TPM®

VALIDATION AND EQUATING OF THE TWO VERSIONS OF THE SF-36 FIVE-ITEM MENTAL HEALTH INDEX

CAROLINA S. FELLINGHAUER
SWISS PARAPLEGIC RESEARCH (SPF)
UNIVERSITY OF LUCERNE

CHRISTINE FEKETE
SWISS PARAPLEGIC RESEARCH (SPF)

MARTIN W. G. BRINKHOF
BIRGIT PRODINGER
SWISS PARAPLEGIC RESEARCH (SPF)
UNIVERSITY OF LUCERNE

CARLA SABARIEGO
LUDWIG-MAXIMILIAN UNIVERSITY OF MUNICH

Two versions of the Five-Item Mental Health Index (MHI-5) exist differing in their number of response options. Score sufficiency of the MHI-5 has not been evaluated yet. The aims of this study are to test metric properties of these two MHI-5 versions and to equate them using three different methodologies. The two versions of the MHI-5 were assessed in two Swiss surveys. These were equated with a linear rescaling approach and two Rasch-based score equating methodologies: a mean anchoring and a multi-group analysis. Metric properties and score invariance across methodologies are investigated with a stratified analysis by gender, age, and health conditions. The MHI-5 versions show reliability, unidimensionality, local item independence, and fit. Mean scores varied depending on the equating methodology applied and were consistently higher with linear rescaling. However, the relative differences in mean scores were comparable across strategies. MHI-5 has robust metric properties in the general and a disease-specific population. Although equating with linear rescaling may be used, a Rasch-based approach is generally superior with regards to the reliability of the resulting person ability estimates.

Key words: Mental Health Index; Psychometrics; Test equating; Rasch analysis; Scaling.

Correspondence concerning this article should be addressed to Carolina Fellinghauer, Swiss Paraplegic Research, Guido A-Zäch Strasse 4, CH-6207 Nottwil, Switzerland. Email: carolina.fellinghauer@paraplegie.ch

Effective policy making, planning, and evaluation for targeted public health interventions requires up-to-date epidemiological estimates. In that sense, accurate measurement is indispensable, as reliable and robust measurement instruments improve the quality of estimates and guarantee sound quantitative analysis for comparisons within and between populations. Moreover, aggregation of data from different sources, for instance different surveys, is often necessary to enrich and refine available information. Consequently, the quality of epidemiological information not only relies on the quality of the applied measurement instruments but also on the aggregation strategy applied to harmonize data from different sources.

The Five-Item Mental Health Index (MHI-5), one of the subscales of the 36-item Short Form Health Survey (SF-36) (Ware, 2000a) is a commonly used screening instrument to measure mental health in surveys and clinical research. This five item subscale can be used as stand-alone tool to collect information on general mental health states and has been applied as a screening instrument for mental disorders (Means-Christensen, Arnau, Tonidandel, Bramson, & Meagher, 2005). Its empirical validity and reliability is widely supported (Ware, 2000a) and it can reliably be used to assess mood and anxiety disorders (Cuijpers, Smits, Donker, ten Have, & de Graaf, 2009). While the response options of the MHI-5 Version 1 (Ware & Sherbourne, 1992) are rated on a 6-point Likert scale, Version 2 uses a 5-point scale (Ware, 2000b). Interestingly, both versions are still widely used and sum scores of the two MHI-5 versions can be aligned by rescaling them on a metric ranging from 0 to 100 (Ware, Kosinski, & Dewey, 2001). A strong assumption when adding up items to a sum score is that all items have equivalent measurement precision and hierarchical ordered response options (Bond & Fox, 2007). These important assumptions can and should be tested with a modern test theory approach.

Metric studies on the MHI-5 which support key modern test theoretical assumptions, such as the interval scale property of the sum score, are lacking so far. Most available metric studies of the MHI-5 apply classical test theory (CTT) methods (Rumpf, Meyer, Hapke, & John, 2001; van Leeuwen, van der Woude, & Post, 2012; Ware, 2000b). While these methods allow to evaluate metric properties or item functioning, they are not appropriate to determine the fulfillment of fundamental measurement assumptions. Whether sum scores have indeed metric properties or not, needs to be confirmed using modern test theory (Petrillo, Cano, McLeod, & Coon, 2015), which provides methods to test metric properties of scales to create interval-scaled scores with cardinal properties (Wright, 1992). Modern test theoretical approaches, such as the Rasch analysis (Rasch, 1960), test if items of a scale measure a single, unidimensional latent trait with sufficient reliability. Further, Rasch analysis allows to determine if items have a good fit, ordered response options, and are pairwise independent from each other (Tennant & Conaghan, 2007). An analysis of the metric properties of the MHI-5 with Rasch analysis is essential to support the interval-scale properties of the 0-100 score as well as the validity of the recommended linear rescaling to equate the two versions. While Rasch analyses have already been applied for the total SF-36 (Hawthorne, Densley, Pallant, Mortimer, & Segal, 2008) or the physical functioning subscale (Haley, McHorney, & Ware, 1994; Kim & So, 2015; Raczek et al., 1998) the evaluation of metric properties of the MHI-5 using Rasch analyses is neither available for general populations nor for populations with physical impairments.

In summary, the two versions of MHI-5 which differ in the number of response options are widely used. However, two major research gaps remain: first, an analysis of metric properties with modern test theory of either version of the MHI-5 are missing, and second, a proper cross-validation of the two versions using a modern measurement methodology for equating scores has not been carried out. The overall aim of this study is thus to evaluate the metric properties of the MHI-5 and to test the equivalence of the 0-100 scores derived from the two MHI-5 versions when cross validated with three different methodologies: a linear score rescaling approach and two Rasch-based score equating methodologies, namely an anchoring on the mean threshold and a multigroup analysis. To assess the invariance of transformed scores, an analysis of the 0-100 MHI-5 scores within gender, age, and health condition groups was performed for both a general and a disease specific population.

METHOD

Data Sources and Sample Characteristics

The MHI-5 ratings from the Swiss Health Survey (SHS) 2012 (BFS Sektion Gesundheit, 2014) and the community survey of the Swiss Spinal Cord Injury (SwiSCI) cohort study (Brinkhof, Fekete, Chamberlain, Post, & Gemperli, 2016; Fekete, Segerer, Gemperli, & Brinkhof, 2015) were used to determine the equivalence of the two versions of the MHI-5. The SHS has routinely been performed every five years since 1992 and has been part of the Swiss census data collection program since 2010. The assessment framework is holistic and dynamic, based on the World Health Organization (WHO) definition of health, including questions about physical health, the ecological, social, and cultural environment, lifestyle, and behavior. The SHS population is a simple stratified random sample of persons above 15 years living in a private household in Switzerland and not being in process of seeking asylum or living in a collective household; the strata being the Swiss cantons (BFS Sektion Gesundheit, 2014). The SHS data collected in 2012 were included in this study to be on a common time line with the SwiSCI survey.

The SwiSCI survey is part of a research program seeking to understand the lived experience of persons with a spinal cord injury (SCI) living in Switzerland. The first SwiSCI community-based survey, which is conducted every five years, took place between September 2011 and March 2013 and included Swiss residents with traumatic and nontraumatic SCI and aged over 16 years. The study population was established based on records from three specialized rehabilitation centers and two national associations for persons with SCI (Brinkhof et al., 2016; Fekete et al., 2015). Exclusion criteria were congenital conditions leading to SCI, neurodegenerative disorders, and new SCI in the context of palliative care. In total, 3,144 persons were eligible and 1,549 participated in the main module of the survey, resulting in a response rate of 49.3% with limited non-response bias on major outcomes. The SwiSCI survey is formally approved by the principal Swiss ethics committee and informed consent was obtained from all survey participants (Brinkhof et al., 2016).

Sociodemographic information of SHS participants, SwiSCI participants, and the total population including frequencies and proportions of the samples by age groups, gender, and selected health conditions, are shown in Table 1. Age group distributions differed between surveys, with a higher percentage of participants in the lowest (< 30) and highest (> 75) age groups in the SHS and a higher percentage of participants between the age of 45 and 60 in SwiSCI. While slightly over half of SHS participants were female, the proportion of females was markedly lower in SwiSCI, confirming the evidence on higher prevalence of spinal cord injuries in men (Shackelford, Farley, & Vines, 1998). The most prevalent health conditions in the SHS were cardiovascular conditions (15.3%), followed by migraine (14.6%), and allergies (14.2%). Incomplete paraplegia was the most frequent diagnosis in the SwiSCI population.

Measure: Mental Health Index (MHI-5)

The SF-36 is one of the most widely used generic measures to assess health and has been used in general and specific populations across various healthcare settings all over the world

TABLE 1
Sample sizes, age, gender, and health conditions: Frequencies and proportions in the Swiss Health Survey (SHS) and the Swiss Spinal Cord Injury (SwiSCI) cohort study

		All	SHS	SwiSCI
Sample sizes	*N* (%)	21,474 (100%)	20,027 (93.3%)	1,447 (6.7%)
Age groups	< 30	3,572 (16.6%)	3,446 (17.2%)	126 (8.7%)
	30-45	5,446 (25.4%)	5,081 (25.4%)	365 (25.2%)
	45-60	6,224 (29.0%)	5,674 (28.3%)	550 (38.0%)
	60-75	4,548 (21.2%)	4,215 (21.0%)	333 (23.0%)
	> 75	1,684 (7.8%)	1,611 (8.1%)	73 (5.1%)
Gender	*N* Female (%)	10,943 (51.0%)	10,538 (52.6%)	405 (28.0%)
Health conditions	*N* Cardiovascular (%)		3,072 (15.3%)	
	N Allergies (%)		2,519 (14.2%)	
	N Depression (%)		1,130 (6.4%)	
	N Cancer (%)		517 (2.9%)	
	N Migraine (%)		2,576 (14.6%)	
	N Asthma (%)		840 (4.7%)	
	N Diabetes (%)		827 (4.7%)	
	N Arthrosis (%)		2,041 (11.5%)	
	N Complete paraplegia (%)			461 (32.1%)
	N Incomplete paraplegia (%)			523 (36.4%)
	N Complete tetraplegia (%)			152 (10.6%)
	N Incomplete tetraplegia (%)			300 (20.9%)

(Scoggins & Patrick, 2009). With a set of 36 questions divided into eight subscales, the SF-36 allows a "global assessment" of eight health components. The MHI-5 subscale assesses the frequency of problems in mental functions such as nervousness, sadness, peacefulness, downheartedness, and happiness. The response options to assess the frequency of mental health problems of the original MHI-5 scale are 1) *all of the time*, 2) *most of the time*, 3) *a good bit of the time*, 4) *some of the time*, 5) *a little of the time*, and 6) *none of the time*. The response option 3) has been removed in Version 2 of the MHI-5. The MHI-5 has been used to assess mental health in the SHS using Version 2 (5-point Likert scale) and in the SwiSCI survey using Version 1 (6-point Likert scale). Table 2 shows the response options as well as frequency and proportion of responses for the MHI-5 in the SHS and SwiSCI surveys.

The MHI-5 has generally been reported to have good reliability (McCabe, Thomas, Brazier, & Coleman, 1996; van Leeuwen et al., 2012) as well as good construct validity (Friedman, Heisel, & Delavan, 2005). The sensitivity in detecting mood disorders is widely supported (Cuijpers et al., 2009; Friedman et al., 2005; Rumpf et al., 2001).

TABLE 2

Response frequencies of the MHI-5 in the Swiss Health Survey (SHS) 2012 and the Swiss Spinal Cord Injury (SwiSCI) cohort study 2012

MHI-5 items	These questions are about how you feel and how things have been with you during the past four weeks.[1] For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past four weeks.							
		All of the time	*Most of the time*	*A good bit of the time*	*Some of the time*	*A little of the time*	*None of the time*	Missing values
MHI-5 nervous	SHS	342 (1.7%)	1,137 (5.7%)		3,849 (19.2%)	5,841 (29.2%)	8,836 (44.1%)	22 (0.1%)
	SwiSCI	6 (0.4%)	34 (2.3%)	156 (10.8%)	335 (23.2%)	5,48 (37.9%)	337 (23.3%)	31 (2.1%)
MHI-5 down	SHS	169 (0.8%)	442 (2.2%)		1,678 (8.4%)	3,382 (16.9%)	14,289 (71.3%)	67 (0.3%)
	SwiSCI	4 (0.3%)	28 (1.9%)	71 (4.9%)	201 (13.9%)	400 (27.6%)	703 (48.6%)	40 (2.8%)
MHI-5 calm	SHS	5,522 (27.6%)	10,432 (52.1%)		2,504 (12.5%)	1,104 (5.5%)	414 (2.1%)	51 (0.3%)
	SwiSCI	134 (9.3%)	670 (46.3%)	278 (19.2%)	194 (13.4%)	114 (7.9%)	30 (2.1%)	27 (1.9%)
MHI-5 depressed	SHS	161 (0.8%)	468 (2.3%)		2,034 (10.2%)	4,353 (21.7%)	12,974 (64.8%)	37 (0.2%)
	SwiSCI	9 (0.6%)	35 (2.4%)	132 (9.1%)	328 (22.7%)	521 (36%)	390 (27%)	32 (2.2%)
MHI-5 happy	SHS	5,890 (29.4%)	10,576 (52.8%)		2,572 (12.8%)	674 (3.4%)	246 (1.2%)	69 (0.3%)
	SwiSCI	101 (7%)	560 (38.7%)	323 (22.3%)	257 (17.8%)	147 (10.2%)	37 (2.6%)	22 (1.5%)

Note. [1]SwiSCI: the last two weeks.

Procedure

The different strategies to cross-validate the two versions of the MHI-5 are illustrated in Figure 1. The assumed quality of the recommended 0-100 linear score rescaling A) to equate the two versions of the MHI-5 is compared to scores obtained with two Rasch-based (Rasch, 1960) scale equating methodologies: B) a mean threshold anchoring and C) a multigroup analysis from the field of probabilistic measurement.

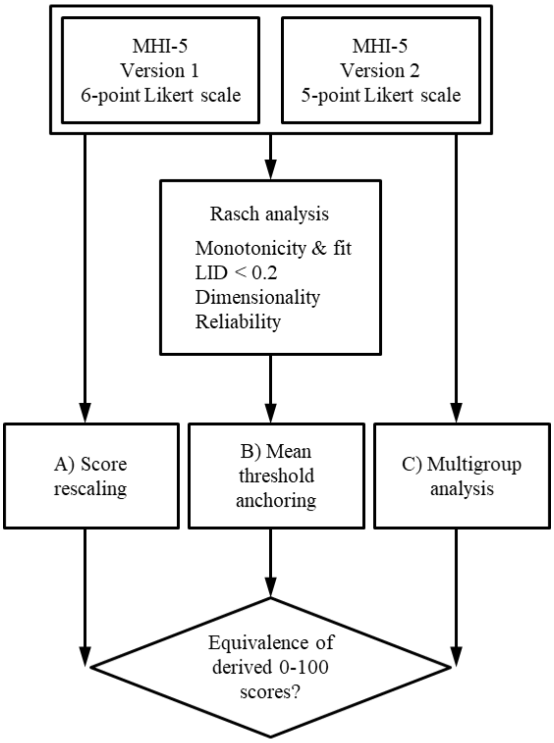

FIGURE 1
Scheme of the methodological approach for the cross-validation of the two MHI-5 versions.

Score Rescaling (A)

Typically, the scores of the five items of the MHI-5 are added up and the resulting total score is than *rescaled*, that is, linearly transformed, to range from 0-100 (Equation 1).

$$\text{Transformed scale} = \left[\frac{(\text{actual raw score} - \text{lowest } possible \text{ raw score})}{possible \text{ raw score range}}\right] \times 10 \qquad (1)$$

The 0-100 transformed scores are expected to be equivalent across versions of the MHI-5, while in fact this transformation assumes metric properties despite the ordinal scaling. However, the rescaling does not produce an interval-scaled metric if the underlying raw score is ordinal.

Rasch Analysis

The Rasch model, a commonly used model within the field of modern test theory, builds on a probabilistic measurement paradigm where the probability of response to an item is a function of the respondents' ability level and the item difficulty. When the fit of the data to the Rasch model is supported, interval-scaled person ability estimates can be calculated and transformed to a 0-100 scale for further statistical analyses (Andrich, Sheridan, & Luo, 2010). An analysis using the Partial Credit Model (PCM; Masters, 1982), a model from the Rasch family, was conducted to accommodate the ordered polytomous rating scale used in the MHI-5. Several model assumptions must be tested to validate the overall fit of the data to the PCM. The scale is expected to measure a single latent trait with enough reliability. The items have to show good fit, ordered response options, enough pairwise independence from each other, and have to represent a unidimensional construct (Tennant & Conaghan, 2007).

Two indices for reliability were taken into account to determine items' model fit: the person separation reliability (PSR; Boone, Staver, & Yale, 2014) and the Cronbach's alpha (Nunnally, 1978), for which values close to 1 indicate perfect reliability. PSR values above .85 are required for using scales for individual measurement and values above .70 for group measurement (Tennant & Conaghan, 2007). Alpha values above .70 demonstrate acceptable reliability (Nunnally, 1978), however, a minimum of .80 is commonly expected. Given the high proportion of participants reporting good mental health, especially in the SHS population, a left-skewed distribution of total scores may occur. It is known that in presence of skewed distributions, the reliability measured with Cronbach's alpha remains more constant than the PSR due to the increase in the error variance in the extreme scores of the latter (Andrich, 2015). In presence of insufficient reliability, a sensitivity analysis of the PSR, using a calibration sample presenting an approximate uniform distribution of the total scores, may be undertaken.

Good fitting items are expected to show infit and outfit values between 0.80 to 1.20 (Wright, Linacre, Gustafson, & Martin-Löf, 1994); the outfit statistic being more outlier sensitive. Values above the cut-off indicate underfit, while values below indicate overfit. Responses to overfitting items are very predictable and are less a threat to measurement than responses to underfitting items (Linacre, 2002).

The thresholds of the response rating scale are expected to be ordered. Response thresholds are the equal probability point between two consecutive response levels, so that the number of thresholds is the number of the scale's response options minus one. Following, the difficulty of response Threshold 1 of an item of the MHI-5 is located between the responses "All of the time" and "Most of the time."

In presence of disordered thresholds, response options are collapsed to obtain monotonic ordering. Disordered thresholds occur when respondents do not consistently discriminate among options of a rating scale for instance due to the influence of another dimension (Rost, Carstensen, & von Davier, 1999).

Higher positive correlations of standardized residuals are indicative of local item dependencies (LID; Yen, 1993). Usually, locally dependent items measure a common aspect of the latent trait addressed by a scale. This study considered correlations of standardized residuals above 0.20 as significant (Marais & Andrich, 2008; Reeve et al., 2007).

Unidimensionality is an important assumption to derive a total score from a set of items. The unidimensionality of the scale is assessed with principal component analysis of the standardized Rasch residuals. First eigenvalues > 2 are considered as substantial and indicative of strong multidimensionality (Raîche, 2005).

For the present study, the items of the MHI-5 from the SHS were first calibrated separately and tested for fit to the Rasch assumptions (Rasch, 1960). Prior to the analysis, the response options of the MHI-5 items with a positive direction were rescored, so that a higher score indicates more mental health difficulties for all the items of the scale. The Rasch analysis then provided logit distributed person ability estimates, which can be transformed to a 0-100 scale. However, to compare ability estimates across surveys, the difficulty of the items needs to be invariant across the two MHI-5 versions. Anchoring is an analytical strategy which allows to keep item difficulty estimates constant across settings. Two different Rasch-based anchoring strategies were applied, namely the mean threshold anchoring and the multigroup analysis.

Mean Threshold Anchoring (B)

The first strategy used the item difficulties found in the SHS as anchor for the difficulties of the SwiSCI items. An analysis with anchored item difficulties enables comparability between different versions of a scale (Dorans, Pommerich, & Holland, 2007). Application of an anchoring strategy across populations entails the inherent assumption that the response to the MHI-5 captures a common latent trait, for example mental health. This study uses the MHI-5 ratings of the SHS data, that is, a large general population sample, to perform an item calibration with Rasch and create valid reference item difficulties. Using the SHS sample item difficulty as anchor permits to gain knowledge on the impact of a condition, such as a SCI, on mental health. The levels of ability may differ across groups, but the latent trait which is measured is expected to be common and similarly understood across groups. Using the same item difficulties in the two samples prevents the item difficulty from drifting, so that on average the items and the MHI-5 scale have the same difficulty in both surveys.

Multigroup Analysis (C)

Using the item difficulties from one sample to anchor the analysis with another sample often results in lower fit for the anchored sample. To avoid forcing sample’s item difficulties to comply with the item difficulties estimated on the anchor sample, a second anchoring strategy is used in this study, namely a multigroup approach (Linden & Hambleton, 1997). In this approach, the items of the MHI-5 from both surveys are analyzed jointly with the PCM, with the expectation that item difficulties optimally fitting the two survey populations can be found. The multigroup approach allows estimating item difficulties which are equally good in both surveys and derive comparable abilities from a common metric.

Evaluation of Equivalence across Different Methodologies

Finally, to investigate the impact of the different score equating methodologies, the transformed 0-100 scores are investigated more closely. Ideally, findings should not vary by equating methodology applied. The differences in mean scores are compared across samples by age, gender, and health conditions focusing on the stability of obtained scores across methodologies. In that sense, the mean MHI-5 scores should not be understood as reference values for any health condition included in the present work as this would require more in-depth analysis of comorbidity patterns and other confounders. The SHS health conditions included in this investigation were cardiovascular conditions, allergies, depression, cancer, migraine, asthma, diabetes, and arthrosis.

Statistical significance in differences of the mean scores across the SHS and SwiSCI sample, the health condition, and gender groups was tested with a *t*-test. ANOVA's *F*-test gives the significance of the differences between age groups, classified according to the guidelines proposed by the International Spinal Cord Society (DeVivo, Biering-Sorensen, New, & Chen, 2011). Given, the large sample size of the SHS sample, highly significant effects in the statistical analysis of the mean differences are expected. However, the focus of the study was to observe the stability of the size of the effects across scale transformation methodologies, that is, MHI-5 score rescaling or Rasch transformation.

The missing values of the MHI-5 were imputed before calculation of the total scores for the linear rescaling and the Rasch analyses (Stekhoven & Buhlmann, 2012). The missing values from participants with less than three responses to the MHI-5 were not imputed and, consequently, these participants were not included in the analysis. A single imputation methodology based on random forests was preferred to the mean imputation to improve accuracy (Hardouin, Conroy, & Sebille, 2011). All analyses were performed with R (R Core Team, 2016), more specifically with the package TAM (Kiefer, Robitzsch, & Wu, 2014) for the Rasch analysis and scale equating.

Results

Rasch Analysis

Table 3A and Table 3B present the results of the different Rasch analyses, including the respective analyses of the two versions of the MHI-5, the quality of the rescaling, and the two Rasch-based anchoring approaches. In summary, all Rasch analyses, either separately per version or jointly with mean difficulty anchoring or a multigroup analysis, supported the good metric properties of the MHI-5.

The Rasch-based reliability of the MHI-5 was good in the SwiSCI sample but insufficient in the SHS sample (PSR = .61). The lower reliability of the SHS survey can be entirely explained by the skewed distribution of total scores, with very high percentages of persons with good mental health. Cronbach's alpha was.77 in the SHS sample, indicating acceptable to good reliability. The sensitivity analysis of the SHS MHI-5 anchored on the difficulty estimates of the complete sample, using a calibration sample with an approximate uniform distribution of the total scores resulted in good reliability (PSR = .81, α =.87).

The infit and outfit statistics supported the good fit of the two versions of the MHI-5 items for both surveys. Independently of the number or response options, the MHI-5 items fulfill the monotonicity assumption, not requiring the collapsing of response options because of disordered thresholds. Residual correlations were far below 0.20, confirming the absence of local item dependencies. The first eigenvalues, close to 1.5 were indicative of unidimensionality in all settings.

TABLE 3A

MHI-5 fit to the Rasch models for the respective analyses in the Swiss Health Survey (SHS) and the Swiss Spinal Cord Injury (SwiSCI) cohort study

Rasch analysis SHS															
	Item fit		Residual correlations					Dimensionality		Item difficulty					
MHI-5 items	Infit	Outfit	MHI-5 nervous	MHI-5 down	MHI-5 calm	MHI-5 depressed	MHI-5 happy	Eigenvalue	PCA	Location	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
MHI-5 nervous	1.11	1.10	1.00	–0.22	–0.24	–0.25	–0.42	1.46	–0.11	–1.65	–3.43	–2.31	–1.04	0.18	
MHI-5 down	0.94	0.91		1.00	–0.30	0.01	–0.21	1.43	–0.51	–2.37	–3.69	–2.83	–1.89	–1.08	
MHI-5 calm	1.03	1.03			1.00	–0.31	–0.14	1.04	0.62	–1.44	–3.21	–2.20	–1.46	1.13	
MHI-5 depressed	0.91	0.86				1.00	–0.17	0.98	–0.49	–2.29	–3.76	–2.87	–1.79	–0.77	
MHI-5 happy	1.06	1.05					1.00	0.08	0.33	–1.69	–3.51	–2.62	–1.67	1.03	
Model fit	PSR	Cronbach's alpha													
Reliability	.61	.77													
Rasch analysis SwiSCI															
	Item fit		Residual correlations					Dimensionality		Item difficulty					
MHI-5 items	Infit	Outfit	MHI-5 nervous	MHI-5 down	MHI-5 calm	MHI-5 depressed	MHI-5 happy	Eigenvalue	PCA	Location	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
MHI-5 nervous	1.14	1.12	1.00	–0.16	–0.13	–0.28	–0.48	1.56	0.50	–1.92	–4.80	–3.54	–2.08	–0.66	1.45
MHI-5 down	0.88	0.86		1.00	–0.30	0.02	–0.28	1.48	–0.29	–2.55	–5.05	–3.52	–2.61	–1.48	–0.07
MHI-5 calm	1.02	1.01			1.00	–0.39	–0.20	0.99	0.53	–0.99	–3.87	–2.30	–1.29	–0.39	2.92
MHI-5 depressed	0.90	0.87				1.00	–0.17	0.93	–0.54	–1.92	–4.50	–3.35	–2.18	–0.78	1.18
MHI-5 happy	1.10	1.10					1.00	0.03	0.31	–0.68	–3.76	–2.07	–0.92	0.11	3.24
Model fit	PSR	Cronbach's alpha													
Reliability	.82	.86													

Note. PSR = person separation reliability; PCA = principal component analysis.

TABLE 3B
MHI-5 fit to the Rasch models for the analyses using B) an anchoring of the mean thresholds or C) a multigroup analysis

Mean treshold anchoring SwiSCI on SHS

	Item fit		Residual correlations					Dimensionality		Item difficulty					
MHI-5 items	Infit	Outfit	MHI-5 nervous	MHI-5 down	MHI-5 calm	MHI-5 depressed	MHI-5 happy	Eigenvalue	PCA	Location	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
MHI-5 nervous	1.01	1.00	1.00	–0.15	–0.13	–0.29	–0.49	1.56	0.59	–1.65	–3.42	–3.01	–2.14	–0.91	1.15
MHI-5 down	0.77	0.77		1.00	–0.30	0.01	–0.28	1.48	–0.19	–2.37	–3.86	–3.25	–2.66	–1.72	–0.41
MHI-5 calm	1.00	0.98			1.00	–0.37	–0.19	0.99	0.43	–1.44	–4.42	–2.62	–1.64	–0.79	2.30
MHI-5 depressed	0.89	0.85				1.00	–0.14	0.93	–0.49	–2.29	–4.98	–3.61	–2.45	–1.14	0.71
MHI-5 happy	1.15	1.15					1.00	0.05	–0.44	–1.69	–6.39	–2.60	–1.37	–0.39	2.31
Model fit	PSR	Cronbach's alpha													
Reliability	.80	.86													

Multigroup analysis SHS

	Item fit		Residual correlations					Dimensionality		Item difficulty					
MHI-5 items	Infit	Outfit	MHI-5 nervous	MHI-5 down	MHI-5 calm	MHI-5 depressed	MHI-5 happy	Eigenvalue	PCA	Location	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
MHI-5 nervous	1.12	1.11	1.00	–0.22	–0.24	–0.25	–0.42	1.45	–0.14	–1.69	–3.54	–2.35	–1.05	0.17	
MHI-5 down	0.95	0.92		1.00	–0.30	0.01	–0.21	1.43	–0.50	–2.41	–3.79	–2.86	–1.90	–1.09	
MHI-5 calm	1.02	1.03			–0.40	–0.31	–0.14	1.05	0.61	–1.43	–3.19	–2.20	–1.45	1.14	
MHI-5 depressed	0.91	0.86				1.00	–0.17	0.98	–0.48	–2.29	–3.74	–2.86	–1.79	–0.77	
MHI-5 happy	1.04	1.04					1.00	0.08	0.35	–1.64	–3.38	–2.58	–1.66	1.04	
Model fit	PSR	Cronbach's alpha													
Reliability	.63	.77													

(Table 3B continues)

Table 3B (continued)

Multigroup analysis SwiSCI

	Item fit		Residual correlations					Dimensionality		Item difficulty					
MHI-5 items	Infit	Outfit	MHI-5 nervous	MHI-5 down	MHI-5 calm	MHI-5 depressed	MHI-5 happy	Eigenvalue	PCA	Location	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
MHI-5 nervous	1.01	1.01	1.00	–0.16	–0.13	–0.29	–0.49	1.56	0.58	–1.69	–3.50	–3.07	–2.17	–0.93	1.13
MHI-5 down	0.78	0.78		1.00	–0.30	0.01	–0.28	1.48	–0.20	–2.41	–3.94	–3.30	–2.69	–1.74	–0.42
MHI-5 calm	1.00	0.98			1.00	–0.38	–0.19	0.99	0.45	–1.43	–4.37	–2.62	–1.64	–0.79	2.31
MHI-5 depressed	0.88	0.85				1.00	–0.14	0.93	–0.49	–2.29	–4.95	–3.60	–2.46	–1.14	0.71
MHI-5 happy	1.14	1.14					1.00	0.05	–0.43	–1.64	–6.14	–2.60	–1.37	–0.39	2.32
Model fit	PSR	Cronbach’s alpha													
Reliability	.63	.86													

Note. PSR = person separation reliability; PCA = principal component analysis.

The Rasch-based equating with B) mean thresholds anchoring or C) multigroup analysis also resulted in good item fit, ordered thresholds, local item independence, and unidimensionality. The reliability of the first anchor analysis, anchoring SwiSCI on the SHS sample, resulted in good reliability (PSR = .80, α = .86). When measured with the PSR, the reliability of the multigroup analysis was lower, as previously, due to the skewed nature of the SHS sample which represented 93.26% of the total sample in the multigroup analysis (PSR = .63; α = .86).

Evaluation of Equivalence across Different Methodologies

Table 4 shows the mean and standard deviation (*SD*) of the MHI-5 scores for each survey, by gender, age group, and methodology, and, when applicable, shows if the difference between the mean-transformed scores is statistically significant. Appendix provides supplemental mean comparisons within different health condition groups by gender and age, respectively, to further investigate the extent of differences and commonalities that the methodologies may produce.

For all analyses, the transformed scores applying the rescaling strategies resulted in higher means for the population or subgroups than the transformed scores generated with the Rasch-based approaches. This difference can easily be explained. While the linear rescaling of the 0-25 scale to a 0-100 scale expands all raw scores with a Factor 4, the rescaling factor from raw to Rasch transformed score varies along the continuum; it is smaller for moderate scores and increases more and more toward the extremes; for example, a raw score of 2 may equal 2 × 4 = 8 when linearly rescaled but 2 × 9.06 = 18.12, (Factor 9 increase), on the transformed logit scale. This has of course an impact on the mean score. In the present analysis, the two Rasch-based approaches did not produce identical but very similar results.

When using the 0-100 scores for further statistical analysis across surveys, age, and gender, the significance of the differences in means corresponded across methodologies. For example, we observed age differences in MHI-5 scores in the SwiSCI population, independently of the applied methodology. Also, independently of the transformation methodology applied, differences in mean MHI-5 scores were significant between surveys and between men and women within surveys. This indicates that the relative change in mean estimates is very similar across methodologies, as shown graphically in Figure 2 for the example of age by survey, using age < 30 as reference.

Finally, in one case differences arose when comparing the surveys by health condition (Appendix). The linear rescaling strategy did not indicate significant differences between the mean scores for cancer in the gender and age comparisons, while the Rasch-based approaches did.

Discussion

The overall aim of this study was to evaluate the metric properties of the MHI-5 and to test the equivalence of the 0-100 scores derived from two versions of the MHI-5 in comparison to two Rasch-based score equating methodologies, using data from a general population health survey (SHS) and a condition specific survey (SwiSCI). All Rasch analyses, either separately per version

TABLE 4

Mean, standard deviation (*SD*) by survey, survey and gender, survey and age group with test for statistical differences in the means for the MHI-5 (0-100 rescaled or transformed Rasch score)

		Method	SHS *Mean (SD)*	SwiSCI *Mean (SD)*				*t*	*p*	
Survey comparison		A) Score rescaling	80.80 (15.79)	71.91 (17.79)				18.50	< .001	
		B) Mean threshold anchoring	68.16 (18.12)	62.84 (14.65)				13.11	< .001	
		C) Multigroup analysis	65.00 (12.87)	61.85 (15.03)				7.76	< .001	
Characteristic	Survey	Method	Female *Mean (SD)*	Male *Mean (SD)*				*t*	*p*	
Gender	SHS	A) Score rescaling	79.01 (16.41)	82.79 (14.82)				–17.16	< .001	
		B) Mean threshold anchoring	65.99 (17.88)	70.58 (18.08)				–18.03	< .001	
		C) Multigroup analysis	63.45 (12.70)	66.71 (12.85)				–18.03	<. 001	
	SwiSCI	A) Score rescaling	68.13 (18.54)	73.38 (17.28)				–4.93	< .001	
		B) Mean threshold anchoring	59.81 (14.22)	64.01 (14.65)				–5.01	< .001	
		C) Multigroup analysis	58.74 (14.60)	63.06 (15.03)				–5.01	< .001	
Characteristic	Survey	Method	Age < 30 *Mean (SD)*	Age 30-45 *Mean (SD)*	Age 45-60 *Mean (SD)*	Age 60-75 *Mean (SD)*	Age > 75 *Mean (SD)*	*F*	*p*	
Age	SHS	A) Score rescaling	80.08 (14.06)	79.61 (15.24)	79.70 (17.03)	83.55 (15.84)	82.80 (15.27)	119.71	< .001	
		B) Mean threshold anchoring	66.19 (16.02)	66.06 (16.67)	67.21 (18.69)	72.41 (19.31)	71.23 (19.16)	293.88	< .001	
		C) Multigroup analysis	63.60 (11.39)	63.50 (11.84)	64.32 (13.28)	68.02 (13.73)	67.18 (13.62)	294.37	< .001	
	SwiSCI	A) Score rescaling	73.46 (16.82)	70.72 (18.35)	71.08 (17.78)	73.98 (17.62)	72.00 (16.80)	1.19	.28	*ns*
		B) Mean threshold anchoring	63.57 (13.60)	61.99 (14.83)	62.04 (14.33)	64.84 (15.27)	62.79 (14.37)	1.95	.16	*ns*
		C) Multigroup analysis	62.60 (13.94)	60.98 (15.21)	61.03 (14.70)	63.90 (15.67)	61.80 (14.74)	1.95	.16	*ns*

Note. *p*-values followed by *ns* (nonsignificant) indicate the absence of sample or groups effect. The level of significance is Bonferroni corrected for repeated measurement ($p < .017$).

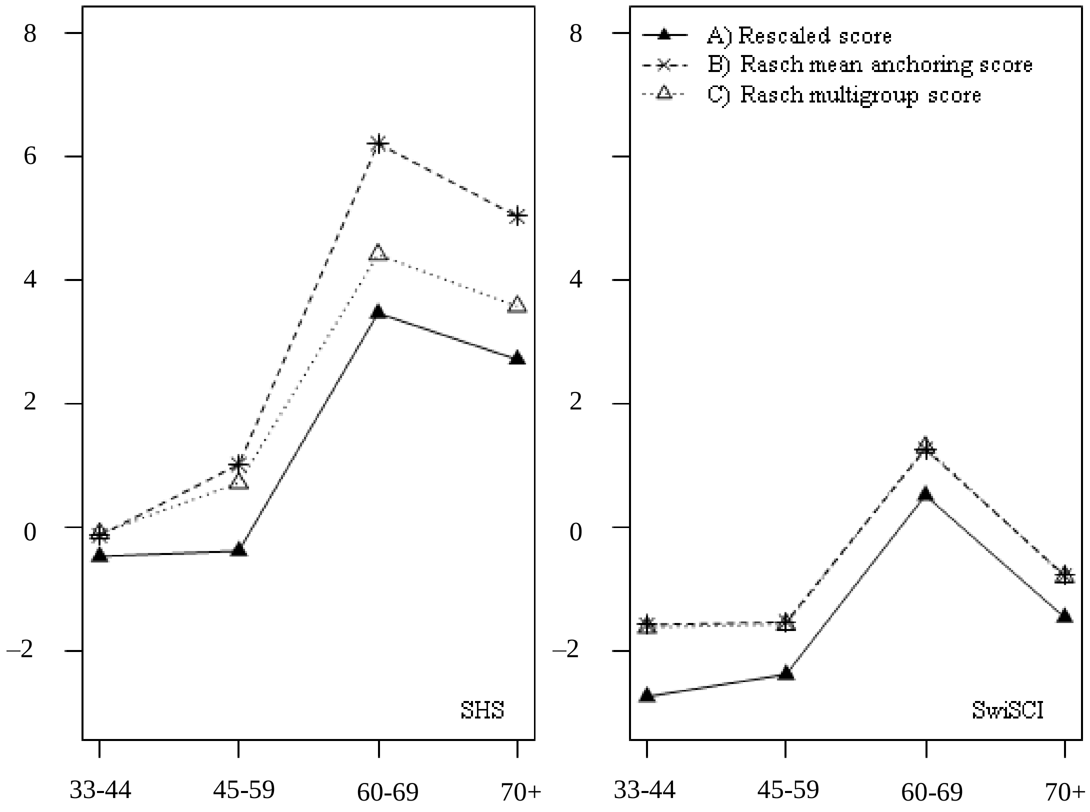

FIGURE 2
Relative change in estimated average MIH-5 score across age groups and by equating method.

or jointly with mean difficulty anchoring or a multigroup analysis, supported the good metric properties of the MHI-5, that is, good fit, ordered thresholds, local item independence, and unidimensionality. The reliability of the MHI-5 was good in the SwiSCI sample. For the SHS sample, the reliability was good when controlling for the skewness of the distribution of the persons' abilities and resulting ceiling effects. This study confirms the comparability of the two MHI-5 versions, showing that to a large extent relative differences using the 0-100 linearly rescaled raw score are similar to what is found with equated Rasch-transformed 0-100 scores.

The analysis of the transformed scores first showed that the 0-100 linear rescaling strategy tends to produce much higher average scores compared to the two Rasch-based strategies. This can be attributed to distributional characteristics of the raw scores in the populations, such as an elevated number of extreme values which receive more weight with the Rasch-based score transformation approaches. Differences, such as in the cancer example, can be entirely attributed to the score distribution, such as more extreme values in one gender or age group receiving higher weights with the Rasch-based approaches than with linear rescaling. However, we observed generally high congruence of the relative differences between sample characteristics across methodologies as all methodologies lead to similar findings in the statistical comparisons across subgroups.

It can be expected that our findings are highly reliable, at least in the present context, as the sample sizes of the SHS and SwiSCI populations are substantial. Also, the good metric properties found for the MHI-5 make the equating of the two versions an almost ideal case. Often, the interoperability of items between different surveys is only partly supported. Different data collection modes, variations in the phrasing of questions, or the use of modified response options present challenges for scale equating. Also, metric limitations such as lack of invariance, small number of anchors, and anchors with a narrow range of difficulty impact on the reliability of the equated scales (Cook & Petersen, 1987; Dorans et al., 2007). In most cases, the use of an equating methodology can only be expected to be reliable in the presence of some common overlapping items that can serve as anchors. In that sense, scales like the MHI-5 with good metric properties will improve the accuracy of the scale equating.

This study has also to be discussed in the light of its limitations. First, anchoring on the mean item difficulties is rarely applied. It can be expected that anchoring the items of two surveys on a common average item difficulty may go along with some loss of measurement precision compared to other methodologies that also fix the items' difficulty thresholds. In this special setting with different rating scales, the mean item difficulties estimated from one dataset were the best available mathematical estimates of the true values to anchor the clinical dataset on. Second, we are conscious that the comparison of the three rescoring methodologies by observing changes and commonalities in the statistical significance of mean scores across subgroups is not the most critical approach. It allows only to gain some insight into the relative magnitude of differences one may expect when using the one or the other score transformation strategy. Third, one must be aware that the MHI-5 is a scale with good metric properties, so that the raw score is expected to be a reliable measure for a person's mental health. In that sense, the results of this study are not at all generalizable to other equating contexts especially when scales with less ideal metric properties are applied.

Finally, the MHI-5 is a screening tool with a high level of standardization shown to efficiently identify depression and anxiety by means of only five screening questions. In public health research, sound and reliable screening tools allowing to detect highly prevalent symptoms and conditions are crucial. In the case of the MHI-5, early detection of high prevalence conditions, such as depression and anxiety, is important so that treatment lags and gaps, which considerably increase the burden for the affected person and for the society, can be prevented .

CONCLUSION

Our study provides evidence that the two versionsof the MHI-5 scale have good metric properties and confirms that the two versions of the MHI-5 can be equated and used for comparisons across surveys in general and disease-specific populations. A linear rescaling strategy to equate the two MHI-5 versions can be supported but may be reliable only with lower proportions of extreme scores. For higher reliability of the resulting person-ability estimates, we recommend using a modern test theoretical approach for equating the versions of the MHI-5.

FUNDING AND ACKNOWLEDGEMENTS

This study has been financed in the framework of the Swiss Spinal Cord Injury cohort study (SwiCI; http://www.swici.ch) supported by the Swiss Paraplegic Foundation. The authors thank the Swiss Federal Office of Statistics for permission to use the data of the 2012 Swiss Health Survey.

REFERENCES

Andrich, D. (2015). Components of variance of scales with a subscale structure using two calculations of coefficient α. *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana*, *52*(2), 6-33.

Andrich, D., Sheridan, B. S., & Luo, G. (2010). RUMM 2030: Rasch unidimensional measurement models [Computer Software]. Perth, Western Australia: RUMM Laboratory.

BFS Sektion Gesundheit. (2014). *Die schweizerische Gesundheitsbefragung 2012 in Kürze: Konzept, Methode, Durchführung* [*The Swiss Health Survey 2012 in short: Concept, methods, implementation*]. Neuenburg, Switzerland: Bundesamt für Statistik.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Person reliability, item reliability, and more. In W. J. Boone, J. R. Staver, & M. S. Yale (Eds.), *Rasch analysis in the human sciences* (pp. 217-234). Dordrecht, The Netherlands: Springer.

Brinkhof, M. W., Fekete, C., Chamberlain, J. D., Post, M. W., & Gemperli, A. (2016). Swiss national community survey on functioning after spinal cord injury: Protocol, characteristics of participants and determinants of non-response. *Journal of Rehabilitation Medicine*, *48*, 120-130. doi:10.2340/16501977-2050

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied psychological measurement*, *11*(3), 225-244.

Cuijpers, P., Smits, N., Donker, T., ten Have, M., & de Graaf, R. (2009). Screening for mood and anxiety disorders with the five-item, the three-item, and the two-item mental health inventory. *Psychiatry Research*, *168*, 250-255. doi:10.1016/j.psychres.2008.05.012

DeVivo, M. J., Biering-Sorensen, F., New, P., & Chen, Y. (2011). Standardization of data analysis and reporting of results from the international spinal cord injury core data set. *Spinal Cord*, *49*, 596-599. doi:10.1038/sc.2010.172

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Fekete, C., Segerer, W., Gemperli, A., & Brinkhof, M. W. (2015). Participation rates, response bias and response behaviours in the community survey of the Swiss Spinal Cord Injury cohort study (SwiSCI). *BMC Medical Research Methodolology*, *15*, 80. doi:10.1186/s12874-015-0076-0

Friedman, B., Heisel, M., & Delavan, R. (2005). Validity of the SF-36 five-item mental health index for major depression in functionally impaired, community-dwelling elderly patients. *Journal of the American Geriatrics Society*, *53*, 1978-1985. doi:10.1111/j.1532-5415.2005.00469.x

Haley, S. M., McHorney, C. A., & Ware, J. E., Jr. (1994). Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *Journal of Clinical Epidemiology*, *47*(6), 671-684.

Hardouin, J. B., Conroy, R., & Sebille, V. (2011). Imputation by the mean score should be avoided when validating a patient reported outcomes questionnaire by a Rasch model in presence of informative missing data. *BMC Medical Research Methodology*, *11*, 105. doi:10.1186/1471-2288-11-105

Hawthorne, G., Densley, K., Pallant, J. F., Mortimer, D., & Segal, L. (2008). Deriving utility scores from the SF-36 health instrument using Rasch analysis. *Quality of Life Research*, *17*, 1183-1193. doi:10.1007/s11136-008-9395-5

Kiefer, T., Robitzsch, A., & Wu, M. (2014). TAM: Test Analysis Modules. Retrieved from http://CRAN.R-project.org/package=TAM

Kim, S. H., & So, W. Y. (2015). Rasch validation of the SF-36 for assessing the health status of Korean older adults. *Journal of Physical Therapy Science*, *27*, 601-606. doi:10.1589/jpts.27.601

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(2), 105-124.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174. doi:10.1007/bf02296272

McCabe, C. J., Thomas, K. J., Brazier, J. E., & Coleman, P. (1996). Measuring the mental health status of a population: A comparison of the GHQ-12 and the SF-36 (MHI-5). *British Journal of Psychiatry*, *169*(4), 516-521.

Means-Christensen, A. J., Arnau, R. C., Tonidandel, A. M., Bramson, R., & Meagher, M. W. (2005). An efficient method of identifying major depression and panic disorder in primary care. *Journal of Behavioral Medicine*, *28*, 565-572. doi:10.1007/s10865-005-9023-6

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value Health*, *18*, 25-34. doi:10.1016/j.jval.2014.10.005

R Core Team. (2016). R: A language and environment for statistical computing. Wien, Austria: R Foundation for Statistical Computing.

Raczek, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., . . . Sullivan, M. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA project. International quality of life assessment. *Journal of Clinical Epidemiology*, *51*(11), 1203-1214.

Raîche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis. *Rasch Measurement Transactions*, *19*(1), 1012.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish National Institute for Educational Research.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care*, *45*, S22-31. doi:10.1097/01.mlr.0000250483.85507.04

Rost, J., Carstensen, C. H., & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? [Are the Big Five Rasch-scaleable?]. *Diagnostica*, *45*, 119-127. doi:10.1026//0012-1924.45.3.119

Rumpf, H. J., Meyer, C., Hapke, U., & John, U. (2001). Screening for mental health: Validity of the MHI-5 using DSM-IV Axis I psychiatric disorders as gold standard. *Psychiatry Research*, *105*(3), 243-253.

Scoggins, J. F., & Patrick, D. L. (2009). The use of patient-reported outcomes instruments in registered clinical trials: Evidence from ClinicalTrials.Gov. *Contemporary Clinical Trials*, *30*, 289-292. doi:10.1016/j.cct.2009.02.005

Shackelford, M., Farley, T., & Vines, C. L. (1998). A comparison of women and men with spinal cord injury. *Spinal Cord*, *36*(5), 337-339.

Stekhoven, D. J., & Buhlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112-118. doi:10.1093/bioinformatics/btr597

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism*, *57*, 1358-1362. doi:10.1002/art.23108

van Leeuwen, C. M., van der Woude, L. H., & Post, M. W. (2012). Validity of the mental health subscale of the SF-36 in persons with spinal cord injury. *Spinal Cord*, *50*, 707-710. doi:10.1038/sc.2012.33

Ware, J. E., Jr. (2000a). *SF-36 Health Survey: Manual and interpretation guide*. Lincoln, RI: QualityMetric.

Ware, J. E., Jr. (2000b). SF-36 health survey update. *Spine*, *25*(24), 3130-3139.

Ware, J. E., Jr., Kosinski, M., & Dewey, J. E. (2001). *How to score Version 2 of the SF-36 health survey : Standard & acute forms* (3rd ed.). Lincoln, RI: QualityMetric.

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473-483.

Wright, B. D. (1992). Raw scores are not linear measures: Rasch vs. Classical test theory CTT comparison *Rasch Measurement Transactions*, *6*(1), 208.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Yen, W. (1993). Scaling peformance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

APPENDIX

Statistical comparison of the mean the MHI-5 (0-100 rescaled or transformed multigroup analysis) across health conditions and age or gender groups

Health condition	Method	Mean female	Mean male	*t*	*p*		Mean < 30 yrs.	Mean 30-45 yrs.	Mean 45-60 yrs.	Mean 60-75 yrs.	Mean > 75 yrs.	*F*	*p*	
Cardio-vascular	Score rescaling	79.14	82.36	–5.36	< .001		75.77	76.27	78.13	82.83	82.23	52.14	< .001	
	Multigroup analysis	63.89	66.90	–6.18	< .001		60.42	61.09	63.39	67.41	66.13	55.43	< .001	
	Mean threshold anchoring	66.61	70.84	–6.18	< .001		61.73	62.66	65.89	71.56	69.76	55.44	< .001	
Allergies	Score rescaling	77.14	80.95	–6.04	< .001		79.01	79.26	77.47	79.73	79.72	0	.98	*ns*
	Multigroup analysis	61.82	64.67	–5.86	< .001		62.54	63.09	62.37	64.76	64.56	5.22	.02	*ns*
	Mean threshold anchoring	63.69	67.70	–5.86	< .001		64.71	65.48	64.47	67.82	67.54	5.19	.02	*ns*
Depression	Score rescaling	62.50	62.64	–0.10	.92	*ns*	64.17	61.52	60.85	63.78	69.92	1.88	.17	*ns*
	Multigroup analysis	52.97	53.24	–0.33	.74	*ns*	53.15	52.32	52.12	54.35	57.47	5.18	.02	*ns*
	Mean threshold anchoring	51.23	51.60	–0.33	.74	*ns*	51.49	50.31	50.03	53.18	57.57	5.18	.02	*ns*
Cancer	Score rescaling	76.28	80.05	–2.37	.02	*ns*	74.62	76.92	75.94	78.37	81.06	4.32	.04	*ns*
	Multigroup analysis	61.49	65.06	–2.89	< .001		59.49	61.38	61.54	63.87	65.06	6.61	.01	
	Mean threshold anchoring	63.23	68.24	–2.89	< .001		60.42	63.07	63.29	66.58	68.25	6.60	.01	
Migraine	Score rescaling	75.78	78.63	–4.10	< .001		77.09	77.14	75.84	77.36	75.29	0.83	.36	*ns*
	Multigroup analysis	60.67	63.11	–4.74	< .001		61.18	61.47	60.99	63.22	62.54	2.11	.15	*ns*
	Mean threshold anchoring	62.07	65.50	–4.73	< .001		62.79	63.20	62.52	65.66	64.69	2.10	.15	*ns*
Asthma	Score rescaling	75.60	80.88	–4.54	< .001		79.45	77.44	75.49	78.90	79.54	0	.96	*ns*
	Multigroup analysis	61.03	64.84	–4.30	< .001		62.86	61.71	61.40	64.11	64.89	2.31	.13	*ns*
	Mean threshold anchoring	62.58	67.94	–4.30	< .001		65.16	63.54	63.10	66.91	68.01	2.30	.13	*ns*
Diabetes	Score rescaling	79.31	83.01	–3.15	< .001		79.83	78.99	76.18	83.66	84.01	16.19	< .001	
	Multigroup analysis	64.03	67.50	–3.62	< .001		63.41	62.55	62.04	68.12	68.00	21.61	< .001	
	Mean threshold anchoring	66.79	71.69	–3.62	< .001		65.93	64.71	64.00	72.55	72.38	21.61	< .001	

(Appendix continues)

Appendix (continued)

Health condition	Method	Mean female	Mean male	*t*	*p*		Mean < 30 yrs.	Mean 30-45 yrs.	Mean 45-60 yrs.	Mean 60-75 yrs.	Mean > 75 yrs.	*F*	*p*	
Arthrosis	Score rescaling	77.09	79.22	–2.57	.01		72.94	75.69	74.25	80.02	80.86	37.94	< .001	
	Multigroup analysis	62.30	64.58	–3.61	< .001		59.74	60.73	60.59	64.96	65.09	39.08	< .001	
	Mean threshold anchoring	64.37	67.57	–3.61	< .001		60.76	62.15	61.96	68.10	68.29	39.09	< .001	
Complete paraplegia	Score rescaling	71.24	74.74	–1.98	.05	*ns*	71.35	73.94	73.56	75.75	70.77	0.70	.40	*ns*
	Multigroup analysis	60.77	63.87	–2.11	.04	*ns*	61.18	62.59	63.00	65.03	60.46	1.26	.26	*ns*
	Mean threshold anchoring	61.78	64.81	–2.11	.04	*ns*	62.19	63.56	63.95	65.94	61.49	1.25	.26	*ns*
Incomplete paraplegia	Score rescaling	66.42	71.89	–3.03	< .001		75.71	67.67	68.06	73.69	69.30	0.29	.59	*ns*
	Multigroup analysis	57.73	62.02	–3.01	< .001		64.65	59.41	58.29	64.05	59.71	0.28	.60	*ns*
	Mean threshold anchoring	58.82	63.00	–3.01	< .001		65.55	60.46	59.37	64.98	60.76	0.28	.59	*ns*
Complete tetraplegia	Score rescaling	68.50	74.38	–1.36	.18	*ns*	68.00	72.15	75.79	73.50	75.20	1.52	.22	*ns*
	Multigroup analysis	59.85	63.54	–0.99	.33	*ns*	57.53	62.39	64.93	62.71	62.78	0.99	.32	*ns*
	Mean threshold anchoring	60.89	64.49	–0.99	.33	*ns*	58.63	63.37	65.84	63.68	63.75	1.00	.32	*ns*
Incomplete tetraplegia	Score rescaling	67.49	73.18	–2.58	.01		75.25	69.69	69.75	72.49	77.56	0.28	.60	*ns*
	Multigroup analysis	57.74	63.34	–3.13	< .001		63.84	60.03	60.38	62.80	66.80	0.72	.40	*ns*
	Mean threshold anchoring	58.83	64.29	–3.13	< .001		64.78	61.07	61.41	63.76	67.66	0.72	.40	*ns*

Note. *p*-values followed by *ns* (nonsignificant) indicate the absence of sample or groups effect. The level of significance is Bonferroni corrected for repeated measurement ($p < .017$).