Original Article

SAGE Open Medicine

SAGE Open Medicine
3: 2050312115585041

DOI: 10.1177/2050312115585041
smo.sagepub.com
SAGE

Factor structure and measurement invariance of the Health Education Impact Questionnaire: Does the subjectivity of the response perspective threaten the contextual validity of inferences?

Gerald R Elsworth[1], Sandra Nolte[1,2] and Richard H Osborne[1]

Abstract

Objective: On-going evidence is required to support the validity of inferences about change and group differences in the evaluation of health programs, particularly when self-report scales requiring substantial subjectivity in response generation are used as outcome measures. Following this reasoning, the aim of this study was to replicate the factor structure and investigate the measurement invariance of the latest version of the Health Education Impact Questionnaire, a widely used health program evaluation measure.

Methods: An archived dataset of responses to the most recent version of the English-language Health Education Impact Questionnaire that uses four rather than six response options (N = 3221) was analysed using exploratory structural equation modelling and confirmatory factor analysis appropriate for ordered categorical data. Metric and scalar invariance were studied following recent recommendations in the literature to apply fully invariant unconditional models with minimum constraints necessary for model identification.

Results: The original eight-factor structure was replicated and all but one of the scales (Self Monitoring and Insight) was found to consist of unifactorial items with reliability of ⩾0.8 and satisfactory discriminant validity. Configural, metric and scalar invariance were established across pre-test to post-test and population sub-groups (sex, age, education, ethnic background).

Conclusion: The results support the high level of interest in the Health Education Impact Questionnaire, particularly for use as a pre-test/post-test measure in experimental studies, other pre–post evaluation designs and system-level monitoring and evaluation.

Keywords

Health program evaluation, construct validation, measurement invariance, patient self-report measure, Health Education Impact Questionnaire

Date received: 25 April 2014; accepted: 26 March 2015

Introduction

Construct validation of descriptive and causal interpretations derived from measurements in education, health and the social sciences is an on-going and responsive process, requiring the generation of new evidence to support emerging conclusions. 'Validity is a property of inferences'. Not '… a property of designs or methods, for the same design may contribute to more or less valid inferences under different circumstances' (p. 34).[1] Similarly, responses to a measurement instrument may vary with a change in the research context such that '…

[1]Public Health Innovation, Population Health Strategic Research Centre, Deakin University, Melbourne, VIC, Australia
[2]Medical Clinic, Department of Psychosomatic Medicine, Charité – Universitätsmedizin Berlin, Berlin, Germany

Corresponding author:
Richard H Osborne, Public Health Innovation, Population Health Strategic Research Centre, Deakin University, Burwood Campus, 221 Burwood Highway, Melbourne, VIC 3125, Australia.
Email: richard.osborne@deakin.edu.au

each interpretation of the scores needs to be validated …' by a '… program of research to support the … application of the tool in relation to an increasing range of interpretations …' (p. 2)[2] (see also Moss[3,4] and references therein). Paralleling the use of the concept in relation to the importance of contextual factors in evaluating health equity interventions,[5] we think of 'contextual validity' in healthcare measurement as documenting, describing and understanding the extent and limits to which an instrument (questionnaire, rating scale, etc.) will yield consistent and valid interpretations in the varying contexts and purposes for which it is administered.

One important change in the measurement context encountered in the health sciences is the use of a patient self-report questionnaire for baseline (pre-test) and follow-up (post-test) measurement in the evaluation of a health promotion or health education intervention. A phenomenon known as 'response shift' is arguably a common occurrence.[6] Response shift entails a possible change in respondent perspective engendered by the educational or social context of the intervention and may produce various qualitatively different changes in the appraisal process during the generation of item responses.[7–9] These changes in appraisal and response can threaten the validity of any inference about change in a construct that is measured, for example, by a multi-item composite scale. Furthermore, it has been argued that data derived from measures that require an increasing amount of subjective personal judgement in the generation of a response (so-called perception-based and evaluation-based measures) will be most vulnerable to response shift bias.[10,11]

An analogous phenomenon to response shift may also be present when comparisons are made across respondent groups. Different life experiences across age groups, males and females, cultural and educational background and so on may engender differing perspectives on the meaning of questionnaire items and consequent frameworks for responding that may, in turn, generate systematic differences in response and consequently factor structure across groups.[12] Hence, the concept of contextual validity is of critical concern in both longitudinal and cross-sectional measurement in healthcare evaluation.

From a measurement perspective, the concept of response shift is closely related to longitudinal measurement invariance, or, more specifically, factorial invariance when measurement invariance is conceptualised within a factor analytic framework.[13] When comparisons between factor or scale-score means and construct interrelationships across time, respondent groups or settings are based on composite scales, it is assumed that the measurement structure is unchanged, that is, each item continues to make an invariant contribution to the target construct.[14–16] If invariance assumptions are violated, the validity of these comparisons is threatened.

Health Education Impact Questionnaire

The Health Education Impact Questionnaire (heiQ) is a perception- and evaluation-based measure that was developed 10 years ago to be a user-friendly, relevant and psychometrically sound instrument for the comprehensive evaluation of patient education programs and activities.[17] The present version (Version 3) measures eight constructs by multi-item composite scales. The English-language heiQ has been used in Australia, Canada, Great Britain, New Zealand, Singapore and the United States, translated into 20 other languages and applied across a wide range of evaluation studies from national and regional quality management systems to experimental trials.[18] The heiQ was chosen for this study as its widespread use (particularly in longitudinal studies evaluating the short- and medium-term impact of chronic disease self-management programs), and the comprehensive range of constructs measured justifies careful and on-going construct validation. The heiQ scales, number of items in each of the scales, a brief description of the construct being measured and a sample item for each scale are listed in Table 1. (The heiQ is copyright to Deakin University, Australia. Information on how to access the full questionnaire for research, course evaluation and translation into languages other than English is available on the heiQ website (http://www.deakin.edu.au/health/research/phi/heiQ.php)).

The heiQ was developed following a grounded approach that included the generation of a program logic model for health education interventions and concept mapping workshops to identify relevant constructs.[17] Based on the results of the workshops, 69 candidate items were written and tested on a construction sample of 591 respondents drawn from potential participants of patient education programs and persons who had recently completed a program. The 69 items were reduced to a 42-item questionnaire measuring eight constructs and again tested on a replication sample of 598 respondents drawn from a broader population of attendees at a general hospital outpatient clinic and community-based self-management programs. Both confirmatory factor analysis (CFA) and item response theory (IRT) were used for item selection and scale refinement. In the revisions leading to Version 3, the number of Likert-type response options was reduced from six to four on advice from users (they are now *strongly disagree, disagree, agree, strongly agree* with *slightly* options removed) and the number of items was reduced to 40.

The heiQ is scored as eight separate scales using simple summation and dividing the summed score by the number of items such that the total score has the same potential range as an individual item (1–4). Thus, higher scores on all scales except emotional distress (ED) are regarded as a desirable outcome of a health education program. Scores on the ED scale are typically not reversed such that lower scores are regarded as a positive outcome.

The general factor structure of the original version of the heiQ was replicated by Nolte and colleagues[19,20] who investigated its factorial invariance[21–23] in the context of response shift bias across a traditional pre–post design as well as across a post-test compared with a retrospective pre-test ('then-test') design. Nolte's results supported the

Table 1. heiQ Version 3: scale names, acronyms, number of items and construct descriptions.

Scale	Acronym	Number of items	Construct description
Health-directed activities	HDA	4	This construct relates to a tangible change in lifestyle, specifically related to healthful behaviours such as exercise and relaxation/ recreation (e.g. 'On most days of the week, I do at least one activity to improve my health (e.g. walking, relaxation, exercise)')
Positive and active engagement in life	PAEL	5	This construct covers motivation to be actively engaged in life-fulfilling activities (e.g. 'I am doing interesting things in my life')
Emotional distress	ED	6	This construct measures overall negative affect including worry, depression and anger (e.g. 'I often worry about my health')
Self-monitoring and insight	SMI	6	This construct captures the individuals' ability to monitor their condition, and their physical and/or emotional responses that lead to insight and appropriate actions to self-manage (e.g. 'I carefully watch my health and do what is necessary to keep as healthy as possible')
Constructive attitudes and approaches	CAA	5	This construct aims to measure how individuals view the impact of their condition(s) on their life (e.g. 'I do not let my health problems control my life')
Skill and technique acquisition	STA	4	This construct aims to capture the knowledge-based skills and techniques that persons acquire (or re-learn) to help them cope with symptoms and health problems (e.g. 'When I have symptoms, I have skills that help me cope')
Social integration and support	SIS	5	This construct aims to capture the positive impact of social engagement and support that evolves through interaction with others (e.g. 'If I need help, I have plenty of people I can rely on')
Health service navigation	HSN	5	This construct covers an individual's understanding of and ability to interact with a range of health organisations and health professionals, including confidence and ability to communicate with healthcare providers to get needs met (e.g. 'I communicate very confidently with my doctor about my healthcare needs')

heiQ: Health Education Impact Questionnaire.

stability of the factor structure across measurement occasions and questionnaire formats (configural invariance) and the metric and scalar invariance of the heiQ when used in the traditional pre–post design. While, in this design, approximately 10% of items were found to show some form of non-invariance from pre-test to post-test, Nolte[20] concluded that '… group level response shifts were not strong enough in any of the datasets to threaten the validity of comparing actual pretest with posttest data …' (p. 118). However, factorial invariance was less clearly supported when the heiQ was used in the then-test design where approximately one-third of the heiQ items exhibited some form of non-invariance.

Given the wide application of the heiQ and its role in making clinical, program and policy decisions, further validation of its measurement structure using pre-test and post-test data is warranted. Furthermore, conclusions about the differences between scale-score means in longitudinal or cross-sectional designs are only justifiable if invariance of factor loadings and, particularly, item intercepts (or thresholds) is confirmed.[24–26] Using a large independent sample, this article presents analyses of the 40 heiQ items retained in Version 3 where the simplified four ordinal response options are used. We seek to add further rigour and validity to the investigation of program impact and group differences when using the heiQ by addressing configural, metric (or 'weak') and scalar (or 'strong') factorial invariance[15,16,27] over time and across important population sub-groups (sex, age, education, language spoken at home and country of birth).

We thus tested the hypotheses that the originally proposed structure of the heiQ was replicated with the revised response options and reduced item number, and that the measurement properties of the scales were sufficiently invariant to justify valid comparison of factor or scale-score means and interrelationships. The initial focus was to test the hypothesis that the specified clusters of items had acceptable unidimensionality, discriminant validity and reliability. Unidimensionality is a fundamental and necessary condition for assigning meaning to constructs measured by composite scales.[28–30] It is defined as the existence of a single latent trait (variable) underlying each hypothesised item cluster[30,31] and thus as a properly specified independent clusters measurement model having acceptable fit to the data.[32,33] Subsequently, we investigated configural, metric and scalar invariance across time and population sub-groups. Configural invariance entails the demonstration of consistent item clusters as identified by the pattern of zero (or near-zero) and non-zero factor loadings across groups or time points while, similarly, metric invariance entails equality of factor loadings and scalar invariance equality of item intercepts (or alternatively item thresholds if

the data are ordered categorical and analysed using a weighted least-squares approach).[23,25]

Methods

Data

A dataset containing responses from all programs that utilised a data management website at both pre-test and post-test for the period July 2007–December 2012 was used. While the majority of respondents were participants in Australian chronic disease self-management programs run in hospitals, community health facilities or complementary care providers, data from a small number of similar programs in Canada were also included. After removal of records from those who made no response to more than 50% of the heiQ items at either pre-test or post-test, 3221 cases were available for analysis.

These data were gathered by a large number of individual healthcare organisations for their own monitoring and evaluation purposes using an 'opt-in' consent process. The de-identified data were provided to the heiQ research team specifically for on-going validation studies only. Some archived data were also gathered as part of a pilot health education quality assurance study funded by the Australian Government Department of Health and Ageing. Ethics approval for the use of these data for scale validation purposes was obtained from the University of Melbourne Human Research Ethics Committee.

Statistical approach

In re-examining the factor structure and measurement invariance of the heiQ, both unrestricted (frequently labelled 'exploratory factor analysis' – EFA) and restricted factor analyses (CFA) were employed in a complementary manner, taking advantage of the exploratory structural equation modelling (ESEM) routine in Mplus[34] for the unrestricted analyses.

The complementary use of EFA/ESEM and CFA can be very instructive for scale validation.[35] By specifying that each item should load on only one factor and constraining all 'non-target' factor loadings to exactly 0 in the form of a strictly specified 'independent clusters' model, CFA is frequently problematic for the analysis of multi-item multi-scale questionnaires.[36–39] In particular, model fit may not reach acceptable standards and, even if it does, inter-factor correlations may be upwardly biased and lead to spurious conclusions about construct interrelationships. Additionally, particularly in large models, the incremental use of modification indices (MIs) to improve model fit can be confusing and potentially misleading. It is frequently recommended that parameters set to 0 in an initial model should be freely estimated on the basis of large MIs only if this is 'theoretically justified',[40] but this is a loosely interpreted caveat. A disadvantage of CFA can therefore be that 0 loadings are 'forced' on non-target factors even though associations may exist. Hence, while a finding of arguably acceptable fit for multi-factor CFA models may appear to support the conclusion of independent item clusters, the results may conceal salient evidence that associations that appear as high inter-factor correlations are better interpreted as cross-loadings indicating factorial complexity of items. To address this threat to model validity (and hence a clear conclusion of configural stability and invariance in the present version of the heiQ), a combination of ESEM and CFA was used with the expectation that the results would be consistent, and thus, evidence in support of the hypothesised structure would be strengthened.[35]

Model estimation and fit

The mean and variance-adjusted weighted least-squares (WLSMV) estimator, suitable for the analysis of ordered categorical data, was used for all ESEM and CFA. WLSMV provides robust standard errors and a robust mean- and variance-adjusted chi-square fit statistic[40] (designated χ^2_{WLSMV} herein). Mplus also provides various 'close-fit' statistics: the comparative fit index (CFI), the Tucker-Lewis index (TLI) and the root mean square error of approximation (RMSEA). (Mplus also reports the weighted root mean residual for analyses using the WLSMV estimator. This statistic is, however, regarded as experimental and Mplus currently advises that it not be used (LK Muthén, Mplus discussion list, 4 January 2013.)) Mplus utilises, by default, a pairwise approach to missing data with WLSMV estimation.[41]

As the sample size for this study was large, the primary focus for model acceptance was the extent of model fit (and misfit) indicated by the indices of close fit. Indicative threshold values for these were CFI $\geqslant$ 0.95, TLI $\geqslant$ 0.95 and RMSEA $\leqslant$ 0.06, while a value of $\leqslant$0.08 for the RMSEA was taken to indicate a 'reasonable' fit.[42–44]

Factor rotation in ESEM

A wide range of rotation options for ESEM are available in Mplus. A rotation approach that is designed to provide an approximation to Thurstone's[45] original conception of 'simple structure' that allows for possible multi-factorial items is oblique Geomin.[35,46] The default epsilon value of 0.01 for four or more factors was used.[46] While an independent clusters solution was hypothesised, oblique Geomin was chosen to provide evidence for the factorial complexity of the items should such complexity be indicated.

Configural, metric and scalar invariance

Factorial invariance of the heiQ was investigated following recent advocacy for a refocus of the usual statistical approach and development of a revised methodology.[47] The investigation of metric and scalar invariance is predicated on the

demonstration of satisfactory configural invariance. Given the potential hazards in the use of CFA alone to establish configural invariance discussed above, both ESEM and CFA were used for this stage of the investigation.

To test the hypothesis that an eight-factor model was a satisfactory fit to the correlations between the 40 items of the heiQ, irrespective of the specific configuration of the factors.[48] ESEM analyses were conducted, one for the pre-test data and one for the post-test. From six to eight factors were extracted in each analysis. The ESEM analyses were followed by validation of the specific multi-factor configuration by fitting and examining the results of CFA and ESEM (Geomin rotation) models to the pre-test and post-test data separately.

Following the demonstration that the hypothesised eight-factor model was a satisfactory fit to the data, metric and scalar invariance were investigated using full eight-factor CFA and ESEM models and scale-by-scale analyses. Typically, metric and scalar invariance are investigated by fixing factor loadings and item intercepts (or thresholds) to equality across groups or time in a hierarchical manner.[14,21] But, it has been argued by Raykov et al.[47] that the structural equation models used in this approach have significant limitations, and, in general, do not provide a complete and unconditional statistical assessment of either metric or scalar invariance. As the scalar invariance model is typically not nested within that for metric invariance, a statistical test of whether the additional constraints result in a meaningful reduction in fit is not possible. Furthermore, as the metric invariance model normally requires that the loading of one factor indicator (e.g. questionnaire item) be fixed to 1.0 in each group, a complete test of the equality of factor loadings is not possible[47] (pp. 955–956). Hence, it is recommended that at present, metric and scalar invariance be investigated using only an unconditional model with a complete set of constraints for both metric and scalar invariance but minimum constraints necessary for model identification. Multi-factor CFA and ESEM models were fitted across sex, age, education level, country of birth and home-language groups separately using the CONFIGURAL and SCALAR 'convenience features' available in Mplus 7.1 (see Version 7.1 Mplus Language Addendum available at http://www.statmodel.com/) to achieve the minimal constraints necessary for identification required by the Raykov and Marcoulides approach. (The Mplus 'convenience features' resulted, for the CONFIGURAL model, in the factors in both groups being identified by setting one loading to 1.0, while all other loadings and the factor variances were freely estimated. Also, the scale factors were set to 1.0, while all item thresholds were estimated. For the SCALAR model, factors were similarly identified by setting one factor loading in each scale to 1.0, while the other loadings were constrained to be equal across groups and factor variances were free. Factor means, however, were fixed to zero in the reference group only and were freely estimated in the comparison group. All item thresholds were constrained to be equal across groups, while the scale factors were fixed to 1.0. The Delta parameterisation was used for both analyses.) The chi-square difference test appropriate for WLSMV estimation (provided by the DIFFTEST) (see p. 5 of Version 7.1 Mplus Language Addendum (available at http://www.statmodel.com/).) was used to assess the change in model fit between the configural and scalar models only. For across-occasion measurement invariance, an analogous model was tested in which pre-test factor means were fixed to 0 and factor variances to 1.0 but were free to vary at post-test. Additionally, the longitudinal character of the model was taken into account by freely estimating correlations between parallel item residuals across time.

Reliability

Reliability was assessed using the Mplus coding for composite scale reliability developed by Raykov.[33,49] Composite scale reliability is defined as the ratio of true variance to total variance in a homogeneous cluster of test items and is obtained as a robust maximum likelihood estimate of this ratio. While Cronbach alpha can be seriously biased if the test items are not at least tau-equivalent (i.e. have, in practice, equal factor loadings) and in the presence of correlated residuals, the maximum likelihood estimator of composite scale reliability is, instead, consistent and unbiased.[33] Cronbach alpha is, however, also presented for possible comparison with the results of similar scale validation studies.

Discriminant validity

Discriminant validity of the heiQ constructs was studied by inspecting the size of the inter-factor correlations in both CFA and ESEM results[50] and by comparing the inter-factor shared variance estimates with the average variance extracted (AVE) by each factor involved.[51,52]

Results

Replicating the structure and reliability of heiQ Version 3

Model fit statistics for the ESEM analyses to establish the number of factors are shown in the upper part of Table 2. According to the close-fit criteria, all but the six-factor model at pre-test satisfied all thresholds for a good fit. In corresponding 'scree' plots of the eigenvalues of the two correlation matrices, there were six eigenvalues >1.0 while the plot for the post-test data showed a clear discontinuity between the eigenvalues of the eighth and ninth factors. It was concluded that eight factors, as hypothesised, would be satisfactory for subsequent investigation of the factorial structure of the data, minimising potential problems with underfactoring.[53]

Table 2. Fit statistics for exploratory factor analyses (ESEM) and CFA of pre-test and post-test heiQ data separately.

Model	χ^2_{WLSMV}	d.f.	*p*	CFI	TLI	RMSEA (90% CI)
ESEM Pre-test 6 Factors	5616.48	555	<0.0000	0.961	0.945	0.053 (0.52–0.54)
ESEM Pre-test 7 Factors	3776.87	521	<0.0000	0.975	0.962	0.044 (0.43–0.45)
ESEM Pre-test 8 Factors	2602.28	488	<0.0000	0.984	0.974	0.037 (0.035–0.038)
ESEM Post-test 6 Factors	4759.79	555	<0.0000	0.969	0.956	0.048 (0.47–0.50)
ESEM Post-test 7 Factors	3086.12	521	<0.0000	0.981	0.971	0.039 (0.38–0.40)
ESEM Post-test 8 Factors	2357.36	488	<0.0000	0.986	0.978	0.034 (0.033–0.036)
CFA Pre-test 8 Factors	8391.71	712	<0.0000	0.940	0.935	0.058 (0.057–0.059)
CFA Post-test 8 Factors	6935.57	712	<0.0000	0.953	0.949	0.052 (0.051–0.053)

ESEM: exploratory structural equation modelling; CFA: confirmatory factor analysis; heiQ: Health Education Impact Questionnaire; d.f.: degrees of freedom; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; CI: confidence interval.

Validation of the specific configuration of these eight factors was then conducted by fitting eight-factor CFA models to the pre-test and post-test data separately and tabulating and examining the standardised factor loadings from the eight-factor ESEM analysis.

Fit statistics for the two eight-factor CFA independent cluster models are shown in the lower part of Table 2. While model fit did not reach the 'satisfactory fit' thresholds established above, they suggest a closer fit than is frequently found with similar self-report psychological data.[38] As can be seen in the upper-right triangle of Table 3, however, inter-factor correlations are high in a number of instances (particularly between self-monitoring and insight (SMI) and skill and technique acquisition (STA) at both pre-test and post-test). As these high inter-factor correlations may be the result of the 'forced' zero cross-loadings in the CFA, ESEM analyses were examined to further investigate the possibility that some items may be factorially complex, thus questioning the homogeneity of the scales.

As anticipated, model fit was considerably improved when cross-loadings were not fixed precisely to 0 in the eight-factor ESEM analyses (see the appropriate rows in the upper part of Table 2). Given the very good fit of the ESEM models, the potential for model improvement by including correlated item residuals was not explored.[50]

Table 4 shows the factor pattern for the Geomin obliquely rotated solutions for pre-test and post-test data separately with the order of the factors in the raw output rearranged to correspond to the hypothesised heiQ factors. It can be seen that the unrestricted ESEM analyses resulted in factor patterns that, while showing some evidence of factorial complexity, corresponded well with the hypothesised structure based on the original scales.[17] First, there was clear evidence of at least moderate loadings of all hypothesised factors on their target items and no evidence of any substantial factorial complexity in the constituent items for six of the a priori heiQ factors, namely, health-directed activities (HDA), positive and active engagement in life (PAEL), emotional distress (ED), constructive attitudes and approaches (CAA), skill and technique acquisition (STA) and health service navigation (HSN) (i.e. all hypothesised factor loadings were $\geqslant$0.4 while there were no secondary loadings on the constituent items $\geqslant$0.3). Furthermore, for one additional scale (social integration and support (SIS)), factorial complexity was found for only one item (Item 45) at both pre-test and post-test and all hypothesised loadings were $\geqslant$0.4. The factor pattern for the SMI items was, however, somewhat more complex. For this scale, all but one item appeared factorially complex with one or two loadings $\geqslant$0.3 from non-hypothesised factors. In all but one instance, these non-target loadings were higher than the respective target loading; four of the factorially complex items appeared in post-test heiQ data, while two factorially complex items were found in pre-test heiQ data, with Item 21 being complex in both pre- and post-test data.

Correlations between the latent variables were typically considerably smaller in the ESEM analyses than those in the CFA (Table 3). As Marsh et al.[54] point out, the '… inappropriate imposition of zero factor loadings …' (p. 472) on non-target items in CFA '… usually leads to distorted factors with positively biased factor correlations …'. The median absolute (with ED reflected) inter-factor correlation in the ESEM analyses for the pre-test data was 0.38 (range: 0.06–0.60) and for the post-test 0.41 (range: 0.09–0.65) compared with parallel results for the CFAs of 0.61 (0.29–0.83) and 0.67 (0.30–0.88). The maximum inter-factor correlation of 0.65 in the ESEM results is well below the threshold of 0.80–0.85 that is frequently recommended as indicating poor discriminant validity[55] (p. 131). The inter-factor correlations between the factor pair that was identified as potentially confounded in the ESEM (SMI with STA) were 0.24 at pre-test and 0.39 at post-test in the ESEM analysis compared with 0.83 at pre-test and 0.88 at post-test in the CFA (Table 3, values in bold type). Additionally, the correlations between SMI and ED (−0.34 and −0.37 in the CFA) were very small but marginally significant in the ESEM analysis (at pre-test: −0.06 (95% confidence interval (CI) = −0.10 to −0.01) and at post-test: −0.09 (95% CI = −0.15 to −0.05)).

Discriminant validity was also investigated by calculating the AVE by each of the heiQ factors in both the CFA and ESEM analyses at pre-test and post-test and comparing these values to the appropriate estimates of the shared variance

Table 3. Factor correlations in the CFA (upper right) and ESEM (lower left) analyses.

Scale	HDA		PAEL		ED		SMI		CAA		STA		SIS		HSN	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
HDA			0.58	0.68	−0.29	−0.31	0.55	0.66	0.49	0.54	0.47	0.53	0.40	0.48	0.33	0.40
PAEL	0.49	0.55			−0.56	−0.51	0.69	0.78	0.83	0.83	0.66	0.73	0.62	0.68	0.52	0.56
ED	−0.22	−0.24	−0.39	−0.37			−0.34	−0.37	−0.64	−0.60	−0.43	−0.41	−0.43	−0.41	−0.29	−0.30
SMI	0.25	0.34	0.29	0.34	−0.06	−0.09			0.67	0.77	**0.83**	**0.88**	0.59	0.63	0.75	0.77
CAA	0.38	0.44	0.60	0.59	−0.57	−0.53	0.23	0.27			0.75	0.81	0.74	0.78	0.63	0.70
STA	0.38	0.43	0.46	0.51	−0.33	−0.29	**0.24**	**0.39**	0.52	0.65			0.70	0.71	0.73	0.77
SIS	0.34	0.36	0.46	0.45	−0.32	−0.32	0.18	0.15	0.54	0.65	0.51	0.54			0.71	0.71
HSN	0.28	0.32	0.37	0.39	−0.21	−0.20	0.40	0.43	0.47	0.57	0.55	0.61	0.57	0.57		

Estimated correlations between the SMI and STA factors in the contrasting CFA and ESEM analyses are given in bold. The ESEM analysis yields considerably lower estimates, arguably a result of allowing non-target loadings to be estimated rather than fixed precisely to 0.
CFA: confirmatory factor analysis; ESEM: exploratory structural equation modelling; HDA: health-directed activities; PAEL: positive and active engagement in life; ED: emotional distress; SMI: self-monitoring and insight; CAA: constructive attitudes and approaches; STA: skill and technique acquisition; SIS: social integration and support; HSN: health service navigation.

between each pair of constructs. The presence of sufficient discriminant validity between the constructs is demonstrated when the shared inter-factor variance is less than the AVE of each of the factors involved.[51,52] By this criterion, in the CFAs, there was evidence of insufficient discriminant validity between SMI and PAEL, CAA, STA and HSN and also PAEL and CAA at both pre-test and post-test. In the ESEM analysis, there was evidence of insufficient discriminant validity between SMI and HSN at pre-test and between SMI and HDA, SMI and STA and SMI and HSN at post-test (full results are available in Supplementary Table 1). Taking the results together and considering the likely over-estimation of inter-factor correlations in the CFAs, the results suggest that the discriminant validity of the SMI construct from HSN and STA in particular may not be fully established.

Composite scale reliability with 95% confidence intervals (italicised) based on robust standard errors and, for comparison with other studies, Cronbach α (in parenthesis) for the scales in Version 3 estimated from the pre-test data are as follows – HDA: 0.83/*0.82–0.84* (0.83), PAEL – 0.83/*0.82–0.84* (0.83), ED: 0.86/*0.86–0.87* (0.86), SMI: 0.74/*0.72–0.76* (0.74), CAA: 0.88/*0.87–0.89* (0.87), STA: 0.80/*0.78–0.81* (0.80), SIS: 0.88/*0.88–0.89* (0.88) and HSN: 0.85/*0.84–0.86* (0.85). All reliability estimates were ⩾0.8 with the exception of that for SMI.

Configural invariance

The complementary CFA and ESEM analyses described above replicated the eight-factor structure of the heiQ and the homogeneity of seven of the scales, thus clearly establishing the basis for a detailed investigation of the across-time and across-group invariance of the scales. The factorial identity and homogeneity of the SMI scale, however, was not so clearly established, but it was retained for the invariance analyses to seek further information on its psychometric performance.

To establish configural invariance across time, a 16-factor CFA was conducted with no cross-loadings and with correlated residuals allowed only between identical items at pre-test and post-test. To identify the model, factor variances were set to 1.0 at pre-test and post-test, while factor loadings and item intercepts were freely estimated as were all inter-factor correlations. Fit statistics for this model were as follows: $\chi^2_{WLSMV} = 16213.12$, 2920 degree of freedom (d.f.), $p < 0.0000$, RMSEA = 0.038 (90% CI = 0.037–0.038), CFI = 0.942 and TLI = 0.937. While the CFI and TLI are (marginally) below acceptable threshold values, the RMSEA is within the threshold for a good fit. As the CFI and TLI tend to demonstrate worse fit in models with large numbers of measured variables and thus do not function well under these conditions,[56] (p. 349) and as the model is very tightly specified (zero cross-loadings and only identical-item correlated residuals), this analysis provided initial support for the hypothesis of satisfactory longitudinal configural invariance for the heiQ. (A similar ESEM model was also fitted to the data for comparison to the CFA model. This was specified in a similar manner such that the pre-test items were restricted to load only on pre-test factors but cross-loadings were allowed and, similarly, post-test items were restricted to load only on post-test factors. As anticipated, allowing for cross-loadings improved model fit considerably so that the close-fit statistics were well within acceptable limits ($\chi^2_{WLSMV} = 5622.47$, 2728 d.f., $p < 0.0000$; CFI = 0.987, TLI = 0.985, RMSEA = 0.018 (90% CI = 0.017–0.019).)

Similar CFAs (but without allowing estimation of any residual correlations) were conducted across groups formed by sex, age (split at the median, 63 years), education (year 10 or less, above year 10), country of birth (Australia vs. overseas) and language spoken at home (English vs other) for the pre-test and post-test data separately. For these analyses, models were identified using the CONFIGURAL specification in Mplus 7.1 described in section 'Methods'. The results are shown in Table 5.

Table 4. Standardised factor loadings for two eight-factor ESEM analyses of the pre-test and post-test heiQ data – Geomin rotation ($\varepsilon = 0.01$) (N = 3221).

Item	HDA		PAEL		ED		SMI		CAA		STA		SIS		HSN	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
1	0.73	0.60														
12	0.80	0.80														
18	0.81	0.75														
25	0.82	0.80														
3			0.68	0.69												
6			0.51	0.51												
9			0.81	0.77												
14			0.59	0.53												
20			0.52	0.47												
5					0.63	0.69										
8					0.67	0.70										
17					0.78	0.85										
19					0.70	0.74										
24					0.87	0.84										
28					0.79	0.74										
4							0.38	0.26								*0.32*
7							0.45	0.25				*0.40*				
16							0.49	0.50								
21							0.36	0.33			*0.54*	*0.50*				
23	*0.32*						0.25	0.31								
27							0.43	0.37				*0.30*				
35									0.47	0.58						
42									0.69	0.75						
44									0.62	0.64						
47									0.76	0.77						
49									0.61	0.65						
31											0.55	0.60				
33											0.66	0.68				
34											0.74	0.76				
38											0.50	0.54				
30													0.88	0.84		
36													0.78	0.75		
39													0.64	0.74		
43													0.81	0.81		
45													0.43	0.43	*0.33*	*0.30*
32															0.74	0.85
37															0.82	0.81
40															0.72	0.64
41															0.65	0.73
46															0.69	0.79

All loadings ⩾0.3 shown with those not hypothesised in italics; hypothesised loadings <0.3 also shown (underlined). Items and factor loadings are arranged according to the hypothesised structure.

ESEM: exploratory structural equation modelling; heiQ: Health Education Impact Questionnaire; HDA: health-directed activities; PAEL: positive and active engagement in life; ED: emotional distress; SMI: self-monitoring and insight; CAA: constructive attitudes and approaches; STA: skill and technique acquisition; SIS: social integration and support; HSN: health service navigation.

All models met the threshold for a good fit indexed by the RMSEA while values for the CFI and TLI either satisfied the threshold for good fit or were very close to it. Given the requirement that the models contained no cross-loadings or correlations between item residuals, these results suggest a quite satisfactory fit of the eight-factor configural model across the selected groups at both pre-test and post-test. It should be noted, however, that the numbers of cases in the country-of-birth and home-language analyses are unbalanced. Chen[57] has pointed out that 'unequal sample sizes … might affect changes in goodness of fit indices …' (p. 469). In a Monte Carlo study, Chen[57] showed that in across-group

Table 5. CFAs of two-group eight-factor models of the heiQ for pre-test and post-test separately with factor loadings and item thresholds freely estimated testing for configural invariance.

Grouping; time	N	χ^2_{WLSMV}	d.f.	*p*	CFI	TLI	RMSEA
Sex; pre-test	Female = 1829; male = 1250	8219.45	1424	<0.0000	0.947	0.942	0.056 (0.055–0.057)
Sex; post-test	Female = 1829; male = 1250	6946.80	1424	<0.0000	0.959	0.955	0.050 (0.049–0.051)
Age; pre-test	Younger = 589; older = 629	4270.16	1424	<0.0000	0.941	0.935	0.057 (0.055–0.059)
Age; post-test	Younger = 589; older = 629	3871.28	1424	<0.0000	0.956	0.952	0.051 (0.053–0.057)
Education; pre-test	Year 10 or less = 1043; >year 10 = 1941	7695.25	1424	<0.0000	0.949	0.944	0.054 (0.053–0.056)
Education; post-test	Year 10 or less = 1043; >year 10 = 1941	6896.89	1424	<0.0000	0.958	0.954	0.051 (0.051–0.052)
Country of birth; pre-test	Aust = 2454; O'seas = 767	8045.39	1424	<0.0000	0.951	0.946	0.054 (0.053–0.055)
Country of birth; post-test	Aust = 2454; O'seas = 767	6769.80	1424	<0.0000	0.965	0.962	0.048 (0.047–0.049)
Home-language; pre-test	English = 2980; other = 241	6568.45	1424	<0.0000	0.960	0.956	0.047 (0.046–0.049)
Home-language; post-test	English = 2980; other = 241	5356.73	1424	<0.0000	0.974	0.971	0.041 (0.040–0.043)

CFA: confirmatory factor analysis; heiQ: Health Education Impact Questionnaire; d.f.: degree of freedom; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation.

tests of invariance of factor loadings, item intercepts and residual variances, estimated changes in the CFI and RMSEA (among other fit indices) were reduced when sample sizes were unequal and therefore '… invariance tests are more likely to fail to detect invariance' (p. 499). (Parallel eight-factor ESEM configural invariance models were also estimated. With one minor exception, the close-fit indices for these models were within the thresholds established for this study. The least well-fitting model was for country of birth at pre-test where the TLI was marginally below the threshold of 0.95. Close-fit indices for this model were RMSEA = 0.054 (90% CI 0.053–0.055), CFI = 0.951, TLI = 0.946.)

Longitudinal and across-group metric and scalar invariance of the heiQ

To investigate across-time metric and scalar invariance, a 16-factor CFA model was fitted to the pre-test and post-test data combined. In this model, factor loadings and item thresholds were constrained to be equal, while only the residuals of item pairs across pre-test and post-test were allowed to be correlated. All inter-factor correlations were estimated. For identification, factor means were fixed to 0 at pre-test and factor variances to 1.0. Means and variances were freely estimated at post-test. Fit statistics for the 16-factor longitudinal CFA model were as follows: $\chi^2_{WLSMV} = 26001.26$, 3069 d.f., $p < 0.0000$, RMSEA = 0.048 (90% CI = 0.048–0.049), CFI = 0.900, TLI = 0.897. While the CFI and TLI are below acceptable threshold values, the RMSEA is well within the threshold for a good fit. As the model was, again, very tightly specified, this analysis provided initial support for the hypothesis of satisfactory across-time measurement invariance for the heiQ. It was, however, followed up by fitting longitudinal models to the indicators of each of the eight constructs separately, specified in a similar manner to the eight-factor model. Fit statistics and the ranges of the inter-item polychoric correlations and standardised factor loadings for these models are shown in Table 6. Following established criteria for the CFI, TLI and RMSEA, it is apparent that these single-scale measurement invariance models all fitted the data well, supporting the hypothesis of an acceptable level of across-time measurement invariance across all scales.

A final set of CFAs addressed the question of metric and scalar invariance across five salient demographic groups: sex, age, educational level, country of birth and language spoken at home. These models utilised the SCALAR model command in Mplus 7.1. Model fit statistics are shown in Table 7. The results for the Mplus chi-square difference tests for comparison of the chi-square estimates derived from the configural compared with the scalar models are also shown. Fit was clearly satisfactory (RMSEA < 0.06; CFI and TLI > 0.95) for all models with the exception of those for age at pre-test where both the CFI and TLI values were below the recommended thresholds and (very marginally) for sex and education at pre-test. The chi-square difference tests results were variable, with some not significant (NS) but the majority $p < 0.05$. Chi-square difference tests are, however, known to be dependent on sample size in a similar manner to chi-square tests of model fit in a single group,[58] and as the sample size for this study was very large, they should be interpreted with some caution. The generally very good fit of the scalar models across groups provides substantial support for the invariance of the heiQ scales across sex, education and ethnic background. Metric and scalar invariance across age groups may be less well established as already noted in the test of configural invariance. (Parallel ESEM analyses were conducted for scalar invariance over pre-test to post-test and across the socio-demographic groups. The fit of the longitudinal ESEM model was very similar to the fit of the parallel CFA model in that the CFI and TLI were below the acceptable thresholds while the RMSEA was clearly <0.05 ($\chi^2_{WLSMV} = 19583.82$, 2845 d.f., $p < 0.0000$; RMSEA = 0.043 (90% CI = 0.042–0.043), CFI = 0.927, TLI = 0.919). All across-group models (including those for age) for both pre-test and post-test fitted the data very well on all 'close-fit' criteria. The results suggest that when the factorial complexity of the items observed in a small number of the heiQ scales (notably SMI) is allowed for, the heiQ scales are acceptably invariant across pre-test to post-test and clearly

Table 6. Inter-item correlations, factor loadings and fit statistics for CFAs of eight separate longitudinal (pre-test and post-test) models for the heiQ scales testing for metric and scalar invariance.

Scale	CFI	TLI	RMSEA	Range of inter-item polychoric correlations	Range of factor loadings
HDA	0.992	0.992	0.049 (0.044–0.055)	Pre-test 0.58–0.67 Post-test 0.56–0.68	Pre-test 0.75–0.85 Post-test 0.74–0.83
PAEL	0.991	0.992	0.041 (0.037–0.046)	Pre-test 0.52–0.68 Post-test 0.57–0.69	Pre-test 0.74–0.85 Post-test 0.76–0.87
ED	0.991	0.991	0.045 (0.042–0.049)	Pre-test 0.41–0.73 Post-test 0.47–0.71	Pre-test 0.66–0.84 Post-test 0.66–0.84
SMI	0.967	0.969	0.049 (0.046–0.053)	Pre-test 0.36–0.51 Post-test 0.33–0.57	Pre-test 0.62–0.71 Post-test 0.59–0.77
CAA	0.995	0.996	0.039 (0.034–0.043)	Pre-test 0.63–0.75 Post-test 0.65–0.77	Pre-test 0.78–0.86 Post-test 0.79–0.86
STA	0.989	0.989	0.054 (0.049–0.060)	Pre-test 0.54–0.74 Post-test 0.54–0.75	Pre-test 0.69–0.86 Post-test 0.71–0.88
SIS	0.994	0.995	0.048 (0.044–0.052)	Pre-test 0.62–0.78 Post-test 0.63–0.79	Pre-test 0.76–0.90 Post-test 0.77–0.90
HSN	0.989	0.990	0.055 (0.051–0.059)	Pre-test 0.56–0.74 Post-test 0.66–0.76	Pre-test 0.78–0.84 Post-test 0.82–0.86

CFA: confirmatory factor analysis; heiQ: Health Education Impact Questionnaire; d.f.: degree of freedom; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; HDA: health-directed activities; PAEL: positive and active engagement in life; ED: emotional distress; SMI: self-monitoring and insight; CAA: constructive attitudes and approaches; STA: skill and technique acquisition; SIS: social integration and support; HSN: health service navigation.

Table 7. CFAs of two-group eight-factor models of the heiQ with factor loadings and item thresholds fixed to be equal across demographic sub-groups testing for metric and scalar invariance.

Grouping; time	N	'Close-fit' statistics for scalar model			Chi-square difference test scalar against configural		
		CFI	TLI	RMSEA	χ^2_{WLSMV}	d.f.	*p*
Sex; pre-test	Female = 1829; male = 1250	0.950	0.949	0.053 (0.052–0.054)	219.70	104	0.000
Sex; post-test	Female = 1829; male = 1250	0.962	0.961	0.047 (0.046–0.048)	147.79	104	0.003
Age; pre-test	Younger = 589; older = 629	0.941	0.940	0.055 (0.053–0.057)	221.62	104	0.000
Age; post-test	Younger = 589; older = 629	0.959	0.958	0.052 (0.050–0.053)	122.64	104	0.102
Education; pre-test	Year 10 or less = 1043; >year 10 = 1941	0.951	0.949	0.052 (0.051–0.053)	281.29	104	0.000
Education; post-test	Year 10 or less = 1043; >year 10 = 1941	0.960	0.959	0.048 (0.047–0.049)	234.22	104	0.000
Country of birth; pre-test	Aust = 2454; O'seas = 767	0.954	0.953	0.050 (0.049–0.051)	158.02	104	0.001
Country of birth; post-test	Aust = 2454; O'seas = 767	0.968	0.967	0.045 (0.044–0.046)	133.48	104	0.027
Home-language; pre-test	English = 2980; other = 241	0.963	0.962	0.044 (0.043–0.045)	144.93	104	0.005
Home-language; post-test	English = 2980; other = 241	0.977	0.976	0.038 (0.037–0.039)	128.59	104	0.051

CFA: confirmatory factor analysis; heiQ: Health Education Impact Questionnaire; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; d.f.: degree of freedom.

invariant across important socio-economic groups. The model fit statistics for the across-group ESEM analyses are available in the supplementary material (Supplementary Table 2).)

Discussion and conclusion

While patient self-report questionnaires are often used to investigate change in healthcare interventions, their contextual validity, including cross-sectional and longitudinal measurement invariance, is infrequently investigated. Additionally, many such questionnaires comprise items and scales that entail a high level of personal subjective judgement from respondents. The absence of a clear demonstration of measurement invariance when evaluating change and across-group differences threatens the validity of interpretations and conclusions derived from the use of these scales. In this article, using

recently developed factor analytic approaches, we demonstrated measurement invariance of the heiQ. This is an important finding as the heiQ has become widely used to make program and policy decisions – decisions that affect patient care, program implementation and program funding.

Among the principal reasons for the extensive application of the heiQ is that it yields timely and understandable information about the impact of self-management interventions across a variety of chronic conditions.[18] Given this widespread use in different contexts, it is incumbent on the scale developers to provide a framework within which the validity of inferences drawn from the instrument can be supported. While most patient self-report questionnaire development studies provide initial evidence of reliability, factor structure and (possibly) concurrent or predictive validity, on-going research is required to provide rigorous support for the increasing range of inferences drawn from these instruments.[2] When used to assess change across time as well as outcomes across a diverse range of patient groups, the rigorous investigation of their contextual validity is particularly necessary.

Following recent arguments,[35,36,38,54] ESEM was used in this article in combination with CFA to substantiate the hypothesised eight-factor structure of the heiQ. Additionally, the CFA method of investigating configural, metric and scalar invariance applied was recently reviewed and recommended.[47] While there is an extensive literature on invariance testing extending over the past three decades and a consensus that factor analysis provides an appropriate and powerful approach, there remains considerable controversy about the specific CFA (or, indeed, ESEM) models that are most appropriate. The advocated model uses minimal restrictions for identification but full equality constraints for both metric and scalar invariance.[47] If this model yields a satisfactory fit to the data, the way is clear to make valid inferences about possible differences between factor- or scale-score means across groups or time and about possible interrelationships between the invariant construct measures.

The eight-factor structure and configural invariance of the 40-item version of the heiQ were clearly replicated with items consistently aligning well with their hypothesised target construct over the period of a self-management intervention and across salient demographic groups. Furthermore, metric and scalar invariance across time and over demographic groups was well established with the caveats (a) that invariance across age groups may warrant further investigation and (b) that the analyses across country of birth and home language may have reduced sensitivity to detect invariance due to the unbalanced numbers in the compared groups. This finding of metric and scalar invariance is particularly important given that the heiQ items are largely 'perception-based' or 'evaluation-based' where the amount of personal judgement involved in generating a response is large and, particularly for 'evaluation-based' items, the subjectivity of the criteria used to make these judgements is such that comparisons across time and persons may be particularly problematic.

The finding of satisfactory factor structure and psychometric properties for the heiQ in this study also supports the decision to use a simplified four-option response set in later versions of the questionnaire. Both pre-test and post-test factor structures and the reliability of all scales have now been replicated in the analysis of the website data (four response options) and the Nolte[20] study (six response options).

The reliability of the 6-item SMI scale has been found to be consistently lower than that of the other scales[17,20] while, in this study, its discriminant validity was less clearly supported. The factorial complexity of the items in this scale as seen in the ESEM analyses may be contributing to the lower reliability and lack of discriminant validity; however, as the scale measures a construct that is central to a conception of self-management, we believe it should continue to be used with caution while the construct is investigated further. It is interesting to note that the items of the SMI scale that show factorial complexity appear, in most part, to be related to the STA scale. There is quite possibly a strong, perhaps iterative, causal relationship between the constructs measured by these two scales, with the results of self-monitoring and consequent awareness of progression of a chronic condition leading to the person actively seeking new strategies and skills to improve their condition (note that the most strong multi-factorial SMI item is Item 21 – When I have health problems, I have a clear understanding of what I need to do to control them – an item that connotes a clear action orientation to addressing the health problem). This possible causal relationship may lead to a confounding of some items of the SMI scale with the STA and other constructs (HDA, HSN) with consequent cross-loadings and lowered discriminant validity, particularly for those respondents who score high on it.

The poorer model fit and low factor loadings of the SMI scale may also suggest that there are two underlying constructs that are being brought together in the scale: (a) self-monitoring and (b) consequent insight and understanding of, for example, triggers of flare-up of the chronic condition. Further research might explore these issues through in-depth qualitative interviews with individuals scoring at different levels on these two scales and the development and psychometric testing of additional items that could identify the separate constructs. However, despite the factorial complexity of some of its constituent items, the SMI scale shows a satisfactory level of across-time measurement invariance; hence, summed scores on the scale are comparable from pre-test to post-test in the study of self-management education interventions.

Despite the caveats associated with the SMI scale, this study supports the high level of interest in the use of the English-language version of the heiQ, particularly as a pre-test/post-test measure in experimental studies, other pre-test/post-test evaluation designs and system-level monitoring and evaluation. Positive psychometric evaluations of French, German and Japanese translations of the heiQ have been reported[59–61] and independent studies of translations into Danish, Dutch, Canadian French, Italian and Norwegian are

underway, providing support for its use across a wide range of languages, cultures and healthcare systems, and opportunities to establish extensive cross-cultural measurement invariance and contextual validity.

Acknowledgements

The authors wish to acknowledge the very helpful comments of two anonymous reviewers of the original submission.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

This work was partially funded by the Australian Government Department of Health and Ageing as a component of a pilot health education quality assurance study. R.H.O. is a recipient of a National Health and Medical Research Council of Australia Senior Research Fellowship #1059122.

References

1. Shadish WR, Cook TD and Campbell DT. *Experimental and quasi-experimental designs for generalised causal inference*. Belmont, CA: Wadsworth Cengage Learning, 2002.
2. Buchbinder R, Batterham R, Elsworth G, et al. A validity-driven approach to the understanding of the personal and societal burden of low back pain: development of a conceptual and measurement model. *Arthritis Res Ther* 2011; 13: R152.
3. Moss PA. Shifting conceptions of validity in educational measurement: implications for performance assessment. *Rev Educ Res* 1992; 62: 229–258.
4. Moss PA. Reconstructing validity. *Educ Res* 2007; 36: 470–476.
5. Phillips K, Muller-Clemm W, Ysselstein M, et al. Evaluating health inequity interventions: applying a contextual (external) validity framework to programs funded by the Canadian Health Services Research Foundation. *Eval Program Plann* 2013; 36: 198–203.
6. Osborne R, Hawkins M and Sprangers MAG. Change of perspective: a measurable and desired outcome of chronic disease self-management intervention programs that violates the premise of preintervention/postintervention assessment. *Arthritis Rheum* 2006; 55: 458–465.
7. Schwartz CE and Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med* 1999; 48: 1531–1548.
8. Rapkin BD and Schwartz CE. Toward a theoretical model of quality-of-life appraisal: implications of findings from studies of response shift. *Health Qual Life Outcomes* 2004; 2: 1–12.
9. Oort FJ. Using structural equation modeling to detect response shifts and true change. *Qual Life Res* 2005; 14: 587–598.
10. Nolte S, Elsworth GR, Newman S, et al. Measurement issues in the evaluation of chronic disease self-management programs. *Qual Life Res* 2012; 22: 1655–1664.
11. Schwartz CE and Rapkin BD. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health Qual Life Outcomes* 2004; 2: 2–16.
12. Jöreskog KG. Simultaneous factor analysis in several populations. *Psychometrika* 1971; 36: 409–426.
13. King-Kallimanis BL, Oort FJ and Garst GJA. Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *AStA: Adv Stat Anal* 2010; 94: 139–156.
14. Nolte S and Elsworth GR. Factorial invariance. In: Michalos AC (ed.) *Encyclopedia of quality of life and well-being research*. Dordrecht: Springer, 2013, pp. 2146–2148.
15. Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 1993; 58: 525–543.
16. Meredith W and Teresi JA. An essay on measurement and factorial invariance. *Med Care* 2006; 44: S69–S77.
17. Osborne R, Elsworth G and Whitfield K. The Health Education Impact Questionnaire (heiQ): an outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. *Patient Educ Couns* 2007; 66: 192–201.
18. Osborne R, Batterham R and Livingston J. The evaluation of chronic diseases self-management support across settings: the international experience of the Health Education Impact Questionnaire quality monitoring system. *Nurs Clin North Am* 2011; 46: 255–270.
19. Nolte S, Elsworth GR, Sinclair A, et al. Tests of measurement invariance failed to support the application of the 'then-test'. *J Clin Epidemiol* 2009; 62: 1173–1180.
20. Nolte S. *Approaches to the measurement of outcomes of chronic disease self-management interventions using a self-report inventory*. Melbourne, VIC, Australia: School of Global Studies, Social Science and Planning, RMIT University, 2008.
21. Millsap RE and Meredith W. Factorial invariance: historical perspectives and new problems. In: Cudeck R and MacCallum RC (eds) *Factor analysis at 100: historical perspectives and future developments*. Mahwah, NJ: Lawrence Erlbaum Associates, 2007, pp. 131–152.
22. Millsap RE and Olivera-Aguilar M. Investigating measurement invariance using confirmatory factor analysis. In: Hoyle RH (ed.) *Handbook of structural equation modeling*. New York: The Guilford Press, 2012, pp. 380–392.
23. Millsap RE and Yun-Tien J. Assessing factorial invariance in ordered-categorical measures. *Multivar Behav Res* 2004; 39: 479–515.
24. Steenkamp JB and Baumgartner H. Assessing measurement invariance in cross-national consumer research. *J Consum Res* 1998; 25: 78–90.
25. Widaman KF, Ferrer E and Conger RD. Factorial invariance within longitudinal structural equation models: measuring the same construct across time. *Child Dev Perspect* 2010; 4: 10–18.
26. Steinmetz H. Analysing observed composite differences across groups: is partial measurement invariance enough? *Meth Eur J Res Behav Soc Sci* 2013; 9: 1–12.
27. Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Med Care* 2006; 44: S78–S94.
28. Anderson JC and Gerbing DW. Some methods for respecifying measurement models to obtain unidimensional construct measurement. *J Marketing Res* 1982; 19: 453–460.

29. Anderson JC and Gerbing DW. Structural equation modeling in practice: a review and recommended two-step approach. *Psychol Bull* 1988; 103: 411–423.
30. Hattie J. Methodology review: Assessing unidimensionality of tests and items. *Appl Psych Meas* 1985; 9: 139–164.
31. McDonald RP. *Test theory: a unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, 1999.
32. Ping RA Jr. On assuring valid measures for theoretical models using survey data. *J Bus Res* 2004; 57: 125–141.
33. Raykov T. Scale construction and development using structural equation modeling. In: Hoyle RH (ed.) *Handbook of structural equation modeling*. New York: The Guilford Press, 2012, pp. 472–492.
34. Muthén LK and Muthén BO. *Mplus user's guide*. 7th ed. Los Angeles, CA: Muthén & Muthén, 1998–2012.
35. McDonald RP. Semiconfirmatory factor analysis: the example of anxiety and depression. *Struct Equ Modeling* 2005; 12: 163–172.
36. Asparouhov T and Muthén B. Exploratory structural equation modeling. *Struct Equ Modeling* 2009; 16: 397–438.
37. Marsh HW, Liem GAD, Martin AJ, et al. Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): new approaches to key substantive issues in motivation and engagement. *J Psychoeduc Assess* 2011; 29: 322–346.
38. Marsh HW, Muthén B, Asparouhov T, et al. Exploratory structural equation modeling, integrating CFA and EFA: application to students' evaluations of university teaching. *Struct Equ Modeling* 2009; 16: 439–476.
39. Marsh HW, Nagengast B, Morin AJS, et al. Construct validity of the multidimensional structure of bullying and victimization: an application of exploratory structural equation modeling. *J Educ Psychol* 2011; 103: 701–732.
40. Byrne BM. *Structural equation modeling with Mplus: basic concepts, applications and programming*. New York: Routledge, 2012.
41. Asparouhov T and Muthén B. *Weighted least squares estimation with missing data*. Los Angeles, CA: Muthén & Muthén, 2010, http://www.statmodel.com/
42. Browne MW and Cudeck R. Alternative ways of assessing model fit. In: Bollen KA and Long JS (eds) *Testing structural equation models*. Newbury Park, CA: SAGE, 1993, pp. 136–162.
43. Yu C-Y. *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Los Angeles, CA: Education, University of California, Los Angeles, 2002.
44. West SG, Taylor AB and Wu W. Model fit and model selection in structural equation modeling. In: Hoyle RH (ed.) *Handbook of structural equation modeling*. New York; London: The Guilford Press, 2012, pp. 209–231.
45. Thurstone LL. *Multiple factor analysis*. Chicago, IL: University of Chicago Press, 1947.
46. Browne MW. An overview of analytic rotation in exploratory factor analysis. *Multivar Behav Res* 2001; 36: 111–150.
47. Raykov T, Marcoulides GA and Cheng-Hsien L. Measurement invariance for latent constructs in multiple populations: a critical view and refocus. *Educ Psychol Meas* 2012; 72: 954–974.
48. Mulaik SA and Millsap RE. Doing the four-step right. *Struct Equ Modeling* 2000; 7: 36–73.
49. Raykov T. Reliability if deleted, not 'alpha if deleted': evaluation of scale reliability following component deletion. *Br J Math Stat Psychol* 2007; 60: 201–216.
50. Reichenheim M, Souza W, Coutinho ESF, et al. Structural validity of the Tonic Immobility Scale in a population exposed to trauma: evidence from two large Brazilian samples. *PLoS ONE* 2014; 9: e94367.
51. Farrell AM. Insufficient discriminant validity: a comment on Bove, Pervan, Beatty, and Shiu. *J Bus Res* 2010; 63: 324–327.
52. Fornell C and Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. *J Marketing Res* 1981; 18: 39–50.
53. Fabrigar LR, Wegener DT, MacCallum RC, et al. Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods* 1999; 4: 272–299.
54. Marsh HW, Lüdtke O, Muthén B, et al. A new look at the big five factor structure through exploratory structural equation modeling. *Psychol Assess* 2010; 22: 471–491.
55. Brown TA. *Confirmatory factor analysis for applied research*. New York: The Guilford Press, 2006.
56. Kenny DA and McCoach DB. Effect of the number of variables on measures of fit in structural equation modeling. *Struct Equ Modeling* 2003; 10: 333–351.
57. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct Equ Modeling* 2007; 14: 464–504.
58. Cheung GW and Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct Equ Modeling* 2002; 9: 233–255.
59. Epstein J, Osborne RH, Elsworth GR, et al. Cross-cultural adaptation of the heiQ questionnaire: experimental study showed expert committee, not back-translation, added value. *J Clin Epidemiol* 2015; 68: 360–369.
60. Schuler M, Musekamp G, Faller H, et al. Assessment of proximal outcomes of self-management programs: translation and psychometric evaluation of a German version of the Health Education Impact Questionnaire (heiQ™). *Qual Life Res* 2013; 22: 1391–1403.
61. Morita R, Arakida M, Osborne RH, et al. Adaptation and validation of the Japanese version of the Health Education Impact Questionnaire (heiQ-J) for the evaluation of self-management education interventions. *Jpn J Nurs Sci* 2013; 10: 255–266.