RESEARCH ARTICLE

Psychometric Evaluation of the Brunei-Malay SF-36 version 2 Health Survey

Mohammed M Alhaji[1]*, Nur Hanisah Johan[2], Sharima Sharbini[1], Mas RW Abdul Hamid[1], Muhammad A M Khalil[2], Jackson Tan[2], Lin Naing[1], Nik A A Tuah[1,3]

Abstract

Objectives: To culturally adapt the Short Form Health-36 version 2 (SF-36v2) into the Brunei-Malay context and determine its reliability and validity for measuring health-related quality of life (HRQOL) in healthy individuals and patients with chronic kidney disease in Brunei Darussalam. **Methods:** An iterative multistep strategy involving setting up a bilingual expert panel, pretesting, text revision and back translation was used to prepare the Brunei-Malay SF-36v2 as an adaptation from the Malaysian-Malay SF-36v2. The Brunei-Malay SF-36v2 was then self-administered to a sample of healthy individuals (n=95) and predialysis chronic kidney disease outpatients (n=95) resident in Brunei. The mean (SD) age of the participants was 46.6 (17.8) years. **Results:** Data completion rate was 100% with minimal floor effects (≤0.21) in all the 8 domains and >15% ceiling effects in 3 of the 8 domain scales. Cronbach's alpha was >0.70 for all the 8 domain scales. Scaling success was 100% for convergent validity, with 100% item discriminant validity for all domain scales except Social Functioning (94%), Mental Health (85%) and General Health (85%). Principal component analysis of the two-factor dimension explained 68% overall variance and accounted for 81% reliable variance, but the exact SF-36 two-factor summary constructs in the standard algorithm were not replicated in the Bruneian population. **Conclusions:** The Brunei-Malay SF-36v2 is a valid and reliable instrument for measuring HRQOL in healthy individuals and patients with chronic kidney disease in Brunei. The summary scales should, however, be interpreted with caution. Further studies should be carried out to assess additional psychometric properties of the Brunei-Malay SF-36v2.

Keywords: HRQOL- SF-36- validity- reliability- Asia

Introduction

The Medical Outcome Study (MOS) Short Form (SF-36) health survey is the most widely used generic patient-reported outcome measure (PROM) tool for assessing functional status or health-related quality of life (HRQOL) in individuals with disease conditions (Cons et al., 2000; Mark, 2011; Ware, 2000). Its reliability and validity have also been reported in several culturally-different populations (Mark, 2011).

The SF-36 uses 35 multidimensional items to assess health perception in 8 domains namely: Physical Functioning (PF), Role Limitations due to Physical Health (RP), Role Limitations due to Emotional Problems (RE), Vitality (VT), Mental Health (MH), Social Functioning (SF), Bodily Pain (BP) and General Health (GH) (Mark, 2011; McHorney et al., 1993; Ware, 2000). Two summary scales, the Physical Component Summary (PCS) and Mental Component Summary (MCS), are derivable from the 8 domain scales. While PCS is derived from PF, RP, BP and GH, Mental Component Summary (MCS) is derived from VT, SF, RE and MH domains (Ware, 2000).

Despite its applicability and adaptability to diverse population and languages, noticeable dimensional differences in SF-36 have been reported (Hann and Reeves, 2008). The differences have been attributed to cultural and lexical differences existing between populations (Tseng et al., 2003), thereby typically necessitating the need for validating health survey instruments before use in 'new' populations (Hobart et al., 2001).

Considering that the SF-36 had not been previously validated in Brunei Darussalam, this study was carried out to culturally adapt the SF-36v2 into Brunei-Malay context and test its reliability and validity for measuring HRQOL in healthy individuals and patients with chronic kidney disease (CKD).

[1]PAPRSB Institute of Health Sciences, University Brunei Darussalam, [2]RIPAS Hospital, Bandar Seri Begawan, Brunei Darussalam, [3]Primary Care and Public Health Department, Imperial College London, United Kingdom. *For Correspondence: mohalhaj20@ yahoo.com.sg

Materials and Methods

Expert Panel (for adaptation and revision)

An English-Malay bilingual panel comprising of 5 professionals in Public Health, Epidemiology, Microbiology, Malay Literature, and Social Care revised the Malaysian Malay SF-36v2 (standard 4-week recall version) for context and lexical adaptation into Brunei-Malay. Face validity was determined, semantics and sentence structure were also reviewed. Problematic words and phrases that are not applicable in Brunei context were replaced with contextually synonymous terms, preferring contextual equivalence over literal equivalence (Bullinger et al., 1998; WHO, 2017). The panel also reviewed the comments from the pretesting of the Brunei-Malay SF-36v2 (Figure 1).

Pretesting
Participating Pretest (Focus Group Discussion)

The revised SF-36(i) was then put forth for the pretest. Eight adults Malay residents in Brunei who were able to read and write Malay were purposively invited from the general public to participate in the participatory pretesting (Barribeau et al., 2012). A Focus Group Discussion (FGD) session, facilitated by 2 trained facilitators, was convened to obtain feedback on sentence structure and general understandability of the SF-36 (i) (Tarrant et al., 2014). The participants were each given a print copy of the SF-36 (i) and asked to make comments while the discussion was ongoing. Participants' feedbacks were tape-recorded and transcribed verbatim (Tarrant et al., 2014). Their written comments were also retrieved and reviewed by the expert panel, from which SF-36 (ii) was drafted.

Undeclared Pretesting

The SF-36 (ii) was then self-administered to randomly selected adults (n=12), as would in a standard study (Barribeau et al., 2012). For this part, additional open-ended questions on comprehensibility and level of difficulty encountered in the Brunei-Malay SF-36v2 were added. The feedback was further reviewed by the panel and the final SF-36 (iii) (referred as Brunei-Malay SF-36v2 or simply SF-36) was drafted for onward validation in a sample of healthy and CKD population in Brunei Darussalam. This version was back-translated to the English language, proofread by a third party bilingual expert, and compared with the original English version of SF-36v2 to ensure that original context was maintained (WHO, 2017) (Figure 1).

Sampling, Data Collection for Validation

One-ninety (n=190) adult participants (95 healthy individuals and 95 predialysis CKD patients) self-administered the Brunei-Malay SF-36v2. The healthy participants were purposively recruited from the general population and the CKD patients were randomly enrolled from Rimba Dialysis Center (RDC) Clinic between July and August 2017. Heterogeneous sample was considered, as suggested (Gie Yong and Pearce, 2013).

Item Scoring and Calibration

Ten items (1, 6, 7, 8, 9a, 9e, 9d, 9h, 11b and 11d) were reverse-coded, with adjusted calibrations in 2 items (GH1 and BP2) (Newvine, n.d). Scale raw scores were computed by simple summation, without standardization, from their derivative item scores. The computed raw scale scores were then converted to a 0-100 linear scale using the formula "(Actual raw domain score - Lowest possible raw score) / (Possible raw score range) * 100" (Newvine, n.d).

Psychometric Statistical Analysis

Data were analyzed using the Statistical Package for the Social Sciences (SPSS) Version 16.0. Internal consistency reliability of the Brunei-Malay SF-36 was determined from Cronbach's alpha (α) of each of the 8 domains and item-scale correlation (corrected for overlap). Spearman's correlation was used to determine the correlation between items and scales. Factor analysis was carried out using principal component analysis (PCA) and orthogonal rotation (Varimax) to confirm the hypothesized two-component internal construct validity of the SF-36. Mean with standard deviation (SD), median with interquartile range (IQR), frequency (n) and percentage (%) were shown to describe the sample characteristics and responses. Independent t-test was used to compare SF-36 domain scores of the sample based on health status. Statistical difference was set at $P<0.05$ significance.

Ethical Consideration

The eligibility criteria were: being an adult (>=18 years); residing in Brunei; able to read and understand Brunei-Malay; and a willingness to participate via written consent. Participants' information were collected anonymously and kept confidentially, and participation was entirely voluntary. This study was reviewed and approved by the Medical and Health Research Ethics Committee (MHREC), Ministry of Health, Brunei Darussalam.

Results

Cultural and Lexical Adaptation

Of the 36 item-questions, item 1, 2, 3, 4, 4a, 5, 8, and 9 have either been rephrased or some of their words replaced with contextual equivalents. Answer options for item 2, 3a-j and 11b-d were also either rephrased or replaced. Example, the word "anda" in the Malaysian-Malay SF-36v2 was replaced with its Brunei-Malay equivalent, "awda" in the Brunei-Malay SF-36v2.

All participants (n=12) in the undeclared pretest understood the content of the Brunei-Malay SF-36v2, 9 (75.0%) found it easy, 2 (16.7%) found it average and 1 (8.3%) found it difficult to understand.

Sample Characteristics

The median age of the 'healthy' individuals (n=20) who participated in the pretest (participatory and

Table 1. Sample Demographic Profile (n=190)

	Pretest (n=20)		Validation (n=190)	
	n (%)	Mean (SD)	n (%)	Mean (SD)
Pretesting				
Participatory	8 (40.0)			
Undeclared	12 (60.0)			
Age in years		30.0 (20.0)*⁺	-	46.6 (17.8)
≤45	14 (70.0)		86 (45.3)	
>45	6 (30.0)		104 (54.7)	
Gender (n, %)				
Male	9 (45.0)		92 (48.4)	
Female	11 (55.0)		98 (51.6)	
Marital Status				
Not married	12 (60.0)		69 (36.3)	
Currently married	8 (40.0)		121 (63.7)	
Highest Qualification				
None	2 (10.0)		13 (6.8)	
Primary	3 (15.0)		25 (13.2)	
Secondary	6 (30.0)		97 (51.0)	
Tertiary	9 (45.0)		56 (29.5)	
Health Status				
Healthy	20 (100)		95 (50.5)	
CKD	–		95 (50.5)	

SD, Standard Deviation; *Median (IQR); ⁺positively skewed

undeclared) was 30.0 years (IQR 20.0), age range 18–69 years. The majority (n=12, 60%) was, at the time of the data collection, not married; 11 (55%) were female, and 9 (45%) had above-secondary school education.

A total of 190 individuals (95 from the general public, 95 patients with predialysis CKD) aged between 18–84 years, mean (SD) 46.6 (17.8) years, participated in the validation exercise. The majority (54.7%) were aged above 45 years; were female (51.6%); and were educated up to secondary school level (51.0%) (Table 1).

Brunei-Malay SF-36v2 Data description and quality
The data completion rate was 100% as each respondent's questionnaire was inspected for completeness, post self-administration. Minimal floor effects (≤0.21%) for all the 8 scales were seen. Ceiling effects were also minimal but moderately pronounced for Role-Emotional (25.3%), Physical Functioning (23.2%) and Bodily Pain (21.5%) domains. The sample had the highest mean score (SD)

in BP (83.1/100, SD 25.8) and least mean score in RP (60.7/100, SD 23.5) domain scales. Three domain scales (PF, SF, and BP) had negatively skewed mean scores, with highest median scores seen in the PF domain (85, IQR 35.0). CKD patients had lower mean scores than healthy individuals in 5 domains (PF, RP, RE, BP, and GH) and difference were statistically significant ($p < 0.05$) in 3 domains (PF, BP, and GH). The scores were similar for 3 domains (SF, MH, and VT) between the two groups (Table 2).

Item Internal Consistency
Items' SD within each domain scale were close and comparable. Example, the range of SD for PF, BP, and RE items were 0.50–0.80; 1.32–1.47; 1.18–1.23, respectively. The scaling success for convergent validity, measured as corrected item correlation with own scale r>=0.4 (item internal consistency), was 100% for all the 8 scales/domains (Table 3).

Table 2. Description of the Brunei-Malay SF-36v2 Domain Scales (n=190)

Scale	PF	RP	RE	SF	BP	MH	VT	GH
Items (n)	10	4	3	2	2	5	4	5
Responses (%)	100	100	100	100	100	100	100	100
Floor (%)	1.1	2.1	2.1	0.5	0.5	0.5	1.1	0.5
Ceiling (%)	23.2	13.7	25.3	21.1	21.6	11.1	6.3	3.2
Mean (SD)ᵂ	85.0 (35.0)*	60.7 (26.8)	64.0 (27.6)	62.5 (37.5)*	77.5 (56.6)*	70.4 (18.5)	62.4 (20.4)	65.3 (22.5)
Mean (SD)ᴬ	67.9 (27.9)	58.9 (27.4)	63.2 (28.6)	68.5 (23.7)	78.4 (31.6)	71.2 (18.3)	63.6 (21.4)	59.9 (21.7)
Mean (SD)ᴮ	82.1 (21.5)	62.4 (26.3)	64.8 (26.6)	68.0 (23.4)	87.6 (31.5)	70.1 (18.7)	61.2 (19.3)	70.8 (22.2)

SD, Standard deviation; *Median (Interquartile range), negatively skewed; ᵂwhole (n=190); ᴬCKD (n=95); ᴮHealthy (n=95)
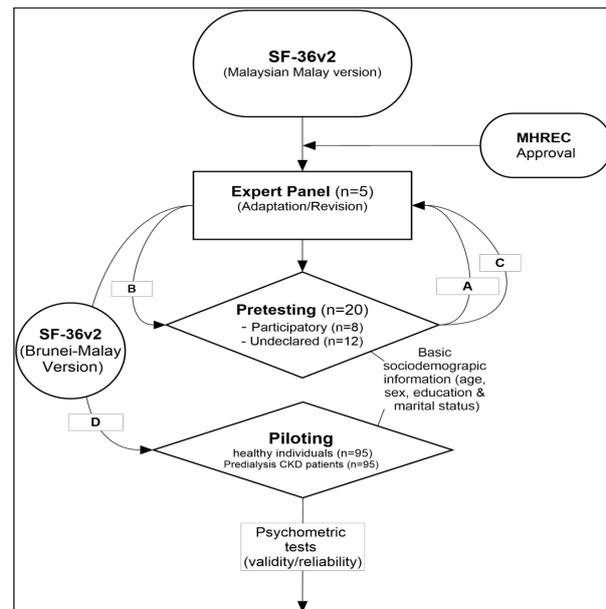
Figure 1. Schematic Outline of the Study. (A) Revision after participatory pretest, (B) Administered for undeclared pretest, (C) Revision after undeclared pretest, (D) Final version administered for validation.

Item Discriminant Validity

Item scores from Spearman's correlation were significantly better correlated with items of the same scale than with other scales, with the exception of 13 items (1 in SF, 6 in MH, 6 in GH). In other words, scaling success was 100% for all domains except SF (94%), MH (85%) and GH (85%). The exceptions were: SF1 correlated more with RE scale than with its scale (SF); MH1-5 with GH scale, and MH5 with VT scale; GH1-5 with MH scale, and GH5 with VT scale. Overall, 267 items (95.4%) of the total 280 items (100%) showed acceptable discriminant validity (Table 3).

Scale Internal Consistency Reliability

The internal consistency reliability of the 8 scales were all above the recommended minimum value (0.70). The least α score was in SF domain (0.72) and the highest (0.92) in PF domain scale. The internal consistency reliability remained acceptable even when the sample was disaggregated by gender, marital status and health status (Table 4).

Factor Analysis (SF-36v2 Summary Scales)

The two hypothesized components or constructs (PCS and MCS) were extracted in PCA using orthogonal rotation (Varimax). The data met all the set statistical assumptions for carrying factor analysis: determinant score, which checks for multicollinearity/singularity was 0.018 and greater than the recommended minimum value of 0.00001; Bartlett's Test of Sphericity confirmed that our data showed a patterned relationship among the variables as seen from the correlation matrix (p<0.001); Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) was 0.88 and the minimum MSA from the diagonal element of the anti-correlation matrix was 0.82, both were greater than the recommended minimum value of

Table 3. Item Description and Test of Scaling Assumptions for Brunei-Malay SF-36v2 (n=190)

Scale	Item description			Correlation between: items and scales (range)		Scaling Success Success/total (%)	
	Item (n)	Mean (range)	SD (range)	Item internal consistency	Item discriminant validity	Discriminant Validity**+	Convergent Validity*
Physical Functioning	10	2.12–2.83	0.50–0.80	0.54–0.76	0.18–0.45	80/80 (100)	10/10 (100)
Role–Physical	4	3.37–3.50	1.16–1.18	0.76–0.81	0.27–0.66	32/32 (100)	4/4 (100)
Role–Emotional	3	3.48–3.67	1.18–1.23	0.73–0.84	0.36–0.68	24/24 (100)	3/3 (100)
Social Functioning	2	3.71–3.73	1.03–1.10	0.56	0.29–0.57	15/16 (94.0)	2/2 (100)
Bodily Pain	2	4.25–4.40	1.32–1.47	0.75	0.35–0.54	16/16 (100)	2/2 (100)
Mental Health	5	3.66–4.00	0.87–1.08	0.51–0.68	0.20–0.82	34/40 (85.0)	5/5 (100)
Vitality	4	3.22–3.72	0.97–1.06	0.52–0.69	0.29–0.56	32/32 (100)	4/4 (100)
General Health	5	3.23–4.01	1.01–1.36	0.43–0.65	0.26–0.81	34/40 (85.0)	5/5 (100)

*Item correlation with own domain scale >0.4; **Item significant correlation with own scale greater than with other scales; +Spearman's correlation; SD, standard deviation; P<0.05 for all correlation coefficients

Table 4. Brunei-Malay SF-36v2 Domains' Cronbach's Alphas (n=190)

	Gender			Marital Status		Health Status	
	Whole	Male	Female	Single	Married	CKD	Healthy
Scale	(n=190)	(n=92)	(n=98)	(n=69)	(n=121)	(n=95)	(n=95)
Phys. Functioning	0.92	0.92	0.92	0.93	0.92	0.90	0.92
Role-Physical	0.91	0.92	0.89	0.93	0.89	0.90	0.91
Role-Emotional	0.90	0.91	0.88	0.95	0.87	0.89	0.91
Social Functioning	0.72	0.74	0.70	0.74	0.71	0.73	0.71
Bodily Pain	0.86	0.82	0.88	0.87	0.85	0.87	0.85
Mental Health	0.80	0.82	0.76	0.82	0.77	0.81	0.79
Vitality	0.80	0.82	0.76	0.84	0.79	0.79	0.82
General Health	0.81	0.84	0.78	0.80	0.82	0.85	0.74

Table 5. Scale Validity and Correlation with Principal Components (rotated) for Brunei-Malay SF-36v2 (n=190)

| | SF-36v2 (U.S) | | | SF-36v2 (Brunei-Malay) | | | | | | | | |
| | Standard (n=4016)* | | | Whole (n=190) | | | CKD (n=95) | | | Healthy (n=95) | | |
Scale	PCS	MCS	h^2	PCS	MCS	h^2	PCS	MCS	h^2	PCS	MCS	h^2
PF	++	-	0.85	0.64	0.36	0.53	0.71	0.36	0.64	0.57	0.38	0.47
RP	++	+	0.87	0.90	0.20	0.84	0.90	0.19	0.84	0.89	0.20	0.82
RE	+	++	0.70	0.83	0.31	0.78	0.81	0.33	0.77	0.85	0.26	0.79
SF	+	++	0.60	0.49	0.59	0.58	0.32	0.66	0.53	0.67	0.50	0.70
BP	++	+	0.71	0.40	0.64	0.57	0.27	0.64	0.48	0.49	0.66	0.68
MH	-	++	0.74	0.22	0.82	0.72	0.15	0.84	0.73	0.36	0.75	0.69
VT	+	++	0.68	0.26	0.82	0.74	0.38	0.77	0.74	0.19	0.87	0.79
GH	+	+	0.88	0.26	0.77	0.65	0.24	0.74	0.60	0.29	0.81	0.74
				Variance explained								
Variance	75%			68%			71%			67%		
Reliable	85%			81%			84%			80%		

*U.S General population, 2009 (Mark, 2011); ++strong association ($r \geq 0.70$), +moderate-substantial association ($0.30 < r < 0.70$), -Weak association ($r < 0.30$) (McHorney et al., 1993); Bold shows highest loading correlation in a domain scale

0.5 (Gie Yong and Pearce, 2013; Williams et al., 1996).

The rotated summary components (PCS and MCS) in the whole sample (n=190) explained 68% of the overall variance and accounted for 81% reliable variance of all the 8 domains and for more than 50% in each domain scale. Disaggregating the sample by health status (healthy and CKD), the rotated summary components (PCS and MCS) explained 71% (in healthy) and 67% (in CKD) of the overall variance, and accounted for 84% (in healthy) and 80% (in CKD) reliable variance of all the 8 domains in each group, and for at least 47% in each domain scale of the two groups (Table 5).

The PF (r=0.64), RP (r=0.90) correlated more strongly with component 1 (PCS) while MH (0.82) and VT (r=0.82) correlated more strongly with component 2 (MCS). Moderate correlation of the SF (r=0.49) with the PCS component, in addition to its strong correlation with the MCS (r=0.59) component, was seen. RE (r=0.83) correlated more strongly with PCS (instead of MCS) and moderately (r=0.31) with MCS; BP (r=0.64) correlated strongly with MCS (instead of PCS) and moderately (r=0.40) with PCS; and GH (r=0.77) correlated more strongly unto MCS (Table 5).

Discussion

Our study has confirmed the validity and reliability of the SF-36v2 (Brunei-Malay) for measuring HRQOL in Brunei Darussalam. The study had 100% response rate, and minimal floor and ceiling effects for all 8 SF-36 domain scales. The wide sociodemographic spectrum of our study sample (e.g. the healthy and sick, the young and old) aided in minimizing the floor and ceiling effects.

On testing scaling assumptions, the Brunei-Malay SF-36v2 item internal consistency (100%), convergent validity (100% scaling success), item discriminant validity (overall 95.4% scaling success) and scale internal consistency reliability ($\alpha > 0.70$) in all the 8 SF-36 domains) were generally satisfactory (Cho and Kim, 2015; Tavakol and Dennick, 2011) and comparable to values reported for SF-36 validations in France (Leplège et al., 1998), Bangladesh (Feroz et al., 2012) and elsewhere. Such failures are considered minor when the majority of items (>90%) fully support the concepts being tested (Hobart et al., 2001), as in our study (with 95.4% success rate).

The exact SF-36 summary constructs (PCS and MCS) in the standard summary algorithm were not replicated in the Bruneian population. Essentially, the pattern of

the factor loadings of the 8 domains on the summary constructs was as hypothesized for 5 domain scales (PF, RP, SF, MH and VT) and contrary for 3 domain scales (RE, GH and BP). The pattern remained unchanged even when the sample was disaggregated by health status and when oblique rotation method, as suggested by Han and Reeves (2008), was applied.

This was not unexpected as other psychometric studies of the SF-36 have reported similar findings (Suzukamo et al., 2011). In Malaysia, for example, GH correlated more strongly with PCS instead of near-equal correlations with both PCS and MCS, and VT correlated more with PCS instead of correlating with MCS (Sararaks et al., 2005). Similar nonconformities of PCS and MCS summary scoring were reported in Japan (Fukuhara et al., 1998), in Taiwan (Tseng et al., 2003), in Australia (Tucker et al., 2010), and in people with some disease conditions such as multiple sclerosis (Hobart et al., 2001), and 'functional somatic syndrome' (Schröder et al., 2012). A study conducted in Australia using the 1995 National Health Survey dataset (n=18,141) concluded that the PCS and MCS are correlated, resulting in occasional cross-correlation in the domains (Tucker et al., 2010). Furthermore, the difference in the PCS and MCS summary constructs, especially between Asians and Europeans, have been attributed to variations in cultures and customs (Sararaks et al., 2005).

Meanwhile, these discrepancies had resulted in the emergence of several alternative scoring algorithms for summary constructs being suggested such as the three-component models developed in Japan (Suzukamo et al., 2011) and elsewhere (Keller et al., 1998), the reduced 5-scale oblique model (Hann and Reeves, 2008), and the "SF-36 Total/Global/Overall Scores" (Lins and Carvalho, 2016).

The differences do not discredit the validity and reliability of the SF-36 to adequately measure HRQOL across diverse populations, rather, they point to the inadequacy of the standard scoring algorithm to cater for populations other than the US population (Hann and Reeves, 2008; Tucker et al., 2010). In situation where the hypothesized summary constructs (PCS and MCS) are not confirmed, studies have suggested that summary construct should be interpreted with caution (Hann and Reeves, 2008; Rannou et al., 2007) or country-specific 'algorithm' or 'scoring coefficients' be considered (Fukuhara et al., 1998; Tucker et al., 2010). In any case, however, the individual domain scales should always be reported (Hann and Reeves, 2008; Ware, 2000).

This study is not without its limitations. First, the sample size was only "fairly large enough" for factor analysis according to some studies (Gie Yong and Pearce, 2013; Williams et al., 1996). Second, the translation and cultural adaptation of health survey instruments almost always result in unavoidable 'modifications' that might have profound effects on their original contents (Leplège et al., 1998). However, no extensive alteration was carried out as only incomprehensible words or phrase were rephrased or replaced with contextual equivalents.

In conclusion, our study confirmed the validity and

reliability of the SF-36v2 (Brunei-Malay version) for measuring HRQOL in Brunei Darussalam. The standard PCS and MCS scoring as hypothesized in the standard summary scoring algorithm should be interpreted with caution. Further studies should be carried out on a more varied sample distribution to confirm these findings and to explore other psychometric evaluations such as test-retest reliability.

Acknowledgments

References

Barribeau P, Bonnie B, Jeff C, et al (2012). Survey Research. Writing@CSU. Retrieved from https://writing.colostate.edu/guides/guide.cfm?guideid=68.

Bullinger M, Alfonso J, Apolone G, et al (1998). Translating health status questionnaires and evaluating their quality: The IQOLA project approach. *J Clin Epidemiol*, **51**, 913–23.

Cho E, Kim S (2015). Cronbach's coefficient alpha. *Organizational Res Methods*, **18**, 207–30.

Coons SJ, Rao S, Keininger DL, Hays RD (2000). A comparative review of generic quality-of-life instruments. *Pharmacoeconomics*, **17**, 13–35.

Feroz AHM, Islam MN, Ten Klooster PM, et al (2012). The Bengali Short Form-36 was acceptable, reliable, and valid in patients with rheumatoid arthritis. *J Clin Epidemiol*, **65**, 1227–35.

Fukuhara S, Ware JE, Kosinski M, Wada S, Gandek B (1998). Psychometric and clinical tests of validity of the Japanese SF-36 Health Survey. *J Clin Epidemiol*, **51**, 1045–53.

Gie Yong A, Pearce S (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutor Quant Methods Psychol*, **9**, 79–94.

Hann M, Reeves D (2008). The SF-36 scales are not accurately summarised by independent physical and mental component scores. *Qual Life Res*, **17**, 413–23.

Hobart J, Freeman J, Lamping D, Fitzpatrick R, Thompson A (2001). The SF-36 in multiple sclerosis: why basic assumptions must be tested. *J Neurol Neurosurg Psychiatry*, **71**, 363–70.

Keller SD, Ware JE, Bentler PM, et al (1998). Use of structural equation modelling to test the construct validity of the SF-36 Health Survey in ten countries: Results from the IQOLA Project. *J Clin Epidemiol*, **51**, 1179–88.

Leplège A, Ecosse E, Verdier A, Perneger TV (1998). The French SF-36 health survey: translation, cultural adaptation and preliminary psychometric evaluation. *J Clin Epidemiol*, **51**, 1013–23.

Lins L, Carvalho FM (2016). SF-36 total score as a single measure of health-related quality of life: Scoping review. *SAGE Open Med*, **4**, 205031211667172.

Mark M (2011). User's manual for the SF-36v2 health survey (3rd ed.). Lincoln: Quality metric incorporated. Retrieved from www.QualityMetric.com.

McHorney CA, Ware JE, Raczek AE (1993). The MOS 36-item short-form health survey (SF-36): II. Psychometric and

clinical tests of validity in measuring physical and mental health constructs. *Med Care*, **31**, 247–63.

Newvine K (n.d.). Using the SAS® System to Score the SF-36® Health Survey. Cooperative Studies Program Coordinating Center, 963–968. Retrieved from http://www.sascommunity.org/sugi/SUGI94/Sugi-94-168 Newvine.pdf.

Rannou F, Boutron I, Jardinaud-lopez M, et al (2007). Should aggregate scores of the medical outcomes study 36-item short form health survey be used to assess quality of life in knee and hip osteoarthritis? A national survey in primary care. *Osteoarthritis Cartilage*, **15**, 1013–8.

Sararaks S, Azman AB, Low LL, et al (2005). Validity and reliability of the SF-36: The Malaysian context. *Med J Malaysia*, **60**, 163–79.

Schröder A, Oernboel E, Licht RW, Sharpe M, Fink P (2012). Outcome measurement in functional somatic syndromes: SF-36 summary scores and some scales were not valid. *J Clin Epidemiol*, **65**, 30–41.

Suzukamo Y, Fukuhara S, Green J, et al (2011). Validation testing of a three-component model of Short Form-36 scores. *J Clin Epidemiol*, **64**, 301–8.

Tarrant C, Angell E, Baker R, et al (2014). Developing and piloting the questionnaires (pilots 1 and 2). In health services and delivery research. Southampton (UK): NIHR Journals Library, pp 1–18.

Tavakol M, Dennick R (2011). Making sense of Cronbach's alpha. *Int J Med Edu*, **2**, 53–5.

Tseng HM, Lu JF, Gandek B (2003). Cultural issues in using the SF-36 Health Survey in Asia: results from Taiwan. *Health Qual Life Outcomes*, **1**, 72.

Tucker G, Adams R, Wilson D (2010). New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires. *Qual Life Res*, **19**, 1069–76.

Ware JE (2000). SF-36 health survey update. *Spine*, **25**, 3130–9.

WHO (2017). WHO | Process of translation and adaptation of instruments. Retrieved July 27, 2017, from http://www.who.int/substance_abuse/research_tools/translation/en/.

Williams B, Onsman A, Brown T (1996). Exploratory factor analysis: A five-step guide for novices. *J Emerg Prim Health Care*, **19**, 42–50.