Personality and Individual Differences 66 (2014) 74–78

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Investigating invariant item ordering in the Mental Health Inventory: An illustration of the use of different methods

Roger Watson [a,*], Wenru Wang [b], David R. Thompson [c], Rob R. Meijer [d]

[a] The University of Hull, Hull HU6 7RX, UK
[b] National University of Singapore, Republic of Singapore
[c] Australian Catholic University, Melbourne, Australia
[d] University of Groningen, Groningen, Netherlands

ARTICLE INFO

Article history:
Received 25 November 2013
Received in revised form 11 March 2014
Accepted 15 March 2014
Available online 12 April 2014

Keywords:
Item response theory
Mokken scaling
Invariant item ordering
Mental health
Coronary heart disease

ABSTRACT

Invariant item ordering is a property of scales whereby the items are scored in the same order across a wide range of the latent trait and across a wide range of respondents. In the package 'mokken' in the statistical software R, the ability to analyse Mokken scales for invariant item ordering has recently been available and techniques for inspecting visually the item response curves of item pairs, have also been included. While methods to assess invariant item ordering are available, there have been indications that items representing extremes of distress in mental well-being scales, such as suicidal ideation, may lead to claiming invariant item ordering where it does not exist. We used the Mental Health Inventory to see if invariant item ordering was indicated in any Mokken scales derived and to see if this was being influenced by extreme items. A Mokken scale was derived indicating invariant item ordering. Visual inspection of the item pairs indicated that the most difficult item (suicidal ideation) was located far from the remaining cluster of items. Removing this item lowered invariant item ordering to an unacceptable level.

1. Introduction

Invariant item ordering (IIO) is a property of scales whereby items are scored in the same order by all respondents at all levels of the latent trait being measured (Ligtvoet, 2010). As stated by Ligtvoet (2010, p. 8) 'IIO is a strong requirement in measurement practice' and that researchers 'do not realize that an ordering relationship that holds at the aggregation higher level of mean item scores does not automatically generalize to the lower level of individual subjects' (p. 1). As such, IIO can be considered an exacting but important property of scales (Sijtsma & Junker, 1996) and the extent to which IIO holds in a set of items can be examined using methods that fall under item response theory (IRT) including parametric methods such as Rasch scaling (Meijer, Sijtsma, & Smid, 1990) and the non-parametric method of Mokken scaling analysis (MSA). IRT methods are able to relate, meaningfully, the score on a scale with the score on the latent trait as that score can be related to a specific set of items (Watson et al., 2011).

1.1. Mokken scaling

Mokken scaling is a non-parametric application of IRT (Watson et al., 2011). It is non-parametric in the sense that, unlike parametric IRT models such as Rasch, no assumptions are made about the shape of the curve describing the relationship between the score on a latent trait and the probability of obtaining that score (Meijer et al., 1990)—the item characteristic curve (ICC)—other than that individual ICCs are monotone and that sets of ICCS do not intersect (Mokken & Lewis, 1982). Mokken scaling can be considered to be a stochastic version of Guttman scaling. It is stochastic in the sense that the relationship between the score on a latent trait and the probability of obtaining that score is not deterministic, unlike the relationship in Guttman scaling which only catered for dichotomous responses and could not accommodate items that did not exactly fit the Guttman model. Mokken scaling may be considered to have advantages over Rasch scaling in that, while it is a rigorous method, it is less demanding in terms of its criteria for retaining items in scales and, therefore, more conservative of items (Meijer et al., 1990). In certain situations where order rather than precision is required, for example in the measurement of activities of daily living or attitudes, as opposed to situations where precision is required, for example in the measurement of achievement,

* Corresponding author. Tel.: +44 1482464525; fax: +44 1142269790.
E-mail address: r.watson@hull.ac.uk (R. Watson).

http://dx.doi.org/10.1016/j.paid.2014.03.017
0191-8869/

Mokken scaling is both adequate and preferable as it tends to retain more information about the latent trait by retaining more items.

MSA adheres to all the assumptions of IRT such as: unidimensionality in a set of scaled items whereby all the items purport to measure a single attribute (Sick, 2010); local stochastic independence of items in a scale whereby the score on any item is a result of its relationship to the latent trait and not to any of the other items in the scale (Watson et al., 2011); monotone homogeneity whereby the score on an item increases continually as the latent trait increases (Mokken & Lewis, 1982); and IIO, as defined above. To attain IIO in MSA requires that ICCs do not intersect (Deary, Watson, Booth, & Gale, 2013). In addition, IIO requires that item step response functions (ISRFs) do not intersect. ISRFs are analogous to ICCs in that they represent the relationship between the score on the latent trait and the score on each of the response options for each item; the number of ISRFs for an item are related to the response options by $n - 1$ where n = the number of response options.

In common with other IRT methods—and with Guttman scaling from which it was derived—Mokken scaling is used to generate hierarchical scales where items are incorporated into scales and ordered according to difficulty. In this sense, 'difficulty' refers to the likelihood of endorsing an item in a questionnaire. Such scales are useful in that endorsement of items is ordered by the difficulty of the items such that endorsement of a particular item implies endorsement of those remaining items which are less difficult to endorse but not, necessarily, those remaining items which are more difficult to endorse. For example, in a questionnaire inquiring about attitudes to abortion, a person who endorses and item, with regard to abortion, that 'It is a woman's right to choose to have an abortion' is more likely to endorse an item 'Life does not begin at conception'. However, someone endorsing the latter question may not necessarily endorse the former question and someone who does not endorse the former question is very unlikely to endorse the latter. In this way, the score on a set of items in a hierarchical scale is a measure of the presence of the latent trait in such a way that the score is more meaningful as it can be related to a specific item or set of items.

With the development of the package 'mokken' in the online public domain statistical software R (van der Ark, 2007) it is possible to analyse polytomous Mokken scales for IIO. Until this development there was some confusion in the literature about the nature of IIO in these scales (Meijer, 2010; Sijtsma, Meijer, & Van der Ark, 2011; Watson & Deary, 2010). Since the advent of the package 'mokken' in R and the concomitant development of methods to investigate IIO, the application of these methods to Mokken scales has been demonstrated (Ligtvoet, van der Ark, Marvelde, & Sijtsma, 2010) and there have also been some warnings about misinterpreting data which appear to show IIO (Meijer & Egberink, 2012). For example it is possible to conclude that a scale shows IIO where the majority of items are very close together and even intersecting—an indication of weak or non-existent IIO—but where a single ICC is far away from the remaining items and is accounting for the apparent IIO. It is advised, therefore, to plot the IRCs for item pairs, to inspect these visually and to observe for IRC closeness and intersection and/or 'outlying' items in terms of their distance from the remaining items. It is also possible—not available in R—to plot IRCs together (Meijer & Egberink, 2012) to provide a single graph of the relationships between all IRCs in a scale. It has been observed that 'extreme' items often represent extremes in the scales; for example, in scales measuring mental well-being, items measuring suicidal ideation may be scored much higher or lower (depending on the scoring system) than remaining, more general, items about mental well-being. In this sense, 'extreme' items means those that have mean item scores which place them a long way conceptually from the remaining items in scale. For example, scales designed to measure psychological morbidity often contain an item related to suicidal ideation; clearly this could be considered 'extreme' in every sense of the word, but it has also been observed that these items act as anchoring items at the higher end of difficulty in the scale. Likewise, these scales are often anchored at the least difficult end of the scale by items related to very mild levels of psychological distress and these least difficult items could also lead to the same phenomenon of exaggerated IIO. Examples of scales, recently analysed using Mokken scaling, which show anchoring at the most difficult end of the scale include the 30-item General Health Questionnaire (Watson, Deary, & Shipley, 2008), the CORE-OM (Bedford, Watson, Lyne, et al., 2010) and the DSSI (Bedford, Watson, Henry, Crawford, & Deary, 2010).

The primary aim of this study was to analyse the Mental Health Inventory (MHI) for the existence of Mokken scales and, specifically, IIO. A secondary aim was to investigate if specific items at the extreme ends, in terms of difficulty, were contributing to the IIO.

2. Methods

2.1. Participants

A total of 204 patients with coronary heart disease were recruited from the cardiac outpatient clinics of two public hospitals in the city of Xi'an in the People's Republic of China. Inclusion criteria were: having coronary heart disease; and being able to read simplified Chinese characters. Exclusion criteria were: having no evidence of past psychiatric illness; and no severe related morbidity. Mean age was 63.1 years (SD = 11.8), 139 (68.1%) of the participants were male. The majority were married (n = 185, 90.7%) and 157 (76.9%) were educated until at least the secondary school level which, in China, means either middle school for 3 years or high school of 3 years between the ages of 13 and 18 years.

2.2. Materials

The Mental Health Inventory is a 38-item measure of psychological distress and well-being, developed for use in general population distress (Veit & Ware, 1938) (Table 1). The MHI was originally developed by Veit and Ware (1938) on a sample of 5089 participants. Factor analysis demonstrated that the MHI had a higher order structure between two correlated factors of psychological distress and well-being and a lower order structure of five factors related to anxiety, depression, emotional ties, general positive affect, and loss of behavioural emotional control. A Chinese Mandarin version of the MHI (CM:MHI)—used in this study—was developed from the original English version through a rigorous forward–backward translation process (Liu, Chow, Lau, He, & Wang, in press).

2.3. Procedure

Ethical approval was sought and attained from the Ethics Committees of the aforementioned hospitals and consent was obtained from each. Completing the CH-MHI took approximately 20 min.

2.4. Analysis

Data were analysed using the automated item selection procedure to allocate items to putative Mokken scales in the data. The automated item selection procedure is an iterative procedure in R which allocates items to Mokken scales on the basis of those with the best scaling properties first and then adding items until the