Personality and Individual Differences 45 (2008) 333–335

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Short Communication

A hierarchy of items within Eysenck's EPI

Roger Watson [a,*], Beverly Roberts (nee Shipley) [b], Alan Gow [b], Ian Deary [b]

[a] School of Nursing and Midwifery, The University of Sheffield, Sheffield S10 2TN, UK
[b] Department of Psychology, The University of Edinburgh, Edinburgh, UK

ARTICLE INFO

Article history:
Received 17 December 2007
Received in revised form 8 April 2008
Accepted 28 April 2008
Available online 10 June 2008

Keywords:
Personality
Eysenck Personality Inventory
Neuroticism
Mokken scaling
Latent traits
Psychometrics
Item–response theory

ABSTRACT

Based on the recent finding of a hierarchical scale for Neuroticism in the NEO-Five Factor Inventory, a further personality inventory: the Eysenck Personality Inventory was analysed using the Mokken Scaling Procedure for hierarchical scales. Items from two dimensions of the Eysenck Personality Inventory – Neuroticism and Extraversion – produced reliable hierarchical scales of 12 and five items, respectively, for all subjects but only a scale for Neuroticism for males and females separately. The Neuroticism items ran from items expressing mild to more extreme worry and a single hierarchical scale for Neuroticism, usable with males and females, is suggested. The utility of hierarchical scales in personality measurement is discussed in terms of furthering the theoretical understanding of personality traits and also their practical applications.

1. Introduction

A recent paper in *Personality and Individual Differences* (Watson, Deary, & Austin, 2007) demonstrated that some items within personality trait scales form a hierarchy. Specifically, items from the Neuroticism scale of the NEO-Five Factor Inventory (NEO-FFI; Costa & McRae, 1992) formed a hierarchy whereby participants pooled from several large UK studies ($n = 1028$) more readily endorsed that they were 'stressed' and 'going to pieces' than that they were 'hopeless' and wanted someone to 'solve problems' for them, with other items intervening between these. In the terminology of hierarchical scales, items which are less readily endorsed are referred to as having greater 'difficulty' (Watson et al., 2007) and, thereby, represent a greater amount of the latent trait being measured. In the case of the Neuroticism scale of the NEO-FFI, greater difficulty means higher levels of Neuroticism.

Hierarchies among scale items represents a different – complementary and useful – way of looking at the relationship between item scores and scale total scores, which is the purpose of methods emanating from item response theory. In this way, individuals' scores on a questionnaire – 'the observable data' (Hulin, Drasgow, & Parsons, 1983, p. 14) – can be related to the underlying theoretical construct that is being measured. In this way, scales become more theoretically interesting and useful because of the ways in which individuals 'progress' through a latent trait where individual items in an inventory actually represent the extent to which the latent trait is present.

The method of choice for investigating the possibility of hierarchical scales in multivariate data is Mokken scaling (Mokken & Lewis, 1982). Mokken scaling is based upon item–response theory (IRT), part of the set of modern test theory – as opposed to classical test theory (factor analysis and internal consistency) – and is often compared with its companion technique, Rasch scaling (van Schur, 2003). Mokken scaling has advantages over Rasch scaling because, whilst rigorous in its selection of items for inclusion in scales, it is less restrictive in its assumptions (Meijer, Sijtsma, & Smid, 1990). Both Mokken and Rasch scaling assume local stochastic independence of items and neither tolerate violations of monotone homogeneity and double monotonicity (Meijer et al., 1990). However, Rasch scaling makes more stringent assumptions in line with the fact that it orders items on a ratio or difference basis whereas Mokken scaling does this on an ordinal basis. Nevertheless, according to Meijer et al. (1990), unless very sophisticated applications such as item banking and equating scores across populations is required, Mokken scaling is adequate and, due to its less restrictive assumptions, more inclusive of items. Finally, unlike both classical test theory and Rasch scaling which use 'top down' procedures for item selection, whereby the best set of items conforming to the theory are selected and items conforming least are deleted, Mokken scaling works through a 'bottom up' procedure whereby items conforming best to the assumptions of the model (one pair, initially) are selected and other items then clustered around these (van Schur, 2003).

* Corresponding author. Tel.: +44 (0) 1142269624; fax: +44 (0) 1142229790.
E-mail address: roger.watson@sheffield.ac.uk (R. Watson).

0191-8869/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2008.04.022

The application of Mokken scaling has been greatly facilitated by the development of a computer programme for the analysis: the Mokken Scaling Procedure (MSP; Sijtsma, Debets, & Molenaar, 1990). The details of Mokken scaling and its application using the MSP to data from the NEO-FFI are explained by Watson et al. (2007). Briefly, the MSP provides a series of diagnostics which allows the investigator to establish whether reliable hierarchies of items exist in a multivariate data set and, thereafter, to check that these items have both monotone homogeneity and are doubly monotonous. This establishes that the score on an item increases as the score on the latent trait increases and that the item–response curve for the items that are included in the hierarchical scale do not intersect. The extent to which a set of items is hierarchically scalable is given by Loevinger's coefficient of homogeneity (H) which provides a measure of how often items are found relative to one another in a group of individuals responding to a set of items. A Loevinger's coefficient $\geqslant 0.3$ is considered to indicate the presence of a hierarchical scale with $H \geqslant 0.4$ indicating a strong scale. These values are considered to have stood the test of time over the years since Mokken scaling has been commonly in use (van Schur, 2003). Violations of monotone homogeneity and double monotonicity are detected by the *Crit* diagnostic in the MSP. *Crit* is a combined statistic generated by the MSP which indicates the extent to which H falls below an acceptable level, how many violations of the Mokken model there are and the size of these violations (van Schur, 2003). Items with values of $Crit \geqslant 80$ should be discarded and values $\leqslant 40$ are ideal.

Given the identification of a hierarchical scale for Neuroticism in the NEO-FFI it is important to discover whether or not such hierarchies exist in other widely-used personality inventories. The present study investigates the hierarchical nature of the items within the Neuroticism and Extraversion scales of the Eysenck Personality Inventory (EPI; Eysenck & Eysenck, 1964).

2. Methods

2.1. Participants

2.1.1. EPI

The 9003 (3905 men, 5098 women) participants completing the Eysenck Personality Inventory (EPI) came from the Health and Lifestyle Survey (HALS), which is a UK nationwide sample survey of community-dwelling adults resident in England, Scotland, and Wales. Mean age of the sample was 45.9 years (SD = 17.7). Full details of the study can be seen elsewhere (Cox et al., 1987; Shipley, Weiss, Der, Taylor, & Deary, 2007). Initial sample selection and interviewing took place between 1984 and 1985, where 12,254 addresses were randomly chosen from UK electoral registers. One individual aged 18 years or over was chosen from each household, yielding 9003 interviews (Cox et al., 1987; Shipley et al., 2007). When compared with the 1981 Census the study population was judged to be a reasonably good representation of the general population (Cox et al., 1987). Information was collected at home in three stages. The first stage consisted of an interview questionnaire collecting socio-demographic, health behaviour and health status data. The second stage involved a physiological examination. For the final stage, participants completed a self-report questionnaire that assessed personality and psychiatric status. This was completed at leisure and returned in the post. The EPI (Eysenck & Eysenck, 1964) is a self-report personality inventory consisting of 57 items measuring Extraversion (24 items) and Neuroticism (24 items). The response format of the EPI's items involves marking 'yes' or 'no' to each statement. Nine of the 57 questions formed a lie scale. Scores range from 0 to 24 for each personality trait with higher scores representing higher Neuroticism or Extraversion. The questionnaire was completed at home and returned by post. Test–retest reliabilities of the scale based on normal samples are excellent at 0.84 for Neuroticism and 0.88 for Extraversion with a time lapse of one year between test and retest (Eysenck & Eysenck, 1964). The limitations of performing secondary analysis on this sample are acknowledged; essentially, the data used in the present analysis were not gathered for the purpose of Mokken scaling analysis and, of course, are several years old. However, we consider that neither of these limitations will have a significant effect on the outcomes generated for the present study.

2.2. Statistical analysis

Data from participants who completed the EPI were entered into an SPSS version 15.0 database. These data were saved in a tab-delimited form with the 'write variable names to spreadsheet' option deselected. Any individuals with missing data were then removed from the database to prepare the data for analysis using the MSP, leaving 5795 participants who completed the EPI. Mokken scaling was carried out on resulting data sets (all participants and all items) using the MSP version 5.0 for Windows (Molenaar & Sijtsma, 2000). The procedure for running the MSP and selecting items has already been described by Watson et al. (2007). The distribution characteristics of Mokken scales are not established. Therefore, formal power analysis cannot be performed for estimation of sample size. However, in multivariate analysis, adequacy of sample size is normally judged by the ratio of items to subjects and, in the present study, this ratio was in the hundreds. We consider that the sample size – being unusually large, in the thousands – in the present study was more than adequate to detect any scaling properties of the EPI.

3. Results

The EPI produced two Mokken scales for all the data as shown in Table 1. One scale was composed entirely of 12 of the EPI's 24 items related to Neuroticism. The other scale of EPI items was composed entirely of five of the EPI's 24 items related to Extraversion. Both scales demonstrated acceptable Mokken scaling parameters of scalability ($H > 0.4$), reliability ($\rho \approx 0.7$) and probability

Table 1
Mokken scale of EPI for all subjects checked for violations of monotone homogeneity and double monotonicity

EPI Item (or short paraphrase of item)	Mean	*H*
Neuroticism scale[b]		
43. Have nightmares	1.11	0.33
35. Get attacks of shaking	1.12	0.38
47. Nervous person	1.23	0.42
26. Think you are tense	1.24	0.41
23. Often troubled by guilt	1.29	0.37
52. Feel inferior	1.30	0.38
2. Need friend to cheer you up	1.39	0.37
40. Worry about awful things that may happen	1.45	0.38
9. Feel 'just miserable' sometimes	1.45	0.36
16. Feelings easily hurt	1.60	0.45
50. Easily hurt when people fault your work	1.60	0.42
14. Worry about things shouldn't have done	1.63	0.46
Extraversion scale[c]		
10. Do anything for a dare	1.11	0.44
53. Get some life into dull party	1.35	0.57
29. Quiet when with other people[a]	1.52	0.41
27. Other people think lively	1.56	0.49
51. Hard to enjoy lively party[a]	1.70	0.53

[a] These items are reverse scored.
[b] Scale $H = 0.40$; $p < 0.001$; $\rho = 0.82$.
[c] Scale $H = 0.49$; $p < 0.001$; $\rho = 0.69$.

($p < 0.05$). Items in the Neuroticism scale run from mild expressions of Neuroticism such as worrying and being easily hurt at the lower end of the scale to having attacks of shaking and having nightmares at the higher end of the scale. Items in the Extraversion scale run from liveliness at the low end of the scale to being the 'life and soul of a party' and 'doing anything for a dare' at the higher end. The EPI data were run separately for males and females and this produced a Neuroticism scale with 13 items for males and 14 items for females. There were no further scales when male and female data were run separately. Across the three hierarchical Neuroticism scales (for all subjects, men, and women) there were seven items in common, all appearing in the same order in both male and female scales and these are shown in Table 2. This final 7 item scale has a high coefficient of scalability ($H = 0.40$) and acceptable reliability ($\rho = 0.71$) and probability ($p < 0.0014$).

4. Discussion

Mokken scaling has again proved to be a useful analytical technique for demonstrating that personality traits, at least in the case of the EPI, can be measured hierarchically. These novel findings complement the more usual factor-analytic approaches to the traits. It is also interesting to note, in the case of the EPI, that items from two personality dimensions form hierarchies and that there was no overlap of items between dimensions in the hierarchical scales. It was already known that the developers of the EPI produced robust scales in which the relationship between items and latent traits is distinct. It appears that, as with the NEO-FFI (Watson et al., 2007), they have also, and perhaps inadvertently, produced scales in which the items related to these latent traits are hierarchical. In reporting these results we acknowledge that any item pool from which unidimensional scales is constructed may have multiple dimensions and, in agreement with Van Abswoude, Vermunt, Hemker, and van der Ark (2004), we recognise the drawback (p. 332) that the "item construction and scale selection may not find the dominant underlying dimensionality of the responses to a set of items". Further clustering procedures are recommended. Nevertheless, the present scale is a strong one ($H = 0.40$), meaning that there were minimal violations of Guttman scalability in the data, i.e. the relative ordering of items by participants. A crucial test of any such scale is its face validity in psychological terms: i.e., does it make sense? We believe that the ordering of the EPI items as extracted in the present study makes psychological sense in terms of the relative ordering of the items that constitute neuroticism, and that the formal statistical testing of this ordering adds new validating information to Eysenck's scales.

Table 2
Final Mokken scale for neuroticism

EPI item (or short paraphrase of item)	Mean	H
43. Have nightmares	1.11	0.33
35. Get attacks of shaking	1.12	0.38
47. Nervous person	1.23	0.42
26. Think you are tense	1.24	0.41
2. Need friend to cheer you up	1.39	0.37
40. Worry about awful things that may happen	1.45	0.38
16. Feelings easily hurt	1.60	0.45

Scale $H = 0.40$; $p < 0.0014$; $\rho = 0.71$; Mean = 9.13; Standard deviation = 1.76; Skewness = 0.69; Kurtosis = −0.23.

In addition to providing measures of the extent to which a latent trait is present, the added value of Mokken scaling is that the presence of particular aspects of personality traits also indicates what other aspects of the trait are likely to be present. For example, with reference to Neuroticism, if someone is plagued by feelings of guilt then they are also likely to feel inferior and to have their feelings easily hurt. We also describe a strong Mokken scale with items that are common to both males and females which demonstrates the same properties, in terms of ordering of items, as the scale from the combined male and female data. Therefore, we present a scale that is applicable, equally, to males and females. It should be noted that the scale we present does not contain any items related to physical aspects of Neuroticism that were included in the EPI such as 'shaking', 'palpitations' and 'aches and pains' and, indeed, by Eysenck, Eysenck, and Barrett (1985) in a revised shorter derivative scale, the Eysenck Personality Questionnaire. It appears from this analysis that Extraversion, as measured by the EPI, may not be a trait applicable to both men and women; both display extraversion but the scales in men and women are comprised of different items. This may be an artefact of the present analysis and requires further investigation.

Mokken scaling adds theoretical information and possibly practical utility to personality scales. It is theoretically valuable because it finds a reliable hierarchy of phenomenology (mostly self-reported) on trait dimensions. It tells us more about what the trait 'feels' like at different points on the scale. It could be practically valuable too. Adaptive testing is useful in cognitive assessment. If more work were done to find robust hierarchies of items on personality scales, then adaptive testing of personality traits could be possible also.

References

Costa, P. T., & McRae, R. R. (1992). *Revised NEO personality inventory and NEO-five-factor inventory: Professional manual*. Odessa FL, USA: Psychological Assessment Resources, Inc.

Cox, B. D., Blaxter, M., Buckle, A. L. J., Fenner, N. P., Golding, J. F., Gore, M., et al. (1987). *The health and lifestyle survey: A preliminary report*. London: Health Promotion Trust.

Eysenck, S. B., & Eysenck, H. J. (1964). An improved short questionnaire for the measurement of extraversion and neuroticism. *Life Sciences, 305*, 1103–1109.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences, 6*, 21–29.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Introduction to item response theory. In C. L. Hulin, F. Drasgow, & C. K. Parsons (Eds.), *Item response theory* (pp. 13–74). Homewood, Illinois: Dow Jones-Irwin.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and Rasch approach to IRT. *Applied Psychological Measurement, 3*, 283–298.

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. (2000). *MSP5 for Windows*. Groningen: iec ProGAMMA.

Shipley, B. A., Weiss, A., Der, G., Taylor, M. D., & Deary, I. J. (2007). Neuroticism, extraversion and mortality in the UK Health and Lifestyle Survey: 21 year prospective cohort study. *Psychosomatic Medicine, 69*, 923–931.

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken scaling analysis for polychotomous items: Theory, a computer programme and an empirical application. *Quality and Quantity, 24*, 171–188.

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

van Schur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*, 139–163.

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 Differences, 43*, 1460–1469.