Qual Life Res (2011) 20:1005–1010
DOI 10.1007/s11136-010-9838-7

BRIEF COMMUNICATION

How useful are the SF-36 sub-scales in older people? Mokken scaling of data from the HALCyon programme

Gita D. Mishra · Catharine R. Gale · Avan Aihie Sayer · Cyrus Cooper · Elaine M. Dennison · Lawrence J. Whalley · Leone Craig · Diana Kuh · Ian J. Deary · The HALCyon Study Team

Accepted: 20 December 2010 / Published online: 12 January 2011

Abstract

Purpose To evaluate two psychometric properties of SF-36, namely unidimensionality and reliability.

Methods The data are from three cohorts in the HALCyon collaborative research programme into healthy ageing: Aberdeen Birth Cohort 1936 ($n = 428$), Hertfordshire Ageing Study ($n = 358$) and Hertfordshire Cohort Study ($n = 3{,}216$). The Mokken scaling model was applied to each sub-scale of SF-36 to evaluate unidimensionality as indicated by scalability. The lower bound for internal consistency reliability was determined by Cronbach's alpha.

Results All six sub-scales of SF-36, with the exception of general health (GH) and mental health (MH), demonstrated strong scalability ($0.5 \leq H < 1$). The results were consistent across all 3 cohorts. Both GH and MH showed medium scalability ($0.4 \leq H < 0.55$), although individual items 'sick easier..', 'as healthy as..' and 'expect to get worse' of the GH sub-scale and 'nervous', 'happy' in the MH sub-scale had low scalability ($H < 0.4$) in the oldest cohort (aged 73–83). Cronbach's alphas for all sub-scales were between 0.70 and 0.92.

Conclusions The unidimensionality and reliability of the sub-scales of SF-36 are sufficient to make this a useful measure of health-related quality of life in older people. Caution is needed when interpreting the results for GH and MH in the oldest cohort due to the poor unidimensionality.

Keywords SF-36 · Psychometric properties · Unidimensionality · Reliability · Cronbach's alpha · Mokken scaling

Introduction

The Medical Outcomes Study (MOS) Short-Form General Health Survey (SF-36) is a self-reported multidimensional measure of general health status or quality of life [1]. It produces eight scales of health status: physical functioning (PF), role limitation caused by physical health problems (RP), bodily pain (BP), general health perceptions (GH), vitality (VT) for energy levels and fatigue, social functioning (SF), role limitation because of emotional problems (RE) and mental health (MH). Lower scores on these scales reflect poorer health. The number of items in each sub-scale varies from two (SF and BP) to ten (PF), and therefore, sub-scales such as SF may exhibit more measurement errors than those with more items. Nevertheless, the SF-36 has been tested for use in the United Kingdom [2] and shown considerable evidence for construct validity and reliability of the scores.

G. D. Mishra (✉) · D. Kuh
MRC Unit for Lifelong Health and Ageing, University College London, 33, Bedford Place, London WC1B 5JU, UK
e-mail: g.mishra@nshd.mrc.ac.uk

C. R. Gale · A. A. Sayer · C. Cooper · E. M. Dennison
MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK

C. R. Gale · I. J. Deary
Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, University of Edinburgh, Edinburgh, UK

L. J. Whalley
Geriatric Medicine Unit, University of Edinburgh, Royal Victoria Hospital, Edinburgh, UK

L. Craig
Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK

The suitability of SF-36 for use in older people is the subject of ongoing discussion due to the relatively poor levels of item response, including the inability to compete SF-36 with increasing age and the relevance of the items to older people [2–4]. Limited research has been done on variations of the psychometric properties of the scores in older people. For instance, if a scale is found to have hierarchical properties, then this indicates that items are ordered relative to one another along the latent trait being measured [5]. This property can also be interpreted as measuring the extent to which items in the scale are measuring the same trait or property (unidimensionality). Researchers have sometimes inappropriately used Cronbach's alphas to represent unidimensionality [6, 7] when it is known that alpha can take a high value even when the set of items measures several unrelated latent constructs [8, 9]. A non-parametric item response model, known as Mokken scaling, can provide a useful alternative to determine whether hierarchical scales exist in the collection of items and hence obtain a measure of 'unidimensionality' [10]. A detailed description of hierarchical properties and the assumptions underlying of Mokken scaling is provided elsewhere [5, 11].

The other characteristic of interest is whether the measurements are precise or internally consistent (reliability). This measures the variation of the test results if the test is repeated under comparable circumstances. Data from the Dutch National Study have been used to examine these properties, but this was for a wide age range of adults using a Dutch version of the scale [11].

HALCyon–Healthy Ageing across the Life Course—is a collaborative research programme using data from up to nine UK narrow age range cohorts to examine how factors across the life course influence healthy ageing in older people. We used data from three of these cohorts to evaluate the unidimensionality and internal consistency (reliability) of the SF-36 sub-scales.

Methods

This study uses data from the Hertfordshire Ageing Study, the Hertfordshire Cohort Study and the Aberdeen Birth Cohort 1936 [12–14] .

The Hertfordshire Ageing Study (HAS)

From 1911 to 1948, details of each birth in Hertfordshire, United Kingdom including birthweight, was recorded by the attending midwife and held in central registers. The National Health Service Central Register was used to trace singleton infants born to married mothers. Those still living in Hertfordshire who had been born between 1920 and 1930 were invited to take part in research into life course influences on ageing [12]. Of 1,428 people invited to participate in the initial study in 1994–1995, 824 (58%) agreed to a home interview and 717 attended a clinic for further assessments. In 2003–2005, a follow-up study was carried out when the participants were aged between 72 and 83 years. In total, 359 men and women (60% of those surviving) were interviewed, during which 349 completed the SF-36.

The Hertfordshire Cohort Study (HCS)

In 1998–2004, men and women born in Hertfordshire between 1931 and 1939 and still living in the county were recruited to a new, larger cohort study in order to evaluate interactions between the genome, the intrauterine and early postnatal environment, and adult diet and lifestyle in the aetiology of chronic disorders in later life [13]. Of 6,099 men and women approached, 3,225 (53%) agreed to be interviewed at home. As part of the interview, 3,215 people completed the SF-36.

The Aberdeen Birth Cohort 1936 (ABC)

On 4 June, 1947, as part of the Scottish Mental Survey, all children born in 1936 who attended school in Scotland sat a test of mental ability, a version of the Moray House Test No. 12. Records of these tests on the 70,805 children who took part were preserved. In 1999–2001, 567 of these people who were living in the Aberdeen area were invited to take part in a study of life course influences on cognitive ageing [14]. Of the 506 (89%) who took part, 429 completed the SF-36.

Statistical methods

The Mokken scaling model, a non-parametric model, was applied to each sub-scale of SF-36 in order to evaluate unidimensionality. The extent to which a set of items is scalable is given by the Loevinger's coefficient (H), which is a measure of how well the set of items meet the hierarchical criteria of Mokken scales. *H* will be calculated for individual items in terms of the number of items that violate hierarchical assumptions relative to other items, and an overall *H* can be calculated for a set of items. The lower bound for reliability was determined by Cronbach's alpha [11]. Mokken scaling was performed using the module *MSP* in STATA 10.

Results

Table 1 presents summary statistics of each of the three British cohorts. HAS participants were, on average,

Table 1 Mean (SD) and median scores SF-36 sub-scales in three British cohorts

	Hertfordshire Ageing Study			Hertfordshire Cohort Study			Aberdeen Birth Cohort 1936		
	Males (n = 208)	Females (n = 150)	All (n = 358)	Males (n = 1,682)	Females (n = 1,540)	All (n = 3,216)	Males (n = 206)	Females (n = 222)	All (n = 428)
Age in years, mean (SD)	76.2 (2.4)	75.8 (2.1)	76.0 (2.3)	65.6 (2.9)	66.6 (2.7)	66.1 (2.9)	64.6 (0.9)	64.7 (1.0)	64.6 (0.9)
	76.0	76.0	76.0	65.7	66.4	66.1	64.7	64.8	64.7
Sub-scale									
Physical functioning	67.9 (26.6)	59.7 (27.6)	64.5 (27.3)	83.9 (21.1)	75.4 (23.9)	79.8 (22.9)*	76.7 (25.2)	73.9 (24.5)	75.2 (24.8)
	75.0	65.0	70.0	90.0	85.0	90.0	85.0	80.0	85.0
Role physical	63.1 (42.6)	58.8 (42.4)	61.3 (42.5)*	83.1 (32.8)	77.8 (37.0)	80.5 (34.9)	78.5 (37.7)	80.9 (36.5)	79.7 (37.1)
	75.0	75.0	75.0	100	100	100	100	100	100
Bodily pain	71.2 (24.9)	64.0 (25.6)	68.2 (25.4)	75.6 (23.3)	70.2 (25.3)	73.0 (24.4)	73.6 (23.9)	72.9 (24.0)	73.2 (23.9)
	74.0	62.0	72.0	84.0	72.0	72.0	74.0	84.0	74.0
General health	63.6 (20.8)	63.9 (20.0)	63.8 (20.5)*	71.3 (19.5)	70.8 (20.3)	71.1 (19.9)	68.9 (20.7)	73.5 (19.2)	71.3 (20.1)
	67.0	67.0	67.0	75.0	75.0	75.0	72.9	77.0	76.0
Vitality	63.0 (19.9)	56.6 (17.4)	60.3 (19.2)*	69.1 (18.6)	63.6 (19.2)	66.5 (19.1)*	69.9 (18.8)	67.6 (19.4)	68.7 (19.1)
	65.0	55.0	60.0	70.0	65.0	70.0	75.0	70.0	73.3
Social functioning	83.8 (23.4)	79.9 (23.8)	82.2 (23.6)*	91.2 (18.0)	87.7 (21.7)	89.6 (19.9)	89.5 (20.5)	88.4 (21.1)	88.9 (20.8)
	100	87.5	100	100	100	100	100	100	100
Role emotion	81.7 (35.2)	72.5 (38.7)	77.8 (36.9)*	93.3 (21.9)	89.4 (27.4)	91.4 (24.8)*	88.3 (28.2)	86.8 (30.2)	87.5 (29.2)
	100	100	100	100	100	100	100	100	100
Mental health	79.8 (14.9)	72.3 (18.6)	76.6 (16.9)*	82.7 (14.6)	77.0 (15.6)	80.0 (15.3)	82.2 (15.2)	79.4 (16.6)	80.9 (16.0)
	84.0	76.0	80.0	88.0	80.0	84.0	88.0	84.0	85.0

* $P \leq 0.01$ from multiple linear regression model for each of the eight sub-scales of SF-36 adjusted for age, sex and cohort (reference category was the Aberdeen Birth Cohort)

Table 2 Scalability (as assessed by Loevinger's H and reliability (as assessed by Cronbach alpha) of SF-36 sub-scales in three British cohorts

Scale	Hertfordshire Ageing Study ($n = 349$)		Hertfordshire Cohort Study ($n = 3{,}215$)		Aberdeen Birth Cohort 1936 ($n = 475$)	
	H	Reliability	H	Reliability	H	Reliability
Physical functioning	0.65	0.92	0.71	0.92	0.70	0.92
Role physical	0.77	0.90	0.80	0.91	0.83	0.92
Bodily pain	0.79	0.83	0.79	0.79	0.82	0.84
General health	0.42	0.74	0.46	0.77	0.54	0.81
Vitality	0.55	0.80	0.59	0.83	0.60	0.83
Social functioning	0.57	0.71	0.71	0.80	0.77	0.85
Role emotion	0.78	0.87	0.77	0.86	0.81	0.87
Mental health	0.42	0.76	0.43	0.77	0.51	0.82

10 years older than the other two cohorts. Compared with the youngest cohort, ABC 1936, HAS participants scored lower on six of the eight sub-scales of the SF-36.

The scalability and reliability coefficients of the sub-scales of SF-36 are given in Table 2. Using established guidelines for interpretation, a Mokken scale of $0.4 \leq H < 0.5$ indicating medium scalability and $0.5 \leq H < 1$ strong scalability [5, 11], then all sub-scales of all cohorts have strong scalability ($H > 0.6$), with the exception of the GH and MH sub-scales ($H < 0.5$) in the Hertfordshire cohorts.

In terms of internal consistency (reliability), the Cronbach's alphas fall between 0.7 and 0.8 for all sub-scales, except for PF (0.92) and RP where they are between 0.90 and 0.92.

Table 3 presents the scalability and reliability of the individual items of SF-36. With the exception of the items in the general health and mental health sub-scales of the oldest cohort (HAS), items in the other sub-scales and for the other cohorts have an H coefficient of greater than 0.4. In particular, across all the cohorts, items in RP, BP and RE exhibited an H coefficient of greater than 0.7. As for the overall results, the Cronbach's alphas fall between 0.7 and 0.8 for all sub-scales, except for PF (greater than 0.9) and RP (between 0.8 and 0.9) and some items in the GH sub-scale (less than 0.7) especially for the oldest cohort (HAS).

Table 3 Scalability (as assessed by Loevinger's H) and reliability (as assessed by Cronbach alpha) of individual SF-36 items in three British cohorts (as assessed by Hertfordshire Ageing Study (*HAS*), Hertfordshire Cohort Study (*HCS*), Aberdeen Birth Cohort 1936 (*ABC*))

Items	HAS		HCS		ABC	
	H	Reliability	H	Reliability	H	Reliability
Physical functioning						
Vigorous activities	0.45	0.93	0.68	0.92	0.53	0.93
Moderate activities	0.69	0.91	0.71	0.91	0.71	0.91
Lifting/carrying	0.67	0.91	0.68	0.91	0.69	0.91
Climbing several flights	0.67	0.91	0.69	0.91	0.70	0.92
Climbing one flight	0.75	0.90	0.75	0.91	0.73	0.91
Bending/kneeling	0.60	0.91	0.63	0.92	0.68	0.92
Walking > 1 mile	0.67	0.91	0.73	0.91	0.72	0.91
Walking several blocks	0.73	0.90	0.76	0.91	0.78	0.91
Walk one block	0.68	0.91	0.77	0.91	0.78	0.91
Bathe, dress	0.65	0.91	0.68	0.92	0.71	0.92
Role physical						
Cut down amount of time	0.79	0.89	0.86	0.90	0.88	0.90
Accomplishes less	0.77	0.85	0.79	0.87	0.81	0.90
Limited kind of work	0.79	0.85	0.79	0.87	0.83	0.89
Had difficulty	0.74	0.88	0.78	0.88	0.80	0.90
Bodily pain						
Had bodily pain	0.79	0.83	0.79	0.80	0.82	0.84

Table 3 continued

Items	HAS		HCS		ABC	
	H	Reliability	H	Reliability	H	Reliability
Pain interfere	0.79	0.83	0.79	0.80	0.82	0.84
General health						
General health is..	0.48	0.67	0.53	0.70	0.63	0.75
Sick easier	0.38	0.72	0.44	0.75	0.53	0.79
As healthy as	0.38	0.71	0.49	0.70	0.51	0.78
Expect get worse	0.34	0.73	0.31	0.79	0.45	0.82
Is excellent	0.51	0.63	0.55	0.67	0.62	0.73
Vitality						
Full of pep	0.57	0.74	0.60	0.78	0.60	0.80
Lot of energy	0.55	0.76	0.59	0.78	0.60	0.80
Worn out	0.54	0.76	0.57	0.79	0.59	0.79
Feel tired	0.54	0.75	0.60	0.77	0.62	0.78
Social functioning						
Social extent	0.57	0.71	0.71	0.84	0.77	0.85
Social time	0.57	0.71	0.71	0.84	0.77	0.85
Role emotion						
Cut down time	0.75	0.81	0.78	0.80	0.79	0.80
Accomplished less	0.86	0.77	0.82	0.76	0.87	0.82
Not careful	0.73	0.85	0.72	0.84	0.79	0.81
Mental health						
Nervous	0.34	0.76	0.37	0.76	0.47	0.79
Down in dumps	0.50	0.68	0.48	0.71	0.57	0.76
Peaceful	0.45	0.71	0.45	0.72	0.53	0.78
Blue/sad	0.48	0.68	0.49	0.70	0.58	0.75
Happy	0.35	0.76	0.41	0.75	0.43	0.82

Discussion

To our knowledge, this is the first report of these psychometric properties of SF-36 using data specifically obtained from older people. We found that with the exception of the GH and MH, which showed medium scalability, all other sub-scales demonstrate strong scalability. The results are consistent across all the three cohorts, providing evidence that items in each of the eight sub-scales have measured the same trait or property. There was no evidence of regional differences between Hertfordshire and Aberdeen in terms of unidimensionality.

Individual items 'I seem to get sick a little easier than other people', 'I'm as healthy as anyone I know' and 'I expect my health to get worse' of the general health sub-scale and the 'have you been a very nervous person', 'have you been a happy person' of the mental health sub-scale have low scalability in the oldest cohort. These items may be modified or removed from the sub-scales for the older population as they may have difficulty in responding to or endorsing these items.

In terms of reliability, the Cronbach's alphas are high enough for research purposes (between 0.7 and 0.8) for all sub-scales, while for physical functioning (0.92) and role limitation caused by physical health problems (0.90, 0.92) they reach levels high enough for clinical application (decisions about individuals). These results broadly concur with those found from a study using a Dutch language version of SF-36 [11], with the exception that we have also found the reliability of the role physical sub-scale to be high enough for clinical application.

Conclusion

The consistency of our findings across three cohorts supports the usefulness of all the sub-scales for research purposes and suggests that the sub-scales physical functioning and role limitation caused by physical health problems could also be used for clinical purposes. Our results suggest that caution is needed when interpreting results for sub-scales with poor unidimensionality. There is some scope, therefore, for modification of items in the general health and mental health sub-scales of SF-36 for older populations (aged 73+).

Acknowledgments The HALCyon collaborative research programme is funded by the New Dynamics of Ageing (NDA) programme, a joint seven-year initiative of five UK Research Councils. Follow-up of the Aberdeen Birth Cohort 1936 was supported by a Wellcome Trust career development award to Professor L Whalley. The Hertfordshire Cohort Study was funded by the Medical Research Council, the British Heart Foundation, Arthritis Research Campaign, the National Osteoporosis Society, Wellcome Trust and the University of Southampton. The Hertfordshire Ageing Study was funded by the Medical Research Council and the University of Southampton.

References

1. Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health survey manual and interpretation guide*. Boston: The Health Institute, New England Medical Center.
2. Brazier, J. E., Harper, R., Jones, N. M., et al. (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ, 305*, 160–164.
3. Hayes, V., Morris, J., Wolfe, C., & Morgan, M. (1995). The SF-36 health survey questionnaire: is it suitable for use with older adults? *Age and Ageing, 24*, 120–125.
4. Parker, S. G., Bechinger-English, D., Jagger, C., Spiers, N., & Lindesay, J. (2006). Factors affecting completion of the SF-36 in older people. *Age and Ageing, 35*, 376–381.
5. Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine, 38*, 575–579.
6. McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.
7. Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.
8. Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.
9. Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2006). Estimating generalizability to a universe of indicators that all have an attribute in common: A comparison of estimators for omega. *Applied Psychological Measurement, 30*, 121–144.
10. Mokken, R. J., & Lewis, C. (1982). Rejoinder to the Mokken scale: A critical discussion. *Applied Psychological Measurement, 6*, 417–430.
11. van der Heijden, P. G., van Buuren, S., Fekkes, M., Radder, J., & Verrips, E. (2003). Unidimensionality and reliability under Mokken scaling of the Dutch language version of the SF-36. *Quality of Life Research, 12*, 189–198.
12. Syddall, H. E., Simmonds, S. J., Martin, H. J., et al. (2010). Cohort profile: The hertfordshire ageing study (HAS). *International Journal of Epidemiology, 39*, 36–43.
13. Syddall, H. E., Aihie, S. A., Dennison, E. M., Martin, H. J., Barker, D. J., & Cooper, C. (2005). Cohort profile: the hertfordshire cohort study. *Int J Epidemiology, 34*, 1234–1242.
14. Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology, 86*, 130–147.