman scale. As the number of observations in the error cell increase H_{ij} decreases. Additionally, there is an item

Table 4 Contingency table showing possible response categories to two items *i* and *j*

	Item *j*	
Item *i*	1	0
1	a (1, 1)	b (1, 0)
0	c (0, 1)	d (0, 0)

scalability coefficient H_i, indicating the scalability of an individual item, and a scalability coefficient H, indicating the scalability of the set of items. A set of items forms a so-called Mokken scale if two conditions are met: (1) for all item pairs, scalability coefficient H_{ij} is greater than zero and (2) scalability coefficient H_i is greater than some a priori chosen criterion c (c usually equals 0·3, but this is up to the researcher). In addition, the following guidelines for H are common: 0·30–0·40 is called a weak scale, 0·40–0·50 is called a medium scale, and >0·50 is called a strong scale Molenaar and Sijtsma (2000).

Polytomous items

Thus far we have been concerned with dichotomous items where the non-intersecting ICCs as implied by the DMM is equivalent to IIO. Mokken scaling has been developed to analyse polytomous (Sijtsma *et al.* 1990, Hemker & Sijtsma 1995) items. The principles remain the same, but the analysis is not simply of ICCs, but of the responses to each of the levels in the items (e.g. 1–5 on a Likert type scale). The resulting relationship between these responses and the score on the latent trait can be represented using item step characteristic curves or item step response functions (ISRFs). These are the responses of each of the steps in the scale. For example, in a Likert scale with five response categories, there are four steps between the five categories, so there are four ISRFs. ISRFs are central to the analysis of polytomous items using Mokken scaling. The procedure for establishing if a set of polytomous items forms a Mokken scale follows the same pattern as that for dichotomous items. Without a detailed explanation, the parameters of H, MHH and DMM should hold for ISRFs and diagnostics exist to establish this in versions of the Mokken scaling procedure that have been developed to process polytomous items.

An important difference between dichotomous items and polytomous items is that for polytomous items, the DMM does not automatically yield items that are invariantly ordered in terms of difficulty. Hence, for polytmous items, establishing the DMM is not enough for establishing IIO (Sijtsma *et al.* 2011). Recently, Ligtvoet *et al.* (2010) developed a method to investigate whether IIO holds. The method should be applied after the conventional Mokken scaling procedure and consists of two steps: first, based on statistical testing, items are detected (and removed if requested) that clearly violate IIO. Second, a diagnostic analogous to H, called Htrans or H^T is used to establish IIO; values of $H^T > 0{\cdot}3$ are considered to indicate a scale with IIO (Ligtvoet *et al.* 2010). The methods for investigating IIO will be described below. Likewise, establishing IIO is possible and the software for these analyses will be described below and some examples of the application of Mokken scaling will be provided.

Software for Mokken scaling

The Mokken scaling procedure (MSP) (Molenaar & Sijtsma 2000) is commercially available for Windows and this software is capable of analysing polytomous items and of investigating scalability, MMH and DMM for dichotomous and polytomous items. It is also capable of analyzing IIO for dichotomous items but not, in its most recent release, of analyzing IIO for polytomous items. However, public domain software called R contains Mokken scaling analysis (van der Ark 2007) and is capable of analysing IIO for polytomous items and generating H^T (van der Ark 2010). For specific details on how to run these software packages, manuals are available with step-by-step guidance. The outputs from MSP and R have some similarities and some differences. MSP is probably more 'user friendly' and produces outputs that are more directly interpretable and which can be imported relatively easily into publications. R, on the other hand, is more sophisticated in its analytical possibilities but, as this is public domain, the output options are less sophisticated. Mokken scaling is only one method for investigating IRT models; other IRT methods, such as Rasch modelling and the Generalized Partial Credit Model will not be described here. However, other software for analyzing items is available and TestGraf, for example, is also public domain software for item analysis (http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html; retrieved 1 October 2010) and can be a useful adjunct for investigating item properties.

Applications of Mokken scaling

Mokken scaling has been applied to a wide range of sets of items in psychology, social science, medicine and nursing; it has been applied to psychological questionnaires (Moorer & Suurmeijer 1994, Watson *et al.* 2007, 2008a,b, Bedford *et al.* 2010a,b, Deary *et al.* 2010, Stewart *et al.* 2010), attitude measurements (Gillespie 1988, Tenvegert *et al.* 1992), quality of life (Moorer *et al.* 2001, Ringdal *et al.* 1999, 2003, Thompson & Watson 2010), disability (Kempen *et al.* 1995, van Boxel *et al.* 1995), nursing (Watson 1996, Dijkstra *et al.*

1999, 2000, Lin *et al.* 2008), psychiatry (de Jong & Molenaar 1987, Meijer & Baneke 2001), pain (von Korff *et al.* 1992), sleep (Kingshott *et al.* 1998) and activities of daily living (Kempen & Suurmeijer 1990, 1991, Suurmeijer & Kempen 1990, Fieo *et al.* 2010) inventories. Specific examples of recent applications include the use of MSP to analyse sets of items in inventories of psychological distress such as the General Health Questionnaire (Watson *et al.* 2008) and the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) (Bedford *et al.* 2010).

Psychological morbidity

The application of MSP to inventories measuring psychological morbidity and personality disturbance show hierarchies of items ranging from relatively mild distress to more extreme forms of distress and these are often anchored at the most difficult end of the scale in items which indicate feelings of worthlessness or suicidal ideation. For example, Mokken scaling of one version the General Health Questionnaire (GHQ), the GHQ-30, shows a hierarchy of items running in terms of difficulty from being unable to face up to problems and being constantly under strain to feeling that life is not worth living (Watson *et al.* 2008a). Mokken scaling of the CORE-OM shows a hierarchy of items running in terms of difficulty from not being able to achieve things and being unable to ignore problems to making plans to end one's life (Bedford *et al.* 2010). Clearly, in these cases, the intervening items are also logically related to the anchoring items in the range.

ADL scales

An ADL (Activities of Daily Living; Townsend 1962) scale, the Townsend Functional Ability Scale (Townsend), has been analysed using MSP (Fieo *et al.* 2010) and this shows a hierarchy of items in terms of difficulty from tying a knot in a piece of string to cutting one's own toe nails indicating that, as ADL function is increasingly impaired, it is fine movements that are affected. The above inventories are self-administered, but MSP has been applied to observational data and one example of this is the development of the EdFED scale (Watson 1994 a,b, 1997, Watson & Deary 1996, 1997, Watson *et al.* 2001, 2002). The EdFED was developed from a set of items designed to measure feeding difficulty in older people with dementia. Early versions of the EdFED were analysed using factor analysis and Guttman scaling and it was shown that there were distinct underlying dimensions to the EdFED and that one of these was related solely to the behavioural aspects of feeding displayed by the older people with dementia. The factor analysis and the Guttman scaling confirmed this and the later application of Mokken scaling further confirmed that this was the case. Subsequent analysis of separate datasets have confirmed both the factor structure of the EdFED and the hierarchical ordering of the behavioural items and this holds for a recent dataset obtained from Taiwan using a translated version of the EdFED scale (Lin *et al.* 2008). The EdFED items form a hierarchy, in terms of difficulty, from a general refusal to eat, up to being unable to swallow and allowing food to fall out of the mouth.

The present study

Therefore, the psychometric properties of the EdFED and Townsend scales have been intensively studied, including the use if IRT. Both demonstrate four of the essential features of Mokken scales but neither – in common with many published Mokken scales with polytomous items – has been investigated for IIO, which can only be implied from the above parameters. We considered that these two short scales were good candidates for a study of IIO that would complement the existing information on these scales, complete the present paper, and illustrate the principles and application of IIO. With the advent of the R programme containing the Mokken scaling procedure and, within that, the ability to check for IIO in polytomous items, it is now possible to analyse the Townsend and the EdfED for IIO.

Methods

The EdFED database came from a study by Lin *et al.* (2008) and comprised data from 477 older people with dementia (mean age 79·5 [SD 9·62] years; 67·7% male) in nine special licensed long-term care facilities in Taiwan. The Townsend database came from a study by Fieo *et al.* (2010) and comprised data from 548 older people aged 79 and 42·3% male. The analysis here is concerned only with items that showed Mokken scaling properties in the previous studies; i.e. six of 11 items in the case of the EdFED and seven of nine items in the case of the Townsend. Data from both sets of items were imported into the R package by converting SPSS databases to *.Rdata files and, using the default settings, checked for violations of IIO and coefficient H^T by running the 'check.iio' procedure for both sets of data.

Results

The results are shown in Table 5. For both scales there are no significant violations of IIO for either set of items; therefore, no items required removal and only one step was required.

Table 5 Violations of IIO and coefficient H^T for the EdFED and the TFAS

EdFED	Step	TFAS	Step
Items*	1	Items*	1
Refusing to eat	0	Tie knot	0
Refusing to open mouth	0	Wash or bathe	0
Turning head away	0	Reach overhead	0
Spitting	0	Get on bus	0
Refusing to swallow	0	Up and down stairs	0
Leaving mouth open	0	Cut toe nails	0
Coefficient H^T	0·27		0·48

*Items are abbreviated.

The EdFED items show H^T of 0·27 and those from the Townsend show H^T of 0·48. The retention of all items suggests that there is evidence of IIO in both sets of items but it is stronger for the Townsend items. According to Ligtvoet *et al.* (2010), H^T may be interpreted in the same way as H meaning that $H^T < 0{\cdot}3$ indicates that item ordering is too inaccurate to be useful and $0{\cdot}4 \le H^T < 0{\cdot}5$ means that accuracy of item ordering is medium.

Discussion

This paper has described the development of Mokken scaling analysis from its roots in the non-stochastic, deterministic scaling method of Guttman through its evolution as a method for the analysis of dichotomous items and polytomous items. The basic precepts of IRT: MMH (unidmensionality, local stochastic independence), DMM and IIO have been considered and how these apply to Mokken scaling and how these compare and contrast with classical test theory. The computational software, principally MSP and R, available for Mokken scaling analysis have been described along with some applications of Mokken scaling in different fields of study in the social sciences.

However, recent exchanges between scaling practitioners (Watson & Deary 2009, Bedford *et al.* 2010, Meijer 2010) have revealed some misunderstanding of the extent to which IIO is both understood and can be implied from the Mokken scaling analysis procedures and software available for polytomous items. In addition, recent additions to the literature (Ligtvoet *et al.* 2010, Sijtsma *et al.* 2011) have demonstrated that the concept of IIO, as it applies to polytomous items, is now much better understood and how, with the advent of new computational software, it can be addressed in sets of polytomous items. Therefore, the original contribution of this paper is to re-visit two existing databases which have previously been analysed using the MSP and where items have been shown to be unidmensional and hierarchical and where ISRFs have been shown to adhere to the MMH and the DMM. The unique contribution of this paper has been to investigate these sets of items for IIO.

IIO was shown to be a safe assumption for items from the Townsend but not for the EdFED, despite there being no violations of IIO in either set of items. This is good news for users of the Townsend and little more needs to be said here – the items extracted by the MSP are hierarchical and do not overlap in the sample of individuals tested. However, what about the EdFED scale? The uniqueness and utility of the EdFED scale have been highlighted (Amella 2002, Amella & Stockdell 2008, Chang & Roberts 2008, Aselage 2010, Aselage & Amella 2010) and it has been recommended by the The Hartford Institute for Geriatric Nursing, New York University, College of Nursing, and the Alzheimer's Association in the USA (http://consultgerirn.org/uploads/File/trythis/try_this_d11_1.pdf; retrieved 10 October 2010). The EdFED shows good psychometric properties (Watson *et al.* 2001, 2002), is responsive to different levels of dementia and has been used as an outcome measure in recent studies of feeding difficulty (Lin *et al.* 2010a,b). Clearly, the EdFED is a useful instrument but the indications here, based on the work of Ligtvoet *et al.* (2010), are that the items lack either discrimination or are located very close on the latent trait of feeding difficulty. Analysis using TestGraf (unpublished) suggests that ICCs are relatively steep and, therefore, discriminating. It is likely, therefore, that the explanation for inaccurate IIO is closeness of items on the latent trait of feeding difficulty in dementia.

Conclusion

Mokken scaling is a useful way of investigating the behaviour of items in scales in response to varying levels of a latent trait. Mokken scaling has recently been enhanced by the introduction of software to enable analysis of IIO. The Townsend, a scale to measure ADL, shows good IIO and the EdFED, a scale to measure feeding difficulty in older people with dementia, does not. Further development of the EdFED scale is required to increase its utility. Possibly, additional items are required and its properties need to be studied further across a range of severity of dementia.

Relevance to clinical practice

Generally, in clinical practice, we require good measurement scales with good psychometric properties including IIO. Sequential functional loss scales are more informative than those simply summing functional loss, with predictive value to the clinician monitoring, for example, an older patient: if the sequence is out of order or accelerated, the need for interventions may be indicated (Daltroy *et al.* 1992). Additionally, examination of the sequence of loss may help characterise adaptations to impairment and differences among subgroups. The Townsend shows good IIO and is likely to be useful in its present form in clinical practice. The EdFED remains, demonstrably, a useful clinical tool but its applicability across a wide range of levels of feeding difficulty and dementia remains problematic. In addition, clearly, replication of the EdFED study is necessary.

Acknowledgements

This study was unfunded.

Contributions

Study design: RW, RF, L-CL; data collection and analysis: RF, L-CL, RW, IJD and manuscript preparation: RW, LAvA, RF, RM.

Conflict of interest

None.

References

Alagumalai S & Curtis D (2005) Classical test theory. In *Applied Rasch Measurement: A Book of Exemplars* (Alagumalai S, Curtis D & Hungi eds). Springer, Dordrecht, pp. 1–14.

Amella EJ (2002) Resistance at mealtimes for persons with dementia. *The Journal of Nutrition, Health & Aging* **6**, 117–22.

Amella EJ & Stockdell R (2008) Edinburgh Feeding in Dementia Scale: Developing a plan of care for mealtimes. *American Journal of Nursing* **108**, 46–54.

van der Ark LA (2007) Mokken scale analysis in R. *Journal of Statistical Software* **20**, 1–19.

van der Ark LA (2010) Getting started with Mokken scale analysis in R. http://spitswww.uvt.nl/~avdrark/research/mokkenstart.pdf (retrieved 1 October 2010).

Aselage M (2010) Measuring mealtime difficulties: eating, feeding and meal behaviours in older adults with dementia. *Journal of Clinical Nursing* **19**, 621–631.

Aselage M & Amella EJ (2010) An evolutionary analysis of mealtime difficulties in older adults with dementia. *Journal of Clinical Nursing* **19**, 33–41.

Balazs K & de Boeck P (undated) Detecting local item dependence stemming for minor dimensions. Technical Report 0684 IAP Statistics Network, Interuniversity Attraction Pole. http://www.stat.ucl.ac.be/Iapdp/tr2006/TR0684.pdf (retrieved 30 September 2010).

Bannigan K & Watson R (2009) Reliability and validity in a nutshell. *Journal of Clinical Nursing* **18**, 3237–3243.

Bedford A, Watson R, Lyne J, Tibbles J, Davies F & Deary IJ (2010a) Mokken scaling and principal components analysis of the CORE-OM in a large clinical sample. *Clinical Psychology and Psychotherapy* **17**, 51–56.

Bedford A, Watson R, Henry JD, Crawford JR & Deary IJ (2010b) Mokken scaling analyses of the Personal Disturbance Scale (DSSI/sAD) in large clinical and non-clinical samples. *Personality and Individual Differences* **50**, 38–42.

van Boxel Y, Roest FHJ, Bergen MP & Stam HJ (1995) Dimensionality and hierarchical structure of disability measurement. *Archives of Physical Medicine and Rehabilitation* **76**, 1152–1155.

Chang C-C & Roberts BL (2008) Feeding difficulty in older people with dementia. *Journal of Clinical Nursing* **17**, 2266–2274.

Daltroy LH, Logigian M, Iversen MD & Liang MH (1992) Does musculoskeletal function deteriorate in a predictable sequence in the elderly? *Arthritis Care Research* **5**, 146–150.

Deary IJ, Wilson JA, Carding PN, MacKenzie K & Watson R (2010) From dysphonia to dysphoria: Mokken scaling shows a strong, reliable hierarchy of voice symptoms in the Voice Symptom Scale questionnaire. *Journal of Psychosomatic Research* **68**, 67–71.

Dijkstra A, Buist G, Moorer P & Dassen T (1999) Construct validity of the Nursing Care Dependency Scale. *Journal of Clinical Nursing* **8**, 380–388.

Dijkstra A, Brown L, Havens B, Romeron TI, Zanotti R, Dassen T & van den Heuvel W (2000) An international psychometric testing of the care dependency scale. *Journal of Advanced Nursing* **31**, 944–952.

Fieo R, Watson R, Deary IJ & Starr JM (2010) A revised activities of daily living/instrumental activities of daily living instrument increases interpretive power: theoretical application for functional tasks. *Gerontology* **56**, 483–490.

Finch M, Kane RL & Philip I. (1994) Developing a new metric for ADLs. *Journal of the American Geriatric Society* **43**, 877–884.

Fisher WP & Fisher AG (1993) Applications of Rasch analysis to studies in occupational therapy. *Physical Medicine and Rehabilitation Clinics of North America New Developments in Functional Assessment* **4**, 551–569.

Gillespie M, Tenvengert EM & Kingsma J (1988) Using Mokken methods to develop robust cross-national scales: American and West-German attitudes toward abortion. *Social Indicators Research* **20**, 181–203.

Hambleton RK & Swaminathan H (1985) *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing, Boston.

Hambleton RK, Swaminathan H & Rogers HJ (1991) *Fundamentals of Item Response Theory*. Sage, Newbury Park, CA.

Hemker BT & Sijtsma K (1995) Selection of unidimensional scales from a multidimensional item bank in the polytmous Mokken IRT model. *Applied Psychological Measurement* **19**, 337–352.

Hulin LH, Drasgow Y & Parsons CK (1983) *Item Response Theory: Application to Psychological Measurement*. Dow Jones-Irvin, Homewood, IL.

de Jong A & Molenaar IW (1987) An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatric Research* **21**, 137–149.

Katz E (1988) In memoriam: Louis Guttman, 1916–1987. *Public Opinion Quarterly* **52**, 240–242.

Kempen GIJM & Suurmeijer ThPBM (1990) Depression, loneliness, physical disability and the utilization of professional home care among older adults. *International Journal of Health Sciences* **1**, 271–276.

Kempen GIJM & Suurmeijer ThPBM (1991) Factors influencing professional home care utilization among the elderly. *Social Science and Medicine* **32**, 77–81.

Kempen GIJM, Myers AM & Powell LA (1995) Hierarchical structure in ADL and IADL: analytical assumptions and applications for clinicians and researchers. *Journal of Clinical Epidemiology* **48**, 1299–1995.

Kingshott R, Douglas N & Deary I (1998) Mokken scaling of the Epworth Sleepiness Scale items in patients with the sleep apnoea/hypopnoea syndrome. *Journal of Sleep Research* **7**, 293–294.

von Korff MV, Ormel J, Keefe FJ & Dworkin SF (1992) Grading the severity of chronic pain. *Pain* **50**, 133–149.

Ligtvoet R, van der Ark LA, te Marvelde JM & Sijtsma K (2010) Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement* **70**, 578–595.

Lin L-C, Watson R, Lee Y-C, Chou Y-C & Wu S-C (2008) Edinburgh Feeding Evaluation in Dementia (EdFED) scale: cross-cultural validation of the Chinese version. *Journal of Advanced Nursing* **62**, 116–123.

Lin L-C, Huang Y-J, Su S-G, Watson R, Tsailk BW-J & Wu S-C (2010a) Using spaced retrieval and Montessori-based activities in improving eating ability for residents with dementia. *International Journal of Geriatric Psychiatry* **25**, 953–959.

Lin L-C, Watson R & Wu S-C (2010b) What is associated with low food intake in older people with dementia? *Journal of Clinical Nursing* **19**, 53–59.

Meijer RR (2010) A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): how to investigate whether personality items form a hierarchical scale? *Personality and Individual Differences* **48**, 502–503.

Meijer RR & Baneke JJ (2001) Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods* **9**, 354–368.

Meijer RR & Sijtsma K (1990) Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement* **14**, 283–298.

Menzel H (1953) A new coefficient or scalogram analysis. *Public Opinion Quarterly* **17**, 268–280.

Mokken RJ & Lewis C (1982) A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement* **6**, 417–430.

Molenaar IW & Sijtsma K (2000) *MSP5 for Windows: a program for Mokken Scale Analysis for Polytomous Items.* iec ProGAMMA, Groningen.

Moorer P & Suurmeijer TBPM (1994) A study of unidimensionality and cumulativeness of the MOS Short-Form General Health Survey. *Psychological Reports* **74**, 467–470.

Moorer P, Suurmeijer ThBPM, Foets M & Molenaar IW (2001) Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Quality of Life Research* **10**, 637–645.

Nandakumar R (1994) Assessing dimensionality of a set of item responses-comparison of different approaches. *Journal of Educational Measurement* **31**, 17–35.

Niemöller K & van Schuur W (1983) Stochastic models for unidimensional scaling: Mokken and Rasch. In *Data Analysis and the Social Sciences* (McKay D, Schofield N & Whitely P eds). Frances Pinter, London, pp. 120–170.

Nunnally JP (1978) *Psychometric Theory*. McGraw-Hill, New York.

Ringdal K, Ringdal GI, Kaasa S, Bjordal K, Wisløff BF, Sundstrøm S & Hjermstad MJ (1999) Assessing the consistency of psychometric properties of the HRQoL scales within EORTC QLQ-C30 across populations by means of the Mokken scaling model. *Quality of Life Research* **8**, 25–43.

Ringdal GI, Jordhøy MS & Kaasa S (2003) Measuring quality of palliative care: psychometric properties of the FAMCARE scale. *Quality of Life Research* **12**, 167–176.

Rust J & Golombock S (1999) *Modern Psychometrics*. Routlegde, London.

Schuessler K (1971) *Analyzing Social Data: A Statistical Orientation*. Houghton Mifflin Company, Boston.

van Schuur WH (2003) Mokken scale analysis: between the Guttman scale and parametric item response theory. *Political Analysis* **11**, 139–163.

Sijtsma K & Junker BW (1996) A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology* **49**, 79–105.

Sijtsma K & Molenaar IW (2002) *An Introduction to Nonparametric Item Response Theory*. Sage Publications, Thousand Oaks.

Sijtsma K, Debets P & Molenaar IW (1990) Mokken scaling analysis for polychotomous items: theory, a computer programme and an empirical application. *Quality & Quantity* **24**, 173–188.

Sijtsma K, Meijer RR & van der Ark LA (2011) Mokken scale analysis as time goes by: an update for scaling practitioners. *Personality and Individual Differences* **50**, 31–37.

Stewart ME, Watson R, Clark A, Ebmeier KP & Deary IJ (2010) A hierarchy of happiness? Mokken scaling analysis of the Oxford Happiness Inventory. *Personality and Individual Differences* **48**, 845–848.

Stouffer SA, Guttman L, Suchman EA, Lazarsfeld PF, Star SA & Clausen JA (1950) *Measurement and Prediction*, Vol. 4. Princeton University Press, Princeton, NJ.

Suurmeijer ThPBM & Kempen GIJM (1990) Behavioural changes as an outcome of disease: the development of an instrument. *International Journal of Health Sciences* **1**, 189–194.

Tenvegert E, Gillespie MW, Kingma J & Klasen H (1992) Abortion attitudes, 1984–1987–1988: effects of item ordering and dimensionality. *Perceptual and Motor Skills* **74**, 627–642.

Thompson DR & Watson R (2010) Mokken scaling of the Myocardial Infarction Dimensional Assessment Scale (MIDAS). *Journal of Evaluation in Clinical Practice* **17**, 156–159.

Townsend P (1962) *The Last Refuge*. Routledge & Kegan Paul, London.

Watson R (1994a) Measuring feeding difficulty in patients with dementia: developing a scale. *Journal of Advanced Nursing* **19**, 257–263.

Watson R (1994b) Measuring feeding difficulty in patients with dementia: replication and valiadation of the EdFEd Scale#1. *Journal of Advanced Nursing* **19**, 850–855.

Watson R (1996) Mokken scaling procedure (MSP) applied to feeding difficulty in elderly people with dementia. *International Journal of Nursing Studies* **33**, 385–393.

Watson R & Deary IJ (1994) Measuring feeding difficulty in patients with dementia: multivariate analysis of feeding problems, nursing interventions and feeding difficulty. *Journal of Advanced Nursing* **20**, 283–287.

Watson R & Deary IJ (1996) Is there a relationship between feeding difficulty and nursing intervention in elderly people with dementia? *NT Research* **1**, 44–45.

Watson R & Deary IJ (1997) Feeding difficulty in elderly people with dementia: confirmatory factor analysis. *International Journal of Nursing Studies* **34**, 405–414.

Watson R & Deary IJ (2009) Reply to: A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): how to investigate whether personality items form a hierarchical scale? *Personality and Individual Differences* **48**, 504–505.

Watson R, Green S & Legg L (2001) The Edinburgh Feeding Evaluation in Dementia Scale #2 (EdFED#2): convergent and discriminant validity. *Clinical Effectiveness in Nursing* **5**, 44–46.

Watson R, MacDonald J & McReady T (2002) The Edinburgh Feeding Evaluation in Dementia Scale #2: inter- and intra-rater reliability. *Clinical Effectiveness in Nursing* **5**, 184–186.

Watson R, Deary I & Austin E (2007) Are personality trait items reliably more or less 'difficult'? Mokken scaling of the NEO-FFI. *Personality and Individual Difference* **43**, 1460–1469.

Watson R, Deary IJ & Shipley B (2008a) A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine* **28**, 575–579.

Watson R, Roberts(nee Shipley) B, Gow A & Deary IJ (2008b) A hierarchy of items within Eysenck's EPI. *Personality and Individual Differences* **45**, 333–335.

Wilson M (2005) *Constructing Measures: An Item Response Modelling Approach*. Erlbaum, Mahwah, NJ, p. 187.