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

*Corresponding author. Tel: +441482464525. E-mail address: r.watson@hull.ac.uk

Key words:

Item response theory

Mokken scaling

Invariant item ordering

Mental health

Coronary heart disease

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

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.

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

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, van der Ark, Lin, Fieo, Deary, & Meijer, 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, Sijtsma, & Smid, 1990). In

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

certain situations where order rather than precision is required, for example is the measurement of activities of daily living or attitudes, as opposed to situations where precision is required, for example in the measurement of achievement, 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 difficulty 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

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

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 misintepreting 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

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

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, Lynne, Tibbles, Davies, & Deary, 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 three years or high school of three years between the ages of 13 to18 years. Ethical permission was obtained from the participating hospitals.

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, 1983). The MHI was originally developed by Veit and Ware (1983) on a sample of 5089 participants. Factor analysis demonstrated that the MHI had a higher

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

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 et al., 2013).

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 minutes.

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 scalability falls below a lowerbound value. Items which have poor scaling properties are excluded from the scales. Thereafter, the scales were analysed for unidimensionality using Loevinger's coefficient (H)) and is a measure of the extent of Guttman violations in the data (Watson et al. 2012; H is a measure of the extent to which items are always scored in the same way in the data and MSA provides Hi (item H) and Hij (item pair H) and the strength of a scale can be assessed as follows (Watson et al. 2012):

H>0.3 indicates a weak scale

H>0.4 indicates a moderate scale

H>0.5 indicates a strong scale.

Monotone homogeneity was evaluated using the 'crit' statistic generated by package 'mokken' in R. Item with 'crit' values <40 are considered to be acceptable (Molenaar, & Sijtsma, 2000). Reliability (Rho) was

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

Table 1 Mokken scale 1 from the Mental Health Inventory (n=204)

Item	Label	Mean	H(SE)
22	How much of the time, during the past month, were you able to relax without difficulty?	3.24	0.41(0.048)
1	How happy, satisfied or pleased have you been with your personal life during the past month?	3.15	0.42(0.040)†
13	During the past month, how much of the time have you felt tense or “high-strung”?	2.79	0.51(0.035)
35	How often during the past month did you find yourself trying to calm down?	2.79	0.39(0.049)†*
18	How much of the time, during the past month, have you felt emotionally stable?	2.75	0.44(0.042)
27	How often, during the past month, have you felt so down in the dumps that nothing could cheer you up?	2.73	0.47(0.045)
24	How often, during the past month, did you feel that nothing turned out for you the way you wanted it to?	2.69	0.41(0.044)
11	How much of the time, during the past month, have you been a very nervous person?	2.67	0.42(0.040)
30	During the past month, how much of the time have you been moody or brooded about things?	2.62	0.45(0.045)
19	How much of the time, during the past month, have you felt downhearted and blue?	2.59	0.54(0.030)
36	During the past month, how much of the time have you been in very low spirits?	2.56	0.56(0.032)
33	During the past month, have you been anxious or worried?	2.49	0.47(0.039)
29	During the past month, how much of the time have you felt restless, fidgety, or impatient?	2.49	0.45(0.041)
32	During the past month, how often do you get rattled, upset of flustered?	2.47	0.37(0.048)†*
3	How often did you become nervous or jumpy when faced with excitement or unexpected situations during the past month?	2.41	0.39(0.047)
25	How much have you been bothered by nervousness, or your “nerves”, during the past month?	2.35	0.56(0.032)
9	Did you feel depressed during the past month?	2.21	0.50(0.034)
20	How often have you felt like crying, during the past month?	1.94	0.33(0.057)†
21	During the past month, how often have you felt that others would be better off if you were dead?	1.80	0.34(0.053)†
28	During the past month, did you think about taking you own life?	1.40	0.43(0.056)

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

$H=0.44$(SE 0.029); Rho=0.91; $H^T=0.34$; ✝- items where 95% CI includes 0.3; *- items violating IIO

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

evaluated where Rho indicates 'the degree of stability of a respondent's test score across independent replications of a test administration' (van der Ark, 2007, p. 13); Rho>0.7 is considered to indicate a reliable scale. Invariant item ordering was evaluated using H^T (H trans) which is analogous to H and is a measure of IIO whereby:

H^T>0.3 indicates weak IIO

H^T>0.4 indicates moderate IIO

H^T>0.5 indicates strong IIO.

Item pairs were plotted for visual inspection of proximity and intersection and the standard errors (SE) for H, Hi and Hij were generated, thereby allowing the estimation of 95% confidence intervals (CIs) (Kuijpers, van der Ark, & Croon, 2013). For H and Hi the 95% CIs should not include the lowerbound threshold for H (usually referred to as *c*) of 0.3, otherwise the item should be excluded from the scale, and for the Hij the CIs should not include zero or the item pair should be inspected to decide which of the items should be excluded to improve the strength of the scale. Finally, the probability of obtaining a Mokken scale can be estimated using a Bonferroni corrected method for the multiple iterations in the procedure; the default setting for probability of obtaining a Mokken scale in package 'mokken' in R is 0.05.

As a final step, we also investigated whether there were persons for whom the item ordering did not apply. From a diagnostic point-of-view this may be important information because these persons cannot be scaled (Meijer & Sijtsma, 2001) and researcher should be very careful in interpreting the total score for these persons. We calculated the normed Guttman errors (Emons, 2008) using the R Package PERFIT (Tendeiro & Meijer, 2014).

3. Results

Two Mokken scales were derived from the present data incorporating 33 items in total; only one scale showing initial IIO is considered here; 8 items were excluded from both scales by the automated item selection procedure because they did not scale due to Hi<0.30; this is below the lowerbound value and their

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

inclusion in a scale would have lowered the overall Hs to < 0.30. No items included in the scales violated monotone homogeneity. The Mokken scale considered here includes 20 items mainly related to emotional stability and is moderate (Hs=0.44) in strength with weak IIO (H^T=0.34). The scale can be considered unidimensional, according to the criteria for Mokken scaling, because of the automated item selection procedure allocation to the scale and the fact that Hs>0.40. Five items may be excluded on the basis that their 95% CIs included the lowerbound *c*=0.30. Several Hij had 95% CI<0 and in all cases these item pairs included an item where the 95% CIs included the lowerbound c=0.30. In addition, two of the above items violated IIO, therefore, it is advisable to remove these items and re-analyse the data to see if the IIO improves. However, removing these items did not improve IIO.

There is a sensible hierarchy of items running from very general and relatively mild feelings related to relaxation (e.g. "How much of the time, during the past month, were you able to relax without difficulty?"; "How often during the past month did you find yourself trying to calm down?"), through more specific feelings of anxiety (e.g. "How much of the time, during the past month, have you been a very nervous person?"; "During the past month, have you been anxious or worried?"), to frank expressions of depression and suicidal ideation (e.g." Did you feel depressed during the past month?"; "How often have you felt like crying, during the past month?"; "During the past month, how often have you felt that others would be better off if you were dead?"; "During the past month, did you think about taking you own life?").

With specific reference to the topic of their paper, inspection of the item pair plots (Figure 1) shows that item 28 ("During the past month, did you think about taking you own life?") is located far away from all of the remaining items at the most difficult end of the scale suggesting that it may be responsible for the apparent IIO in this scale. In fact, removing item 28 and re-analysing lowers H^T to 0.23, which indicates very weak IIO.

We inspected the item score patterns with the 5% normed Guttman errors and found some interesting score patterns. For example, an extreme pattern was the pattern of person 40. When we order the items to

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

decreasing mean score, the item score pattern was [1,2,2,5,0,1,1,1,1.1,1,2,2,2,2,2,1,1,1,4,0], thus it is strange that the most popular item is answered with a 1 and one of the least popular items with a 4. This may, for example, point to not completing the questionnaire seriously or to an atypical mental health condition.

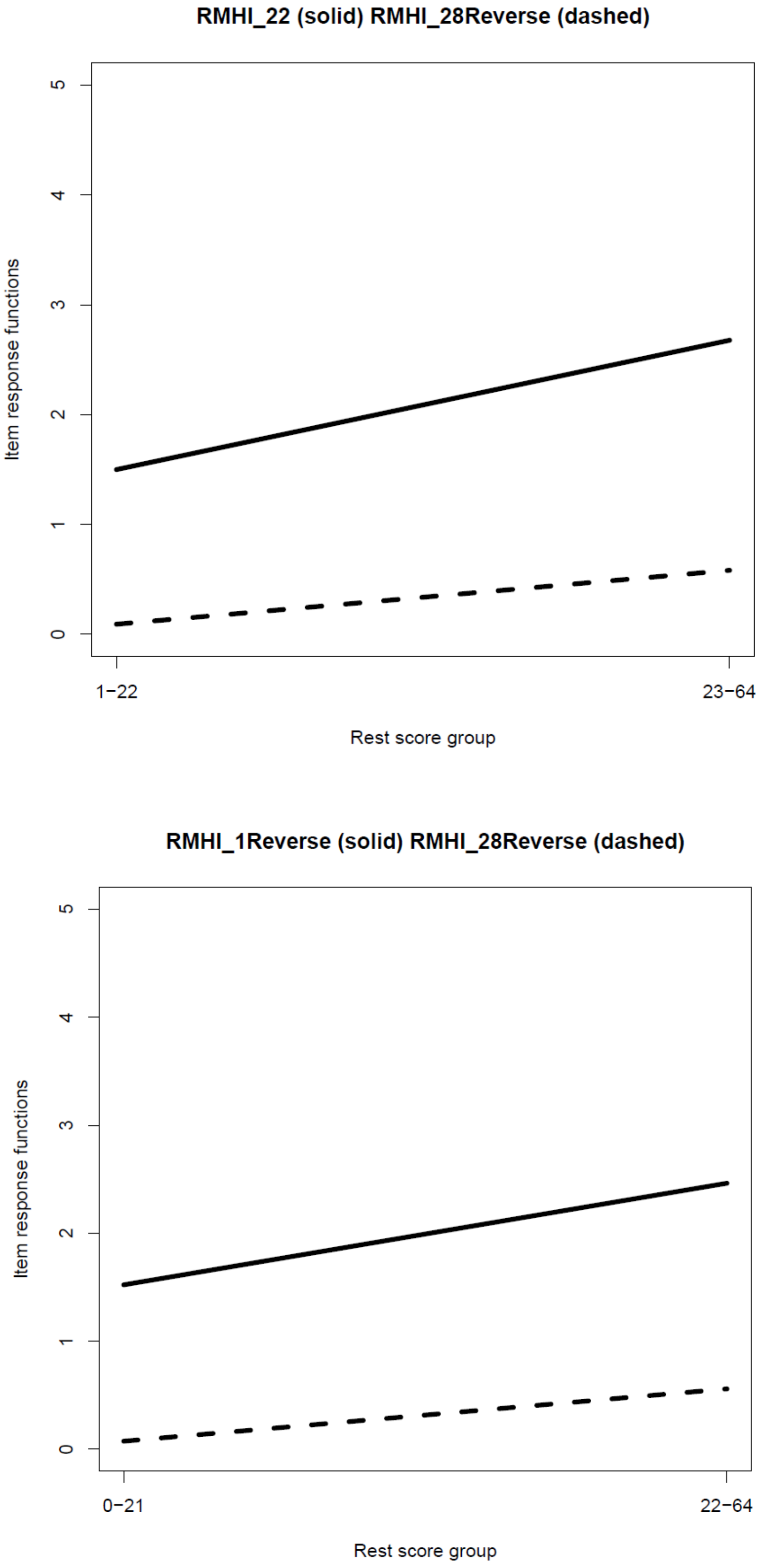
4. Discussion

We have demonstrated using MSA that a translation of the CH-MHI contains a strong Mokken scale, initially showing weak IIO. The scale also contains a sensible hierarchy of items indicating that this could be a useful measure of emotional stability. Clearly, the sample size in our study is relatively small but given the scale H is > 0.4 it is probably adequate (Straat, 2012). However, sample size estimation for Mokken scales is still a relatively new aspect of this field. Due to the inclusion of an extremely difficult item in the MHI: one indicating suicidal ideation; and with the warnings of how such an item could exaggerate or mislead about the existence of IIO in Mokken scales (Meijer, & Egberink, 2012), we inspected the item-pairs visually and confirmed that one of these items was, in fact, 'outlying' in terms of distance from the remaining items. Removing the item did, indeed, lower IIO from weak, but acceptable (H^T>0.30), to a level (H^T<0.30) where IIO could not be claimed for the resulting scale. The initial assessment of IIO was, therefore, illusory and this serves to demonstrate the value of careful analysis and consideration of each item in a Mokken scale before claims of IIO are made.

Clearly, we would not necessarily advocate the removal of items on suicidal ideation from scales measuring psychological distress. This extreme expression of despair is an important indicator of the level of distress and serious psychological morbidity may be missed if the items are not available for clinical assessment. However, in any procedure involving IIO, the removal of extremely difficult (or extremely easy) items for analysis should be considered. In this way, the quality of the remaining items can be evaluated and, if necessary, improvements made to the psychometric properties of the scale by removing or modifying the remaining items.

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

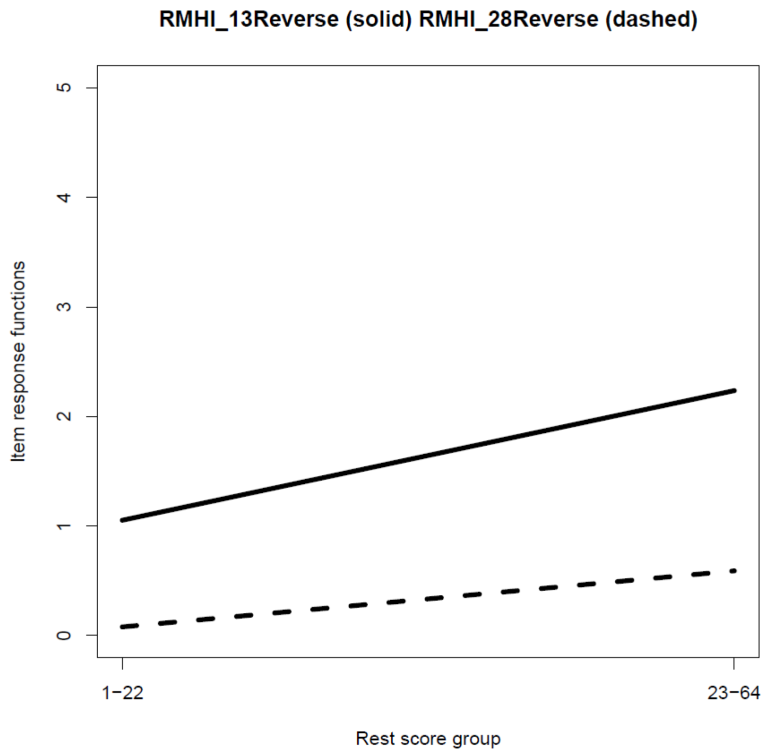
Figure 1
Selected item pair plots demonstrating the extreme nature of item 28 (During the past month, did you think about taking your own life?)

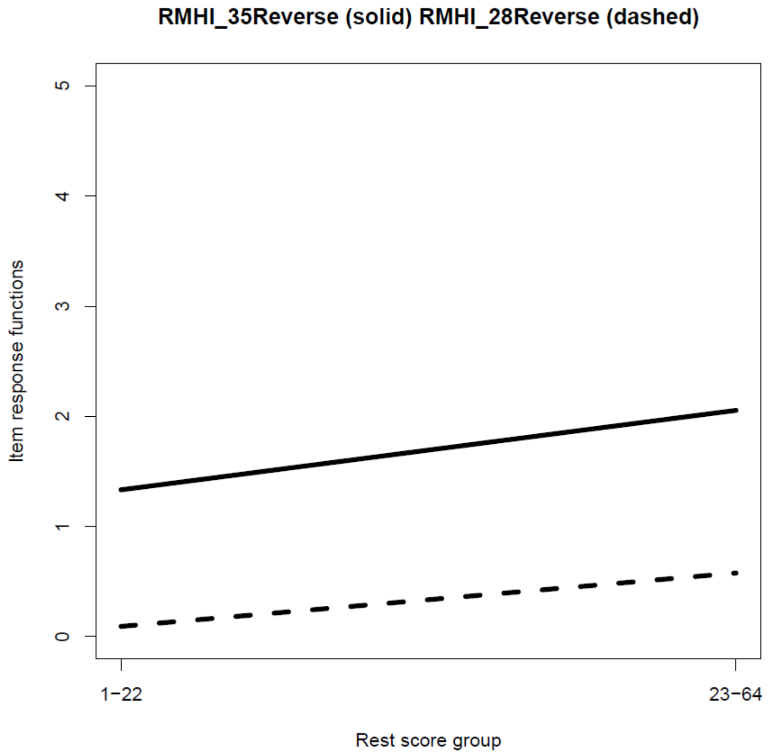


RMHI_22 (solid) RMHI_28Reverse (dashed)

RMHI_1Reverse (solid) RMHI_28Reverse (dashed)

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/





© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

The implications of this study, alongside another major study of the same phenomenon (Meijer, & Egberink, 2012), indicates that caution should be exercised in making claims about IIO. The study also implies that IIO is an elusive, but nevertheless crucial property of Mokken scales and greater effort needs to be invested in item selection and scale design to produce scales that are robust across wide ranges of latent traits and across a wide range of respondents. Finally, the study implies that some previous work (Bedford, et al., 2010; Cosco, Doyle, Watson, Ward, McGhee, 2012; Deary, et al., 2013) where claims of IIO have been made, or which preceded the ability to analyse Mokken scales for IIO should be reviewed and reassessed using these new graphical techniques available in package 'mokken' in R. It is hoped that the ability to plot multiple IRCs in R will also soon be available.

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

References

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

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

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

Cosco, T.D., Doyle, F., Watson, R., Ward, M., McGee, H. (2012). Mokken scaling analysis of the Hospital Anxiety and Depression Scale in individuals with cardiovascular disease, *General Hospital Psychiatry*, 34, 167-172.

Deary, I.J., Watson, R., Booth, T., & Gale, C.R. (2013) Does cognitive ability influence responses to the Warwick-Edinburgh Mental Well-Being Scale? *Psychological Assessment*, 25, 313–318.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224-247.

Kuijpers, R.E., van der Ark, L.A., & Croon, M.A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models, *Sociological Methodology*, 43, 42–69.

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

Ligtvoet, R. (2010). *Essays on invariant item ordering*, GildeprintDrukkerijen, Enschede.

Ligtvoet, R.,van der Ark, L.A., Marvelde, J.M., & Sijtsma, K. (2010). Investigating an Invariant Item Ordering for Polytomously Scored Items. *Educational and Psychological Measurement*, 70, 578–595.

Liu, M.., Chow, A., Lau, Y., He, H-G., & Wang, W. (In press) Psychometric testing of the Chinese Mandarin version of the Mental Health Inventory among Chinese patients with coronary heart disease in Mainland China. *International Journal of Nursing Practice,* Accepted for publication on 2 August 2013.

Meijer, R.R. (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, R.R., & Egberink, I.J.L. (2012). Investigating invariant item ordering in personality and clinical scales: some empirical findings and a discussion, *Educational and Pychological Measurement,* 72, 589–607.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meijer, R., Sijtsma, K., & Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT, *Applied Psychological Measurement*, 14, 283-298.

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

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows*, Groningen: iec ProGAMMA.

Sick, J. (2010) Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, 14:2, 23-29.

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

Sijtsma, K. Meijer, R.R., Van der Ark, L.A. (2111). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, 50, 31-37.

Straat, H. (2010). *Using scalability coefficients and conditional association to assess monotone momogeneity*, Ridderprint BV, Ridderkerk.

Tendeiro, J. N. & Meijer, R.R. (2014). PERFIT, a package for calculating person-fit statistics. (in preparation).

Veit, C.T., & Ware, J.E. (1938).The structure of psychological distress and well-being in general populations, *Journal of Consulting and Clinical Psychology*, 51, 730-742.

Watson, R., van der Ark, L.A., Lin, L-C., Fieo, R., Deary, I.J., & Meijer, R.R. (2011). Item response theory: how Mokken scaling can be used in clinical practice, *Journal of Clinical Nursing*, 21, 2736–2746.

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

Watson, R., & Deary, I.J. (2010). 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., Deary, I.J., Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine*, 28, 575-579.

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/