HEALTH PSYCHOLOGY AND BEHAVIORAL MEDICINE
2018, VOL. 6, NO. 1, 203–213
https://doi.org/10.1080/21642850.2018.1505520

Taylor & Francis
Taylor & Francis Group

What are the minimal sample size requirements for Mokken scaling? An empirical example with the Warwick- Edinburgh Mental Well-Being Scale

Roger Watson [a], Iris J. L. Egberink [b], Lisa Kirke [a], Jorge N. Tendeiro [b] and Frank Doyle [c]

[a]Faculty of Health and Social Care, University of Hull, Hull, UK; [b]Department of Psychometrics and Statistics, University of Groningen, Groningen, The Netherlands; [c]Division of Population Health Sciences (Psychology), Royal College of Surgeons in Ireland, Dublin, Ireland

ABSTRACT

Purpose: Sample size in Mokken scales is mostly studied on simulated data, reflected in the lack of consideration of sample size in most Mokken scaling studies. Recently, [Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement*, *74*, 809–822] provided minimum sample size requirements for Mokken scale analysis based on simulation. Our study uses real data from the Warwick-Edinburgh Mental Well-Being Scale ($N = 8463$) to assess whether these hold.

Methods: We use per element accuracy to evaluate the impact of sample size, with scaling coefficients and confidence intervals around scale, item and item pair scalability coefficients.

Results: Per element accuracy, scalability coefficients, and confidence intervals around scalability coefficients are sensitive to sample size. The results from Straat et al. were not replicated; depending on the main goal of the research, sample sizes ranging from > 250 to > 1000 are needed.

Conclusions: Using our pragmatic approach, some practical recommendations are made regarding sample sizes for studies of Mokken scaling.

ARTICLE HISTORY
Received 3 May 2018
Accepted 15 July 2018

KEYWORDS
Mokken scaling; item response theory; per element accuracy; scalability; confidence intervals; sample size

Introduction

The use of Mokken scaling to analyse the properties of scale items has increased (Straat, van der Ark, & Sijtsma, 2014) and, concomitantly, methods surrounding this method have also improved. From its early beginnings, when it was often described as a stochastic version of Guttman scaling for binary items (Mokken & Lewis, 1982), sample size estimation was considered impossible and the investigation of invariant item ordering (IIO) was difficult (Molenaar, Sijtsma, & Boer, 2000). However, since then, the advent of Mokken scaling for polytomous items (Sijtsma, Debets, & Molenaar, 1990), modelling

CONTACT Roger Watson r.watson@hull.ac.uk Faculty of Health and Social Care, University of Hull, Hull, HU6 7RX, UK
Supplemental data for this article can be accessed http://dx.doi.org/10.1080/21642850.2018.1505520.

of sample size requirements (Straat et al., 2014), and the calculation of standard errors and, thereby, confidence intervals (CIs) for scalability coefficients (Kuijpers, van der Ark, & Croon, 2013) has been enabled. These methods have demonstrated the increasing utility and sophistication of Mokken scaling.

One aspect of Mokken scaling that is still largely unstudied is sample size determination and according to Straat et al. (2014) this is important for two reasons. First, to prevent the capitalisation on the chance of observing a Mokken scale when none exists due to small sample sizes. Second, considering the limited time and resources available to researchers, obviating the necessity of obtaining unnecessarily large sample sizes. Sample size estimation has previously been a matter of speculation in Mokken scaling. One estimate of sample size set the minimum for Mokken scaling at 400 (Meijer & Baneke, 2001), but it has been shown that sample sizes for published papers with Mokken scaling range from 133 to 15,022 (Straat et al., 2014). Furthermore, the nonparametric nature of Mokken scales and lack of information on the distribution properties of Mokken scales has made sample size estimation difficult. Clearly there has been little agreement on sample size, and further work is needed to establish pragmatic guidelines for sample size estimation for Mokken scale analysis.

Recently, Straat et al. (2014) provided minimum sample size requirements for Mokken scale analysis for scales consisting of one or more constructs, for short and long scales, and for poorly and highly discriminating items/scales. Their results are based on a simulation study. However, it is unknown whether these guidelines hold in general for different latent traits and with real data. Therefore, in this study the guidelines from Straat et al. (2014) are applied to real data of the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS), to investigate whether their minimum sample size requirements lead to similar results when using real data. Straat et al. (2014) used per element accuracy (PEA) as the dependent variable to investigate the correct allocation of items to scale(s). Besides using PEA as the dependent variable, however, the scalability coefficient (H_s) and the 95% confidence intervals around H_s, H_i and H_{ij} (Loevinger's coefficient for item pairs) should arguably also be considered as these will be of significant interest to researchers.

After introducing Mokken scaling, a short overview of relevant studies on sample size requirements is provided.

Mokken scaling

Mokken scaling was derived from Guttman scaling (Stouffer et al., 1950) and is an improvement on that method as it is stochastic (Mokken & Lewis, 1982) – rather than deterministic – in describing the relationship between trait scores and item responses. This brings Mokken scaling under the umbrella of item response theory (IRT) models (Watson et al., 2012) which includes, for example, the Rasch model and the partial credit model (Embretson & Reise, 2000). However, Mokken scaling is described as nonparametric because, unlike parametric models, no assumptions are made about the shape of the (non-decreasing) item response function (IRF) – the curve relating the level of the latent trait being measured and the probability of obtaining a score on a particular item (Meijer, Sijtsma, & Smid, 1990). Given the recent controversies in the literature about the uncritical use of alpha to indicate dimensionality in health psychology (Crutzen & Peters, 2017), and a psychometric tutorial that includes Mokken scaling in

this issue (Dima, 2018), the promotion of such methods, and indeed adoption of appropriate samples sizes, is badly needed.

The first, less strict and often applied Mokken model is Mokken's model of monotone homogeneity (MMH), which is based on three assumptions: unidimensionality; the assumption of local stochastic independence of items; and monotone homogeneity of items (Meijer & Baneke, 2001). Unidimensionality is assessed by estimating the scalability of a set of items and this is done on the basis of the extent of Guttman errors in a dataset; in other words, the extent to which some item pairs are not ordered in the same way relative to one another. For example, for two items *i* and *j*, if *j* represents more severity or 'difficulty' of the latent trait being measured than *i*, then we would always expect *i* to be endorsed more readily than *j*. The number of times *j* is endorsed more readily than *i*, or endorsed equally to *i*, are errors and these are used to calculate Loevinger's coefficient (H_s). By convention, the strength of a scale is considered as being weak ($.30 \leq H_s < .40$), moderate ($.40 \leq H_s < .50$) or strong ($.50 \leq H_s \leq 1.00$) (Hemker, Sijtsma, & Molenaar, 1995). Local stochastic independence of items means that individual item scores should be independent of one another and are a function only of the latent trait being measured and not of some other relationship between the items in a scale (van Alphen, Halfens, Hasman, & Imbos, 1994). Local stochastic independence is violated if the score on any item in a scale is dependent on the score on another item (Watson, Thompson, & Wang, 2014). Monotonicity is the property whereby the relationship between an item score and a score on the latent trait are non-decreasing; i.e. as the score on the latent trait increases, the score on an item continuously increases (Meijer & Baneke, 2001) and can be estimated using Mokken scaling methods. Mokken also described a second and more restricted Double Monotonicity Model (Mokken, 1971, 1997), to which the assumption of non-intersecting IRFs (i.e. invariant item ordering [IIO]) applies.

In addition, a method of calculating standard errors for scalability coefficients (total scale, items and item pairs) in Mokken scale analysis has now been derived and this permits the calculation of confidence intervals for scalability coefficients for individual items, item pairs and scales (Kuijpers et al., 2013). Calculating confidence intervals allows us to see if the scalability coefficient for an item and scale excludes (or includes) the lower bound level of estimation (*c*), a predetermined constant (Van Abswoude, Vermunt, Hemker, & van der Ark, 2004) usually set at .30. If item confidence intervals include the lower bound level of estimation then the item can be discarded. If scale confidence intervals include the lower bound level of estimation one does not have sufficient evidence to claim that the scale is at least a weak Mokken scale (i.e. $.30 \leq H_s < .40$). Calculating confidence intervals for the scalability coefficients of item pairs allows us to see if that coefficient excludes (or includes) zero, since under the monotone homogeneity model $H_{ij} \geq 0$. If the confidence interval for a pair of items includes zero, then one of the items can be discarded.

Previous research on sample size requirements

Using a Monte Carlo simulation method, Ligtvoet, van der Ark, te Marvelde, and Sijtsma (2010) studied the effect of varying a range of parameters on invariant item ordering (IIO) – the extent to which items are scored in the same order by all respondents at all levels of the latent trait. The parameters included the minimum level of violations of IIO, the item

discrimination, number of items, number of answer categories and sample size. The sample sizes tested were 200, 433 and 800; 433 related to a real-data sample and the minimum and maximum sample sizes were simulated. The results showed that increasing sample size improved the sensitivity of correctly identifying Mokken scales with IIO with no effect on the specificity of rejecting those that did not. Along with sample size, increasing the minimum level of acceptable violations and item discrimination also increased sensitivity while item discrimination, number of items and number of answer categories increased specificity. Since it is not possible to manipulate item discrimination in reality, and increasing test length and response categories may make questionnaires harder for people with lower levels of literacy (Wolf, Bennett, Davis, Marin, & Arnold, 2005), it seems that more precise information about the effect of sample size on Mokken scales is needed.

Recently, Straat et al. (2014) investigated the effect of sample size in Mokken scaling using the concept of per element accuracy (PEA). PEA is a term derived from factor analysis (Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005) and refers to the extent to which items load correctly on putative factors. Since Mokken scaling is capable of detecting multiple dimensions in databases and allocating items to subscales, Straat et al. (2014) used PEA (mediocre [> .8], adequate [> .90], good [> .95], and excellent [> .99]) as an outcome to investigate the correct allocation of items to scales under different simulated conditions for sets of 10 and 20 items; two latent traits were assumed, the correlation between the latent traits was varied and the scalability coefficient of items (H_i) also varied. Both H_i and the correlation between the latent traits had an effect on sample size requirements under this simulation. According to Straat et al., for $H_i = .32$ and a correlation of .30 between the latent variables, a sample of 750 will achieve adequate PEA whereas, at $H_i = .42$ and correlation of .30 between the latent variables, to achieve adequate PEA a sample of 50 would be required. The number of items in the analysis had a minimal effect on the sample size requirement. Essentially, therefore, the larger the value of H_i the smaller the required sample size (Straat et al., 2014); where H_i is close to the lower bound level (denoted '*c*') of .30, required sample sizes are in the thousands compared with being in the hundreds for higher values of H_i. In this light, the difference between the lower bound value of c and H_i is analogous to an effect size.

Therefore, while the effects of H_i on sample size requirement in relation to item partitioning into subscales and also IIO has been studied (Ligtvoet et al., 2010; Straat et al., 2014), this work has mainly used simulation and not real samples (although Ligtvoet et al. did relate their data to one real sample of $n = 433$). It is not possible to manipulate levels of item discrimination, which they studied, or varying parameters other than sample size. Little work has considered the effects of selecting varying sample sizes using real data. Also, no work to date on sample size has considered parameters such as confidence intervals for items and item pairs. Such work would be of pragmatic value to researchers, allowing the estimation of sample sizes required for future work, and would be more generalisable to a given population than simulated data. We, therefore, attempt to fill this gap in the literature.

The present study used a large dataset of the Warwick Edinburgh Mental Wellbeing Scale (WEMWBS) items derived from real – not simulated – publicly available data previously demonstrated to show good Mokken scaling properties in a single scale for the sample (Deary, Watson, Booth, & Gale, 2013; Stochl, Jones, & Croudace, 2012). The

size of the database ($N = 8643$) exceeds the minimum sample size generated by the recent simulation study of Straat et al. (2014) and the scalability coefficient $H_s = .48$, which is considered to be a moderate scale (i.e. $.40 \leq H_s < .50$), provide the opportunity to derive a definitive set of Mokken scaling items and also to take a set of samples of varying sizes to test the effect – empirically – of sample size on different outcome measures. Specifically, the present study aims to test the effect of varying sample sizes in a large database of real data showing a single Mokken scale on the PEA, H_s, and the 95% confidence intervals around H_s, H_i and H_{ij}. IIO is outside the scope of this study; for a real data application of IIO, see Meijer and Egberink (2012). We use this work to provide pragmatic recommendations for researchers.

Materials and methods

The WEMWBS

The WEMWBS is a 14-item questionnaire using a series of positive questions enquiring about mental well-being such as 'I've been feeling relaxed' and 'I've been thinking clearly' scored on a five-point scale from 'None of the time' to 'All of the time' (Tennant et al., 2007). The WEMWBS shows good psychometric properties and has been demonstrated to be a unidimensional scale (Deary et al., 2013; Stochl et al., 2012).

Participants

For this study, we used the responses of 7510 participants from a total sample of 8643; 1123 cases were excluded due to missing data. The sample was all 50 years of age with 3,623 males and 3,887 females. The methods of data collection and the demographics of this sample have been described previously by Deary et al. (2013); this was a secondary analysis of a dataset in the public domain made available from The National Child Development Study (1958 cohort) followed up in 2008–2009 at 50 years of age.

Procedures

Data were converted from SPSS to a format suitable for analysis in *R* using package 'foreign' in *R* and then analysed using package 'mokken' (van der Ark, 2007). The data and the R script used to generate the output are freely available at https://osf.io/j27te.

Based on the minimum sample size requirements provided by Straat et al. (2014) in Table 2 of their article, in our case (i.e. $H_i \approx .42$, $r(\theta_1, \theta_2) = 1$ since the WEMWBS consists of one scale, $10 < J < 20$, namely $J = 14$) a minimum sample size of 50 is recommended for mediocre and adequate PEA and a minimum sample size of 250 is recommended for good and excellent PEA. Therefore, we started by taking random samples of size $n = 50$ and 250 from the dataset to study the effect of sample size on scale properties, like PEA, the scalability coefficient (H_s), and the 95% confidence intervals around H_s, H_i and H_{ij}. Based on those results and the sample sizes used in Straat et al. (2014), random samples of size $n = 500$, 600, 750 and 1000 were also taken from the dataset. A bootstrapping procedure was used, in which a thousand random samples were taken, with replacement, at each sample size.

To establish the number of Mokken scales in the data and to obtain the PEA, each sample was analysed using both the automated item selection procedure (procedure AISP) and procedure genetic algorithm (procedure GA) in *R*, since these different searching algorithms can provide different results. The default settings were used in each algorithm. Furthermore, using the TEST option, for all 14 WEMWBS items their different scalability coefficients were obtained (H_s, H_i and H_{ij}) together with their standard errors. The obtained standard errors were used to calculate the 95% confidence intervals for H_s, H_i and H_{ij} coefficients. We proceeded on the assumption that adequate sample size would be achieved at perfect PEA and where neither H_s, nor H_i, nor H_{ij} had 95% confidence intervals which included their respective lower bound values. The number of confidence intervals that included the respective lower bound was averaged across the 1,000 replications, for each scalability coefficient at each sample size.

Ethics Statement

This is a secondary analysis of public domain data from the 1958 Child Development Study and ethical approval was obtained for the original study – initiated in 1958 – by the Centre for Longitudinal Studies, Institute of Education, University College London, UK: http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=724&sitesectiontitle=National+Child+Development+Study (accessed 1 January 2018). Informed consent was obtained from all individual participants included in the study.

Results

Supplementary Table 1 shows the individual item distribution which indicated little skew among the items. Table 1 shows the effect of varying sample size on H_s in terms of the value of H_s and the number of times the 95% confidence intervals around it is below the lower bound of .30. The value of H_s for the total sample with 95% confidence intervals is also provided. The results show that H_s is only slightly affected by the sample size, given the larger standard deviation for $n = 50$ compared to the other sample sizes. The effect of increasing sample size on the 95% confidence intervals is, generally, to narrow these thereby increasing confidence in the strength of the scales. This effect can already be obtained by using a sample size of 250 (or more) instead of 50, since none of the lower bounds of the 95% confidence intervals is below .30.

Table 2 shows the effect of varying sample size on the H_i coefficients and their 95% confidence intervals for each item. The (mean) value of the H_i coefficients become

Table 1. Mean and standard deviation of H_s coefficients and the number of times the lower bound of their 95% confidence intervals was below .30 for the thousand replications per different sample size.

Sample size	Mean H_s	SD H_s	# H_s with 95% CI < .3
50	.49	.07	173
250	.49	.03	0
500	.48	.02	0
600	.48	.02	0
750	.48	.02	0
1000	.48	.02	0

Note: For total sample ($n = 7510$), $H_s = .48$; 95% CI = .47–.49.

Table 2. Mean and standard deviation of H_i coefficients and the number of times the lower bound of their 95% confidence intervals was below .30 for the thousand replications per different sample size.

	$n = 50$		$n = 250$		$n = 500$		$n = 600$		$n = 750$		$n = 1000$	
Item	Mean H_i	# < .3	Mean H_i	# < .3	Mean H_i	# < .3	Mean H_i	# < .3	Mean H_i	# < .3	Mean H_i	# < .3
1	.45 (.11)	592	.45 (.05)	139	.45 (.03)	10	.45 (.03)	7	.45 (.03)	1	.45 (.02)	0
2	.48 (.10)	442	.48 (.04)	37	.48 (.03)	0	.47 (.03)	1	.47 (.03)	0	.47 (.02)	0
3	.51 (.09)	339	.51 (.04)	6	.50 (.03)	0	.50 (.03)	0	.50 (.02)	0	.50 (.02)	0
4	.35 (.12)	879	.34 (.05)	869	.33 (.04)	828	.33 (.03)	782	.34 (.03)	763	.33 (.03)	715
5	.41 (.10)	718	.41 (.05)	298	.41 (.03)	82	.41 (.03)	62	.41 (.03)	20	.41 (.02)	6
6	.51 (.09)	311	.51 (.04)	4	.50 (.03)	0	.50 (.03)	0	.50 (.02)	0	.50 (.02)	0
7	.55 (.09)	192	.54 (.04)	1	.54 (.03)	0	.54 (.02)	0	.54 (.02)	0	.54 (.02)	0
8	.59 (.07)	46	.58 (.03)	0	.58 (.02)	0	.58 (.02)	0	.58 (.02)	0	.58 (.02)	0
9	.48 (.09)	422	.47 (.04)	17	.47 (.03)	0	.47 (.03)	0	.47 (.02)	0	.47 (.02)	0
10	.59 (.07)	43	.58 (.03)	0	.58 (.02)	0	.58 (.02)	0	.58 (.02)	0	.58 (.02)	0
11	.47 (.10)	509	.47 (.05)	62	.47 (.03)	3	.47 (.03)	0	.47 (.03)	0	.47 (.02)	0
12	.43 (.11)	658	.42 (.05)	250	.42 (.03)	27	.42 (.03)	15	.42 (.03)	3	.42 (.02)	0
13	.48 (.09)	450	.47 (.04)	16	.47 (.03)	0	.47 (.03)	0	.47 (.02)	0	.47 (.02)	0
14	.59 (.07)	55	.58 (.03)	0	.58 (.02)	0	.58 (.02)	0	.58 (.02)	0	.58 (.02)	0

Note: Standard deviations are between brackets, # < .3 = number of times the lower bound of the 95% CI was below .3, H_i values are based on TEST for all 14 items, H_i values for $n = 1000$ are equal to H_i values for total sample ($n = 7510$).

Table 3. Mean number of times the lower bound of the 95% confidence intervals of H_{ij} was below 0 and results with regard to PEA (both AISP (left) and GA (right)) for the thousand replications per different sample size (i.e. number of times PEA was smaller than .8, minimal observed PEA value in the 1000 replications, number of times PEA was classified as mediocre, adequate, and excellent).

Sample size	Mean # 95% CI < 0	PEA ≤ .8	Min PEA	# Mediocre	# Adequate	# Excellent
50	18.64 (10.6)	83/78	.50	164/160	319/332	434/430
250	0.46 (0.96)	0/0	.86	2/2	233/233	765/765
500	0.01 (0.08)	0/0	.93	0/0	178/178	822/822
600	0.01 (0.08)	0/0	.93	0/0	155/155	845/845
750	0 (0)	0/0	.93	0/0	104/104	896/896
1000	0 (0)	0/0	.93	0/0	93/93	907/907

Note: Due to the total number of WEMWBS items (i.e. 14) , the qualification 'good' is never applicable (i.e. 13/14 = .93; which is qualified as 'adequate'), mediocre: .8 < PEA ≤ .9, adequate: .9 < PEA ≤ .95, excellent: PEA > .99 (due to the total number of items, 'excellent' always means PEA = 1).

more stable (i.e. lower standard deviation) when increasing sample size. The largest improvement is between $n = 50$ and $n = 250$, beyond $n = 250$ there is little effect on the values of H_i and the reported standard deviation. With increasing sample size, the number of items for which the 95% confidence intervals of H_i includes the lower bound of .30 decreases, reaching zero for almost every item at the largest sample size ($n = 1000$), except for the items 4 and 5. However, those items have the lowest H_i values (i.e. $H_4 = .33$ and $H_5 = .41$).

Table 3 shows the effect of varying sample size on the 95% confidence intervals for H_{ij} and the PEA (for both AISP and GA). With a sample size of $n = 750$ or more for all item pairs the confidence intervals did not include zero, meaning that all H_{ij}s are significantly greater than zero. Beyond sample size $n = 500$, PEA is adequate to excellent. With $n = 750$ and $n = 1000$, PEA is excellent in the majority of the thousand replications.

Discussion

The aim of this paper was to investigate the effect of sample size on PEA and confidence intervals around scaling coefficients using Mokken scaling in a large WEMWBS database with a unidimensional scale. The study is original in that it approaches the effect of sample size on Mokken scaling empirically. Our reference point for this study – both its conception and interpretation – is the study by Straat et al. (2014). However, we acknowledge that a simulation study, where the correlation between two dimensions/scales was one of the manipulated variables in the study, with a single scale study may not be directly comparable. Nevertheless, the pioneering nature of this work and the availability of a sufficiently large database has provided a baseline for real data studies of the influence of sample size on Mokken scaling parameters.

The results with $n = 50$ and 250 showed that we were not able to confirm the minimum sample size required indicated by Straat et al. (2014). Our main observation is, perhaps unsurprisingly, that all outcome measures improved as the sample size was increased. However, there was a differential effect of increasing sample size on the scale properties. H_i, H_s and 95% confidence intervals around H_s are least sensitive to sample size, meaning that their values are the least stable (i.e. highest standard deviation) for $n = 50$; thereafter, increasing sample size quickly led to diminishing returns with respect to this scale property. Confidence intervals around individual item scalability coefficients were more

sensitive to sample size than those around item pair scalability coefficients, only showing no 95% confidence intervals including the lower bound value at $n = 1000$ (with the exception of items 4 and 5 with $H_4 = .33$ and $H_5 = .41$). PEA was very poor at the lowest sample size but a sample of $n = 750$ was required to achieve excellent PEA in the majority of the thousand replications. Therefore, we observe that even where H_i is quite high that – for the WEMWBS at least – larger sample sizes than those indicated by Straat et al. (2014) may be required. Specifically, we would recommend minimum sample sizes of $n = 600$ to achieve at least adequate PEA, and minimal H_i lower bound violations. However, with reference to other outcome measures, particularly scalability of a set of items regardless of PEA, a minimum sample size of $n = 250$ is required. Therefore, based on the present study we propose, where only one cluster of items was identified as a Mokken scale and where the strength of the scale was moderate that, depending on the specific interests of the person conducting a Mokken scaling analysis, the following sample sizes should apply:

- $n \geq 250$ sufficient to establish scalability of the whole scale (H_s)
- $n \geq 500$ sufficient to establish scalability between items pairs (H_{ij}) with minimal lower bound violations
- $n \geq 750$ sufficient to establish adequate per element accuracy (PEA) with minimal lower bound violations of item scalability (H_i)
- $n > 1000$ possibly required to eliminate lower violations of item scalability (H_i)

The extent to which these guidelines are applicable to other scales remains to be tested and should be the focus of further research as outlined below. It should be noted that we did not eliminate all lower violations of H_i for items 4 and 5. The poor performance of item 4 may necessitate its removal from the scale in general use – provided this does not lead to construct underrepresentation. It was also possible that item 4 was affecting the overall performance of the scale and, particularly, item 5 for which all lower violations of H_i were not eliminated. Therefore, we tested the scale in the absence of item 4. The results without item 4 are like the results with item 4 included in the scale, as shown in the Supplementary Tables 2–4. There is one difference, namely PEA is already adequate with a sample size of $n = 250$. However, this does not lead to different guidelines, since for adequate PEA combined with minimal lower bound violations of H_i, a sample size of $n = 750$ would be sufficient, with or without item 4.

This study has demonstrated using real data that Mokken scaling properties are, indeed, sensitive to sample size. For the first time, the specific effect of sample size on the 95% confidence intervals around Loevinger's scalability coefficients has been demonstrated and recommendations for sample sizes for Mokken scaling of the WEMWBS have been made based on these real data. Whether the above recommendations hold for other similar scales remains to be tested.

Therefore, future directions for work on sample size should focus on further real data sets and on data sets with more than one cluster of items. The data in the present study were concerned with a single psychological construct. However, Mokken scales have been observed in other types of data, for example, functional data such as activities of daily living scales, opinion scales and behaviour scales and the effect of sample size on other latent traits, with different levels of correlation among these traits and behaviours. Moreover, the WEMWBS was selected for this study because of its excellent and

previously demonstrated Mokken scaling properties. While it remains to be tested, it is possible to speculate that alternative sets of items with poorer Mokken scaling properties – for example lower item and overall scale scalability, and/or multiple dimensionality, with cross-loading items – may require even larger sample sizes to achieve stability across an acceptable range of parameters.

The use of alternative psychometric methods in health psychology has been illustrated in recent papers (Crutzen & Peters, 2017; Dima, 2018). We feel that the adoption of alternative scaling techniques, such as Mokken scaling, may be especially important in this area given that popular theoretical constructs are typically assessed by very few items (e.g. Theory of Planned Behaviour constructs; Doherty, Dolan, Flynn, O'Carroll, & Doyle, 2017), with a wide range of response formats (Johnston, Benyamini, & Karademas, 2016), and the underlying assumption of classical test theory – that all items are equivalent when assessing constructs, is unlikely to be true. While not the aim of the present paper, the interested reader is encouraged to see a summary of these and other advantages that have been published previously in this issue (Dima, 2018).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Roger Watson http://orcid.org/0000-0001-8040-7625
Iris J. L. Egberink http://orcid.org/0000-0003-0191-2733
Lisa Kirke http://orcid.org/0000-0002-1480-0645
Jorge N. Tendeiro http://orcid.org/0000-0003-1660-3642
Frank Doyle http://orcid.org/0000-0002-3785-7433

References

Crutzen, R., & Peters, G.-J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, *11*, 242–247.

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.

Dima, A. L. (2018). X-raying concepts: A 6-step protocol for scale validation in applied health research. *Health Psychology and Behavioral Medicine*, (in press).

Doherty, S., Dolan, E., Flynn, J., O'Carroll, R. E., & Doyle, F. (2017). Circumventing the "ick" factor: A randomized trial of the effects of omitting affective attitudes questions to increase intention to become an organ donor. *Frontiers in Psychology*, *8*, 821.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*, 337–352.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality and overdetermination. *Educational and Psychological Measurement*, *65*, 202–226.

Johnston, M., Benyamini, Y., & Karademas, E. C. (2016). Measurement issues in health psychology. In Y. Benyamini, M. Johnston, & E. C. Karademas (Eds.), *Assessment in health psychology* (pp. 320–334). Gottingen: Hogrefe.

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

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

Meijer, R. R., & Baneke, J. J. (2001). Analysing psychopathology items: A case for non-parametric item response theory modelling. *Psychological Methods*, *9*, 351–368.

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 Psychological Measurement*, *72*, 589–607.

Meijer, R. 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. (1971). *A theory and procedure of scale analysis*. Den Haag: Mouton/De Gruyter.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York: Springer.

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

Molenaar, I. W., Sijtsma, K., & Boer, P. (2000). *MSP5 for windows: A program forMokken scale analysis for polytomous items*. Groningen: iec ProGAMMA.

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

Stochl, P., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied researchers. *BMC Medical Research Methodology*, *12*, 275. Retrieved from http://www.biomedcentral.com/1471-2288/12/74

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., & Clausen, J. A. (1950). *Measurement and prediction*. Princeton, NJ: Princeton University Press.

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement*, *74*, 809–822.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., … Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes*, *5*, 63. Retrieved from http://www.hqlo.com/content/5/1/63

Van Abswoude, A. A. H., Vermunt, J. K., Hemker, B. T., & van der Ark, L. A. (2004). Mokken scaling analysis using hierarchical clustering procedures. *Applied Psychological Measurement*, *28*, 332–354.

van Alphen, A., Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing*, *20*, 196–201.

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

Watson, R., Thompson, D. R., & Wang, W. (2014). Violations of local stochastic independence exaggerate scalability in Mokken scaling analysis of the Chinese mandarin SF-36. *Health and Quality of Life Outcomes*, *12*, 417. Retrieved from http://www.hqlo.com/content/12/1/149

Watson, R., van der Ark, L. A., Lin, L.-C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing*, *21*, 2736–2746.

Wolf, M. S., Bennett, C. L., Davis, T. C., Marin, E., & Arnold, C. (2005). A qualitative study of literacy and patient response to HIV medication adherence questionnaires. *Journal of Health Communication*, *10*(6), 509–517.