OVERCOMING THE SCORING PROBLEM WITH SF-36 COMPONENT SUMMARY SCORES – A METHOD THAT WORKS

Graeme Tucker

Faculty of Health Sciences

Discipline of Medicine

University of Adelaide

South Australia

Australia

A thesis submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy, January 2017

For my family

Table of Contents

Abstract

In this thesis I will discuss the shortcomings of the statistical methods used to derive scoring coefficients for the physical and mental health component summary scores of the Medical Outcomes Study SF-36 and SF-12 health status scales. I will propose an alternative statistical method for generating scoring coefficients for these scales, and produce scoring coefficients for both version 1 and version 2 of the SF-36 and SF-12. I will then demonstrate the superior measurement properties of summary scores generated using my method compared to the proprietary scoring, and discuss the limitations of the SF-36 author's contention regarding international comparisons using this instrument. The study is articulated through several international peer reviewed publications, which provide a progressive story of the body of research. The papers themselves follow on from a wider discussion of the failure of the methods recommended for scoring the physical and mental component summary scores of the SF-36 and SF-12 and provide the argument for alternative scoring methods.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed ..... Date 20/5/17

Acknowledgements

This thesis would not have been possible without the support of my employer, The Epidemiology Branch of the South Australian Department of Health.

It would not have happened at all without the support, nagging and encouragement I received from one of my supervisors, Professor David Wilson. I also wish to acknowledge the assistance and encouragement I received from my other supervisor, Professor Robert Adams, and Dr. Sarah Appleton for reviewing the thesis draft.

Finally, I wish to thank my family for their patient support and encouragement over this seemingly endless process.

Abbreviations

ABS	Australian Bureau of Statistics
ADF	Asymptotically Distribution Free, also known as Weighted Least Squares
AIDS	Acquired Immune Deficiency Syndrome
AQoL	Assessment of Quality of Life
BMI	Body Mass Index
BRFSS	Behavioural Risk Factor Surveillance System
CDC	Centers for Disease Control
CESD	Centre for Epidemiological Studies Depression scale
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CLD	Selim's Chronic Lung Disease index
DOI	Digital Object Identifier
DWLS	Diagonally Weighted Least Squares
EDSS	Expanded Disability Status Scale
EFA	Exploratory Factor Analysis
EQ-5D	EuroQOL Quality of Life instrument
GHQ	General Health Questionnaire
HALex	Health Activities Limitation index
HIV	Human Immunodeficiency Virus
HRQoL	Health Related Quality of Life
HUI	Health Utilities Index
HUI2	Health Utilities Index mark 2
HUI3	Health Utilities Index mark 3
IQOLA	International Quality of Life Assessment
IRT	Item Response Theory
K10	Kessler 10 item anxiety and depression scale
MCS	Mental Component Summary
ML	Maximum Likelihood
MOS	Medical Outcomes Study
nan	Not a number
NHS	National Health Survey
NIH	National Institutes of Health

NNFI	Non-Normed Fit Index, also known as Tucker Lewis Index
NTWLS	Normal Theory Weighted Least Squares
PCA	Principal Components Analysis
PCS	Physical Component Summary
PROMIS	Patient Reported Outcomes Measurement System
QOLS	Quality Of Life Scale
QWB	Quality of Wellbeing Scale
RAND36	Rand Corporation 36 item health survey
RMSEA	Root Mean Square Error of Approximation
SEM	Structural Equation Modelling
SF-12	Medical Outcomes Study shorter Short Form 12 item health survey
SF-20	Medical Outcomes Study Short Form 20 item health survey
SF-36	Medical Outcomes Study Short Form 36 item health survey
SRMR	Standardized Root Mean Square Residual
TLI	Tucker Lewis Index, also known as Non-Normed Fit Index
ULS	Unweighted Least Squares
US	United States
USD	United States Dollars
V1	Version 1
V2	Version 2
WHO	World Health Organisation
WHOQOL BREF	World Health Organisation Quality of Life Brief Instrument
WHOQOL100	World Health Organisation Quality of Life Instrument
WLS	Weighted Least Squares, also known as Asymptotically Distribution Free

CHAPTER 1 - The origins and importance of the SF-36 Health Related Quality of Life Instrument and Aims of This Thesis.

Chapter Content

- Early Developments and Importance of the SF-36.
- Basic Problems With the SF-36
- The Widespread Use of the SF-36 and its Applications.
- Fundamental Problems Identified in Using the Instrument.

Early Developments and Importance of the SF-36.

The SF-36 and the shorter form SF-12 Health Related Quality of Life questionnaires originated from the Medical Outcomes Study (MOS) conducted by the RAND Corporation in the US in the early 1990's. This was a four-year study designed to test the effects of specific characteristics of patients, providers and health systems on outcomes of care. The study had both a cross-sectional component (n=20,000) and a longitudinal component (n=2546)

The salient characteristics of the MOS related to a broadened conceptual framework of health assessment via a patient self-reported perspective. The study claimed that in addition to measuring the biological state of the patient it also provided a wider context in measuring the cost of care and satisfaction with care. The MOS also asked patients how well they were doing in their everyday activities, how they felt and how they rated their health. The introduction of such patient reported outcome measures was a major advance in the conceptual framework of health assessment and pointed to sharpened instrumental measures of health that would provide the opportunity to monitor the results of health care and health systems under different systems of care, different people, different methods of payment, different medical specialities, different styles of doctor-patient interaction and different rates of use of health care resources [1]. The Medical Outcomes Study was designed firstly, to assess whether variations in

patient outcomes were explained by differences in care, speciality, clinician style and technical ability and secondly, to develop tools for the routine monitoring of patient outcomes in medical practice. It was hypothesised that the MOS methodology could be applied across the country, in different systems (both public and private), with the aim of evaluating health outcomes, improving care and, in addition, providing new ways to assess the quality of health services, develop policy and plan use of resources. Since the early 1970's there had been a shift in what constituted health outcomes from traditional mortality and morbidity rates to outcomes of patient functioning and the performance of daily activities, how patients felt and a growth in general assessment of their own health status [1]. The methodological advances in measuring health status along these new dimensions of health assessment, provided by the MOS, were seen as a major step forward [2]. The MOS provided new health indicators and was the first large-scale study to provide health assessment of patients with different physical and psychiatric conditions. One component of these new patient reported outcome measurements was the assessment of health related quality of life. The MOS constructed the MOS 36 item Short Form Health Survey (SF-36) questionnaire to measure health status. This was a thirty-six item questionnaire which was designed to assess eight health scales (vitality, physical functioning, bodily pain, general health perceptions, physical role functioning, social functioning, emotional role functioning, and mental health) and two summary scales (Physical Component Summary (PCS) and Mental Component Summary (MCS)). The SF-36 questionnaire provided data on measures of functioning and well being at the same time as providing the ability to control for disease severity and other important variables [3]. In this study a shorter version of the SF-36, the SF-20, was also used to assess health-related quality of life and the study results provided strong support for the reliability and construct validity of the instrument [3]. This was later followed by further development of the SF-36 from material that had been included in the MOS long form survey. In essence the MOS provided a new philosophy of health measurement that led to a more comprehensive understanding of health status. It also demonstrated the efficacy of using self-administered questionnaires. This was essentially a new era of health assessment.

The original version of the SF-36 came from the Medical Outcomes Study of the Rand Corporation [3] and was followed by the development of a commercial version by a group of researchers led by JE Ware Jnr [4].

A second version of the SF-36, the Rand-36 item health questionnaire, also came from the same MOS source. These questionnaires are identical, the only difference between the RAND 36 and the SF-36 is the scoring. The scoring of the general health and pain scales is different between the two versions. These differences are summarised by Hays et al [5]. For the summary scales, the SF-36 uses an orthogonal rotation of a principal components decomposition, promoted by J E Ware et al.[6] The RAND 36 uses an oblique rotation of the same principle components decomposition, allowing for a correlation between physical and mental health, and is promoted by R D Hays and colleagues. These latter researchers also produced new T-scores. The SF-36 and the Rand 36-Item Health Survey 1.0 correlate 0.99 using data from the MOS longitudinal panel study [5]. Since the conduct of the MOS a group of the original researchers have produced a further commercial version of the SF-36 (SF-36 version 2) [7].

The reasons behind the two original versions identified above of this 36 item health scale lie with the Rand Corporation's policy to provide unrestricted access to instruments for research purposes, and measures developed by Rand have traditionally been placed in the public domain and are available licence free. For the purposes of the research presented here this is an important difference given the different scoring methods promoted in the Rand version. In the research to follow I will show the differences in effect produced by different scoring methods for the summary scales and the importance of allowing for a correlation between physical and mental health. This will be underpinned using data from major population studies.

The efficacy of promoting the SF-36 as a standardised generic measure of health status, suitable for comparing people's health status firstly across regions, systems, cultures and age groups to provide results that could be used for evaluation and policy purposes, then later allow comparisons between nations of health related quality of life status are other claims that will be investigated in this thesis.

Through the production of a number of health publications later discussed I examine two major hypotheses that address the central essence of these claims.

The first hypothesis is that the orthogonal scoring methods employed by the developers of the SF-36 component summary scores produce poor quality scores that conflict with the sub-scale scores.

The second hypothesis examines the credibility of using the United States (US) scoring coefficients and recommended methods to score the SF-36 data of other countries for cross country comparisons of health status.

Basic Problems with the SF-36

The SF-36 is a high quality and widely used health related quality of life instrument, with 8 sub-scales. These sub-scale scores are not statistically optimal but they are adequate and have never been challenged in the literature. The sub-scale scores are an algebraic manipulation of unit weighted additive scales. They would be more accurate if item weights based on Confirmatory Factor Analysis or Item Response Theory analysis were used in their calculation to avoid the approximation of unit weighting. Because the basic structure and design of the SF-36 is so strong, these sub-scale scores are widely considered to produce robust, useable scores. Where a problem has been identified is in the proprietary scoring of the component summary scores for physical and mental health. This stems from use of scoring methods that do not allow for the correlation between physical and mental health [6]. The work in this thesis demonstrates a superior scoring approach that works and corrects for discrepancies that occur with the recommended scoring methods. Secondly, what the SF-36 cannot do under my recommended scoring approach is provide an accurate comparison between country/language groups, because of the need to re-base the scale in each population group studied. The use of US scoring coefficients to standardise the data across countries for the purpose of cross country comparisons as recommended is invalid, since the method used to derive these scoring coefficients is flawed. However, as will be shown later, my approach provides the opportunity for valid comparisons of subgroups within and across each country.

The Widespread Use of the SF-36 and its Applications

The reliability and external validity of the results produced by the SF-36, using the recommended scoring methods, is an important issue globally, given its intention of being able to produce measures of health status that assess health outcomes and may lead to evidence for investment in health services, basing the investments on results of SF-36 research. In this thesis, reliability and validity of the instrument is not questioned, given the robust results of studies that have been conducted across countries producing very high reliability and validity estimates [8-10]. What is in question is the use of the recommended scoring methods for the PCS and MCS component summary scores given the inherent errors in their construction. Application of the recommended scoring methods could prove costly for any country or authority in which the instrument is used and, in addition, the use of US scoring coefficients to standardise results of studies conducted in other countries may lead to spurious comparative inter-country assessment and consequential investments.

The problem is not a minor one. In the current research it was found that the SF-36 Version 1 [3], released in 1988, has been widely used. A search of PubMed (November 2016) identified 16,083 references covering its use in many countries. Of these, many are local translation or validation studies examining its psychometric properties. For example, there were 683 Australian studies, including several validation studies of the Australian version of the SF-36 Version 1 [11-13]. In a major update on the SF-36 in 2000, Ware described the instrument as a generic health measure, as opposed to one that targets a specific age, disease or treatment group and stated that because of its generic nature it had been useful in comparing general and specific groups and populations assessing the relative burden of disease. Ware outlined the changes and updates achieved in version 2 of the SF-36, the assumptions underlying the scale construction and scoring. Most notably he pointed out its translation for use in more than 40 countries in the International Quality of Life Assessment (IQOLA) project [14]. At that point in time the IQOLA project involved translation for use in cost utility studies in 15 countries, in addition to over 100 health care delivery organisations in the US [15]. In this latter paper it was clearly stated that policy makers and health care managers were looking at health care outcomes to achieve best value for their money, as were clinical investigators evaluating pharmaceuticals and

technologies. It was also pointed out that the SF-36 was not specific to any age or treatment group, thus allowing comparisons of the relative burden of different diseases and the benefits of treatment. Given the promoted generic nature of the SF-36, the relative burden of different diseases or conditions and the relative benefits of any treatments could be compared via this instrument and allow for more targeted health investment and treatment initiatives and more precise applications of cost-benefit and cost-effectiveness measures. Expectations from the instrument were high when Aaronson stated that general population norms would facilitate the interpretation of scale scores and make it possible to estimate the relative burden of various medical and psychiatric conditions in each country. [15] It is understandable, however, that a clear driver of the enthusiasm to use this innovation was the potential to reduce health service cost.

Fundamental Problems Identified in Using the Instrument

In this thesis I will deal with the major issues involved in creating the error that affects summary scores to produce conflicting results for some population groups. The publications which form the core substance of the thesis deal with the following errors and propose corrective measures.

- The developers have used an orthogonal decomposition of physical and mental health, and an orthogonal rotation of the solution, so that physical and mental health measures are not correlated. Whilst this approach may be mathematically attractive, it ignores the real life correlation that exists between physical and mental health [16-22]. More accurate scores are obtained for physical and mental health components when these scores are correlated. This fact was recognised at the very early stages of the development of the SF-36 by Professor Ron Hayes, who promoted an oblique (correlated) solution in the RAND36 instrument [5], which as previously stated is identical to the SF-36 apart from the scoring algorithm.
- The developers used an exploratory factor analysis (EFA) of the eight sub-scale scores to produce factor score weights. An EFA of all 35 data items of the SF-36 would be expected to produce a superior result, but there is no guarantee that the solution of the EFA would have the same factor structure as the theoretical structure of the SF-36.

- The developers use unit weighted sub-scale scores rather than a weighted score, which is inherently more accurate. The sub-scale scores are not continuous despite their appearance after they have been manipulated as specified in the scoring manual. They are still ordinal variables with a finite number of possible values. The developers have used Pearson correlations in the analysis of these data (in an EFA), whereas the nature of the data being ordinal infers that polychoric correlations are likely to produce a superior result.
- The developers have produced sub-scales with 3 items (role emotional) and 2 items (bodily pain, social functioning). To generate a weighted sub-scale score based on one factor, however, a congeneric CFA model requires four items, so there is a difficulty producing weighted scores for these sub-scales using CFA.

"The SF health surveys are the most widely used tools in the world for measuring patient reported outcomes." [23] Despite their widespread use, the physical and mental health component summary scores (PCS and MCS) are based on an orthogonal model that ignores the real world correlation between physical and mental health. It also ignores the real world research literature underpinning this correlation [16-22]. As a result this model is flawed and produces PCS and MCS scores that conflict with the sub-scale scores of the eight domains of health. Use of a correlated model using the data items as input overcomes this problem.

Because of the importance of quality of life and the large scale studies occurring worldwide it is important that instruments such as the SF-36 'get it right' methodologically. The demonstration of valid summary scores for the SF-36 provided in this thesis allows others using the instrument to revisit their data and use the methods shown to re-analyse it. This improvement in accuracy is important in decisions about the allocation of tax payer's money to varying health priorities and programs. The provision of valid component summary scores for the SF-36 removes an impediment to its use.

CHAPTER 2 - The Main Points Supported by the Literature in the Case Against Using the SF-36 Recommended Scoring Methods and the Use of United States Scoring Algorithms to Make Inter Country Comparisons of Health Status.

Chapter Content

- Introduction
- The Widespread Use of the SF-36 and the shorter SF12 Health Related Quality of Life Questionnaires.
- Inconsistencies that Arise Using the Recommended Scoring Methods.
- The Correlation between physical and mental health.
- Statistical Challenges in Publishing the Research Arguments.

Introduction

The aim of this section is to show there is a strong case, supported by the literature and large-scale representative population studies, for not using the recommended scoring methods because of analytical errors produced and the impact on study results. In addition, a case is made for not using United States scoring algorithms to make cross- country comparisons of SF-36 scores.

The Widespread Use of the SF-36 and the shorter SF12 Health Related Quality of Life Questionnaires.

Since its promotion as a valid and reliable measure of health related quality of life, the SF-36 and the shorter SF-12 questionnaires have seen widespread use in many countries worldwide. The acceptance and use of the instrument by the international research and surveillance communities is extensive. Given the extent of use and the potential for conclusions drawn to affect decisions regarding

research and health services resource allocation, which may also have high opportunity cost, this thesis provides a valid line of inquiry in questioning the recommended scoring methods. If resources are likely to be misdirected as a result of conclusions drawn based on the scoring methods, it is appropriate that research communities and health authorities are made aware of the problem and at the very least have the opportunity to re-consider their data.

The SF-36 has also been promoted internationally for the purposes of cross country comparisons of health [24] and has been used to make inter-country comparisons of health [25]. This latter International Quality of Life Assessment (IQOLA) Project compared the impact of chronic conditions using the SF-36 in eight major countries. The conditions compared included allergies, arthritis, congestive heart failure, chronic lung disease, hypertension, diabetes and ischaemic heart disease. It was concluded from the study that health related quality of life measures were useful in characterising the global burden of disease. However, I argue in this thesis that this assertion only applies if the scoring and comparison methods used are valid ways to make these comparisons. It is also noteworthy that prior to the IQOLA Project few studies and no substantive international comparisons of the impact of chronic conditions had been made. This is also likely to have been a factor in the international acceptance of the instrument (i.e. filling a research gap).

Inconsistencies that Arise Using the Recommended Scoring Methods.

The issue of inconsistency between SF-36 sub-scales and summary scores is supported by a number of authors. Simon et al [26] first identified that although Ware et al had made sound theoretical argument for the use of orthogonal rotation methods in analysing SF-36 data [27], the approach encountered significant practical difficulties.

It was shown that in a primary care study the physical and bodily pain sub-scales make modest contributions to the mental summary score and that observed improvements in physical functioning and physical role could produce scores indicating worsened mental health. Simon et al, [26] went on to show that

positively scored physical sub-scales were completely offset by large changes in negatively scored mental health sub-scales. A mathematical anomaly was created as a consequence of negative scoring coefficients used in the computation of summary components.

Wilson, Parsons and Tucker [28] used the Australian National Health Survey [29] providing 18,492 respondents to analyse SF-36 data using the recommended orthogonal methods and an alternative structural equation model approach (SEM). In the orthogonal analysis, the sub-scales that made up the MCS were significantly lower for the 70+ age group, compared to younger age groups, yet the computed MCS summary score was significantly higher. Similar anomalies were observed for different population medication groups in addition to the age anomalies. This large representative population study underscored the problem first identified by Simon, but provided a larger population study sample, which allowed for greater confidence in the generalisability of results showing that a negative Z score when multiplied by a negative coefficient in the orthogonal analyses resulted in a positive inflation of the related component summary score [26]. In 2001 Taft used the Swedish SF-36 National Normative Sample providing data on 8930 respondents aged 15-93 and concluded that the recommended scoring methods produced illogical results.[30] Taft showed that in the upper range the PCS was primarily measuring aspects of mental health (57% of variance) and the MCS was primarily measuring aspects of physical health (65% of variance). He identified the scoring algorithm as responsible for the problem through simulation analyses.. Ware essentially dealt with Taft's criticisms by denying that they were valid, and from what seems like a misreading/misunderstanding of what the Taft paper was saying [31] Ware produced arguments rebutting points that Taft et al had not made, and failed to deal effectively with Taft's criticisms [32].

Norveldt et al [33] studied the performance of the physical and mental summary scales of SF-36, SF-12, and RAND-36. The scales were evaluated by comparing the scores of a cohort of 194 Multiple Sclerosis patients with general population data and using the Expanded Disability Status Scale (EDSS) and the Incapacity Status Scale-mental as criterion variables for physical functioning and mental health. They found that the SF-36 and SF-12 mental health summary scales appeared to overestimate mental health in people with Multiple Sclerosis [33].

The Correlation between physical and mental health.

As has been repeatedly pointed out in the present research, the major putative problem with the recommended scoring methods, in addition to the algebraic problem already identified, is that they do not allow for a correlation between physical and mental health in creating the summary scores; an issue that is not consistent with the health literature [16-22] in which there is strong support for this correlation. The Royal College of Psychiatrists are unanimous in their agreement that poor physical health can cause poor mental health and vice versa [22]. They go on to argue that "when we look at our health we should look at the whole subject of 'mind, body and spirit' in which all elements are connected and in which each feature affects the other" [22]. The following researchers deal with aspects of this problem.

The basic aims of Taft's research [30] were threefold. First, to examine the relationship between SF-36 sub-scale scores and PCS/MCS scores. Second, to examine the relationship between PCS and MCS scores and the magnitudes of their potential effects on each other. Third, to examine the implications of the above relationships on interpreting research findings. For the first aim Taft concluded that "the PCS and MCS scoring procedures may be likened to a seesaw, where below average physical health sub-scales weight up mental health while above average scores weight down mental health. Likewise below average mental health sub-scales increase PCS scores, while above average scores decrease PCS scores." In relation to the second objective he first points out that in an orthogonal analysis the two dimensions are by definition uncorrelated. He then showed correlations between PCS and MCS were highest and significant in scores above the range of expected values. For the third objective he showed from a regression analysis of PCS/MCS that 57% of the variance in the PCS was accounted for by negatively weighted mental health sub-scales, while 65% of the MCS was explained by physical health sub-scales. His overall conclusion reached from these analyses was physical and mental health are indeed dependent. Despite his findings Taft did not discuss an alternative scoring method for the summary scales.

Farivar et al [34] estimated SF-36 and SF-12 summary scores using a correlated (oblique) physical and mental health exploratory factor analysis model. They concluded that "Correlated physical and mental health summary scores for the SF-36 and SF-12 derived from an obliquely rotated factor solution should be used along with the uncorrelated summary scores. The new scoring algorithm can reduce inconsistent results between the SF-36 scale scores and physical and mental health summary scores reported in some prior studies." [34].

Hann and Reeves [35] performed a secondary analysis of two large-scale data sets that utilised the SF-36: Health Survey for England 1996 and the Welsh Health Survey 1998. They used confirmatory factor analysis to compare hypothetical orthogonal and oblique factor models, and exploratory factor analysis to derive data-driven models for condition-specific subgroups. They found that oblique models gave the best fit to the data and indicated a considerable correlation between PCS and MCS. They recommended that users of the SF-36 adopt the oblique model for calculating PCS and MCS. They also found that an oblique five-scale model provided a more universal factor structure without loss of predictive power or reliability [35].

Anagnostopoulos et al [36] compared the two higher order factor structures of the Short-Form 36 (SF-36) Health Survey, using exploratory factor analytic methods and structural equation modelling (SEM). They found that "Exploratory factor analysis supported the existence of two principal components that are the basis for summary physical and mental health measures. SEM showed that models assuming that physical and mental health are correlated provided a better fit to the data than models assuming independence between physical and mental health." [36].

Fleishman [37] conducted a study analysing nationally representative U.S. data, which provided 53,399 observations for the SF-12v2 in 2003–2005.The study derived new summary scores based directly on SF-12v2 data and compared the new summary scores to the standard ones. In addition to the standard SF-12V2 scoring algorithm, summary scores were generated using exploratory factor analysis (EFA) with orthogonal and oblique rotation, principal components analysis (PCA), and confirmatory factor analysis (CFA), with correlated and independent factors. Changes in summary scores derived using orthogonal

rotation of components or factors were not consistent with changes in sub-scales, whereas changes in summary scores derived using oblique rotation were more consistent with patterns of change in sub-scales. They recommended using summary scores based on a correlated CFA model [37].

Prior to my published research no PCS and MCS scoring algorithms for Australian data had been published based on a correlated model of physical and mental health. In this research I will provide Australian scoring coefficients for the SF-36 and SF-12 PCS and MCS scores for both version 1 and version 2, using a correlated model in a confirmatory factor analysis (a structural equation modelling approach). I will demonstrate these coefficients provide valid scores which do not conflict with the SF-36 sub-scales, and demonstrate the superior measurement properties of this approach through comparisons of scoring algorithms in multiple datasets. I will address the question of international comparisons using my approach compared to the proprietary scoring.

Statistical Challenges in Publishing the Research Arguments.

Given the complexity of decisions made in relation to CFA analysis, the following methodological explanations, which helped to guide statistical decision making in my research are provided. First, Rigdon & Ferguson [38] have shown that Maximum Likelihood (ML) estimation based on a polychoric correlation matrix is insufficient to correct for the problems associated with the type of data in this study. For this reason weighted least squares (WLS) estimation is preferred. Further, Mindrilla [39] concluded that Diagonally Weighted Least Squares (DWLS) is superior to ML for the analysis of ordinal data. Nye & Drasgow [40] consider that WLS and DWLS are both from the Asymptotically Distribution Free (ADF) family of estimators, and require similar large size samples. They investigated sample sizes from 400 to 1600. Flora & Curran [41] contradict this paper, concluding that DWLS (they call it robust WLS) is superior to WLS in almost all situations, especially when the model is complex or the sample is small (n=100). The largest sample size they considered was 1000. Forero et. al [42] compared unweighted least squares (ULS) and diagonally weighted least squares (DWLS) as alternatives to WLS for estimating Confirmatory Factor Analysis (CFA) models with ordinal indicators in a Monte Carlo study, and concluded that

ULS was preferable, but if this did not converge then DWLS should be used, even in small samples (they examined sample sizes of 200. 500, and 2000). WLS was eliminated from consideration due to the requirement for very large sample sizes.

For maximum likelihood estimation of multivariate normal data, fit measure cut-offs have been set out by Hu and Bentler [43] as: Root Mean Square Error of Approximation (RMSEA) =0.06; Standardised Root Mean Square Residual (SRMR) =0.08; Tucker Lewis Index (TLI) = 0.95; Comparative Fit Index (CFI) = 0.95. TLI is also known as the Non-Normed Fit Index (NNFI). Nye & Drasgow [40] concluded that the fit measures and cut-offs in use for ML estimation of multivariate normal data do not apply to ADF estimators. They based their proposals for interpretation of fit measures on DWLS estimators of dichotomous indicators in CFA via tetra choric correlations. They used Monte Carlo computer simulation to study the effects of model misspecification, sample size, and non-normality on fit indices generated from DWLS estimation on dichotomous data. The study consisted of a 3 (model misspecification) by 3 (degree of non-normality) by 3 (sample size) design. This is based on simulations of sample sizes of 400, 800, and 1600, using values of 0, 0.5, and 1.75 for skewness, and 0, 1.0, and 3.75 for kurtosis. The reader is indirectly invited to extend the results to ordinal data and polychoric correlations, but this is an assumption. They have set out how to calculate cut-offs for fit measures for different situations (i.e. different levels of skewness, kurtosis, sample size, and required type I error rates). They only considered positive skewness in their calculations. They found that CFI & TLI were almost always near 1, and did not provide any discrimination regarding the fit of these models. Therefore, they recommend judging fit for these models based on their calculated cut-offs for RMSEA and SRMR. Flora & Curran [41] found that ‘‘there were few to no differences found in any empirical results as a function of two category versus five category ordinal distributions.’’ This conclusion supports the generalisation of Nye & Drasgow’s work from tetra choric to polychoric correlations. They also found that DLWS produced more accurate estimates of the model chi-square, and therefore all of the fit measures that are based on it. In WLS estimation, the ‘‘inflation of the test statistic increases Type I error rates for the chi-square goodness-of-fit test, thereby causing researchers to reject correctly specified models more often than expected.’’. In this sense, Flora and Curran argue the opposite of Nye & Drasgow, [40] who proffer the advice that goodness-of-fit criteria need to be tightened up to avoid accepting

inadequate models. Given the conflicting advice regarding the necessary stringency of fit measures, I decided on the basis of the evidence provided to accept the Hu and Bentler recommendations for fit measure cut offs for models using maximum likelihood estimation.

To test the equality of the factor score coefficients across countries, I fitted a multiple-group model with all parameters in both groups constrained to be equal, and an unrestricted multiple group model with all parameters independently estimated. I needed to perform Chi-squared tests for the differences between these models. The most up to date approach to this problem is set out by Satorra and Bentler [44]. In LISREL, the Satorra–Bentler Chi-squared corrects for the non-normality of the data by applying a scaling factor to the normal theory weighted least squares Chi-squared (NTWLS). When using unweighted least squares estimation, there is no maximum likelihood Chi-squared produced for the models analyzed. The scaling factors are therefore applied to the NTWLS Chi-squared for the relevant models in computing the new scaled difference test set out by Satorra and Bentler [44]. The use of the NTWLS Chi squared for the calculation of the new scaled difference test is entirely consistent with advice provided by Bryant and Satorra [45] who point out that LISREL users should use the NTWLS estimates rather than maximum likelihood (ML) estimates in calculating scaling factors.

CHAPTER 3 - Methods

Chapter Content

- Statistical methods
- Other issues regarding the scoring of the physical and mental health summary scales
- Why Confirmatory Factor Analysis is better than Exploratory Factor Analysis

Statistical Methods

The developers of the SF-36 used the standardised sub-scale scores as input into an Exploratory Factor Analysis (EFA) to generate scoring coefficients for the physical and mental health summary scores [6]. I considered it preferable to use a full measurement model (i.e. base the model on the data items not the sub-scales) in a Confirmatory Factor Analysis (CFA) to allow the calculation of scoring coefficients based on all the available information. Unlike the developers, I also allowed for the real life correlation of physical and mental health in the calibration of the model.

My first peer reviewed publication was Tucker, G., Adams, R., & Wilson, D. (2010). New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires. Quality of Life Research, 19(7), 1069–1076. [46]

When I analysed SF-36 version 1 data [46] from the National Health Survey conducted by the Australian Bureau of Statistics [29], the sample size was 18,141. Conventional wisdom at that time was that with such a large sample size it is preferable to use Asymptotically Distribution Free (ADF) estimation, also known as Weighted Least Squares (WLS). I published my results in this version 1 paper in April 2010, but this study was originally developed in 2008. Forero et al published in 2009 [42], and despite journal reviewer scrutiny and the iterations of draft manuscripts prior to publishing, Forero's paper was not identified. I was

therefore unaware of that work when I published my first paper [46]. I was also unaware of the Hu & Bentler paper describing cut-offs for fit measures [43], despite this being published in 1999. Since that time I have come to understand that the solution accepted for version 1 of the SF-36 was inadequate, with a SRMR= .2455. Other fit indices produced in the version 1 model were CFI=.9945 and TLI=.9941, RMSEA=.032, with a probability of close fit = 1.000, which were all very good. At that time I therefore didn't know enough about estimation methods, fit measures and indices and primarily used RMSEA to adjudicate fit, and ignored SRMR. This solution produced scales that were demonstrated to be much superior to the proprietary scoring anyway, in the same manner as the other peer reviewed publications contained in Chapter 6. Thankfully, a PhD is a learning exercise, and peer review of my second paper on version 2 of the SF-36 [47] led me to appropriate references regarding estimation methods and fit measures [38-43].

Following the original publication I have fit the models on the same data using DWLS for SF-36 since ULS did not converge, and using ULS for SF-12, and calculated scoring coefficients for these models. The scoring coefficients and scores generated by this model for SF-36 were different to the results in my first publication [46]. The coefficients calculated using these models are reproduced below: They can be regarded as updated scoring coefficients for the SF-36 Version 1.

Table 1 – SF-36 Physical and Mental Component Summary scoring coefficients based on DWLS estimation for a correlated model

	PCS	MCS
A1	0.9679	0.0311
A3A	0.0019	-0.0004
A3B	0.0131	0.0000
A3C	0.0019	-0.0008
A3D	0.0056	0.0000
A3E	0.0037	0.0004
A3F	0.0000	-0.0004
A3G	0.0000	-0.0004
A3H	2.2417	0.0763
A3I	0.0168	-0.0019
A3J	0.0000	-0.0004
A4A	0.8374	0.0262
A4B	1.1358	0.0353
A4C	2.3256	0.0741
A4D	2.4413	0.0763
A5A	0.3021	0.3239
A5B	0.2126	0.2195
A5C	0.0690	0.0737
A6	0.5110	0.5393
A7	-0.0019	0.0019
A8	2.7881	0.0854
A9A	0.3338	0.3494
A9B	0.0466	0.0482
A9C	0.1100	0.1162
A9D	0.0709	0.0748
A9E	0.3021	0.3220
A9F	0.1361	0.1428
A9G	0.2089	0.2214
A9H	0.0615	0.0653
A9I	0.2145	0.2286
A10	0.5483	0.5837
A11A	0.3991	0.0129
A11B	0.5819	0.0186
A11C	0.3040	0.0095
A11D	1.4621	0.0456
Constant	-8.0548	-7.5507

Table 2 – SF-12 Physical and Mental Component Summary scoring coefficients based on ULS estimation for a correlated model

	PCS	MCS
A1	2.1429	0.4206
A3B	2.1703	0.4261
A3D	0.8712	0.1707
A4B	3.6902	0.7252
A4C	3.8031	0.7470
A5B	0.7291	2.8406
A5C	0.2157	0.8385
A8	3.9675	0.7798
A9D	0.1917	0.7470
A9E	0.6658	2.5947
A9F	0.3509	1.3657
A10	1.1947	4.6528
Constant	-8.6998	-6.6780

Table 3 sets out the differences in PCS and MCS scores generated by the two analyses.

Table 3 – Differences in SF-36 summary scores produced by my published solution [46] and a re-analysis using DWLS estimation to derive scoring coefficients.

	N	Minimum	Maximum	Mean	Std. Deviation
PCS difference	18141	-6.26	6.48	0.0008	1.4306
MCS difference	18141	-2.19	1.95	-0.0009	0.4552

The differences in PCS scores ranged from -6.26 to 6.48, and in MCS from -2.19 to 1.95. The correction to the estimation method had a greater effect on the PCS score than the MCS score. The mean differences for both scores were very close to zero by definition. 95% of the differences in PCS scores were contained in the range (-2.8031,2.8047), and 95% of differences in MCS scores were in the range (-.8930,.8912). The effect for these scores was therefore small using criteria set out by Cohen[48], apart from the few more extreme differences which were moderate to large.

Table 4 – Correlations between SF-36 summary scores calculated using my published solution [46] and coefficients based on DWLS estimation.

		Correlations			
		PCS published	PCS DWLS	MCS published	MCS DWLS
	Pearson				
PCS published	Correlation	1.00	0.99	0.872	0.86
	Sig. (2-tailed)		0.000	0.000	0.000
	Pearson				
PCS DWLS	Correlation	0.99	1.00	0.86	0.85
	Sig. (2-tailed)	0.000		0.000	0.000
	Pearson				
MCS published	Correlation	0.87	0.86	1.00	1.00
	Sig. (2-tailed)	0.000	0.000		0.000
	Pearson				
MCS DWLS	Correlation	0.86	0.85	1.00	1.00
	Sig. (2-tailed)	0.000	0.000	0.000	

The PCS scores for the published model and the DWLS model had a correlation of 0.990, and the MCS scores 0.999. However, high correlations do not guarantee equivalence, as I pointed out in my 4th paper re international comparisons [49].

My second peer reviewed publication was Tucker G. R., Adams, R. J., & Wilson, D. H. (2013). Observed agreement problems between Sub-scales and summary components of the SF-36 version 2-an alternative scoring method can correct the problem. Plos One 8(4):e61191. doi: 10.1371/journal. pone.0061191. http://www.plosone.org/article/info%3Adoi%2F10.
1371%2Fjournal.pone.0061191. [48]

My Second publication analysed version 2 of the SF-36 and SF-12 [48]. The estimation method I used was Unweighted Least Squares (ULS), or Diagonally Weighted Least Squares (DWLS) if ULS did not converge, as recommended by Forero et. Al. [42]. I used RMSEA and SRMR for assessment of fit. Indices CFI & TLI are restricted to be near unity when using ULS/DWLS [40] and so were not considered.

LISREL does not produce scoring coefficients for second level factors and AMOS does, so in both papers I used the formulae provided by AMOS to produce scoring

coefficients from the outputs of the model fit using LISREL based on polychoric correlations.

My third peer reviewed publication was Tucker G, Adams R, Wilson D. (2014) Results from Several Population Studies Show That Recommended Scoring Methods of the SF-36 and the SF-12 May Lead to Incorrect Conclusions and Subsequent Health Decisions. Quality of Life Research 23:2195-2203
DOI: 10.1007/s11136-014-0669-9 [50]

My third publication more thoroughly demonstrated the superiority of my scoring approach over that of the developers, in multiple datasets [50]. My approach consistently achieved improved, more consistent scores. This is true despite the fact that I now recognise that the CFA model on which the Version 1 scoring coefficients was based was inadequate.

My fourth peer reviewed publication was Graeme Tucker, Robert Adams, David Wilson (2016) The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36 Quality of Life Research 25(2), 267-274. DOI: 10.1007/s11136-015-1083-7 [49]

My fourth publication addressed the question of valid International comparisons of SF-36 component summary scores. It is a vexed question, but I believe my approach provides the opportunity for valid comparisons of **subgroups** within each country and between countries [49].

Other issues regarding the scoring of the SF-36 summary scales

The original scoring coefficients for SF-36 V1, produced by the developers, were derived on representative population data gathered in 1991 [6]. The data for norming of SF-36V2 was gathered in 1998, however the coefficients for the generation of summary scores in Version 2 were retained from Version 1 , i.e. they are based on 1991 data. I argue that there have been significant changes in population health over this period and one only needs to cite the obesity epidemic in USA and most other Western democracies, which would have had a significant effect on physical, and/or mental health, since 1991 [51]. It would seem therefore,

an update of the coefficients for the calculation of summary scores was required for version2. The population norms for SF-36 V2 were updated in 2009.

The developers assert that there is no effect on the z transformed sub-scale scores from the changes to the instrument in version 2, and therefore no need to adjust the scoring coefficients calculated for version 1. I argue that the changes in the instrument to both question wording and response options requires a re-calibration of the scoring coefficients for the component summary scores, because the changes in the instrument cause changes in the correlations between items. In my analyses I have recalibrated the SF-36 V2 because of the changes to the instrument, and I argue that this action adds to the superiority of my statistical methods over the recommended methods. It should also be noted that in my publications I have also demonstrated the superior measurement properties of my approach against Hawthorne's [10] updated orthogonal coefficients based on the same dataset that we used to derive the CFA coefficients. Using the original US (or even Australian) scoring coefficients for SF-36 V1, as the developers advocate for US data, would have produced even worse results for the orthogonal scoring approach.

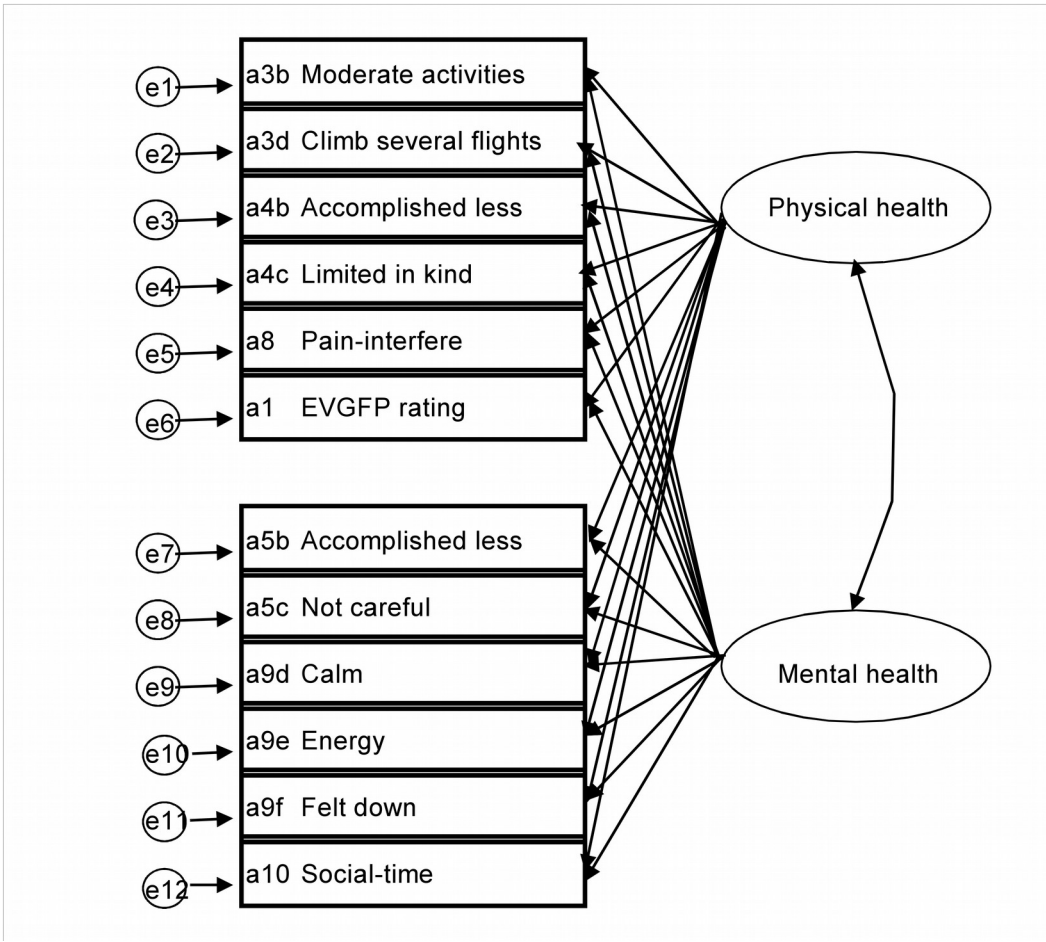
For V2 of the SF-12, the developers have retained the notion of producing sub-scale scores for the known sub-scales, however, their “sub-scale scores” are based on either one or two data items, and are very likely to be inferior to scales based on more items. The SF-12 can only be expected to produce approximate component summary scores, it should not be used to produce approximate sub-scale scores or profiles.

Why Confirmatory Factor Analysis (CFA) is better than Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis (EFA)

EFA is a saturated model. There is a path from every latent variable to every manifest variable. The diagram below demonstrates this for an EFA of the theoretical structure of the SF-12, including a correlation between physical and mental health. (see e.g. Farivar et. al. [34]).

Figure 1 – Theoretical structure of the SF-12 in an EFA

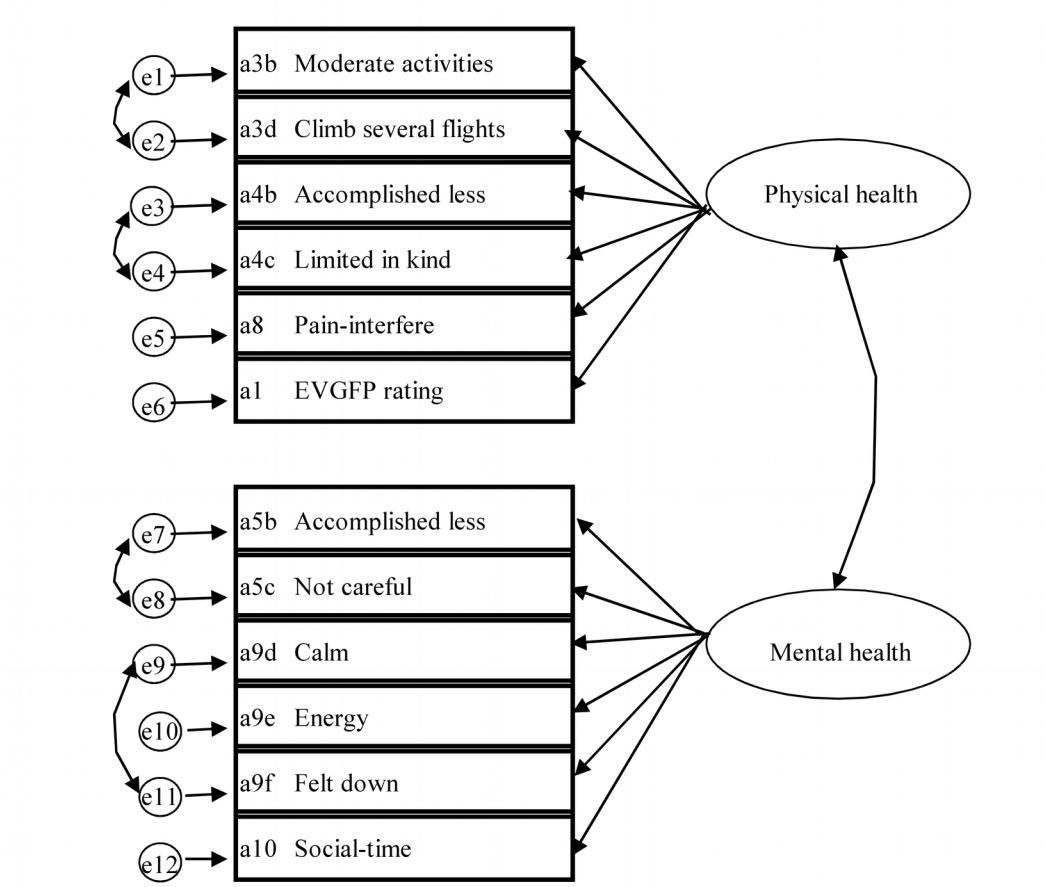


EFA is conducted on the Pearson correlation matrix. The basic assumption of EFA is that all of the correlations observed between the manifest variables are caused by the latent variable (i.e. unmeasurable factor). EFA is an exploratory technique, it seeks to generate possible factor structures to explain the data observed. EFA offers the choice of different factor extraction methods, and factor rotation methods. In particular, EFA offers orthogonal (independent factors) or oblique (correlated factors) rotation techniques.

Confirmatory Factor Analysis (CFA)

A CFA model is not saturated. Only important paths with substantive theoretical meaning are included in the model, so the model itself is much more parsimonious. Also, in Structural equation models such as these it is necessary to model the errors as well as the data [52]. In the model below, correlated errors are allowed for with SF-12 items from the same sub-scale, since these items would be expected to be more similar to each other than the other items of the scale.

Figure 2 – Theoretical structure of the SF-12 in a CFA



CFA is generally conducted on whatever correlations are most appropriate. For continuous variables these are Pearson correlations, for ordinal variables these are polychoric correlations, and for dichotomous variables there are tetra choric correlations. For a correlation between a continuous and a dichotomous variable a point bi-serial correlation would be calculated, and for a continuous by ordinal variable a poly serial correlation would be used. CFA is a confirmatory technique, it seeks to assess how well the hypothesised factor structure of the model fits the observed data.

Pearson correlations assume variables are normally distributed. Polychoric correlations require the assumption of bi-variate normality.

Both EFA and CFA can produce scoring coefficients for the calculation of indices/measures of the latent factors in the model. The CFA coefficients are based on the appropriate measures of correlation, from a parsimonious model, and will always produce a superior result to coefficients derived by EFA, even if an oblique model which allows for real world correlations between factors is used.

I have used confirmatory factor analyses of the hypothesised structure of the SF-36 and SF-12 including a correlation between physical and mental health on representative Australian population data. The analyses modelled polychoric correlation matrices of the data items recoded where necessary so that a higher score indicates better health for every item. This model was used to generate scoring coefficients for the SF-36 PCS and MCS, and the SF-12 PCS and MCS. The developers of the SF-36 produced unit weighted scores for each sub-scale, rescaled them to a score out of 100, then manipulated these scores to a z-score for each sub-scale using US norms for the data items. The developers used an EFA of the sub-scale z-scores, with an orthogonal rotation, to produce scoring coefficients for the summary scales, which were purported to represent the Physical and Mental health of the subjects.[6] There are a number of problems with this approach. Firstly, the sub-scale scores are not continuous despite their appearance/values. They are algebraic manipulations of unit weighted scales that can only assume a finite number of values. Pearson correlations are therefore inappropriate for their analysis. Secondly, more accurate results would be obtained for sub-scale scores if their sums were weighted (preferably using a congeneric CFA to produce the weights) rather than unit weighted. There is a problem with this too, because a congeneric CFA requires a minimum of four manifest variables, and there are sub-scales with 3 items (role emotional) and 2 items (bodily pain, social functioning) in the SF-36. The EFA model would work better with an oblique rotation rather than an orthogonal rotation. Finally, the EFA would produce a better solution if it was fit on the recoded data items as our CFA was. Fitting the EFA using sub-scale z-scores also sacrifices important information which is lost to the index/score created by the EFA model.

CHAPTER 4 - Other Major Developments in Measuring Quality of Life

Chapter Content

- Introduction
- Significant Quality of Life Instruments Used Extensively in Quality of Life Research
- Given the cost of the SF surveys, are they likely to continue being widely used?

Introduction

Over the last three decades many health related quality of life instruments have been produced to fill a major gap in health research, health evaluation and health planning needs. The domains included in the questionnaires produced, (as the conceptual basis of the instruments), have little agreement with each other across instruments, and this basically reflects the need for different content of health information by different organisations for different purposes. Primarily the instruments produced have covered two basic designs. The first comprise a range of generic questionnaires each covering several study domains. The SF-36 fits into this category of instrument. The second design extended the generic health status information to calculate preference or utility measures which could then be used for economic analysis usually based on a single utility score. The diversity of the generic health information in the many questionnaires developed tends to reflect the idiosyncratic nature of quality of life as it is generally considered to be specific to the individual, but is also understood by those who attempt to capture its nature for specific research or planning purposes. Of all the questionnaires developed it is argued that the SF-36 has been most widely used, however, for contextual purposes the following brief descriptions of other major quality of life instruments is provided here.

Significant Quality of Life Instruments Used Extensively in Quality of Life Research

The Centres for Disease Control (CDC) Health Related Quality of Life (HRQoL) Measures

The US Department of Health and Social Services Centers for Disease Control in partnership with the State and Territorial health agencies conducts US population health surveillance of health related quality of life. Health related quality of life is an important element of the state and national surveys and in support of state activities the CDC contains quality of life expertise and collaborates with academic institutes on survey developments. During the 1990's CDC worked on developing and validating a compact set of quality of life measures suitable for states and other communities. A major component of this has been the "Healthy Days Measure" (https//cdc.gov). These measures address the following domains :

- Would you say that in general your health is excellent, very good, good, fair or poor?
- Now thinking about your physical health, which includes physical illness and injury, how many days during the past 30 days was your physical health not good?
- Now thinking about your mental health, which includes stress, depression, and problems with emotions, how many days during the past 30 days was your mental health not good?
- During the past 30 days, approximately how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?

In addition, ten extra items are included and address: recent pain; depression; anxiety; sleeplessness; vitality; and the cause, duration and severity of current activity limitation the individual is experiencing.

Between 1993 and 2001 these questions were included in the state based Behavioural Risk Factor Surveillance System (BRFSS) [53]. Starting in 2000 the healthy days measures were part of the National Health and Nutrition Examination Survey, to provide a generic quality of life component.

The demonstrated value of these measures and the continuous accumulation of public domain data have resulted in support from the CDC Disability, Women's Health, and Arthritis Programs. The HRQoL measures and data also have been used for research or program planning by CDC's Cardiovascular Health and HIV/AIDS Programs as well as by the Public Health Foundation, the Foundation for Accountability, and several other government and academic programs.

In recent years, several organizations have found these Healthy Days measures useful at the national level for: (1) identifying health disparities, (2) tracking population trends, and (3) building broad coalitions around a measure of population health compatible with the World Health Organization's definition of health.. Extra modules are available to cover topics in more detail. (See http://www.cdc.gov/hrqol/methods.htm)

National Institutes of Health (NIH) PROMIS.

In 2004 the NIH developed the Patient Reported Outcomes Measurement Information System (PROMIS) to cover quality of life research needs in the clinical setting. PROMIS assesses global physical, mental, and social HRQoL. PROMIS comprises a 10-question measure and was developed through, a NIH Roadmap initiative providing an electronic system for the collection of self-reported HRQoL data from diverse populations with a variety of chronic diseases. The PROMIS includes domains on self-rated health, physical HRQoL, mental HRQoL, plus individual questions on fatigue, pain, emotional distress, social activities, and roles. Under the Roadmap Initiative questions have undergone qualitative and quantitative testing in several chronic disease populations and in the general U.S. population. A more recent psychometric evaluation of the PROMIS global health questions identified two global physical and mental health summary scales. The PROMIS has been successful in addressing the lack of standardisation in assessing patient outcomes and the way in which patient outcomes are reported. PROMIS is administered by a computer assisted interview system and access to this is provided to clinical researchers. As a result the system and quality of life data collected has engaged research stakeholders at many levels of patient research interest at both the federal and patient organisation levels. The PROMIS global health measure is scheduled to be administered on the National Health Interview Survey (NHIS) every 5 years (2010, 2015, and 2020). (http://www.healthypeople.gov)

The Health and Activities Limitation Index (HALex)

The HALex is a generic measure of quality of life that can be used to produce quality adjusted life years. It was initially developed for use in the National Health Interview Survey, conducted by the National Center for Health Statistics in the 1980s and 1990s. The questionnaire addresses two domains (perceived health and activity limitation). Through a multi-attribute utility scoring a single utility score can be calculated ranging on a 0(death)-1(best health) continuum. The instrument was developed to monitor changes and track change for US population health 2000. The original version of the questionnaire was specifically designed to be used for telephone interview surveys conducted for the Behavioural Risk Factor Surveillance Survey (BRFSS,) by the Center for Disease Control and Prevention (CDC), for calculating Healthy People 2000 Years of Healthy Life. This measurement instrument focuses on obtaining information on how health problems may inhibit or limit people in performing functions or activities of daily life [54].

The World Health Organisation WHOQOL-100 and the WHOQOL-BREF:

The WHOQOL-100 is a generic Quality of Life survey instrument, developed by World Health Organisation Quality of Life Group and validated for several countries, was launched in 1996. It was developed in fifteen international field centres to provide a generic quality of life instrument that would be applicable cross-culturally. Like many of the other quality of life instruments available, it is recommended for use in epidemiological studies and clinical trials, but given its length the WHQOL-BREF may be more useful. The WHOQOL-BREF comprises 26 items, which measure the following broad domains: physical health, psychological health, social relationships, and environment. The main development aim was to produce an international cross-culturally valid instrument of quality of life for comparative purposes. It assesses the individual's perceptions in the context of their culture and value systems, and their personal goals, standards and concerns. The WHOQOL instruments were developed collaboratively in a number of centres worldwide, and have been widely field-tested [55]. See http://www.who.int/mental_health/publications/whoqol/en/

The Quality of Life Scale (QOLS).

The Quality of Life Scale (QOLS), is a 15 item instrument addressing five domains (material and physical well-being, relationships, community activity,

personal development/fulfilment, recreation and independence). Developed by John Flanagan, an American psychologist, to address chronic illness the instrument is valid for measuring quality of life across patient groups. It is claimed that no other quality of life instrument has been developed with such detailed attention to diversity. Since its development the instrument has been used to collect data from diverse population groups on a range of chronic conditions. It has also been used to measure patient change. A review of the instrument states that the instrument has low to moderate correlation with health status and disease measures [56]. The instrument has been used for a range of chronic conditions and patient change.

The Quality of Well-Being Scale (QWB-SA)

The Quality of Well-Being Scale (QWB) was developed in the 1970's as the first instrument to measure quality of life for assessment of quality adjusted life years It produces an estimate of well-being between 0 (death) and 1 (full functioning) and was based on the General Health Policy Model developed by Kaplan and Anderson [57]. The QWB-SA combines preference-weighted values for symptoms and functioning. Symptoms are assessed by questions that ask about the presence or absence of different symptoms or conditions. Functioning is assessed by a series of questions designed to record the domain of functional limitations over the previous three days, within three separate domains (mobility, physical activity, and social activity). The four domain scores are combined into a total score that provides a numerical point-in-time expression of well-being that ranges from zero (0) for death to one (1.0). The widespread use of the instrument has been low because of its length and difficulty of administration. Nonetheless, it has been well validated, has sound psychometric qualities and has been used over the years to evaluate medical and surgical therapies and chronic conditions [58].

The Health Utilities Index (HUI)

The Health Utilities Index (HUI) comprises a family of generic health status and health related quality of life measures developed at McMaster University in Canada over the last 30 years. This comprises the HUI mark1, mark 2 and mark 3. They provide valid estimates of health status in clinical studies and population norm data from large population studies. The instrument was developed to fill a need for health status and health related quality of life comprising: the experience of the patient; the long term outcomes associated with disease or therapy; the

efficacy and efficiency of treatment or interventions and; the health status of the population. Together the HUI mark1 and mark 2 describe 1,000,000 unique health states and provide a generic health status classification system and a generic health utility scoring system. The instruments have been used extensively in clinical trials and cover many health problems. The HUI mark 2 measures seven domains (sensation, mobility, emotion, cognition, self-care, pain and fertility) . The HUI mark 3 measures six domains (sensation (vision, hearing speech), mobility, emotion, cognition, self-care, pain and fertility) HUI3 has been used in four major Canadian population health surveys, providing extensive data on population norms [59].

The Assessment of Quality of Life (AQoL)

The AQoL (of which there are several versions : the AQoL, the AQoL-4D and AQoL-8D) is a quality of life measure and multi-attribute utility instrument covering 5 dimensions (illness; independent living; social relationships; physical senses; and psychological well-being). This instrument was designed mainly for economic evaluation in terms of cost utility analysis. Thus through the AQoL and other cost utility instruments the impact of quality of life can be costed and used by health planners and administrators to improve population programs and health outcomes. There are several AQoL questionnaires of varying length and, for example, the AQoL 4 takes only 1-2 minutes to complete. There is no licence fee and no cost for downloading and using any of the AQoL instruments or algorithms. Each has a scoring algorithm which combines responses into dimension scores and a single utility score. The instruments can also be scored without utility weights [60].

The EuroQOL (EQ-5D)

The concept of health on which the EQ-5D is based comprises both positive dimensions (well-being) and negative dimensions (illness). The instrument comprises 5 dimensions (mobility, self-care, usual activities, pain discomfort, and anxiety depression) and measures three levels of health in each domain. Each health state can be transformed to a single utility score of quality adjusted life years for economic decision making and cost effectiveness studies. EQ-5D is a widely-used survey instrument for measuring economic preferences for health states. It is one of several such instruments that can be used to determine the quality-adjusted life years associated with a health state. The survey was

developed by the EuroQol Research foundation. [61-63]. It has been widely tested and used in both general population and patient samples and is available in 30 languages. The instrument also comes with a visual analogue scale by which responders can report perceived health status ranging from 0 (worst possible) to 100 (best possible) [64].

Given the cost of the SF surveys, are they likely to continue being widely used?

The SF surveys are now owned by OPTUM. According to their website [23] "The SF Health Surveys are the most widely used tools in the world for measuring patient-reported outcomes," Whilst researchers can use the Rand 36 item questionnaire which is identical to the SF-36 version 1, use of version 1 of the SF-36 is no longer licensed, and a licence to use version 2 is expensive enough to place the instrument out of the financial reach of some researchers. I obtained a quote of $9,125 USD to administer the SF-36V2 once only to 3000 subjects in a population survey in partnership with a government organisation. The quote includes both a component for scoring software, and a report on data quality. Optum has a more sophisticated treatment of missing data than the old mean substitution approach documented in the version 1 scoring manual. This cost may be difficult to meet for some projects, and the myriad of other instruments available to measure the same concepts may be appealing based on cost alone.

CHAPTER 5 - First studies

The early peer reviewed publications referred to in this chapter set the scene for my research. They produced models that were also based on the NHS dataset [29] to assess the validity of the recommended scoring methods on large scale population data. In conducting this early research a number of statistical issues arose. The recoding of data items set out in the SF-36 scoring manual appears to have been motivated by an intention to "linearise" the ordinal variables involved, so that the categories of the variables have values that represent a continuous metric along the real number line. This approach lends legitimacy to the AMOS approach of modelling the variance-covariance matrix or the Pearson correlation matrix. Modelling the variance covariance matrix is an unstandardised analysis which takes account of the scale of the variables, whereas modelling the Pearson correlation matrix provides a standardised analysis. These were the only alternatives provided by AMOS for ordinal data at that time. The modelling approach has been refined since the period of the initial research identified here, most importantly introducing the use of LISREL to analyse polychoric correlations.

These early publications were principally a collaboration between one of my supervisors for this thesis (David Wilson) and myself, also involving other colleagues as appropriate. Wilson and I were the initiators, chief planners, analysts, and writers of the publications. Our research began with two publications reproduced in the Appendix for which I was the chief statistician. I came to this work because of my connection in analysing the population data bases identified in these publications in other epidemiological studies. These publications were written to promote the alternative scoring approach based on a structural equation model as an improvement on the original orthogonal scoring of the developers of the SF-36 and SF-12. We only published the method as we considered it inappropriate to attempt to publish the actual scoring coefficients in an international journal at that time.

These early publications laid the foundation and informed the development of my thesis. The first of the publications was Wilson, D., Parsons, J., & Tucker, G. (2000). The SF-36 summary scales: Problems and solutions. Sozial- und Praventivmedizin, 45, 239–246. [28] This publication examined the consistency

of physical and mental health summary scales with the eight underlying sub-scales of the SF-36, using the approach of the developers (exploratory factor analysis with an orthogonal rotation of a principal components decomposition of the sub-scale scores), an exploratory factor analysis of the sub-scale scores with an oblique rotation which allows for the real world correlation between physical and mental health, and a confirmatory factor analysis (CFA) of the relevant 35 data items of the SF-36 used for scoring the instrument. The CFA was performed in AMOS, which analysed the variance covariance matrix.

Despite the considered shortcomings of my original analysis, in the first publication in the Appendix using structural equation modelling (SEM) as the alternative approach to the recommended orthogonal methods the coefficients produced scores that worked fairly well, certainly better than the proprietary scoring or the alternative of oblique rotation of an EFA solution based on the sub-scale scores. The major outcome was that the scores did not conflict with the underlying sub-scales of the SF-36 when compared by age groups or physical and mental health medication groups, whereas there was conflict between sub-scales and summary scores using the recommended scoring methods. This served to demonstrate the problem and promote the solution. The SEM approach was refined in later papers.

A copy of this publication is provided in the Appendix.

The second of these early peer reviewed publications was Wilson D, Tucker G, Chittleborough C. (2002) Rethinking and rescoring the SF-12. Sozial- und Praventivmedizin, 47, 172-177 [65]. The publication examined the make up of the SF-12, and used the regression methods of the developers to derive a set of items which best explained variation in the Australian National Health Survey dataset. Scores for this variable set were compared to scores based on the established US variable set, and found to be very similar. We concluded that although it is possible to derive a valid Australian version of the SF-12, the US version of the SF-12 should be used for reasons of international consistency, but using item weights derived from structural equation modelling. I have since examined the question of international comparisons more closely in my fourth peer reviewed publication [49], and refined my view on this topic.

A copy of this paper is provided in the Appendix.

CHAPTER 6 – Peer Reviewed Publications

Chapter Content

- Introduction
- Peer Reviewed Publications for the Main Body of Research

Introduction

This thesis by research is based on four publications described and reproduced below.

Peer Reviewed Publications for the Main Body of Research

Despite my statistical reservations previously discussed, the first of these publications provided scoring coefficients for version 1 of the SF-36 and SF-12 health status questionnaires, along with a limited demonstration of the superior performance of these scoring coefficients. The component summary scores were compared to the eight SF-36 sub-scale scores for the original recommended scoring method [6], as well as scores based on coefficients derived from a confirmatory factor analysis /structural equation model. The SF-12 scoring algorithms were similarly based on a structural equation model. The component summary scores calculated using the recommended scoring algorithm conflicted with the sub-scale scores for various age groups and physical and mental health medication groups. These conflicts were resolved using my published scoring coefficients based on the confirmatory factor analysis.

New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires [46]
Graeme Tucker, Robert Adams, David Wilson

Statement of Authorship

Title of Paper	New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires.
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Tucker G, Adams R, Wilson D (2010) New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires. Quality of Life Research, 19(7), 1069-1076.

Principal Author

Name of Principal Author (Candidate)	Graeme Tucker
Contribution to the Paper	Study conception and design, statistical analysis, interpretation of data, manuscript preparation, critical revision of the manuscript, corresponding author
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 23/3/17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Robert Adams
Contribution to the Paper	Supervised development of the work, data interpretation, critical revision of the manuscript.
Signature	Date 18/5/2017

Name of Co-Author	David Wilson
Contribution to the Paper	Supervised development of the work, data interpretation, manuscript preparation, critical revision of the manuscript.
Signature	Date 23/03/2017

Please cut and paste additional co-author panels here as required.

Qual Life Res
DOI 10.1007/s11136-010-9658-9

New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires

Graeme Tucker · Robert Adams · David Wilson

Accepted: 13 April 2010

Abstract
Purpose To compare the relationship of the eight SF-36 v1 subscale scores to the summary scores of the PCS and MCS derived from two different scoring algorithms: one based on the original scoring method (Ware, Kosinski and Keller, SF-36 physical and mental health summary scales: a users manual. The Health Institute, New England Medical Centre, Boston, MA, 1994); and the other based on scoring algorithms that use parameters derived from structural equation modelling. Further, to provide SF-12 scoring algorithms similarly based on structural equation modelling.
Methods The Australian Bureau of Statistics 1995 Australian National Health Survey dataset was used as the basis for the production of coefficients. There were 18,141 observations with no missing data for all eight SF-36 subscales following imputation of data items, and 17,479 observations with no missing data for the SF-12 data items. Data were analysed in LISREL V8.71. Structural equation models were fit to the data in confirmatory factor analyses producing weighted least squares estimates, which overcame anomalies found in the traditional orthogonal scoring methods.
Results Models with acceptable fits to the hypothesised factor structure were produced, generating factor score weighting coefficients for use with the SF-36 and SF-12 data items, to produce PCS and MCS summary scores consistent with their underlying subscale scores.
Conclusions The coefficients generated will score the SF-36 summary PCS and MCS in a manner consistent with their subscales. Previous Australian studies using version 1 of SF-36 or SF-12 can re-score their summary scores using these coefficients.

Keywords SF-36 summary scores · Structural Equation Model · PCS · MCS

G. Tucker (✉) · R. Adams · D. Wilson
Department of Health, Adelaide, SA, Australia
e-mail: Graeme.Tucker@health.sa.gov.au

Introduction

The SF-36 and the shorter form SF-12 health status questionnaires have been used in Australian population and other research studies for many years [3], including a health survey conducted by the national statistical agency in Australia [2]. It has provided powerful insights into the health status of groups and populations across a number of health dimensions [17] and has also been used as an outcome measure in studies [13]. For researchers and policy makers, it provides substantial information. The SF-36 was first validated for Australian use by McCallum [5, 6], and an Australia SF-36 was developed by Sanson-Fisher et al. [10].

The original SF-36 version 1 used a method of scoring based on factor coefficients derived through principle components analyses and orthogonal rotation. This method of scoring the SF-36 has been criticised by Simon et al [12], Wilson and Tucker [18, 19] and Taft et al [14], because it produced subscale and summary scores that were inconsistent with each other, although this view was not shared by the developers of the scales [15]. Because of this body of criticisms, in this study, we propose to re-score the SF-36 using structural equation modelling to overcome this problem. Recently, the SF-36 has been revised and become a commercial product [9], potentially putting it out of reach of many researchers who will find it difficult to replace with another health status instrument that has such

Published online: 04 May 2010

wide ranging dimensionality and summary measures. The revised version (SF-36 V2) has modified the question responses slightly and used the traditional method of scoring.

In this study, we provide further evidence for the failure of the orthogonal scoring methods and have produced scoring coefficients from Australian population data, which overcome the disagreement between summary and subscale scores. These scoring coefficients are based on structural equation modelling methods, and we suggest that these coefficients can be now be used free of charge by Australian researchers using the SF-36 version 1. We suggest that a similar approach can be adopted by researchers in other countries where local data are available to produce scoring coefficients and population norms.

Methods

The 1995 NHS dataset was used as the basis for the production of estimates [2]. This is the most recent Australian National population survey available to us, which included the version 1 SF-36 health status questionnaire. The sample design of the NHS is a self-weighting multistage clustered area sample based on Australian Bureau of Statistics (ABS) census collector districts in which households are selected with equal probability. In this survey $n = 23{,}800$ households were selected and all adults aged 15 or older were interviewed. A subset of $n = 19{,}785$ were asked to complete the SF-36 health status questionnaire. Of those interviewed, $n = 18{,}492$ provided some data on the SF-36. There were 18,141 observations with no missing data for all eight SF-36 subscales following imputation of data items by mean substitution, where more than half the data items in a subscale were not missing, as set out in the SF-36 scoring manual [16].

The items of the SF-36 are set out in Table 1.

A hypothetical factor structure has already been documented for the SF-36 [16]. This formed the basis of the model we evaluated, except that we allowed physical and mental health to be correlated (Fig. 1). It was therefore possible to fit a structural equation model (SEM) to the data in a confirmatory factor analysis. The model fit was the full measurement model, using items re-coded as detailed in the SF36 scoring manual [17], with the exception that integer values of the items were retained so that they could be modelled using polychoric and tetrachoric correlations in LISREL. The above model was fit on 18,141 observations with no missing data for all eight SF-36 subscales. A dataset of this size allowed the use of weighted least squares estimates of model parameters, which were preferred to maximum likelihood estimates because they are unbiased.

Data were analysed in LISREL V8.71. LISREL produces factor score weighting coefficients as an optional output, but does not provide scoring coefficients for second order factors. The Amos package does produce factor score weighting coefficients for second order factors, so the formula used by Amos was applied to the outputs from LISREL to generate the second order factor score coefficients. These are the coefficients used to weight the SF-36 data items to produce PCS and MCS summary scores. The existence of factor score weights for all of the 35 items in the calculation of the summary scores based on the model is explained by the fact that all variables have an effect on both physical and mental health by virtue of the correlation between them, which is allowed for in the model. The formula for factor score weights is given by $W = BS^{-1}$, where W is the matrix of regression weights, S is the matrix of covariances among the observed variables, and B is the matrix of covariances between the unobserved and observed variables [1]. These weights are applied to the recoded questionnaire items, which are all positive, so the problem of negative weights being applied to negative scores as noted by Simon et al [12], Wilson et al [18] and Taft [14] is averted. The fit of the model was assessed primarily using the root mean square error of approximation (RMSEA). An RMSEA value of 0.05 or less indicates a close fit of the model, a value of 0.08 indicates a reasonable error of approximation, and values >0.1 are not acceptable [1].

A similar approach was used to model the SF-12 variables (Fig. 2). A structural equation model was again fit to produce the factor score weights. The data were recoded as per the instructions of the SF-36 scoring manual [17], with the exception that question eight of the SF-36 was recoded according to the instructions where question seven is not answered. This is because question seven is not asked in collecting the SF-12 data items. Any records with missing data for the SF-12 items were excluded from the analysis. This resulted in 17,479 records being available to the analysis. In the model, correlations were allowed among the error terms for items from the same SF-36 subscale, because items from the same subscale could reasonably be expected to be more closely correlated with each other than with the other items of the SF-12. The weighted least squares estimation method was used, because it does not suffer from bias as does the Maximum Likelihood method, and there was enough data to support this method.

The ABS published population norms for the transformed subscale scores from the 1995 National Health Survey [2], and they used the traditional scoring approach of Ware et al [16] to produce factor score weights for the calculation of the Australian SF-36 summary scores. We used these published norms and weights to produce

Table 1 Detailed items of the SF-36

Subscale	Item	Short description	Question
Physical functioning	a3a	Vigorous activities	The following questions are about activities that you might do during a typical day. As I read each item, please tell me if your health now limits you a lot, limits you a little, or does not limit you at all, in these activities
	a3b	Moderate activities	
	a3c	Lift/carry groceries	
	a3d	Climb several flights	
	a3e	Climb one flight	
	a3f	Bend, kneel	
	a3g	Walk kilometre	
	a3h	Walk half a kilometre	
	a3i	Walk 100 m	
	a3j	Bathe, dress	
Role physical	a4a	Cut down time	The next four questions ask about your physical health and your daily activities. Have you …? (Yes/No)
	a4b	Accomplished less	
	a4c	Limited in kind	
	a4d	Had difficulty	
Bodily pain	a7	Pain-magnitude	How much Bodily Pain have you had during the past 4 weeks? (1 = None–6 = Very severe)
	a8	Pain-interfere	During the past 4 weeks, how much did pain-interfere with your normal work, including both work outside the home and housework? (1 = Not at all–5 = Extremely)
General health	a1	EVGFP rating	These first questions are about your health now and your current daily activities. Please try to answer every question as accurately as you can. In general, would you say your health is: (1 = Excellent–5 = Poor)
	a11a	Sick easier	Now I'm going to read you a list of statements. After each one, please tell me if its definitely true, mostly true, mostly false, or definitely false. If you don't know just tell me
	a11b	As healthy	
	a11c	Health to get worse	
	a11d	Health excellent	
Vitality	a9a	Full of life	The following questions are about how you feel and how things have been with you in the past 4 weeks. As I read each statement, please give me the one answer that comes closest to the way you have been feeling. Would you say all of the time, most of the time, a good bit of the time, some of the time, a little of the time or none of the time?
	a9e	Energy	
	a9g	Worn out	
	a9i	Tired	
Social functioning	a6	Social-extent	During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours or groups? Has it interfered: (1 = Not at all–5 = Extremely)
	a10	Social-time	During the past 4 weeks, how much of the time has your physical health and emotional problems interfered with your social activities like visiting friends and relatives? Would you say: (1 = All of the time–6 = None of the time)
Role emotional	a5a	Cut down time	The following three questions ask about your emotions and your daily activities. Have you …? (Yes/No)
	a5b	Accomplished less	
	a5c	Not careful	
Mental health	a9b	Nervous	The following questions are about how you feel and how things have been with you in the past 4 weeks. As I read each statement, please give me the one answer that comes closest to the way you have been feeling…. Would you say all of the time, most of the time, a good bit of the time, some of the time, a little of the time or none of the time?
	a9c	Down in dumps	
	a9d	Calm	
	a9f	Felt down	
	a9h	Happy	

summary scales distributed $N(50,10)$ based on the traditional scoring methods.

Having produced the SEM scoring coefficients, we compared their performance in scoring the SF-36 and SF-12 PCS and MCS with SF-36 summary scores produced by the traditional orthogonal scoring methods, across several age groups and in people according to medication use in an independent dataset (1998 SA Health Omnibus Survey). It was hypothesised that the PCS and MCS scores based on SEM will be in greater agreement with the eight

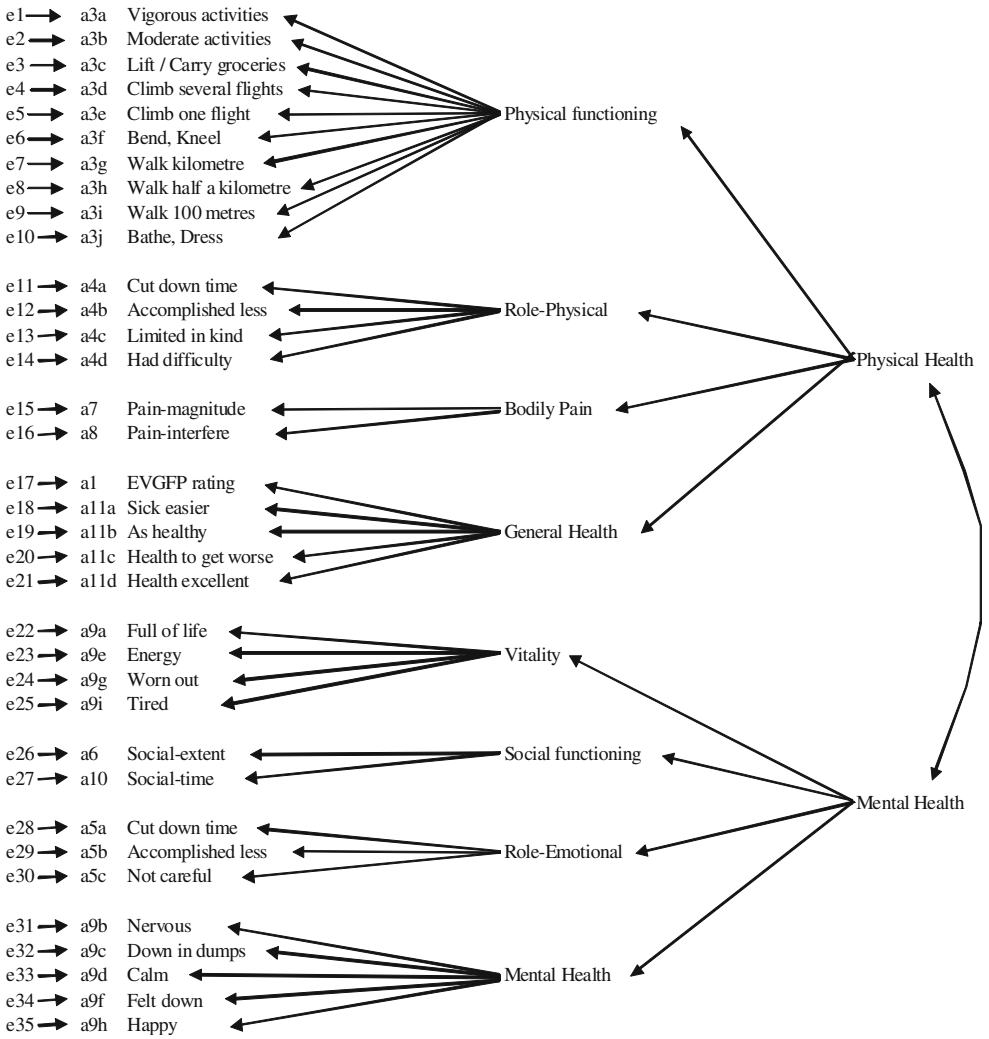


Fig. 1 Hypothesised structure of SF-36 health dimensions and the summary mental (MCS) and physical (PCS) health measures

subscale scores of the SF-36 than will the PCS and MCS scores based on the original scoring system.

Results

The coefficients generated by the SEM analysis for the SF-36 are set out in Table 2. The model had a Minimum Fit Function Chi-square of 10810.2 on 553 degrees of freedom, the size of which is explained by the large sample size. It had an RMSEA of .032 (90% confidence interval .031 to .033), and a probability of close fit of 1.000. The Non-Normed Fit Index was 0.9941, and the Comparative Fit Index was 0.9945.

Using SEM, the file from the National Health Survey produced SF-36 summary scores for the PCS with a mean of 2.7654 and a standard deviation of 0.46136, and for the MCS a mean score of 3.5943 with a standard deviation of 0.61019. To normalise these scores to have a mean of 50 and standard deviation of 10, the calculations that should be applied are:

$$\mathrm{SF36PCS} = 50 + (\mathrm{SF36PCS} - 2.7654) \times 10/0.46136$$
$$\mathrm{SF36MCS} = 50 + (\mathrm{SF36MCS} - 3.5943) \times 10/0.61019$$

The coefficients generated by the SEM analysis for the SF-12 are set out in Table 3. The model had a Minimum Fit Function Chi-square of 2919.8 on 49 degrees of freedom, the size of which is explained by the large sample size. It had an RMSEA of 0.058 (90% confidence interval is 0.056 to 0.060), and a probability of close fit of 0.000. The Non-Normed Fit Index was 0.9679, and the Comparative Fit Index was 0.9762.

Using SEM, the file from the National Health Survey produced SF-12 scores for PCS with a mean of 3.4296 and a standard deviation of 0.59965, and for MCS a mean score of 3.9489 with a standard deviation of 0.69260. To normalise these scores to have a mean of 50 and

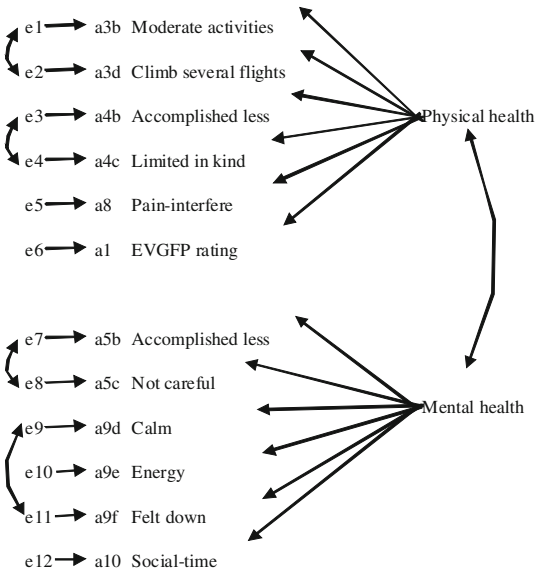


Fig. 2 Hypothesised structure of SF-12 summary mental (MCS) and physical (PCS) health measures

standard deviation of 10, the calculations that should be applied are

$$\text{SF12PCS} = 50 + (\text{SF12PCS} - 3.4296) \times 10/0.59965$$
$$\text{SF12MCS} = 50 + (\text{SF12MCS} - 3.9489) \times 10/0.69260$$

In comparing the effect of orthogonal rotation methods with structural equation modelling, we compared the summary scale scores with their underlying subscale scores for different age groups in Table 4 and for medication groups in Table 5. From the tables, clear discrepancies are apparent between the traditional summary scores and their subscales, which are not evident using scoring coefficients derived from structural equation models.

Table 4 shows the scores for the mental health scales [vitality ($p < 0.001$), social functioning ($p = 0.834$), role emotional ($p = 0.568$) and mental health ($p = 0.818$)] are all higher than average for those aged <30, as are the mental health summary scores (MCS) from SEM coefficients for both SF-36 ($p < 0.001$) and SF-12 ($p = 0.382$), yet the MCS score based on the original orthogonal scoring algorithm is lower than average ($p = 0.406$). Conversely, for those aged 70 or more, three of the four subscale scores are lower than average [VT ($p < 0.001$), SF ($p = 0.002$), RE ($p = 0.374$)], as are the MCS scores from SEM coefficients for both SF-36 ($p < 0.001$) and SF-12 ($p < 0.001$), yet the MCS score based on the original orthogonal scoring method is considerably higher than average ($p < 0.001$). In the interests of balance, it should be pointed out that although three of the four mental health subscale scores are higher than average for those aged 50–69 [SF ($p = 0.719$), RE ($p = 0.192$), MH ($p = 0.248$)], the MCS scores from SEM coefficients for both SF-36 ($p = 0.331$) and SF-12 ($p = 0.064$) are lower than average, and the MCS score based on the original scoring method is higher than average ($p = 0.001$). However, in this case, the difference between the summed subscales scores for mental health for this age group compared to overall is much less than in the other situations, as reflected by the significance probabilities in these comparisons. There were no inconsistencies evident by age for physical health summary scores when compared to their subscales.

Table 2 Australian weighting coefficients for the SF-36

		PCS	MCS
a1	EVGFP rating	0.0282	0.0064
a3a	Vigorous activities	0.0003	−0.0002
a3b	Moderate activities	0.0059	0.0026
a3c	Lift/carry groceries	0.0006	−0.0004
a3d	Climb several flights	−0.0001	0.0001
a3e	Climb one flight	0.0170	−0.0107
a3f	Bend, kneel	0.0004	−0.0003
a3g	Walk kilometre	0.0008	0.0004
a3h	Walk half a kilometre	0.0746	0.0308
a3i	Walk 100 m	0.0009	−0.0005
a3j	Bathe, dress	0.0003	0.0002
a4a	Cut down time	0.0539	0.0105
a4b	Accomplished less	0.0731	0.0164
a4c	Limited in kind	0.2223	0.0497
a4d	Had difficulty	0.1572	0.0343
a5a	Cut down time	0.0250	0.1070
a5b	Accomplished less	0.0131	0.0673
a5c	Not careful	0.0030	0.0136
a6	Social-extent	0.0294	0.1340
a7	Pain-magnitude	0.0004	−0.0002
a8	Pain-interfere	0.0847	0.0188
a9a	Full of life	0.0157	0.0718
a9b	Nervous	0.0026	0.0117
a9c	Down in dumps	0.0076	0.0347
a9d	Calm	0.0041	0.0185
a9e	Energy	0.0178	0.0828
a9f	Felt down	0.0074	0.0340
a9g	Worn out	0.0113	0.0518
a9h	Happy	0.0036	0.0174
a9i	Tired	0.0099	0.0450
a10	Social-time	0.0281	0.1311
a11a	Sick easier	0.0114	0.0026
a11b	As healthy	0.0178	0.0038
a11c	Health to get worse	0.0070	0.0015
a11d	Health excellent	0.0646	0.0137

PCS physical component summary, *MCS* mental component summary

Table 5 dealing with medication use, shows the mental health subscale scores are all lower than average for those taking medications for physical ailments, as are the mental health summary scores (MCS) from SEM coefficients for both SF-36 and SF-12, yet the MCS score based on the original orthogonal scoring method is higher than average. Similarly, the physical health subscale scores are all lower than average for those taking medications for mental health reasons, as are the physical health summary scores (PCS) from SEM coefficients for both SF-36 and SF-12, yet the PCS score based on the original scoring coefficients is higher than average ($p < 0.001$ for all comparisons by medication use commented upon).

Table 3 Australian weighting coefficients for the SF-12

		PCS	MCS
a1	EVGFP rating	0.1031	0.0365
a3b	Moderate activities	0.1166	0.0413
a3d	Climb several flights	0.0383	0.0136
a4b	Accomplished less	0.1516	0.0537
a4c	Limited in kind	0.2833	0.1004
a5b	Accomplished less	0.1832	0.0649
a5c	Not careful	0.0569	0.1905
a8	Pain-interfere	0.0129	0.0432
a9d	Calm	0.0193	0.0645
a9e	Energy	0.0569	0.1905
a9f	Felt down	0.0281	0.0940
a10	Social-time	0.0763	0.2555

PCS physical component summary, *MCS* mental component summary

Discussion

The data presented in this study show that there are inconsistencies associated with the original orthogonal scoring methods. Conversely, structural equation modelling to obtain factor coefficients for each of the SF-36 and SF-12 produces PCS and MCS summaries for both age groups and medication groups that are more consistently aligned with the underlying subscales of the SF-36.

Under the original scoring method for summary scores subtle to moderate declines in physical health (scores) lead to moderate improvements in mental health scores, and vice versa. In practice, this means that small but clinically important changes in summary scores will at best be missed, and at worst misrepresented, using the original scoring algorithms. The improvement in consistency offered by our scoring coefficients overcomes this problem.

Table 4 Comparison of subscale scores and summary scores using various scoring methods, by age groups

	Age				Total
	<30	30–49	50–69	70+	
Unweighted (*n*)	622	1,110	788	488	3,008
Physical functioning	92.8	89.5	80.6	64.7	85.3
Role physical	87.6	82.7	74.6	63.8	79.8
Bodily pain	81.0	77.7	72.2	71.0	76.5
General health	76.5	76.8	71.2	64.7	73.9
Vitality	67.3	64.5	63.7	58.9	64.3
Social functioning	88.6	88.5	88.2	84.0	87.9
Role emotional	88.3	87.1	89.0	86.5	87.8
Mental health	80.1	79.0	80.7	81.6	80.0
SF-36 PCS—USA weights	52.5	51.1	46.9	41.8	49.4
SF-36 MCS—USA weights	52.3	52.2	54.2	55.2	53.0
SF-36 PCS—orthogonal extraction and rotation: using ABS weights based on the NHS	52.8	51.4	47.0	41.8	49.6
SF-36 MCS—orthogonal extraction and rotation: using ABS weights based on the NHS	51.2	51.1	53.0	53.8	51.9
SF-36 PCS—using factor score weights from SEM	52.6	51.6	49.2	45.8	50.6
SF-36 MCS—using factor score weights from SEM	52.2	51.2	50.7	48.4	51.0
SF-12 PCS—using factor score weights from SEM	52.9	51.6	48.8	45.1	50.5
SF-12 MCS—using factor score weights from SEM	52.6	51.3	50.3	47.4	51.0

ABS Australian bureau of statistics, *NHS* National Health Survey, *PCS* physical component summary, *MCS* mental component summary, *SEM* structural equation models, *USA* United States of America

Table 5 Comparison of subscale scores and summary scores using various scoring, by medication status

	Medication taken				Total
	No medication	Physical only	Mental only	Both physical and mental	
Unweighted (*n*)	2,033	780	88	107	3,008
Physical functioning	90.7	71.5	81.8	60.0	85.3
Role physical	86.9	63.3	69.0	39.1	79.8
Bodily pain	81.3	65.2	71.6	49.6	76.5
General health	79.2	61.8	66.1	41.7	73.9
Vitality	67.7	58.3	49.2	41.0	64.3
Social functioning	90.9	83.0	77.1	58.8	87.9
Role emotional	90.6	85.5	63.0	56.5	87.8
Mental health	81.9	78.6	61.3	59.2	80.0
SF-36 PCS—USA weights	52.1	42.3	49.5	37.0	49.4
SF-36 MCS—USA weights	53.6	53.7	42.7	42.6	53.0
SF-36 PCS—orthogonal extraction and rotation: using ABS weights based on the NHS	52.2	42.6	50.6	38.2	49.6
SF-36 MCS—orthogonal extraction and rotation: using ABS weights based on the NHS	52.6	52.2	41.2	40.5	51.9
SF-36 PCS—using factor score weights from SEM	53.2	44.9	45.6	34.7	50.6
SF-36 MCS—using factor score weights from SEM	52.9	47.8	43.0	35.9	51.0
SF-12 PCS—using factor score weights from SEM	53.1	44.7	45.5	34.8	50.5
SF-12 MCS—using factor score weights from SEM	53.0	47.1	43.4	36.0	51.0

ABS Australian bureau of statistics, *NHS* National Health Survey, *PCS* physical component summary, *MCS* mental component summary, *SEM* structural equation models, *USA* United States of America

Five previous investigators have pointed to the scoring anomalies identified in this study [4, 8, 12, 14, 18], however, Ware and Kosinski's [15] re-analysis of datasets failed to corroborate the anomaly. Simon [12], Wilson [18] and Taft [14] have provided reasons why the problem arises. When negative standardised subscale scores are multiplied by negative coefficients the result is positive, producing a summary score that is at variance with the subscale.

We contend that orthogonal rotation methods to obtain factor coefficients are misleading given the inconsistencies observed between the subscale scores and the summary scores. A major question now exists as to whether or not the SF-36 version 2 perpetuates this conundrum, given the orthogonal method of scoring. In addition to the scoring problems raised here, Hawthorne [3], in his Australian study argues that the use of the version 1 weights to score the version 2 SF-36 summary scales is of concern, given that significant differences between the data sets for scoring version 1 and version 2 were observed and that methodological differences including sampling bias were also apparent. Hawthorne, however, did not challenge the additional problem of orthogonal scoring methods.

It should be emphasised that we are not criticising the ability of the SF-36 in any version to produce health status estimates for populations or study groups. The large number of studies conducted around the world since the origin of the SF-36 attests to its importance and construct validity as a health status measure. The main contention is with the method of deriving scoring coefficients for the summary scores. Physical and mental health are correlated, and the use of orthogonal models precludes allowances for that fact in the derivation of the scoring coefficients.

Derivation of the factor coefficients in this study was based on data from the 1995 National Health Survey, the largest Australian population health data set available to us. The coefficients produced will score the version 1 SF-36 summary PCS and MCS in a consistent and reliable way when summary scores consistent with the underlying subscales are desired. They may be used with confidence by Australian researchers whose budget lines do not extend to the commercial version of the SF-36 version 2 or for those studies tracking health issues over time. Researchers from other countries may choose to produce country specific coefficients for use on version 1 using the same approach. The different SEM-based scoring algorithms derived in each country would all provide summary scores from the same factor structure, all with population means of 50 and standard deviations of 10, i.e. all summary scores would have the same scale thus allowing international comparisons.

It should be pointed out that although there may be some improvements in version 2 these are not without their critics [11]. In a major representative German population study Morfeld [7] concluded that although there were gains in psychometric quality and discrimination of version 2 over version 1 this did not justify a preference for version 2. Furthermore, version 1 has already proved popular with Australians having been used in 130 Australian studies [3], which included several validation studies. For each of those studies that used summary scores; however, serious questions are raised as to their consistency as is demonstrated by these analyses.

References

1. Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide*. Chicago: Small Waters Corporation.
2. Australian Bureau of Statistics. (1995) *National Health Survey. SF-36 Population Norms Australia*. Canberra: Australian Bureau of Statistics, Catalogue Number 4399.0.
3. Hawthorne, G., Osborne, R. H., Taylor, A., & Sansoni, J. (2007). The SF-36 Version 2: Critical analysis of population weighting, scoring algorithms and population norms. *Quality of Life Res, 16*, 661–673.
4. Hurst, N. P., Ruta, D. A., & Kind, P. (1998). Comparison of the MOS short form-12 (SF-12) health status questionnaire with the SF -36 in patients with rheumatoid arthritis. *British Journal of Rheumatology, 37*, 862–869.
5. McCallum, J. (1995). The new SF-36 health status measure: Australian validity tests. In *Paper presented to the Health Outcomes and Quality of Life Measurement Conference*. Canberra: National Centre for Epidemiology and Population Health.
6. McCallum, J. (1995). The SF-36 in an Australian sample: Validating a new, generic health status measure. *Australian Journal of Public Health, 19*, 160–166.
7. Morfeld, M., Bullinger, M., Natke, J., & Brahler, E. (2005). Die version 2.0 des SF-36 Health Survey-Ergebnisse einer bevolkerungsreprerasentativen studie. *Sozial- und Praventivmedizin, 50*, 292–300.
8. Nordvedt, M. W., Riise, T., Myhr, K. M., & Nyland, H. I. (2000). Performance of the SF-36, SF-12 and RAND SF36 summary scales in a multiple sclerosis population. *Medical Care, 38*, 1022–1028.
9. Quality Metric Incorporated. (2008). *SF-36 v2TM and SF-12 v2 TM Health Surveys Offer Substantial Improvements*. www.SF-36.org/commnity/SF36V2andSF12V2.shtml. Accessed June 20, 2008.
10. Sanson-Fisher, R. W., & Perkins, J. J. (1998). Adaptation and validation of the SF-36 Health Survey for Use in Australia. *Journal of Clinical Epidemiology, 51*(11), 961–967.
11. Sansoni, J., & Costi, J. (2001). SF-36 Version 1 or Version 2: The need for Australian normative data. In *Proceedings of Health Outcomes 2001: The Odyssey Advances Conference*. Canberra: Australian Health Outcomes Collaboration.
12. Simon, G. E., Revicki, D. A., Grothaus, L., & Vonkorf, M. (1998). SF-36 summary scores. Are physical and mental health truly distinct? *Medical Care, 36*, 567–572.
13. Sorensen, L., Stokes, J. A., Purdie, D. M., et al. (2004). Medication reviews in the community: Results of a randomized, controlled effectiveness trial. *British Journal of Clinical Pharmacology, 58*, 648–664.
14. Taft, C., Karlson, J., & Sullivan, M. (2001). Do SF-36 summary component scores accurately summarise subscale scores? *Quality of Life Research, 10*, 395–404.
15. Ware, J. E., & Kosinski, M. (2001). Interpreting SF-36 summary health measures: A response. *Quality of Life Research, 10*, 405–413.
16. Ware, J. E., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health summary scales: A users manual*. Boston, MA: The Health Institute, New England Medical Centre.
17. Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *The SF-36 health survey manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Centre.
18. Wilson, D., Parsons, J., & Tucker, G. (2000). The SF-36 summary scales: Problems and solutions. *Sozial- und Praventivmedizin, 45*, 239–246.
19. Wilson, D., Tucker, G., & Chittleborough, C. (2002). Rethinking and rescoring the SF = 12. *Sozial- und Praventivmedizin, 47*, 172–177.

Qual Life Res
DOI 10.1007/s11136-010-9686-5

ERRATUM

Erratum to: New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires

Graeme Tucker · Robert Adams · David Wilson

Erratum to: Qual Life Res
DOI 10.1007/s11136-010-9658-9

There were errors in the SF-12 scoring parameters specified in the last paragraph of page 4 in the original publication, and consequently in the two equations following the remainder of the paragraph, on page 5.

The correct values and equations are shown here.

p. 4, last para, line 2: '3.4296' should be '3.2759'
p. 4, last para, line 3: '0.59965' should be '0.54945'
p. 4, last para, line 4: '3.9489' should be '3.8956'
p. 4, last para, line 4: '0.69260' should be '0.67927'

The equations on page 5 should read:

$$\text{SF12PCS} = 50 + (\text{SF12PCS} - 3.2759) \times 10/0.54945$$
$$\text{SF12MCS} = 50 + (\text{SF12MCS} - 3.8956) \times 10/0.67927$$

The online version of the original article can be found under doi:10.1007/s11136-010-9658-9.

G. Tucker (✉) · R. Adams · D. Wilson
Department of Health, Adelaide, SA, Australia
e-mail: Graeme.Tucker@health.sa.gov.au

Published online: 08 June 2010

My second publication addressed the issue of inconsistencies between sub-scale scores and component summary scores using recommended scoring methods of the SF-36 version 2. It established that the previous problems of disagreement between the eight SF-36 Version 1 sub-scale scores and the Physical and Mental Component Summary scores persist in version 2, and went on to provide scoring coefficients for SF-36 and SF-12 version2 to address the problem. The component summary scores calculated using the recommended scoring algorithm conflicted with the sub-scale scores for various age groups and physical and mental health medication groups. These conflicts were again resolved using my published scoring coefficients based on the confirmatory factor analysis.

Observed Agreement Problems between Sub-scales and Summary Components of the SF-36 Version 2 – An Alternative Scoring Method Can Correct the Problem [47]

Graeme Tucker, Robert Adams, David Wilson

Statement of Authorship

Title of Paper	Observed agreement problems between Sub-scales and summary components of the SF-36 version 2-an alternative scoring method can correct the problem.
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Tucker G R, Adams R J, Wilson D H (2013) Observed agreement problems between Sub-scales and summary components of the SF-36 version 2-an alternative scoring method can correct the problem. Plos One 8(4):e61191. doi: 10.1371/journal. pone.0061191. http://www. plosone.org/article/info%3Adoi%2F10. 1371%2Fjournal.pone.0061191.

Principal Author

Name of Principal Author (Candidate)	Graeme Tucker
Contribution to the Paper	Study conception and design, statistical analysis, interpretation of data, manuscript preparation, critical revision of the manuscript, corresponding author.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 23/3/17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Robert Adams
Contribution to the Paper	Supervised development of the work, data interpretation, critical revision of the manuscript.
Signature	Date 18/5/2017

Name of Co-Author	David Wilson
Contribution to the Paper	Supervised development of the work, data interpretation, manuscript preparation, critical revision of the manuscript.
Signature	Date 23/03/2017

Please cut and paste additional co-author panels here as required.

OPEN ACCESS Freely available online

PLOS ONE

Observed Agreement Problems between Sub-Scales and Summary Components of the SF-36 Version 2 - An Alternative Scoring Method Can Correct the Problem

Graeme Tucker[1,2]*, Robert Adams[3], David Wilson[3]

1 SA Department of Health, Adelaide, South Australia, Australia, **2** Discipline of Medicine, University of Adelaide, Adelaide, South Australia, Australia, **3** The Queen Elizabeth Hospital, Woodville, South Australia, Australia

Abstract

Purpose: A number of previous studies have shown inconsistencies between sub-scale scores and component summary scores using traditional scoring methods of the SF-36 version 1. This study addresses the issue in Version 2 and asks if the previous problems of disagreement between the eight SF-36 Version 1 sub-scale scores and the Physical and Mental Component Summary persist in version 2. A second study objective is to review the recommended scoring methods for the creation of factor scoring weights and the effect on producing summary scale scores

Methods: The 2004 South Australian Health Omnibus Survey dataset was used for the production of coefficients. There were 3,014 observations with full data for the SF-36. Data were analysed in LISREL V8.71. Confirmatory factor analysis models were fit to the data producing diagonally weighted least squares estimates. Scoring coefficients were validated on an independent dataset, the 2008 South Australian Health Omnibus Survey.

Results: Problems of agreement were observed with the recommended orthogonal scoring methods which were corrected using confirmatory factor analysis.

Conclusions: Confirmatory factor analysis is the preferred method to analyse SF-36 data, allowing for the correlation between physical and mental health.

Citation: Tucker G, Adams R, Wilson D (2013) Observed Agreement Problems between Sub-Scales and Summary Components of the SF-36 Version 2 - An Alternative Scoring Method Can Correct the Problem. PLoS ONE 8(4): e61191. doi:10.1371/journal.pone.0061191

Editor: Jeremy Miles, Research and Development Corporation, United States of America

Received August 14, 2011; **Accepted** March 9, 2013; **Published** April 12, 2013

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Graeme.Tucker@health.sa.gov.au

Introduction

The SF-36 and the shorter form SF-12 health status questionnaires have been used extensively in international studies to obtain summary measures of health status. The origin of the instruments has an extensive and well-founded methodological history deriving from the Medical Outcomes Study conducted by the RAND Corporation [1]. However, international concern has been raised questioning the validity of the recommended orthogonal scoring methods of Version 1 of the SF-36 to produce Physical and Mental Component Summary scores (PCS & MCS) [2-9]. However, these scoring methods remain in widespread use, indeed they are the default scoring approach around the world. Given the instruments subscales and summary scores are used by national agencies to guide policy [10] and medical authorities to guide treatment and intervention decisions, [11], it is important that questions of validity are addressed to achieve best investment decisions. The creation of Version 2 of the instrument led to a number of refinements to question item response categories, layout and norming of the questionnaire. Data items for the role physical and role emotional items, which contribute substantially to PCS and MCS summary scores were expanded from dichotomous yes/no responses to five point Likert scales. New norms were derived from the 1998 US population, which have since been updated to 2009. [12]. No substantial changes were made to the recommended scoring methods [12], so the question remains as to whether or not the commercial Version 2 still produces summary scores that are at variance with the underlying sub-scale scores [5]. The major putative problem with the recommended scoring methods is they do not allow for a correlation between physical and mental health in creating the summary scores; an issue that is not consistent with the health literature. Epidemiological and clinical studies have shown a strong connection between physical and mental health [13–18]. People with depression often have worse physical health, as well as worse perception of their health [16], a characteristic that would affect their reporting of self-related health. Tucker et al [5], acknowledged this connection in the SF-36 version 1 by demonstrating that the use of the recommended orthogonal scoring methods, which do not allow for the correlation, created important discrepancies between the PCS and MCS and their underlying sub-scale scores, and that this could be corrected by use of confirmatory factor analysis (CFA). Given the extensive use of Version 2 [12] it is important to again compare recommended orthogonal scoring methods with CFA, assess if the problems

found in Version 1 persist and resolve which methods may best analyse Version 2 to produce summary scores consistent with the sub-scales.

A second important question relating to the use of the SF-36 is whether or not cross-country comparisons of health status are valid using the recommended United States (US) factor scoring coefficients in the development of the PCS and MCS. The developers of the SF-36 Version 2 advocate use of US factor score weights in creating the PCS and MCS in other countries [19]. This has the effect of artificially inflating or deflating these components for local decision making, which could confuse investment decisions in health for other countries. Given the potential differences of health status, the distribution of health and the perception of health in different countries, the question arises as to whether or not PCS and MCS scores should be based on country specific weights and, therefore, be free to vary from country to country, in order to accurately reflect the sub scale scores generated. Using US factor score coefficients standardises scores of each country to the US sub-scale score profile [20], which is possibly different to the sub-scale score profile of the country conducting the study. The important question to be answered is whether or not comparisons across countries are best made on the basis of country specific weighting coefficients?

Our aim was to assess whether previous problems of disagreement between the eight SF-36 Version 1 sub-scale scores and the Physical and Mental Component Summary scales (PCS and MCS) persist in version 2 of the instrument. A second study objective is to review the recommended scoring methods for the creation of factor scoring weights and the effect on producing summary scale scores

Methods

Statistical background and methodological issues

In producing the SF-36 component summaries (PCS and MCS) from the SF-36 data there are two main options for rotation of factors. This is done depending on whether or not the investigator believes the factors to be correlated (oblique) or uncorrelated (orthogonal). The recommended scoring methods for the SF-36 are based on orthogonal rotations, but we will argue that this creates data agreement problems and that there is strong support for adopting an oblique approach.

The items of the SF-36 are set out in Table 1.

A hypothetical factor structure has already been documented for the SF-36 [21]. This formed the basis of the model we evaluated, except that we allowed physical and mental health to be correlated (see Figure 1). It was therefore possible to fit a second order confirmatory factor analysis (CFA). The model fit was the full measurement model, using items re-coded as detailed in the SF36 scoring manual [20], with the exception that integer values of the items were retained so that they could be modeled using polychoric and tetrachoric correlations in LISREL V8.7. The above model was fit on 3,014 observations with no missing data for any items. The data produced using the CFA was compared with an analysis using the recommended orthogonal scoring methods [22].

Exploratory factor analysis (EFA) based on z-scores of the sub-scales, employing a principal components (PCA) extraction and an orthogonal rotation of factors was used by the developers to produce the SF-36 scoring coefficients for the component summary scores. This model cannot be directly fit using CFA software as the model is unidentified. However, using MacDonald's "echelon form" [23] where one non-significant path is constrained to zero, fit measures for the EFA model were generated in Stata [24]. It should be pointed out that the EFA model uses Pearson correlations of z-scored normally distributed data for the eight sub-scale scores, whereas the CFA model uses polychoric correlations of the 35 data items involved in the calculation of the SF-36 scores. Also the Akaike Information Criteria (AIC) value from the CFA model fit in LISREL V8.7 [25] is based on the Satorra-Bentler Chi-squared value, and the AIC from the EFA model fit in Stata SE V12 [24] is based on the model chi-square which is -2*log likelihood. To produce a fair comparison of the two models, the AIC was re-calculated for the CFA model based on the value of -2*log likelihood.

Hawthorne et. Al [22]. have published population norms for the transformed subscale scores from the 2004 SA Health Omnibus Survey [26], and they used the traditional scoring approach of Ware et al to produce factor score weights for the calculation of the Australian SF-36 summary scores. We also used these published norms and weights to produce subscale and summary PCS and MCS scales, distributed N(50,100), based on the traditional orthogonal method, for comparison with the CFA, using the 2008 SA Health Omnibus Survey data set.

Given the complexity of decisions made in the process of the CFA analysis the following methodological explanations are provided.

First, Rigdon & Ferguson [27] have shown that Maximum Likelihood (ML) estimation based on a polychoric correlation matrix is insufficient to correct for the problems associated the type of data in this study. For this reason weighted least squares (WLS) estimation is preferred. Further, Mindrilla [28] concluded that Diagonally Weighted Least Squares (DWLS) is superior to ML for the analysis of ordinal data.

Nye & Drasgow [29] consider that WLS and DWLS are both from the Asymptotically Distribution Free (ADF) family of estimators, and require similar large size samples. They investigated sample sizes from 400 to 1600. Flora & Curran contradict this paper, concluding that DWLS (they call it robust WLS) is superior to WLS in almost all situations, especially when the model is complex or the sample is small (n = 100). The largest sample size they considered was 1000 [30].

Forero et. al [31] compared unweighted least squares (ULS) and diagonally weighted least squares (DWLS) as alternatives to WLS for estimating Confirmatory Factor Analysis (CFA) models with ordinal indicators in a Monte Carlo study, and concluded that ULS was preferable, but if this did not converge then DWLS should be used, even in small samples (they examined sample sizes of 200. 500, and 2000). WLS was eliminated from consideration due to the requirement for very large sample sizes.

For our analysis, we have a moderate sample size of 3014. We attempted to use ULS as recommended by Forero et al [31], but this did not converge for the SF-36 model. We therefore chose to use DWLS to fit the model for SF-36. The model for SF-12 converged using ULS.

For maximum likelihood estimation of multivariate normal data, fit measure cutoffs have been set out by Hu and Bentler [32] as: Root Mean Square Error of Approximation (RMSEA) $<= 0.06$, Standardised Root Mean Square Residual (SRMSR) $<= 0.08$, Tucker Lewis Index (TLI) $>= 0.95$, Comparative Fit Index (CFI) $>= 0.95$. TLI is also known as the Non-Normed Fit Index (NNFI).

Nye & Drasgow [29] concluded that the fit measures and cutoffs in use for ML estimation of multivariate normal data do not apply to ADF estimators. They based their proposals for interpretation of fit measures on DWLS estimators of dichotomous indicators in CFA via tetrachoric correlations. They used Monte Carlo computer simulation to study the effects of model misspecification,

Table 1. Detailed items of the SF-36 version 2.

Sub-scale	Item	Short description	Question
Physical	a3a	Vigorous activities	The following questions are about activities that you might do
Functioning	a3b	Moderate activities	during a typical day. As I read each item, please tell me if your
	a3c	Lift/Carry groceries	health now limits you a lot, limits you a little, or does not limit you
	a3d	Climb several flights	at all, in these activities.
	a3e	Climb one flight	1 = Yes, limited a lot
	a3f	Bend, Kneel	2 = Yes, limited a little
	a3g	Walk kilometre	3 = No, no limited at all
	a3h	Walk half a kilometre	
	a3i	Walk 100 metres	
	a3j	Bathe, Dress	
Role	a4a	Cut down time	The following four questions ask you about your physical health
Physical	a4b	Accomplished less	and your daily activities. During the past four weeks, how much
	a4c	Limited in kind	of the time have you.?
	a4d	Had difficulty	1 = All of the time
			2 = Most of the time
			3 = Some of the time
			4 = A little of the time
			5 = None of the time
Bodily Pain	a7	Pain-magnitude	How much bodily pain have you had during the past four weeks?
			1 = None
			2 = Very mild
			3 = Mild
			4 = Moderate
			5 = Severe
			6 = Very severe)
	a8	Pain-interfere	During the past four weeks, how much did pain interfere with your
			normal work, including both work outside the home and
			housework?
			1 = Not at all
			2 = Slightly
			3 = Moderately
			4 = Quite a bit
			5 = Extremely
General	a1	EVGFP rating	These first questions are about your health now and your current
Health			daily activities. Please try to answer every question as accurately
			as you can. In general, would you say your health is:
			1 = Excellent
			2 = Very good
			3 = Good
			4 = Fair
			5 = Poor
	a11a	Sick easier	Now I'm going to read you a list of statements. After each one,
	a11b	As healthy	please tell me if its definitely true, mostly true, mostly false, or
	a11c	Health to get worse	definitely false. If you don't know just tell me.
	a11d	Health excellent	1 = Definitely true
			2 = Mostly true
			3 = Don't know
			4 = Mostly false
			5 = Definitely false

Table 1. Cont.

Sub-scale	Item	Short description	Question
Vitality	a9a	Full of life	The following questions are about how you feel and how things
	a9e	Energy	have been with you in the past four weeks. As I read each
	a9g	Worn out	statement, please give me the one answer that comes closest to the
	a9i	Tired	way you have been feeling. Would you say all of the time, most of
			the time, some of the time, a little of the time or none of the time?
			1 = All of the time
			2 = Most of the time
			3 = Some of the time
			4 = A little of the time
			5 = None of the time
Social	a6	Social-extent	During the past four weeks, to what extent has your physical health
Functioning			or emotional problems interfered with your normal social activities
			with family, friends, neighbours or groups? Has it interfered:
			1 = Not at all
			2 = Slightly
			3 = Moderately
			4 = Quite a bit
			5 = Extremely
	a10	Social-time	During the past four weeks, how much of the time has your
			physical health and emotional problems interfered with your social
			activities like visiting friends and relatives? Would you say:
			1 = All of the time
			2 = Most of the time
			3 = Some of the time
			4 = A little of the time
			5 = None of the time
Role	a5a	Cut down time	The following three questions ask about your emotions and your
Emotional	a5b	Accomplished less	daily activities. During the past four weeks, how much of the time
	a5c	Not careful	have you.?
			1 = All of the time
			2 = Most of the time
			3 = Some of the time
			4 = A little of the time
			5 = None of the time
Mental	a9b	Nervous	The following questions are about how you feel and how things
Health	a9c	Down in dumps	have been with you in the past four weeks. As I read each
	a9d	Calm	statement, please give me the one answer that comes closest to the
	a9f	Felt down	way you have been feeling. Would you say all of the time,
	a9h	Happy	most of the time, some of the time, a little of the time or none of the time?
			1 = All of the time
			2 = Most of the time
			3 = Some of the time
			4 = A little of the time
			5 = None of the time

doi:10.1371/journal.pone.0061191.t001

sample size, and non-normality on fit indices generated from DWLS estimation on dichotomous data. The study consisted of a 3 (model misspecification)×3 (degree of nonnormality)×3 (sample size) design. This is based on simulations of sample sizes of 400, 800, and 1600, using values of 0, 0.5, and 1.75 for skewness, and 0, 1.0, and 3.75 for kurtosis.

The reader is indirectly invited to extend the results to ordinal data and polychoric correlations, but this is an assumption. They

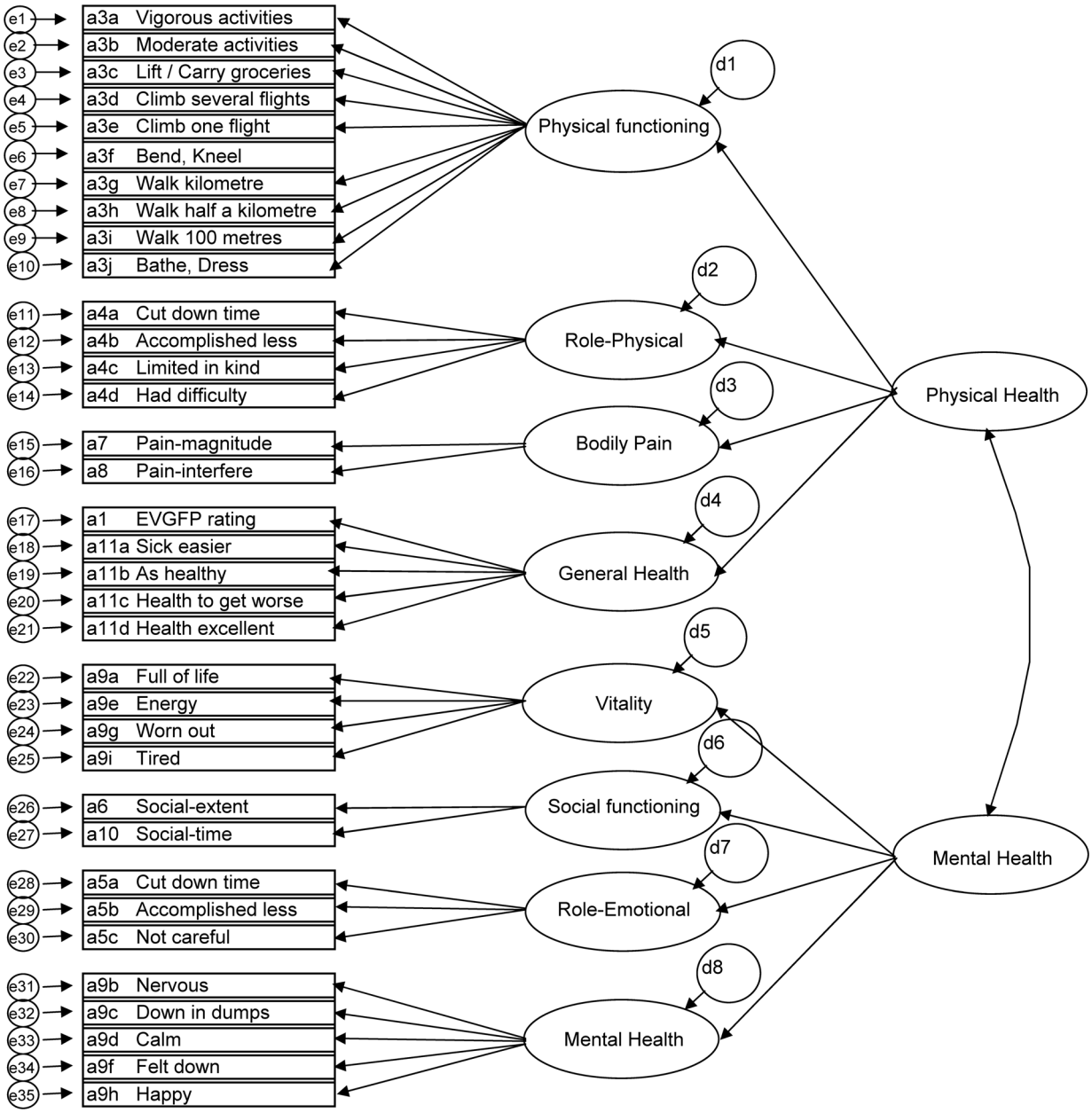


Figure 1. Hypothesised structure of SF-36 Health Dimensions and the Summary Physical (PCS) and Mental (MCS) Health Measures.
doi:10.1371/journal.pone.0061191.g001

have set out how to calculate cutoffs for fit measures for different situations (i.e. different levels of skewness, kurtosis, sample size, and required type I error rates). They only considered positive skewness in their calculations. They found that CFI & TLI were almost always near 1, and did not provide any discrimination regarding the fit of these models. Therefore, they recommend judging fit for these models based on their calculated cutoffs for RMSEA and SRMSR.

Flora & Curran [30] found that "there were few to no differences found in any empirical results as a function of two category versus five category ordinal distributions." This conclusion supports the generalisation of Nye & Drasgow's work from tetrachoric to polychoric correlations. They also found that DLWS produced more accurate estimates of the model chi-square, and therefore all of the fit measures that are based on it. In WLS estimation, the "inflation of the test statistic increases Type I error rates for the chi-square goodness-of-fit test, thereby causing researchers to reject correctly specified models more often than expected.". In this sense, Flora and Curran argue the opposite of Nye & Drasgow, [29] who proffer the advice that goodness-of-fit criteria need to be tightened up to avoid accepting inadequate models.

Nye and Drasgow [29] considered sample sizes up to 1600, and the formulae they provide produce complex roots when applied to our dataset, despite our skewness and kurtosis parameters lying within the ranges used in their simulations. We consider that this is because our sample size is much greater than the experience of their simulations.

Since the Nye and Drasgow [29] formulae fail to provide real valued cutoffs in our dataset, and Flora and Curran [30] argue for

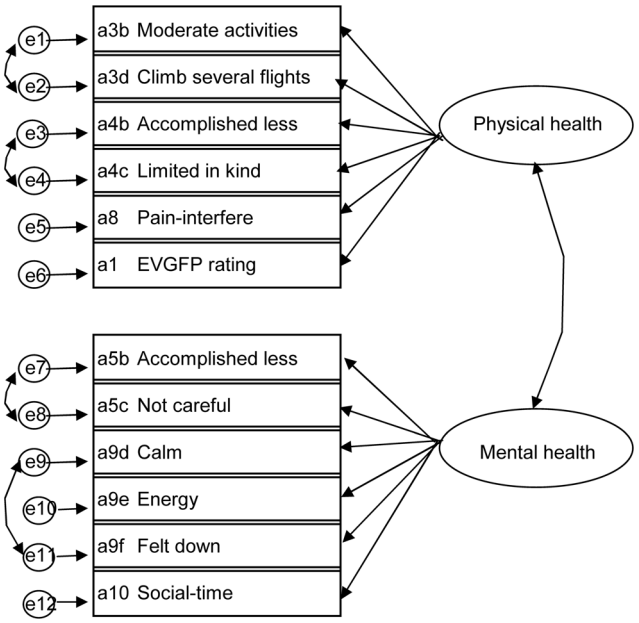


Figure 2. Hypothesised structure of SF-12 Summary Physical (PCS) and Mental (MCS) Health Measures.
doi:10.1371/journal.pone.0061191.g002

less stringent rather than more stringent fit criteria, we are comfortable using the maximum likelihood criteria advanced by Hu and Bentler [32] to assess model fit in this analysis, with the exception that Nye and Drasgow's advice regarding the non-discrimination of the TLI and CFI fit indices is accepted. We have therefore based our acceptance of the model on an RMSEA$<=0.06$ and a SRMSR$<=0.08$.

Statistical analysis

The 2004 South Australian Health Omnibus Survey dataset was used as the basis for the production of scoring coefficients [26]. This is the earliest Australian population survey available which included version 2 of the SF-36 health status questionnaire. In this representative population survey n = 3,014 adults aged 15 years or older were interviewed, all of whom provided full information for the SF-36. This is the same dataset as used by Hawthorne et. al. [22]. The data items were recoded as per the instructions of the SF-36 scoring manual [20].

The confirmatory factor analyses were fit on polychoric correlations in LISREL V8.7 [25] software. The model for SF-36 is a second order confirmatory factor analysis model. Unfortunately LISREL does not produce factor score weights for second order factors. The AMOS package [33] does produce these coefficients, but does not model polychoric correlations. Therefore we applied the AMOS formula for the generation of factor score weights to the outputs provided by LISREL to calculate factor score weights for version 2 of the SF-36. The AMOS formula is given by $W = B\,S^{-1}$ where W is the matrix of factor score weights, S is the fitted variance covariance matrix of the observed variables in the model, and B is the matrix of covariances between the observed and unobserved variables [33]. As pointed out by Joreskog [34] latent variable scores should be independent of the estimation method used to fit the model. The use of this formula satisfies this requirement.

The existence of factor score weights for all of the 35 items in the calculation of the summary scores based on the model is explained by the fact that all variables have an effect on both physical and mental health by virtue of the correlation between them, which is allowed for in the model.

A similar approach was used to model the SF-12 variables (see Figure 2). Models were again fit to produce the factor score weights in a confirmatory factor analysis. The data were recoded as per the instructions of the SF-36 scoring manual [20], with the exception that question eight of the SF-36 was recoded according to the instructions where question seven is not answered. This is because question seven is not asked in collecting the SF-12 data items. This resulted in 3,014 records being available to the analysis. In the model, correlations were allowed among the error terms for items from the same SF-36 sub-scale, because items from the same sub-scale, could reasonably be expected to be more closely correlated with each other than with the other items of the SF-12.

Comparisons of the PCS and MCS mean scores were based on agreement with the underlying subscales for both the orthogonal rotation and CFA. It was postulated that any sub-group summary score that was higher or lower than average should be in statistical agreement with the underlying subscales that contribute to that summary score. For comparison we used four age groups (<30 years, 30–49 years, 50–69 years and 70+ years) and four medication groups (no medication, physical health medication, mental health medication and both physical and mental health medication). Both sets of scores were based on the 2008 SA Health Omnibus Survey data. Since all scores were hypothesised to be distributed normally with a mean of 50 and a standard deviation of 10, comparisons were made assuming equal variances. Mean scores for four age groups and four medication groups were compared with the complementary groups to determine which age and medication groups had scores which were higher or lower than average scores. Similar comparisons were also made for the eight sub-scale scores. For each age and medication group comparisons of summary scores were made with the underlying sub-scale scores using independent groups t-tests. These analyses were carried out using SPSS Version 19 [35].

Results

The traditional orthogonal EFA model had an RMSEA = 0.104, SRMSR = 0.022, CFI = 0.972, TLI = 0.940, and AIC = 58497.72. This can be compared with our CFA model with RMSEA = 0.049, SRMSR = 0.053, CFI = 0.995, TLI = 0.9908, and AIC = 50495.37. From these fit measures it can be seen that the CFA model provides a much superior fit to the data than the EFA model with an orthogonal rotation. We bear in mind the view of Nye and Drasgow [29] that the CFI and TLI are constrained to be near unity in the analysis of polychoric correlations for ordinal data.

Table 3.5 of SF-36 Physical and Mental Health Summary Scales: A User's Manual [21] provides the Pearson product-moment correlations of the sub-scales for the general US population. This table provides sufficient information to test the fit of the original orthogonal EFA model employed by the developers of the scale. Using the same methods as above, the orthogonal EFA of the original US data had an RMSEA = 0.092, SRMSR = 0.028, CFI = 0.971, TLI = 0.938, and AIC = 47130.90. The original US model therefore shows a similar degreee of lack of fit as the same model fit to Australian data by Hawthorne [22].

The coefficients generated by the CFA analysis for the SF-36 are set out in Table 2. The model had a Chi-square of 53511.3 on 551 degrees of freedom, the size of which is explained by the large sample size. The Satorra-Bentler [36] scaled chi-square was 4648.5. The model had an RMSEA of.050 (90% confidence

Table 2. Australian weighting coefficients for the SF-36 version 2.

	PCS	MCS
A3A	0.0258	0.0025
A3B	0.0623	0.0120
A3C	0.0445	0.0025
A3D	0.0680	0.0070
A3E	0.2366	0.0263
A3F	0.0268	0.0031
A3G	0.1044	0.0087
A3H	1.0457	0.1367
A3I	0.1675	0.0262
A3J	0.0169	0.0021
A4A	0.5621	0.0697
A4B	1.2488	0.1658
A4C	1.9280	0.2391
A4D	2.2187	0.2757
A7	0.2566	0.0352
A8	0.9124	0.1121
A1	0.4297	0.0523
A11A	0.1698	0.0212
A11B	0.2881	0.0368
A11C	0.0895	0.0119
A11D	1.2084	0.1553
A9A	0.1653	1.6617
A9E	0.1882	1.8847
A9G	0.0817	0.8083
A9I	0.0839	0.8228
A6	0.1478	1.4785
A10	0.2389	2.4014
A5A	0.1022	1.0694
A5B	0.1017	1.0106
A5C	0.0393	0.3615
A9B	0.0125	0.1300
A9C	0.0503	0.4939
A9D	0.0263	0.2478
A9F	0.0601	0.5955
A9H	0.0288	0.2983
Constant term	−0.1097	−9.6528

doi:10.1371/journal.pone.0061191.t002

interval.048 to.051), a probability of close fit of 0.6522, and a standardised root mean square residual of 0.076. The Non-Normed Fit Index was 0.9904 and the Comparative Fit Index was 0.9911. The estimate of the correlation between physical and mental health was 0.73 ($p<0.001$).

Based on these weights the theoretical range of the SF-36 version 2 PCS is (12.3279,59.6503), and the observed range was (13.5313,59.6503). For the SF-36 version 2 MCS the theoretical range is (5.0138,63.3733), and the observed range was (5.5778,63.3733).

The coefficients generated by the CFA analysis for the SF-12 are set out in Table 3. The model had a Chi-square of 2646.6 on 49 degrees of freedom. The Satorra-Bentler scaled chi-square was

Table 3. Australian weighting coefficients for the SF-12 version 2.

	PCS	MCS
A1	1.3019	0.2044
A3B	1.2625	0.1984
A3D	0.6006	0.0943
A4B	3.0028	0.4730
A4C	2.9809	0.4693
A8	2.0033	0.3157
A5B	0.1863	1.8531
A5C	0.0953	0.9532
A9D	0.0800	0.7996
A9E	0.2422	2.4132
A9F	0.1370	1.3584
A10	0.4964	4.9376
Constant term	0.3833	−9.0891

doi:10.1371/journal.pone.0061191.t003

588.4. The model had an RMSEA of 0.060 (90% confidence interval 0.056 to 0.065), a probability of close fit of 0.000, and a standardised root mean square residual of 0.075. The Non-Normed Fit Index was 0.9874 and the Comparative Fit Index was 0.9906. The estimate of the correlation between physical and mental health was 0.71 ($p<0.001$).

Based on these weights the theoretical range of the SF-12 version 2 PCS is (12.7725,58.6031), and the observed range was (12.7725,58.6031). For the SF-36 version 2 MCS the theoretical range is (4.9811,60.6765), and the observed range was (4.9811,60.6765).

In comparing the effect of orthogonal rotation methods with confirmatory factor analysis we compared the summary scale scores with their underlying sub-scale scores for different age groups in Table 4 and for medication groups in Table 5. From the tables clear discrepancies are apparent between the traditional summary scores and their sub-scales, which are not evident using scoring coefficients derived from confirmatory factor analysis.

Table 4 shows several discrepancies between the summary component scores and their underlying sub-scale scores when scored using orthogonal methods, as set out by Hawthorne [22]. The score for the SF-36 mental health sub-scale for those aged under thirty years is not significantly different to the overall sub-scale average ($p = 0.918$),. The remaining three sub-scale scores that comprise the SF-36 mental component are all significantly higher than average (role emotional ($p = 0.026$), vitality ($p<0.001$), social functioning ($p = 0.005$)), as are the mental component summary scores (MCS) from CFA coefficients for both SF-36 ($p<0.001$) and SF-12 ($p<0.001$), yet the MCS score, based on the original orthogonal scoring algorithm, is significantly lower than average ($p = 0.035$).

For those aged 30–49 years, none of the mental health sub-scales are significantly different to average (vitality ($p = 0.272$), social functioning ($p = 0.650$), role emotional ($p = 0.295$), and mental health ($p = 0.264$)), yet the MCS was significantly lower than average ($p<0.001$) using orthogonal scoring, but there was no significant difference for the SF-36 MCS score using CFA coefficients ($p = 0.561$) or SF-12 using CFA coefficients ($p = 0.294$).

For those aged 50–69 years, three of the mental health scales were not significantly different to average (vitality ($p = 0.120$), role

Table 4. Comparison of subscale scores and summary scores using different scoring methods, by age groups.

	<30	30–49	50–69	70+	Total
n	515	991	939	472	2917
Physical function scale -	54.7	52.7	47.5	39.9	50.0
Aust normed T-score					
Role physical scale - Aust	52.2	50.9	47.2	43.6	49.2
normed T-score					
Bodily pain scale - Aust	52.5	49.0	45.7	45.5	48.5
normed T-score					
General health scale - Aust	51.6	50.4	47.7	46.1	49.3
normed T-score					
Vitality scale - Aust	51.3	49.1	48.9	47.6	49.4
normed T-score					
Social function scale - Aust	50.4	49.3	48.4	48.3	49.2
normed T-score					
Role emotion scale - Aust	49.7	48.6	48.7	48.7	48.9
normed T-score					
Mental health scale - Aust	49.1	48.8	49.0	50.0	49.1
normed T-score					
SF-36–PCS scored using	54.1	51.6	46.6	41.5	49.5
Aust weighted T-score					
SF-36 MCS- scored using	48.3	47.9	49.6	52.0	49.0
Aust weighted T-score					
SF-36 PCS - scored using	52.6	50.8	47.2	43.9	49.3
SEM coefficients					
SF-36 MCS - scored using	51.1	49.3	48.2	47.0	49.1
SEM coefficients					
SF-12 PCS - scored using	52.6	50.8	47.0	43.6	49.2
SEM coefficients					
SF-12 MCS - scored using	51.2	49.5	48.2	47.0	49.2
SEM coefficients					

doi:10.1371/journal.pone.0061191.t004

emotional ($p = 0.466$), and mental health ($p = 0.795$)) and social functioning was significantly lower than average ($p = 0.012$), yet the MCS was significantly higher than average ($p = 0.044$) using orthogonal scoring but significantly lower than average for both SF-36 ($p = 0.003$) and SF-12 ($p = 0.001$) using CFA coefficients.

For those aged 70 years or more, the vitality scale was significantly lower than average ($p<0.001$), whilst the social functioning ($p = 0.083$), role emotional ($p = 0.711$), and mental health score (0.069) were not significantly different to average. The MCS scores from CFA coefficients for both SF-36 ($p<0.001$) and SF-12 ($p<0.001$) were significantly lower than average, yet the MCS score based on the original orthogonal scoring method was significantly higher than average ($p<0.001$). There were no inconsistencies evident by age for physical health summary scores when compared to their subscales.

Similar discrepancies arise in comparison of the component summary scores with their underlying sub-scale scores for those taking medications for either or both physical and mental health conditions. Table 5 shows that for those not taking medications no inconsistencies between sub-scales and summary scores were evident. For those taking medications for physical ailments the vitality ($p<0.001$) and social functioning ($p<0.001$) sub-scales scores were both significantly lower than average, while the role emotional score ($p = 0.155$) and the mental health score ($p = 0.789$) were not significantly different to average. This is consistent with the mental health summary scores (MCS) from CFA coefficients which were significantly lower than average for both SF-36 and SF-12 ($p<0.001$), yet the MCS score based on the original orthogonal scoring method was significantly higher than average ($p<0.001$).

Similarly, three of the physical health subscale scores are significantly lower than average for those taking medications for mental health reasons (role physical ($p = 0.002$), bodily pain ($p<0.001$), and general health ($p<0.001$)), while the physical functioning scale is not significantly different to average ($p = 0.196$). This is consistent with the physical health summary scores (PCS) from CFA coefficients which are significantly lower than average for both SF-36 ($p<0.001$) and SF-12 ($p<0.001$), yet the PCS score based on the original scoring coefficients is not significantly different to average ($p = 0.380$) for PCS calculated using orthogonal methods.

There were no inconsistencies evident for those taking medication for both physical and mental health problems for physical or mental health summary scores when compared to their subscales.

In summary, the CFA produced a superior fit to the SF-36 data, provided acceptable fit measures and solved agreement problems observed in the orthogonal analyses.

Discussion

We raise two points of difference with the developers regarding the development of scoring norms and weights. First, that PCS and MCS summary scores should be based on a model that allows correlation of physical and mental health, to preserve consistency of summary scores with their underlying sub-scales. We thank an anonymous reviewer who has also pointed out that "this issue is probably more of a concern with the SF12 than the SF36. The SF36 generates subscale scores, so users can notice and evaluate the potential problems caused by orthogonally-derived summary scores. But the SF12 generates only summary scores, so the problem will be hidden from users.". Second, that scoring norms and weights should be produced on country specific data, so that all scores are based on the same data items and have the same distributions (normal with mean 50 and standard deviation 10). This is essential for country decision making especially from summary scales for sub-groups, but further in this way all countries will produce T-scores for all sub-scales and summary scales that allow accurate international comparisons, without the need to standardise to USA factor weights

The use of US factor score weights in the calculation of summary scores seems inappropriate for other countries, because the linear combination of z-scored sub-scales using US weights results in the emphasis being placed on those sub-scales which have higher US weights. Hawthorne [22] has analysed Australian SF-36 version 2 data from the 2004 Health Omnibus Survey. His analysis replicated precisely to the methods used by the developers, but included allowances for the production of Australian norms for use in calculating the z-scores for the sub-scales, and for the calculation of Australian factor score weights from an orthogonal EFA. His analysis showed that the factor score weights produced based on Australian data were significantly different to those produced using USA data. None of the USA weights were in the 95% CI of the Australian weights. Thus the profile of locally calculated weights can be very different to the US weights and

Table 5. Comparison of subscale scores and summary scores using different scoring methods, by medication status.

	No medication	Physical only	Mental only	Both	Total
n	1549	1120	95	153	2917
Physical function scale -	53.5	45.6	48.7	41.0	50.0
Aust normed T-score					
Role physical scale - Aust	52.2	45.8	45.9	40.6	49.2
normed T-score					
Bodily pain scale - Aust	51.6	44.9	43.7	38.7	48.5
normed T-score					
General health scale - Aust	52.4	45.8	44.1	40.6	49.3
normed T-score					
Vitality scale - Aust	51.4	47.9	41.9	40.7	49.4
normed T-score					
Social function scale - Aust	51.1	47.8	41.2	40.7	49.2
normed T-score					
Role emotion scale - Aust	50.7	48.5	37.4	38.2	48.9
normed T-score					
Mental health scale - Aust	50.3	49.2	39.0	40.1	49.1
normed T-score					
SF-36–PCS scored using	53.1	44.6	48.5	40.9	49.5
Aust weighted T-score					
SF-36 MCS- scored using	49.8	50.0	37.0	40.3	49.0
Aust weighted T-score					
SF-36 PCS - scored using	52.7	45.6	44.4	39.3	49.3
SEM coefficients					
SF-36 MCS - scored using	51.6	47.4	39.2	37.9	49.1
SEM coefficients					
SF-12 PCS - scored using	52.5	45.6	44.6	39.1	49.2
SEM coefficients					
SF-12 MCS - scored using	51.7	47.5	39.4	38.0	49.2
SEM coefficients					

doi:10.1371/journal.pone.0061191.t005

therefore the summary scores produced by locally produced weights would emphasise different sub-scales than the US weights. This results in the calculation of inaccurate summary scores when using US weights. In principal therefore, calculation of summary scores should be based on locally calculated weights. In the present study we used the Australian norms and factor score weights based on Australian data developed by Hawthorne [22] to produce the component summary scores for the traditional orthogonal scoring method. Table 2 of Hawthorne's paper also demonstrated the shortcomings of applying US norms and weights to Australian data, in that the 95% CI for all subscale T-scores and the MCS T-score excluded 50. So even if we stick to orthogonal analyses there is important and increasing evidence that strictly applying US factor score weights in the creation of summary scores is a problem for local interpretation and use of data. It is argued that the profile of locally calculated weights can be very different, as demonstrated by Hawthorne [22], and often for the valid reasons of differences in health. The aim of measuring health status should primarily be for the production of valid local scores based on country specific norms and not for the primary purpose of standardising to US data for comparison purposes. Further, if we need to compare with the US or with any other country it would best be done on the basis of subscale T-scores and summary scores based on individual data items and local population norms for the creation of factor score weights in a second order confirmatory factor analysis, so that scores are all based on the same data items and have the same distribution.

In fairness to the authors of the SF-36 they have produced a leading generic quality of life instrument and measure and there is little or no criticism about the long-term historical development of question items. The main points of contention are involved in scoring the summary scores. The question which has to be answered by other interested researchers is does the proposed CFA fix the underlying problems identified with the PCS and MCS and should US factor score weights be used for anything other than academic comparison with US data, and not for country specific estimates which may be skewed by US coefficients.

The CFA used in this analysis is based on the original data items and the orthogonal analysis on the underlying subscales. It is argued this is a reasonable comparative approach of the two methods as the data items are used to create the subscales. The main difference in the comparisons is therefore based on the methodological difference of orthogonal or oblique rotation and not on data differences. We argue the oblique rotation method is

an improved way of handling the data. We further argue that the approach recommended by the developers is unsustainable in Australia, and possibly elsewhere, because the factor score weights should be free to vary from country to country in order to accurately reflect the sub-scale scores generated by the SF-36 data in each country. This point is supported by Hawthorne's analysis of the Australian data [22].

We accept that Hawthorne's findings contradict the findings of the IQOLA project [19]. Australia appears to offer divergent results to the other mainly European countries included in the IQOLA study, and we note that these analyses were conducted on different datasets. The critical point is the existence of the dataset that produced Hawthorne's results. Hawthorne's analysis satisfactorily demonstrates the need for an Australian country specific scoring algorithm. The question of the need for country specific scoring algorithtms elsewhere has not been covered by our analysis, and should be the subject of further research.

We are aware that demonstration of the inconsistencies between the sub-scales and the component summary scores in two tables (4 & 5) is not a comprehensive validation of the scoring coefficients, but we suggest there are limits to how much analysis can be squeezed into one paper.

Conclusion

The conclusion of the study is that the problems of agreement between PCS and MCS summary scores and their underlying subscales identified in Version 1 of the SF-36 persist in Version 2. As identified in the Version 1 analyses [4], this occurs when a negative Z-score is multiplied by a negative coefficient, resulting in a positive score. This mathematical difficulty is compounded by the orthogonal method used, and why the authors continue to promote the method in the face of international concerns and a real world correlation between mental and physical health is not clearly understood. In a defence of the SF-36 scoring methods and the instruments accuracy, Ware and Kosinski [37], discuss the question of the PCS and MCS being rotated by orthogonal or oblique methods and ask how much physical health should be in mental health and vice versa. If, however, exploratory factor analysis using maximum likelihood extraction and oblique rotation were used, this would estimate the hypothetical factor structure and the data would determine how much mental health is contained in physical health and vice versa.

In Ware and Kosinski's [37] defence of the SF-36 they also contend "results based on summary measures should be thoroughly compared with the SF-36 profile….," before drawing any conclusions. If we followed this advice for the above analyses of Version 2 data (and also for Version1) we would conclude the disagreement between scales and summary scores is consistent using orthogonal modeling and is based on a mathematical artefact.

Acknowledgments

The classification of those taking physical and mental health medications presented in Table 5 was based on data collected by Professor Robert Goldney and coded by Dr Kerena Eckert, both of the University of Adelaide (at that time). Further, we would like to thank the three anonymous reviewers who passed comment on this paper. We believe their comments and suggestions have greatly improved the final manuscript.

Author Contributions

Conceived and designed the experiments: GT RA DW. Analyzed the data: GT. Wrote the paper: GT RA DW.

References

1. Hays RD, Sherbourne CD, Mazel RM (1993) The RAND 36-Item health Survey. Health Econ. 2: 217–227.
2. Simon GE, Revicki DA, Grothaus L, Vonkor M (1998) SF-36 summary scores. Are physical and mental health truly distinct. Med Care 36: 567–72.
3. Taft C, Karlsson J, Sullivan M (2001) Do SF-36 summary scores accurately summarise subscale scores? Qual Life Res 10: 395–404.
4. Wilson D, Parsons J, Tucker (2000) The SF-36 summary scales: problems and solutions. Soz-Praventivmed 45: 239–246.
5. Tucker G, Adams R, Wilson D (2010) New Australian population scoring coefficients for the old version of the SF-36 & SF-12 health status questionnaires. Qual Life Res 19(7): 1069–76.
6. Farrivar SS, Cunningham WE, Hays RD (2007) Correlated physical and mental health summary scores for the SF-36 and SF-12 health survey. Health Qual Life Outcomes 5: 54.
7. Hann M, Reeves D (2008) The SF-36 summary scales are not accurately summarized by independent physical and mental component scores. Qual Life Res 17: 413–23.
8. Agnastopoulos F, Niakis D, Tountas Y (2009) Comparison between exploratory factor analytic and SEM-based approaches to constructing SF-36 summary scores. Qual Life Res 18: 53–63.
9. Fleishman JA, Selim AJ, Kasiz LE (2010) Deriving SF-12 v2 physical and mental health summary scores: a comparison of different scoring algorithms. Qual Life Res 19(2): 231–41.
10. Hemingway H, Stafford M, Stansfield S, Shipley M, Marmot M (1997) Is the SF-36 a valid measure of change in population health? Results from the Whitehall Study. BMJ 315: 1273–78.
11. Kosinski M, Keller SD, Ware JE Jr, Hatoum HT, Kong SX (1999) The SF-36 as a generic outcomes measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: relative validity of scales in relation to clinical measures of arthritis severity. Med Care 37: (%Suppl): MS23–39.
12. Quality Metric. Available: QualityMetric.com/WhoAreWe/ScientificHeritage/tabid/156/Default.aspx.Accessed 2012 Jan 11 .
13. Goodwin RD (2003) Association between physical activity and mental disorders among adults in the United States. Prev Med 36(6): 698–703.
14. Strohle A, Hofler M, PfisterH, Müller AG, Hoyer J, et al (2007) Physical activity and prevalence and incidence of mental disorders in adolescents and young adults. Psychol Med (11): 1657–66.
15. Alonso J, Lepine J-P (2007) European Study of the Epidemiology of Mental Disorders/Mental Health Disability: A European Assessment in the Year 2000 Scientific Committee. J Clin Psychiatry (Suppl 2): 3–9
16. Collingwood J.The relationship between mental and physical health. Psych Central.Available:http://psychcentral.com/lib/2010/the-relationship-between-mental-and-physical-health.Accessed 2012 Jan 11.
17. Chapman DP, Perry GS, Strine TW (2005) The vital link between chronic disease and depressive disorders. Prev Chron Dis 2(1): 1–10.
18. Katon WJ (2003) Clinical and health services relationships between major depression, depressive symptoms, and general medical illness. Biol Psychiatry 54: 216–226.
19. Ware JE Jr, Gandek B, Kosinski M, Aaronson NK, Apolone G, et al. (1998) The Equivalence of SF-36 Summary Health Scores Estimated Using Standard and Country-Specific Algorithms in 10 Countries: Results from the IQOLA Project. J Clin Epidemiol Vol 51, No 11: 1167–1170.
20. Ware JE, Kosinski MA, Dewey JE (2000) How to score version 2 of the SF-36 health survey. Lincoln: Quality Metric Inc.
21. Ware JE, Kosinski M, Keller SD (1994) SF-36 Physical and Mental Health Summary Scales: a Users Manual. Boston, MA. The Health Institute, New England Medical Centre.
22. Hawthorne G, Osborne RH, Taylor A, Sansoni J (2007) The SF-36 Version2: critical analyses of weights, scoring algorithms and population norms. Qual Life Res 16: 661.73.
23. MacDonald RP (1999) Test Theory: A unified Treatment. Mahwah NJ, Lawrence Erlbaum Associates.
24. Stata Corp (2011) Stata Statistical Software: Release 12. College Station, Texas, StataCorp LP.
25. Joreskog KG, Sorbom D (1996) LISREL User's Reference Guide. Chicago,IL.Scientific Software International
26. Wilson D, Wakefield M, Taylor A (1992) The South Australian Health Omnibus Survey. Health Promot J Austr 2: 47–49
27. Rigdon EE, Ferguson CE (1991) The performance of the Polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. J Marketing Res 1991:28: 491–97
28. Mindrilla D (2010) Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: a comparison of estimation bias with ordinal and multivariate non-normal data. International Journal of Digital Society (IJDS) 1: 60–66.

29. Nye CD, Drasgow F (2011) Assessing Goodness of Fit: Simple Rules of Thumb Simply Do Not Work. Org Res Methods 14: 548–570.
30. Flora DB, Curran PJ (2004) An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychol Methods 9: 466–91.
31. Forero CG, Maydeu-Olivares A, Gallardo-Pujol D (2009) Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. Struct Equ Modeling 16: 625–641.
32. Hu L, Bentler PM (1999) Cuttoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modelling 3: 424–53.
33. Arbuckle JL, Wothke W (1999) Amos 4.0 User's Guide. Chicago,IL.Small Waters Corporation.
34. Joreskog KG. (2000) Latent Variable Scores and Their Uses. Lincolnwood, ILScientific Software International.
35. IBM SPSS Statistics Version 19.
36. Satorra A, Bentler PM (1988) Scaling corrections for chi-square statistics in covariance structure analysis. Proceedings of the Business & Economic Statistics Section of the American Statistical Association, 308–313.
37. Ware JE, Kosinski M (2001) Interpreting SF-36 summary health measures: A response. Qual Life Res 10: 405–413.

My third publication compared the measurement properties of the physical component summary (PCS) and mental component summary (MCS) scores of the SF-36 and SF-12 based on the recommended orthogonal scoring algorithms with the performance of the PCS and MCS scores based on the confirmatory factor analysis coefficients from a correlated model. It demonstrated the superior performance of my published scoring coefficients compared to the proprietary scoring algorithm in three representative population survey datasets. Similar comparisons with similar conclusions were presented for another six datasets in the supplementary material provided with this paper. This publication therefore provided large scale evidence of the problems resulting from use of the recommended scoring algorithms. For calculating the summary scores based on the developers algorithm I used Australian norms published by the Australian Bureau of Statistics for version 1 of the SF-36 [29], and Australian norms and scoring coefficients published by Hawthorne for version 2 of the SF-36 [10]. The US weights were used for the SF-12 PCS and MCS scores based on the orthogonal model for version 1 and version 2 since no Australian norms or coefficients have been published. My published scoring coefficients consistently outperformed the recommended scoring coefficients in every dataset, with higher correlations between the relevant PCS and MCS scores and other variables they were expected to be related to, e.g. age, Body Mass Index, Australian Quality of Life index, Selim's Chronic Lung Disease index, General Health Questionnaire, Center for Epidemiologic Studies Depression scale, and the Kessler 10 item Anxiety and Depression scale [50]

Results from several population studies show that recommended scoring methods of the SF-36 and the SF-12 may lead to incorrect conclusions and subsequent health decisions [50]
Graeme Tucker, Robert Adams, David Wilson

Statement of Authorship

Title of Paper	Results from Several Population Studies Show That Recommended Scoring Methods of the SF-36 and the SF-12 May Lead to Incorrect Conclusions and Subsequent Health Decisions.
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Tucker G, Adams R, Wilson D. (2014) Results from Several Population Studies Show That Recommended Scoring Methods of the SF-36 and the SF-12 May Lead to Incorrect Conclusions and Subsequent Health Decisions. Quality of Life Research 23:2195-2203 DOI: 10.1007/s11136-014-0669-9

Principal Author

Name of Principal Author (Candidate)	Graeme Tucker
Contribution to the Paper	Study conception and design, statistical analysis, interpretation of data, manuscript preparation, critical revision of the manuscript, corresponding author.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 23/3/17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Robert Adams
Contribution to the Paper	Supervised development of the work, data interpretation, critical revision of the manuscript.
Signature	Date 18/5/2017

Name of Co-Author	David Wilson
Contribution to the Paper	Supervised development of the work, data interpretation, manuscript preparation, critical revision of the manuscript.
Signature	Date 23/03/2017

Please cut and paste additional co-author panels here as required.

Qual Life Res (2014) 23:2195–2203
DOI 10.1007/s11136-014-0669-9

Results from several population studies show that recommended scoring methods of the SF-36 and the SF-12 may lead to incorrect conclusions and subsequent health decisions

Graeme Tucker · Robert Adams · David Wilson

Accepted: 6 March 2014 / Published online: 20 March 2014

Abstract
Purpose To compare the measurement properties of the physical component summary (PCS) and mental component summary (MCS) scores of the SF-36 and SF-12 based on the traditional orthogonal scoring algorithms with the performance of the PCS and MCS scored based on structural equation model coefficients from a correlated model.
Methods This study used three large-scale representative population studies to compare the measurement properties of the PCS and MCS scores of the SF-36 and SF-12 with the performance of the PCS and MCS scores based on structural equation models producing coefficients from a correlated model. We assessed the relationships of these scores with selected important mental health measures and chronic conditions from three representative Australian population studies that address clinical conditions of high prevalence and health service importance.
Results Structural equation model scoring methods produced summary scores with higher correlations than the recommended orthogonal methods across a range of disease and health conditions. The problem experienced in using the orthogonal methods is that negative scoring coefficients are applied to negative z-scores for sub-scales, inflating the resulting summary scores. Effect sizes over a half of a standard deviation were common.
Conclusions If health policy or investment decisions are made based on the results of studies employing the recommended orthogonal scoring methods then the expected outcome of such decisions or investments may not be achieved.

Keywords Self-rated health · Health-related quality of life · SF-36/SF-12 · Correlated v orthogonal scoring

Introduction

The SF-36 and the shorter form SF-12 health status questionnaires have been used in Australian population and other research studies for many years [1]. They have provided powerful insights into the health status of groups and populations across a number of health dimensions [2] worldwide and have also been used as outcome measures in studies [3, 4]. For researchers and policy makers, they provide substantial decision-making information. The SF-36 was first validated for Australian use by McCallum [5, 6], and an Australian adaptation was developed by Sanson-Fisher et al. [7].

The original SF-36 (version 1) used a method of scoring for physical component summary (PCS) and mental component summary (MCS) scores based on factor coefficients derived through principle components analyses and orthogonal rotation [8]. This method of scoring the SF-36 has been criticised in the literature over many years [9–18]. The basis of this criticism is that orthogonal scoring methods produce PCS and MCS summary scores that are at variance with underlying sub-scale scores. However, the orthogonal scoring methods remain the default scoring method for this measure. This effect is counter intuitive as it implies that physical and mental health are unrelated to each other. This can be resolved when structural equation modelling (SEM) scoring methods are used that allow

Electronic supplementary material The online version of this article (doi:10.1007/s11136-014-0669-9) contains supplementary material, which is available to authorized users.

G. Tucker (✉) · R. Adams · D. Wilson
School of Medicine, University of Adelaide, Adelaide, SA, Australia
e-mail: Graeme.Tucker@health.sa.gov.au

mental and physical health to be correlated when constructing summary scores [13, 19, 20]. The SF-36 was refined in 1998 to create version 2 of the instrument; however, the orthogonal scoring methods were retained.

In practice, there is ample evidence that when orthogonal scoring methods are used, low mental health scale scores are associated with higher PCS scores, and low physical health scale scores are associated with high MCS scores [12–18]. This is caused by the application of negative weights to negative z-scores for sub-scales in the traditional scoring algorithm, as noted by Simon et al. [12], Wilson et al. [13] and Taft [14]. With the SEM models, factor score weights are applied to the recoded questionnaire items, which are all positive, so the problem is averted.

We have recently demonstrated the problems in scoring the physical and mental health summary scores of the SF-36 for both version 1 and version 2, and published effective solutions [19, 20].

The connection between physical and mental health is also supported by a large literature [21–30]. This literature is ignored in the perseverance with orthogonal scoring by the developers of the SF-36. We have shown that correlated scoring is statistically superior in our previous publications. We now examine whether or not correlated scoring gives a more accurate picture clinically.

In this study, we have used a number of high-quality well-established representative Australian population studies to compare the measurement properties of the PCS and MCS summary scores of the SF-36 and SF-12 based on the traditional orthogonal scoring algorithms with the performance of the PCS and MCS scored based on SEM coefficients from a correlated model. We have assessed the relationships of these scores with selected important mental health measures and chronic conditions from three representative Australian population studies that address clinical conditions of high prevalence and health service importance.

Methods

The PCS and MCS summary scores for the SF-36 and/or SF-12 were calculated for each of the studies used based on the traditional scoring methods, and using the SEM coefficients we have previously published for Australian data [19, 20].

For orthogonal scoring of version 1 of the SF-36, the weights published by the Australian Bureau of Statistics (ABS) based on the 1995 National Health Survey [31], which included the SF-36 for the Australian population, were used. For orthogonal scoring of version 2 of the SF-36, the weights published by Hawthorne et al. [1] were employed. In producing these weights, Hawthorne adhered rigidly to the methods used by the developers of the SF-36, namely an orthogonal rotation of the principal components solution. For orthogonal scoring of both versions of the SF-12, the US weights and scoring methods were applied [32, 33].

Correlations of these scores with various other health measures collected in the population survey datasets were examined. The other measures considered were as follows:

Body mass index (BMI) is a simple index of weight-for-height that is commonly used to classify underweight, overweight and obesity in adults. It is defined as the weight in kilograms divided by the square of the height in metres (kg/m^2). The World Health Organisation [34] defines the major categories of BMI as underweight (<18.5), normal (18.5 to <25.0), overweight (25.0 to <30.0) and obese (≥30.0).

The assessment of quality of life (AQoL) instrument is a validated health-related multi-attribute utility quality of life instrument. [35–37].

The Chronic Lung Disease Index (CLD) is a symptom-based measure of the severity of chronic lung disease. [38] It has been validated locally in a general population sample by Ruffin et al. [39] who found it discriminates between different levels of CLD severity and is significantly correlated with all scales of the SF-36.

The GHQ is a self-administered questionnaire that focuses on two major areas: inability to carry out normal functions and appearance of new and distressing phenomena. Its purpose is to screen for non-psychotic psychiatric disorders. In this analysis, the 12 items version was used [40].

The Center for Epidemiologic Studies Depression (CES-D) Scale is a self-report depression scale for research in the general population [41].

The Kessler 10 scale (K10) is a 10-item questionnaire intended to yield a global measure of distress based on questions about anxiety and depressive symptoms that a person has experienced in the most recent 4-week period [42].

Details on the methodology of each of the population study data sources are listed below. Comparison of estimates was undertaken using the data sets described below.

The population data sets used

South Australian health omnibus survey (SF-36 version 2) [43]

The South Australian health omnibus survey (SAHOS) is a face-to-face population interview survey conducted annually since 1991 for government and non-government organisations responsible for planning and/or servicing the health needs of the South Australian community. The goal

of SAHOS is to collect, analyse and interpret data that can be used to plan, implement and monitor health programs and other initiatives. Data presented here were obtained from the SAHOS during September and October 2008. Within each ABS collector district, a random starting point was selected and 10 households were sampled using a fixed skip interval. In a non-replacement sample, one adult aged 15 years or older, whose birthday was next, was selected for interview in their home by trained health interviewers. The SAHOS methodology has been described in detail elsewhere. The SAHOS is based on a representative household population sample in which data are weighted by age, gender and to the probability of selection in that household.

The questionnaire and methodology for this survey were approved by the South Australian Department of Health Human Research Ethics Committee.

The North West Adelaide Health Study (NWAHS) [44]

The North West Adelaide Health Survey (NWAHS) is a combined population household interview survey and clinical study of adults (age >18 years) in the north-western suburbs of Adelaide, South Australia (regional population 0.6 million). Again, households are selected randomly and one person is selected for telephone interview. They were then recruited to a hospital clinic for more detailed interviewing and clinical assessment on a range of important health parameters. All households in South Australia with a telephone connected and the telephone listed in the current version of the electronic white pages (EWP) were eligible for selection in the sample. The computer-assisted telephone operating system (CATI system) was used to conduct the interviews with at least six call-backs made to the telephone number to interview the selected household member. The person chosen for interview was the person last to have a birthday. Different times of the day or evening were scheduled for each call-back. Data obtained were weighted to the closest census data to provide population representative estimates. Data are again weighted by age, gender and to the probability of selection in the household.

The NWAHS was approved by the North West Adelaide Health Service Ethics of Human Research Committee, and all subjects gave written informed consent.

The Western Australia, Northern Territory and South Australia Survey (WANTS) [45]

An Australian Federal health survey was undertaken in 2000 to compare chronic disease and risk factor estimates in South Australia (SA), the Northern Territory (NT) and Western Australia (WA). All households in these states and territory with a telephone connected and the telephone number listed in the current version of the EWP were eligible for selection in the sample. Random samples were drawn separately for each state/territory, and separate samples were drawn for each of the three geographic regions (metro/rural/remote) of the states/territory.

A CATI system was used to conduct the interviews with at least six call-backs made to the telephone number to interview the selected household member. The person chosen for interview was the person last to have a birthday. Different times of the day or evening were scheduled for each call-back.

A total of $n = 7{,}619$ interviews were conducted (approximately $n = 2{,}500$ in each state/territory). Overall, the response rate was 63.1 %. In SA, $n = 2{,}545$ interviews were conducted with a response rate of 63.8 %. Data were weighted by age, gender and probability of selection in the household, and to the Accessibility and Remoteness Index of Australia (ARIA) regions (metropolitan, rural and remote). Only the SA sample has been used in this analysis.

The questionnaire and methodology for this survey were approved by the South Australian Department of Health Human Research Ethics Committee. We have also examined the means produced by the different scoring approaches to give an indication of the size of the effect. Cohen [46] proposed rules of thumb for interpreting effect size. A small effect is 0.2 of a standard deviation, a medium effect is 0.5 of a standard deviation and a large effect is 0.8 of a standard deviation.

Results

Dataset 2008 health omnibus survey

Version 2 SF-36/SF-12

In Table 1, the 2008 SAHOS population sample comprised 2,824 respondents from 4,614 households contacted (61.2 % participation rate), and the socio-demographic

Table 1 Correlation coefficients

	Age	BMI	AQoL
SF-36 PCS (orthogonal)	−.405^	−.220^	.626^
SF-36 PCS (correlated)	−.298^	−.193^	.714^
SF-12 PCS (orthogonal)	−.384^	−.215^	.600^
SF-12 PCS (correlated)	−.303^	−.194^	.700^
SF-36 MCS (orthogonal)	.095^	−.049#	.506^
SF-36 MCS (correlated)	−.146^	−.150^	.722^
SF-12 MCS (orthogonal)	.066^	−.065#	.495^
SF-12 MCS (correlated)	−.149^	−.143#	.713^

* $p < .05$, # $p < .01$, ^ $p < .001$

distribution of participants corresponded to SA population estimates (Australian Bureau of Statistics. Population by age and sex, Australian states and territories Jun 2007. Canberra: ABS, 2008. (ABS Cat. No. 3201.0.)). Of the 2,824 participants, 1,358 (48.1 %) were male and 2,158 (76.4 %) resided in the metropolitan area.

As expected, age has a significant negative correlation with PCS regardless of how the PCS is scored (Table 1). However, age has a significant positive correlation with MCS using orthogonal scoring, and a significant negative correlation with MCS using correlated scoring. Lower scores in the physical health sub-scales inflate MCS scores and vice versa using the traditional orthogonal scoring, because of the application of negative coefficients to negative z-scores for sub-scales. This is the explanation for the results observed throughout this study.

Physical health scores have a significant negative correlation with BMI, regardless of how they are scored. As noted above, with orthogonal scoring, lower PCS scores are associated with higher MCS scores. Consequently, whilst there is a significant decline in MCS score with increasing BMI for both SF-36 & SF-12 when correlated scoring is used, MCS scores have a weaker relationship with BMI when orthogonal scores are analysed. When compared with another generic quality of life scale (AQOL), there is a higher correlation with AQOL for both PCS and MCS using correlated scoring.

Physical component summary and MCS summary scores are calculated based on a normal distribution, with a mean of 50 and a standard deviation of 10. All differences commented on were statistically significant at the 5 % significance level.

In Table 2, for PCS by age groups, the trends are the same, and SF-36 and SF-12 scores are consistent regardless of how the summary scores are calculated. However, for MCS scores, the trends are in opposite directions as noted in Table 1. This is due to the low PCS scores inflating MCS scores in orthogonal scoring. The difference for SF-36 MCS between orthogonal and correlated scoring is 5.0 in the 70+ age group (i.e. a half a standard deviation). The difference for SF-12 MCS scores in the 70+ age group is 7.2. Disaggregation by age groups provides the most striking demonstration of the differences encountered between orthogonal and correlated scoring of the PCS and MCS summary scores.

For BMI, PCS scores had a similar level and pattern regardless of how they were calculated. The obese group scored higher for MCS with orthogonal scoring, as expected. The difference in MCS scores for the obese group was 1.6 for SF-36 and 4.1 for SF-12.

In general, for AQoL scores, correlated PCS and MCS scores for the SF-36 and SF-12 were more consistent with each other than orthogonal scores. This could be due at least in part to the fact that the SF-36 orthogonal scores have Australian weights derived by Hawthorne et al., whereas the SF-12 orthogonal scoring employs US weights, since no Australian weights have been published. Correlated scoring has Australian weights for both the SF-36 and SF-12.

Dataset NWAHS: stage 2

Version 1 SF-36/SF-12

In Table 3, age again has a significant positive relationship with MCS using orthogonal scoring, and a significant negative correlation with MCS using correlated scoring.

Body mass index does not have a significant correlation with MCS scored using orthogonal methods, but a significant negative correlation using correlated scoring.

Structural equation modelling coefficients produce PCS and MCS scores that are more closely correlated with the Chronic Lung Disease Index than orthogonal scoring provides.

Physical component summary scored using SEM coefficients has a stronger correlation with the GHQ and depression as measured by the CESD than PCS under traditional scoring algorithms. This is to be expected given that poor mental health scores inflate physical health scores using orthogonal scoring.

In Table 4, a similar pattern for age is observed as for Table 2. Also note that the decline in physical health for those aged 70+ is overstated using orthogonal scoring for both the SF-36 and the SF-12. The difference between MCS scores for orthogonal v correlated scoring for SF-36 was 6.6, and for SF-12, it was 7.4.

For BMI, similar scores and patterns for PCS were observed, and again low PCS scores were associated with higher MCS scores using the orthogonal method in the obese group. For MCS, the obese group scored higher with orthogonal scoring, by 3.7 for SF-36 and 4.4 for SF-12.

For CLD, the patterns and scores were roughly equivalent for PCS. Those with moderate or severe CLD have much higher MCS scores with orthogonal scoring than with correlated scoring. Differences are 5.3 for moderate CLD for SF-36, and 6.7 for SF-12. For severe CLD, the differences are 8.2 for SF-36 and 10.7 for SF-12.

GHQ is a screening tool for non-psychotic mental health conditions, and so we expect the MCS scores to be relatively consistent using either scoring method. Because of the poor MCS scores in the GHQ $\geq$3 group, we expect increased PCS scores using orthogonal scoring. We observe that MCS scores are fairly consistent using either scoring method, except that orthogonal scores for SF-12 are slightly higher in the GHQ >3 group. As occurred with version 2 data, orthogonal SF-12 scores are the only scores using US weights, since no Australian weights have been published. PCS scores are higher using orthogonal scoring than

Table 2 Means

	SF-36 V2 PCS (orthogonal)	SF-36 V2 PCS (correlated)	SF-12 V2 PCS (orthogonal)	SF-12 V2 PCS (correlated)	SF-36 V2 MCS (orthogonal)	SF-36 V2 MCS (correlated)	SF-12 V2 MCS (orthogonal)	SF-12 V2 MCS (correlated)
Age								
<30	54.0	52.6	53.3	53.6	48.3	51.1	51.8	51.3
30–49	51.6	50.9	51.5	51.7	47.8	49.3	51.0	49.6
50–69	46.5	47.1	46.5	47.7	49.4	48.0	52.4	48.3
70+	41.5	43.9	41.1	44.3	52.0	47.0	54.3	47.1
BMI								
<18.5	51.6	51.1	50.5	52.1	49.1	50.1	52.9	50.9
18.5–25	51.5	51.1	51.4	52.0	49.1	50.3	52.4	50.5
>25–30	49.5	49.4	49.4	50.0	49.4	49.6	52.3	49.7
>30	45.9	46.3	45.7	47.0	48.1	46.5	51.0	46.9
AQoL score								
<.75	40.6	39.5	40.6	40.0	42.2	39.1	45.8	39.2
.75 to <.85	49.8	50.1	49.9	50.6	49.5	49.7	52.4	50.1
.65 to <.95	52.2	52.8	52.4	53.8	52.0	52.8	54.5	53.2
.95+	56.1	56.1	54.8	57.1	53.1	55.9	56.2	55.9

Table 3 Correlation Coefficients

	Age	BMI	CLD	GHQ	CESD
SF-36 PCS (orthogonal)	−.456^	−.221^	−.359^	−.088^	−.160^
SF-36 PCS (correlated)	−.278^	−.196^	−.422^	−.445^	−.479^
SF-12 PCS (orthogonal)	−.441^	−.213^	−.382^	−.166^	−.231^
SF-12 PCS (correlated)	−.303^	−.195^	−.411^	−.429^	−.477^
SF-36 MCS (orthogonal)	.162^	−.017	−.210^	−.693^	−.628^
SF-36 MCS (correlated)	−.102^	−.136^	−.382^	−.648^	−.632^
SF-12 MCS (orthogonal)	.142^	−.027	−.198^	−.677^	−.622^
SF-12 MCS (correlated)	−.145^	−.146^	−.390^	−.607^	−.613^

* $p < .05$, # $p < .01$, ^ $p < .001$

correlated scoring, the differences being 5.6 for SF-36, and 3.5 for SF-12.

Depressed people have low MCS scores, which produces inflated PCS scores using orthogonal scoring. As a result, PCS scores decline much more rapidly with increasing levels of depression when correlated scoring is employed. The difference for SF-36 PCS scores is 10.2, and for SF-12 PCS scores, it is 7.8. MCS scores are again reasonably consistent except for the orthogonal SF-12 scores.

Dataset WANTS

SF-12 version 1

In Table 5, age is again associated with a decline in physical and mental health using correlated scoring, but with the traditional orthogonal scoring, age has a positive correlation with MCS.

Physical health scores have a significant negative correlation with BMI, regardless of how they are scored. MCS scores have no relationship with BMI when orthogonal scores are analysed, but there is a significant decline in MCS scores with increasing BMI for SF-36 when SEM scoring coefficients are used. The correlation of PCS with K10 is about twice as high with correlated scoring as it is with orthogonal scoring. Again, this is to be expected given that physical and mental health are related.

In Table 6, the decline in physical health with age is again overstated with orthogonal scoring (70+ score is 3.2 lower), and the MCS is 6.4 higher for the 70+ age group.

Body mass index has similar scores and pattern for PCS. For MCS, the obese group scores 3.5 higher with orthogonal scoring.

For Kessler 10, the ≤43 score group (i.e. less psychological distress) has an orthogonal PCS score 2.9 lower

Table 4 Means

	SF-36 V1 PCS (orthogonal)	SF-36 V1 PCS (correlated)	SF-12 V1 PCS (orthogonal)	SF-12 V1 PCS (correlated)	SF-36 V1 MCS (orthogonal)	SF-36 V1 MCS (correlated)	SF-12 V1 MCS (orthogonal)	SF-12 V1 MCS (correlated)
Age								
<30	53.2	51.8	52.5	53.7	49.7	50.9	51.0	51.7
30–49	50.4	50.3	50.1	51.5	50.8	50.5	51.6	50.7
50–69	45.1	47.4	45.3	47.9	52.7	49.7	53.3	49.4
70+	38.8	43.3	39.3	43.8	54.3	47.7	54.4	47.0
BMI								
<18.5	47.4	47.7	47.9	48.4	50.0	48.6	50.1	48.5
18.5–25	49.6	50.5	49.3	51.6	52.1	51.2	52.8	51.3
>25–30	47.4	48.5	47.2	49.4	51.9	50.1	52.7	50.0
>30	44.0	45.8	44.3	46.6	51.8	48.1	52.2	47.8
Selim's index for severity of chronic lung disease								
Mild	47.9	49.2	47.9	50.2	52.2	50.6	52.8	50.5
Moderate	39.3	39.8	39.6	40.5	47.6	42.3	48.9	42.2
Severe	33.4	32.7	32.8	33.8	44.5	36.3	47.6	36.9
GHQ score ≥3								
No	47.5	50.1	47.7	51.1	54.6	52.4	55.0	52.2
Yes	45.3	39.7	44.0	40.5	38.7	37.5	40.8	38.0
CESD depression scale score								
<16—Not depressed	47.8	50.0	47.9	51.0	53.8	51.9	54.3	51.8
16 to <27—Mild depression	44.0	39.6	42.6	40.2	41.1	38.9	43.0	39.2
27+—Severe depression	44.0	33.8	41.9	34.1	29.3	29.6	32.1	29.9

Table 5 Correlation coefficients

	Age	BMI	K10
SF-12 PCS (orthogonal)	−.398^	−.208^	−.266^
SF-12 PCS (correlated)	−.293^	−.192^	−.543^
SF-12 MCS (orthogonal)	.112^	−.004	−.698^
SF-12 MCS (correlated)	−.173^	−.141^	−.713^

* $p < .05$, # $p < .01$, ^ $p < .001$

than correlated scoring, and a MCS score equal for either scoring method. The >52 score group (i.e. more psychological distress) has a PCS 3.0 higher for orthogonal scoring, and an MCS 2.1 higher, when compared with correlated scoring.

Discussion

At the heart of this study, we argue that the appropriate method of analysing the SF-36 is based on the relationship of the two major dimensions of health summarised in physical and mental health. It is argued that this connection cannot be ignored in the choice of analysis method and that data are best analysed using structural equation methods. It is difficult to understand the persistence with orthogonal methods given the overwhelming evidence and scientific opinion on the connections between mental and physical health. This relationship is not only a constant theme in the medical literature, but is a connection that is constantly argued by leading health authorities among whom comprise the World Health Organisation [47], the Kings Fund [48], the academy of Medical Royal Colleges [49] and the World Federation of Mental Health [50]. These, and other authorities, argue for a more substantial and integrated perception and management of physical and mental health conditions based on their own substantive and systematic reviews of the literature. Not only do they argue the range of excess chronic conditions and impaired management of physical and mental health in isolation of each other, but also identify the two way relationship and the underlying mechanisms, or consequences, of the relationship between physical and mental health through poor motivation, loss of opportunity, homelessness, withdrawal and early retirement from the workforce, violence and crime in addition to the day to day chronic gnawing despair and anxieties of the comorbidities.

In all three health studies used in our analyses, it is important to note the contrasting positive and negative

Table 6 Means

	SF-12 V1 PCS (orthogonal)	SF-12 V1 PCS (correlated)	SF-12 V1 MCS (orthogonal)	SF-12 V1 MCS (correlated)
Age				
<30	53.4	53.6	51.9	53.1
30–49	51.1	51.3	51.3	51.1
50–69	46.8	48.2	53.0	50.1
70+	41.9	45.1	54.2	47.8
BMI				
<18.5	50.8	51.4	51.4	51.3
18.5–25	50.7	51.5	52.4	51.8
>25–30	49.1	50.4	53.0	51.3
>30	45.5	45.8	50.5	47.0
Kessler 10 score				
Up to 43	52.5	55.4	56.9	56.9
>43–46	49.7	52.2	55.8	53.8
>46–52	48.9	50.4	53.6	51.5
>52	45.9	42.9	43.6	41.5

correlations produced for each scoring method. Only one correlation direction can be correct. Given that health declines with age we argue that SEM provides the correct coefficient. Using SEM scoring, we also showed stronger correlations for both PCS and MCS with another validated generic health scale (AQOL) that does not use these scoring methods. We expect MCS to be negatively related to BMI due to body image considerations, and its effect on social interactions, as was observed using correlated scoring. The measure of lung disease, the CLD, shows stronger correlation on both the PCS and MCS scales using SEM scoring. This is consistent with much prior evidence that shows a significant effect of lung disease on mental and physical health [51, 52]. With all of the instruments measuring aspects of physical or mental health, stronger correlations in the expected direction were seen with correlated scoring using SEM than with orthogonal scoring.

When examining the mean scores for PCS and MCS, the effect sizes were not inconsequential. Most difference in scores between the orthogonal and correlated alternatives was between 0.5 and 1.0 standard deviations. This effect size is characterised by Cohen as moderate to large. It is also evident that the lower the PCS or MCS scores, the greater the difference between the scoring methods for the respective MCS or PCS score. In other words, using orthogonal scoring, the lower the summary score the greater the distortion in the complementary summary score.

Conclusion

We believe we have demonstrated superior measurement properties of correlated scoring of the SF-36 and SF-12 PCS and MCS when compared to the orthogonal scoring methods preferred by the developers of the scales in studies using Australian representative population samples. Our approach produces summary scores with higher correlations to other health measures. Supplementary analyses were also conducted by gender (not shown) and these lead to the same conclusions as for the general population. There are consistent scoring problems when orthogonal scoring methods are used.

We have restricted ourselves to these results in these datasets in order to provide a focussed and succinct examination of the problem. We have conducted appreciably more analyses on different representative population health datasets that have consistently provided similar support for the superiority of correlated scoring algorithms over the traditional orthogonal scoring. These extra analyses can be made available upon request.

It is important to reiterate the consequences of using the traditional scoring methods. This can lead to incorrect decision-making that could affect either policy or health investment decisions, or both, and given that many studies employing the SF-36 have been for large population groups, it may be important for authorities to ask if decisions made on the basis of orthogonal scoring methods have led to erroneous policy or investment decisions.

Acknowledgments We wish to thank the anonymous reviewer for their helpful suggestion that improved the strength of the arguments presented in this paper.

Ethical standard This paper is based on a secondary analysis of various South Australian survey files. As such, this analysis did not require formal ethics approval; however, all of the original data collections were conducted under ethics approval with the informed consent of the participants.

References

1. Hawthorne, G., Osborne, R. H., Taylor, A., & Sansoni, J. (2007). The SF-36 Version 2: Critical analysis of population weighting, scoring algorithms and population norms. *Quality of Life Research, 16*, 661–673.
2. Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *The SF-36 health survey manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Centre.
3. Sorensen, L., Stokes, J. A., Purdie, D. M., et al. (2004). Medication reviews in the community: Results of a randomized, controlled effectiveness trial. *British Journal of Clinical Pharmacology, 58*, 648–664.
4. Commonwealth Department of Health and Aged Care. (1999). *The Australian coordinated care trials: Background and trial descriptions*. Canberra: Department of Health and Aged Care.

5. McCallum, J. (1995). The new SF-36 health status measure: Australian validity tests. Canberra: National Centre for Epidemiology and Population Health. Paper presented to the Health Outcomes and Quality of Life Measurement Conference.
6. McCallum, J. (1995). The SF-36 in an Australian sample: Validating a new, generic health status measure. *Australian Journal of Public Health, 19*, 160–166.
7. Sanson-Fisher, R. W., & Perkins, J. J. (1998). Adaptation and validation of the SF-36 health survey for use in Australia. *Journal of Clinical Epidemiology, 51*(11), 961–967.
8. Ware, J. E., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health summary scales: A users manual*. Boston, MA: The Health Institute, New England Medical Centre.
9. Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-item health survey 1.0. *Health Economics, 2*, 217–227.
10. Hays, R. D., Prince-Embury, S., & Chen, H. (1998). *RAND-36 health status inventory*. San Antonio, TX: The Psychological Corporation.
11. Hays, R. D., & Morales, L. S. (2001). The RAND-36 measure of health-related quality of life. *Annals of Medicine, 33*, 350–357.
12. Simon, G. E., Revicki, D. A., Grothaus, L., & Vonkorf, M. (1998). SF-36 summary scores: Are physical and mental health truly distinct? *Medical Care, 36*, 567–572.
13. Wilson, D., Parsons, J., & Tucker, G. (2000). The SF-36 summary scales: Problems and solutions. *Sozial-und Präventivmedizin, 45*, 239–246.
14. Taft, C., Karlson, J., & Sullivan, M. (2001). Do SF-36 summary component scores accurately summarise subscale scores? *Quality of Life Research, 10*, 395–404.
15. Farivar, S. S., Cunningham, W. E., & Hays, R. D. (2007). Correlated physical and mental health summary scores for the SF-36 and SF-12 health survey, V. 1. *Health and Quality of Life Outcomes, 5*, 54.
16. Hann, M., & Reeves, D. (2008). The SF-36 scales are not accurately summarised by independent physical and mental component scores. *Quality of Life Research, 17*, 413–423.
17. Anagnostopoulos, F., Niakas, D., & Tountas, Y. (2009). Comparison between exploratory factor-analytic and SEM-based approaches to constructing SF-36 summary scores. *Quality of Life Research, 18*, 53–63.
18. Fleishman, J. A., Selim, A. J., & Kazis, L. E. (2010). Deriving SF-12v2 physical and mental health summary scores: A comparison of different scoring algorithms. *Quality of Life Research, 19*(2), 231–241.
19. Tucker, G., Adams, R., & Wilson, D. (2010). New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires. *Quality of Life Research, 19*(7), 1069–1076.
20. Tucker, G. R., Adams, R. J., & Wilson, D. H. (2013). Observed agreement problems between Sub-scales and summary components of the SF-36 version 2-an alternative scoring method can correct the problem. *Plos One 8*(4):e61191. doi: 10.1371/journal.pone.0061191. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0061191.
21. Chapman, D. P., & Perry, G. S. (2008). Depression as a major component of public health for older adults. *Preventing Chronic Disease 5*(1). http://www.cdc.gov/pcd/issues/2008/jan/07_0150.htm. Accessed 21 May 2013.
22. Chapman, D. P., Perry, G. S., Strine, & T. W. (2005). The vital link between chronic disease and depressive disorders. *Preventing Chronic Disease.* http://www.cdc.gov/pcd/issues/2005/jan/04_0066.htm. Accessed 21 May 2013.
23. Cheok, F., Schrader, G., Banham, D., Marker, J., & Hordacre, A. L. (2003). Identification, course, and treatment of depression after admission for a cardiac condition: Rationale and patient characteristics for the Identifying Depression As a Comorbid Condition (IDACC) project. *American Heart Journal, 146*(6), 978–984.
24. Schrader, G., Cheok, F., Hordacre, A. L., & Guiver, N. (2004). Predictors of depression three months after cardiac hospitalization. *Psychosomatic Medicine, 66*(4), 514–520.
25. Wilson, D. H., Appleton, S. L., Taylor, A. W., Tucker, G., Ruffin, R. E., Wittert, G., et al. (2010). Depression and obesity in adults with asthma: Multiple comorbidities and management issues. *Medical Journal of Australia, 192*(7), 381–383.
26. Sullivan, M. D., O'Connor, P., Feeney, P., Hire, D., Simmons, D. L., Raisch, D. W., et al. (2012). Depression predicts all-cause mortality: Epidemiological evaluation from the ACCORD HRQL substudy. *Diabetes Care, 35*(8), 1708–1715. doi:10.2337/dc11-1791.
27. Lin, E. H., Von Korff, M., Ciechanowski, P., Peterson, D., Ludman, E. J., Rutter, C. M., et al. (2012). Treatment adjustment and medication adherence for complex patients with diabetes, heart disease, and depression: A randomized controlled trial. *The Annals of Family Medicine, 10*(1), 6–14. doi:10.1370/afm.1343.
28. Katon, W. J. (2011). Epidemiology and treatment of depression in patients with chronic medical illness. *Dialogues in Clinical Neuroscienc, 13*(1), 7–23.
29. Davydow, D. S., Katon, W. J., & Zatzick, D. F. (2009). Psychiatric morbidity and functional impairments in survivors of burns, traumatic injuries, and ICU stays for other critical illnesses: A review of the literature. *International Review of Psychiatry, 21*(6), 531–538. doi:10.3109/09540260903343877.
30. Llaneza, P., García-Portilla, M. P., Llaneza-Suárez, D., Armott, B., & Pérez-López, F. R. (2012). Depressive disorders and the menopause transition. *Maturitas, 71*(2), 120–130. doi:10.1016/j.maturitas.2011.11.017.
31. Australian Bureau of Statistics (1995). National Health Survey. SF-36 Population Norms Australia. Canberra: Australian Bureau of Statistics, Catalogue Number 4399.0.
32. Ware, J., Kosinski, M., & Keller, S. (1995). *SF-12: How to score the SF-12 physical and mental health summary scales* (2nd ed.). Boston: The Health Institute, New England Medical Center.
33. Ware J. E. Jr., Kosinski M., Turner-Bowker, D., Sundaram M., Gandeck, B., & Maruish M. E. (2002). User's manual for the SF-12 V2 health survey, Second Edition. *Quality Metric*, 24 Albion Rd, Building 400, Lincoln, RI 02865, USA.
34. http://apps.who.int/bmi/index.jsp?introPage=intro_3.html. Accessed 25 July 2013.
35. http://www.aqol.com.au/. Accessed 25 July 2013.
36. Hawthorne, G. & Richardson, J. (1997). The assessment of quality of life (AQoL) instrument construction, initial validation and utility scaling. *Centre for Health Program Evaluation*, Melbourne (15 pages) (ISBN 1 875677 85 2).
37. Hawthorne, G., Korn, S., & Richardson, J. (2013). Population norms for the AQoL derived from the 2007 Australian National Survey of Mental Health and Wellbeing. *Australian and New Zealand Journal of Public Health, 37*(1), 7–16.
38. Selim, A., Ren, X., Fincke, G., Rogers, W., Lee, A., & Kazis, L. (1997). A symptom-based measure of the severity of chronic lung disease: results from the Veterans Health Study. *Chest, 111*, 1607–1614.
39. Ruffin, R. E., Wilson, D. H., Chittleborough, C. R., Southcott, A. M., Smith, B., & Christopher, D. J. (2000). Multiple respiratory symptoms predict quality of life in chronic lung disease: A population-based study of Australian adults. *Quality of Life Research, 9*, 1031–1039.
40. Goldberg, D. P., & Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine, 9*, 139–145.

41. Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.
42. Kessler, R. C., Andrews, G., Colpe, L. J., Hripi, E., Mrocsek, D. K., Normand, S.-L. T., et al. (2002). Short screening scales to monitor population prevalences and trends in nonspecific psychological distress. *Psychological Medicine, 32*(6), 959–976.
43. Wilson, D., Wakefield, M., & Taylor, A. (1992). The South Australian health omnibus survey. *Health Promotion Journal of Australia, 2*, 47–49.
44. Grant, J. F., Taylor, A. W., Ruffin, R. E., Wilson, D. H., Phillips, P. J., Adams, R. J. T., et al. (2009). Cohort profile: The North West Adelaide Health Study (NWAHS). *International Journal of Epidemiology, 38*, 1479–1486. doi:10.1093/ije/dyn262.
45. WANTS Health West. (2001). *Collaborative health and wellbeing survey design and methodology*. Perth, WA: Western Australian Government.
46. Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Eribaum.
47. Kolappe, K., Henderson, D. C., & Kishare, S. P. (2013). No physical health without mental health: Lessons unlearned? *Bulletin of the World Health Organization, 91*, 3.
48. Naylor, C., Amy Galea, Parsonage, M., McDaid, D., Knapp, M., Fossey, M. (2012). Long term conditions and mental health. http://www.kingsfund.org.uk/publications/long-term-conditions-and-mental-health. Accessed 29 Aug 2013.
49. Academy of Medical Royal Colleges. (2009). *No health without mental health: The alert summary report*. London: Millbank Medical Ltd.
50. World Federation of Mental Health (2010). Mental Health and Chronic Illness. The Need for Continued and Integrated care. www.wfmh.org/2010docs/wmhday2010.pdf.
51. Adams, R. J., Wilson, D. H., Taylor, A. W., Daly, A., Tursan d'Espaignet, E., Dal Grande, E., et al. (2004). Psychological factors and asthma quality of life: A population based study. *Thorax, 59*, 930–935.
52. Jain, A., & Lolak, S. (2009). Psychiatric aspects of chronic lung disease. *Current Psychiatry Reports, 11*, 219–225.

My fourth publication addresses my final point of contention with the developers of the SF-36 regarding their published advice that the US scoring algorithms for SF-36 and SF-12 (version1 or version 2) could be safely applied to data from any country or cultural group without adjustment, and provided a valid method of cross country comparison. This fourth paper examined this proposition for multi-country SF-12 data using a large international database drawn from nine countries, to test equality between Australia and twelve other country/language groups. This involved the fitting of 54 different CFA models, with significance tests where appropriate for differences between models, and the collation of the results. It was found that CFA models with common parameters across countries were rejected on the basis of fit, with an inadequate SRMR, as well as by a chi-squared test of the difference between the restricted (equal factor score coefficients) and unrestricted (separately estimated factor score coefficients) models.

The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36 [49]
Graeme Tucker, Robert Adams, David Wilson

Statement of Authorship

Title of Paper	The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Tucker G, Adams R, Wilson D (2016) The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36 Quality of Life Research 25(2), 267-274. DOI: 10.1007/s11136-015-1083-7

Principal Author

Name of Principal Author (Candidate)	Graeme Tucker
Contribution to the Paper	Study conception and design, statistical analysis, interpretation of data, manuscript preparation, critical revision of the manuscript, corresponding author.
Overall percentage (%)	60%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 23/3/17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Robert Adams
Contribution to the Paper	Supervised development of the work, data interpretation, critical revision of the manuscript.
Signature	Date 18/5/2017

Name of Co-Author	David Wilson
Contribution to the Paper	Supervised development of the work, data interpretation, manuscript preparation, critical revision of the manuscript.
Signature	Date 23/03/2017

Please cut and paste additional co-author panels here as required.

The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36

Graeme Tucker, Robert Adams & David Wilson

Quality of Life Research
An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Research

ISSN 0962-9343

Qual Life Res
DOI 10.1007/s11136-015-1083-7

Springer

Qual Life Res
DOI 10.1007/s11136-015-1083-7

The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36

Graeme Tucker[1,3] · Robert Adams[2] · David Wilson[2]

Accepted: 23 July 2015

Abstract
Purpose To examine the validity of using the same scoring coefficients across countries for the SF-12.
Methods We test the equality of scoring coefficients derived for a contraction of the SF-36, the Short Form 12 (SF-12), using a large international database drawn from nine countries, to test equality between Australia and twelve other country/language groups. First, we checked that the theoretical structure of the SF-12 as set out by Ware and colleagues, but including a correlation between physical and mental health, provided an adequate fit to the data for each country/language group in a confirmatory factor analysis. We then compared Australia to all of these country/language groups in multiple-group models to assess whether a model producing common factor score coefficients provided an adequate fit to the data. We also derived Chi-squared tests for the differences between the restricted and unrestricted models, to test the equality of the factor score coefficients across countries.
Results We found that the theoretical structure of the SF-12, with a correlation between physical and mental health, provides an adequate fit to the data for all country/language groups except Hungary. Further, all the unrestricted multiple-group models provide an adequate fit to the data. In contrast, none of the multiple-group models restricted to common parameters provide an adequate fit to the data. The significance tests confirm that the constraints on parameter values produce significantly different models to the unrestricted models.
Conclusions We conclude that researchers should derive their own country-specific scoring coefficients for physical and mental health summary scores.

Keywords Self-rated health · Health-related quality of Life · SF-36 · SF-12 · International comparisons

Introduction

In previous research, we have shown that orthogonal scoring methods that assume no correlation between physical and mental health in scoring the SF-12 and SF-36 physical (PCS) and mental health (MCS) component summary scores are at variance with the underlying subscale scores. We have shown this problem can be corrected by including a correlation between mental and physical health using structural equation modeling (1–4). In this study, we follow up on our previous recommendations that US scoring coefficients for the physical component summary (PCS) and mental component summary (MCS) scores cannot be used across countries because the orthogonal method used to score the data is flawed [1–4]. We take this argument a stage further to test the equality of scoring coefficients derived for a number of countries using our recommended approach of structural equation modeling to assess the fit of the data [3, 4].

Historically, the Short Form SF-12 and SF-36 self-rated health questionnaires have been used extensively in

Electronic supplementary material The online version of this article (doi:10.1007/s11136-015-1083-7) contains supplementary material, which is available to authorized users.

✉ Graeme Tucker
grtucker@adam.com.au

[1] SA Health, 11 Hindmarsh Square, Adelaide, SA 5000, Australia

[2] The Health Observatory, Discipline of Medicine, The Queen Elizabeth Hospital Campus, University of Adelaide, Adelaide, SA, Australia

[3] Discipline of Medicine, University of Adelaide, Adelaide, SA, Australia

Published online: 28 September 2015

national and international studies. Hawthorne last reported the use of the SF-36 in over 5000 international translation or validation studies [5]. As such it has been an important research instrument for a large number of health studies dealing with clinical and policy issues. However, in recent years, the validity of the physical and mental health summary scores of the instrument has been seriously questioned on the basis of the methods used to score data [3–11]. The paper by Taft et al. prompted an exchange of views between this group and Ware's group [12, 13]. A second major concern with the SF-12 and SF-36 has now been raised by Hawthorne et al. [5] regarding the validity of using the same scoring coefficients across countries as recommended by the developers [14]. Hawthorne et al. [5] used Australian population data to show that recommendations to use US (derived) scoring coefficients in other countries to score their data may be seriously flawed. Hawthorne showed that there were differences between Australian-derived scoring coefficients and US-derived coefficients that produced important differences in seven of the eight SF-36V2 subscales and in the mental health summary scale. This is central to international research purposes questioning, not the structure or value of the instrument, but how it may be used to provide meaningful quality of life assessments outside the USA where it was developed. Hawthorne et al. [5] points out that where common descriptive methods are used across countries, there may be important cultural differences within which a person's health assessment is made. He points out this affects population weights, and scoring coefficients [5]. In this study, we address both of these issues. The research question of this study is whether or not the same scoring coefficients can be used across countries to compare health-related quality of life status. In doing so, we use representative population SF-12 data from nine countries to make comparisons of twelve different country/language groups with an Australian SF-12 population study.

Method

Datasets

Data on the twelve country/language groups were obtained from the Adult Literacy and Life Skills Survey 2003 and 2008 Public Use Microdata file obtained from Stats Canada [15]. "This study comprised a large-scale co-operative effort undertaken by governments, national statistics agencies, research institutions and multilateral agencies. The development and management of the study were coordinated by Statistics Canada and the Educational Testing Service (ETS) in collaboration with the National Center for Education Statistics (NCES) of the United States Department of Education, the Organisation for Economic Cooperation and Development (OECD), the Regional Office for Latin America and the Caribbean (OREALC) and the Institute for Statistics (UIS) of the United Nations Educational, Scientific and Cultural Organisation (UNESCO).

The survey instruments for the studies were developed by international teams of experts with financing provided by the Governments of Canada and the United States. A highly diverse group of countries and experts drawn from around the world participated in the validation of the instruments. Participating governments absorbed the costs of national data collection and a share of the international overheads associated with implementation" [15].

Australian SF-12 version 1 data were drawn from the 2006 Adult Literacy and Life Skills Survey (ALLS) dataset obtained from the Australian Bureau of Statistics (ABS) [16]. "The Adult Literacy and Life Skills Survey (ALLS) was also part of the international study coordinated by Statistics Canada and the Organisation for Economic Co-operation and Development (OECD)… The ALLS collected information from July 2006 to January 2007 from 8988 private dwellings throughout non-remote areas of Australia. The sample design ensured that within each state and territory, each household had an equal chance of selection. Information was obtained from one person aged 15–74 years in the selected household. If there was more than one person of this age, the person interviewed was selected at random" [16].

The ALLS survey was a household survey collection using trained professional interviewers, most of whom had over 2-years experience. Standard collection protocols were employed in each country, and checks for compliance were made by the administrators following data collection. The data were weighted to account for variations in the probability of selection of the respondent and further weighted to population benchmarks in each country/language group.

The SF-12 version 1 was a component of the standard questionnaire collected in this survey. The microdata file therefore contained the SF-12 version 1 data items for all of the twelve country/language groups provided on the Canadian file as well as the Australian file (see Table 1).

We also had available the data from the ABS 1995 National Health Survey [17]. The sample design of the 1995 NHS is a self-weighting multistage clustered area sample based on ABS census collector districts in which households are selected with equal probability. In this survey, $n = 23{,}800$ households were selected and all adults aged 15 or older were interviewed. A subset of $n = 19{,}785$ were asked to complete the SF-36 health status questionnaire. Of those interviewed, $n = 17{,}479$ provided full data for the SF-12. Survey records were weighted to allow for probability of selection and then to population benchmarks. These data were used as a double check on model validity, by comparing the model for the Australian ALLS survey

Table 1 Assessing the fit of the data to the correlated and uncorrelated SF-12 models

Country	Sample size	SF-12 version 1 CFA model—correlated				SF-12 version 1 CFA model—orthogonal			
		SB Chi-square[a]	df	RMSEA	SRMR	SB Chi-square[a]	df	RMSEA	SRMR
Canada (English)	15,615	1079.97	49	0.0367	0.0570	1599.55	50	0.0446	0.2945
Canada (French)	4352	343.53	49	0.0372	0.0525	501.42	50	0.0456	0.3285
Switzerland (German)	1731	75.22	49	0.0176	0.0535	112.38	50	0.0269	0.2970
Switzerland (French)	1637	121.41	49	0.0301	0.0651	158.44	50	0.0364	0.2209
Switzerland (Italian)	1339	126.99	49	0.0345	0.0596	177.27	50	0.0436	0.2968
Italy (Italian)	6362	830.19	49	0.0501	0.0776	1087.38	50	0.0571	0.3757
Norway (Bokmal)	5322	375.38	49	0.0354	0.0624	484.37	50	0.0404	0.2664
Bermuda (English)	2657	253.55	49	0.0397	0.0695	305.54	50	0.0439	0.2428
USA	3360	308.48	49	0.0397	0.0503	537.02	50	0.0539	0.3233
New Zealand	7122	684.65	49	0.0427	0.0657	883.85	50	0.0484	0.2595
Netherlands	5569	396.81	49	0.0357	0.0537	582.60	50	0.0436	0.3147
Hungary	5357	1087.45	49	0.0629	0.0515	1639.9517	50	0.07705	0.2955
Australia (ALLS)	8988	765.83	49	0.0404	0.0544	1177.67	50	0.0501	0.3123

[a] Satorra–Bentler scaled Chi-square

with the 1995 NHS survey, with the null hypothesis being no differences between the groups for the two Australian datasets.

Statistical analysis

In our research, we have promoted use of the theoretical structure of the SF-12 as set out by Ware et al. [18] with the exception that the correlation between physical and mental health be included in the model. As a first step in the analysis, we checked that this factor structure provided an adequate fit to the data for each country/language group using confirmatory factor analysis. We further assessed the fit of these models excluding the correlation between physical and mental health to assess the value of orthogonal solutions (Table 1).

As demonstrated by Ferero et al. [19], the preferred estimation method for ordinal manifest variables is unweighted least squares (ULS), or diagonally weighted least squares (DWLS) if ULS fails to converge. Using this estimation method constrains the usual preferred fit indices (Tucker–Lewis index—TLI and comparative fit index—CFI) to be near unity [20]. We therefore adjudged acceptable fit from a root mean square error of approximation (RMSEA) ≤ 0.06 and a standardized root mean square residual (SRMR) ≤ 0.08 as per Hu and Bentler's recommendations [21].

We then compared Australia to all of these country/language groups in multiple-group models to assess whether a model producing common factor score coefficients provided an adequate fit to the data. The formula for factor score weights is given by $\Phi\Lambda'_x\Sigma^{-1}$, where Φ is the covariance matrix of the common factors, Λ_x is the matrix of loadings and Σ is the model implied covariance matrix of the manifest variables [22]. Therefore, for two groups in a multiple-group model to produce equal factor score coefficients, all the estimated parameters of the groups must be identical. By fitting a multiple-group model with all parameters in both groups constrained to be equal, and an unrestricted multiple-group model with all parameters independently estimated, we were also able to derive Chi-squared tests for the differences between these models, to test the equality of the factor score coefficients across countries (Table 2) [23]. In conducting these comparisons, it must be borne in mind that the SRMR is a group goodness-of-fit measure, not a global goodness-of-fit measure. The SRMR applies separately to each group in the multiple-group model, whereas the RMSEA is a global goodness-of-fit measure and applies to the model as a whole. The SRMR quoted in Table 2 is the maximum of the two groups, as this best reflects the overall fit of the model according to this criterion.

Data were analyzed using LISREL V8.7 [24].

In LISREL, the Satorra–Bentler Chi-squared corrects for the non-normality of the data by applying a scaling factor to the normal theory weighted least squares (NTWLS) Chi-squared. Since we are using unweighted least squares estimation, there is no maximum likelihood Chi-squared for the models analyzed. The scaling factors are therefore applied to the NTWLS Chi-squared for the relevant models in computing the new scaled difference test set out by Satorra and Bentler [23]. The use of the NTWLS Chi-squared for the calculation of the new scaled difference test is entirely consistent with advice provided by Bryant and Satorra [25] who point out that LISREL users should use the NTWLS estimates rather than maximum likelihood (ML) estimates in calculating scaling factors.

Table 2 Assessing the invariance of model parameters between the country/language groups and Australia

Dataset	Compared to Australian dataset—common parameters				Separately estimated multiple-group models				Test of difference between models		
	SB Chi-square[a]	df	RMSEA	SRMR	SB Chi-square[a]	df	RMSEA	SRMR	Chi-square	df	Prob
Canada (English)	2026.87	127	0.0349	0.0876	1896.87	98	0.0386	0.0570	12.95	1	0.0003
Canada (French)	1354.08	127	0.0381	0.1637	1032.38	98	0.0378	0.0525	31.03	1	0.0000
Switzerland (German)	822.21	127	0.0320	0.1167	530.96	98	0.0287	0.0535	29.32	1	0.0000
Switzerland (French)	1106.12	127	0.0381	0.1853	772.61	98	0.0360	0.0651	46.97	1	0.0000
Switzerland (Italian)	1109.23	127	0.0387	0.1428	798.99	98	0.0372	0.0596	41.33	1	0.0000
Italy (Italian)	2616.76	127	0.0505	0.1156	1606.96	98	0.0448	0.0776	108.90	1	0.0000
Norway (Bokmal)	1298.66	127	0.0359	0.0839	1069.57	98	0.0372	0.0624	25.69	1	0.0000
Bermuda (English)	990.56	127	0.0342	0.1184	842.33	98	0.0361	0.0695	14.81	1	0.0001
USA	1277.83	127	0.0383	0.1273	1092.75	98	0.0406	0.0503	20.78	1	0.0000
New Zealand	1508.45	127	0.0368	0.0839	1449.23	98	0.0414	0.0657	7.43	1	0.0064
Netherlands	1398.95	127	0.0371	0.0908	1107.07	98	0.0376	0.0537	32.46	1	0.0000
Hungary	6448.18	127	0.0833	0.1618	1783.48	98	0.0490	0.0515	903.79	1	0.0000
Australia 1995 NHS	3648.3216	127	0.0459	0.1332	3087.3418	98	0.0480	0.0550	78.96	1	0.0000
Randomly divided halves of Australian ALLS data	845.3706	127	0.0355	0.0643	826.80	98	0.0407	0.0571	2.21	1	0.1370

[a] Satorra–Bentler scaled Chi-square

The proprietary scoring for SF-12 version 1 was based on regression analyses to predict the SF-36 PCS and MCS based on the SF-12 data items, which were treated as categorical (i.e., dummy variables were constructed for each level of the variables except one). The coefficients were then applied to the dummy variables, and the products summed to produce a score. In our previous publications, we have recommended scoring coefficients based on the CFA models we have fit. The scoring coefficients we published used the 1995 NHS, but for this paper we have produced scoring coefficients based on the ALLS datasets from the correlated models shown in Table 1 and used them to score the SF-12 PCS and MCS. These coefficients are applied to the recoded (if necessary) items of the SF-12 and summed to produce a score.

Results

In Table 1, we examine the fit of the SF-12 version 1 data to the theoretical model, including a correlation between physical and mental health, and the established orthogonal model, in all country/language groups provided in the ALL survey file, and the Australian National Health Survey dataset.

From Table 1, it is evident that the theoretical structure of the SF-12, when a correlation between physical and mental health is included, provides an adequate fit to the data for all country/language groups, with the possible exception of Hungary where the RMSEA of 0.0629 is marginally greater than 0.06, although the SRMR of 0.0515 is good. In addition, it is also evident that the SF-12 model, excluding the correlation between physical and mental health, does not provide an adequate fit to the data, since the SRMR measures for all country/language groups are much >0.08.

Table 2 provides the fit measures for the restricted and unrestricted multiple-group models comparing each country/language group to Australia in turn. The table also shows the new scaled difference Chi-squared tests [23] of the null hypotheses that each country has the same model parameters as Australia, and as a result, the same factor score coefficients for the SF-12 PCS and MCS.

If restricted models with common parameters are adequate, that indicates that common scoring coefficients for the two country/language groups should also be adequate.

From Table 2, it is evident that all of the unrestricted multiple-group models have an RMSEA ≤ 0.06 and an SRMR ≤ 0.08 and therefore provide an adequate fit to the data, including the Australia/Hungary model. In contrast, none of the multiple-group models, restricted to common parameters, have an SRMR ≤ 0.08, and thus none provide an adequate fit to the data.

The significance tests confirm that the constraints on parameter values produce significantly different models from the freely estimated two group models, as significance probabilities are all <0.007 using the new scaled difference Chi-squared tests [23].

As can be seen from the second last line of Table 2, when the Australian ALLS data were compared to the 1995 NHS survey data, the resulting comparisons were similar to those produced for the twelve country/language groups above. The

Table 3 Comparison of PCS and MCS summary scores calculated using different approaches

Country	Age		SF-12 PCS calculated using Australian coefficients	SF-12 PCS calculated using Hungarian coefficients	SF-12 PCS calculated using orthogonal USA coefficients	SF-12 MCS calculated using Australian coefficients	SF-12 MCS calculated using Hungarian coefficients	SF-12 MCS calculated using orthogonal USA coefficients
Australia	<30	Mean	52.73	53.99	52.94	51.43	52.54	51.30
		SD	7.15	6.38	6.31	8.64	7.46	8.79
	30–49	Mean	50.86	52.41	51.21	49.98	50.77	50.72
		SD	9.23	8.03	8.39	9.79	8.51	9.18
	50–69	Mean	46.74	48.85	46.25	48.84	49.65	51.98
		SD	11.84	10.29	11.39	11.09	9.74	9.63
	Total	Mean	50.00	51.66	50.04	50.00	50.88	51.30
		SD	10.00	8.72	9.45	10.00	8.75	9.24
Hungary	<30	Mean	53.67	55.02	53.73	52.91	54.46	52.48
		SD	5.69	6.12	5.09	8.34	7.69	7.66
	30–49	Mean	48.95	51.05	49.91	48.88	50.37	49.87
		SD	8.66	8.79	8.35	9.62	9.26	8.49
	50–69	Mean	42.27	44.10	42.81	44.74	45.52	48.34
		SD	10.72	11.32	10.58	11.15	10.86	9.33
	Total	Mean	48.19	50.00	48.78	48.71	50.00	50.11
		SD	9.74	10.00	9.43	10.28	10.00	8.69

restricted model was inadequate (RMSEA = 0.0459, SRMR = 0.1332), the unrestricted model provided an acceptable fit (RMSEA = 0.0480, SRMR = 0.0550), and the difference between the two models was statistically significant ($p < 0.0001$). The last line of Table 2 shows that when half the Australian ALLS data were compared to the other half, the restricted model was adequate (RMSEA = 0.0355, SRMR = 0.0591 and 0.0643 for each half), and the models produced a new Chi-squared difference test of 2.21 ($p = 0.1370$), indicating that there was no significant difference between the restricted and unrestricted models.

Researchers should also be curious about how much difference scoring using the US coefficients compared to coefficients derived from local data using our recommended approach makes. From Table 1, it can be seen that Hungary is the country/language group most different to Australia. Accordingly, we calculated scores based on country-specific scoring coefficients generated by the correlated models in Table 1, and also scores based on the original US scoring algorithms. The resulting scores for each approach are shown in Table 3 by country and age group.

We can see from Table 3 that the use of US coefficients results in differences up to roughly 0.49 in PCS and 3.14 in MCS scores for Australian subjects in the 50–69 age group, and differences up to 1.28 in PCS and 2.82 in MCS for the 50–69 age group for Hungarian people. The difference between Australian and Hungarian scores were up to 2.12 in the 50–69 age group for PCS and 1.55 in the <30 age group for MCS.

Discussion

This paper has examined the validity of using US scoring algorithms to make comparisons of health-related quality of life between countries. Original research promoting the ability to make such comparisons asked the question whether or not "a questionnaire (the SF-36) designed as a generic measure of health status in one country can be translated with comparable validity," and concluded that "results confirm the hypothesized relationships between SF-36 items and scales and justify their scoring in each country using standard algorithms" [14]. Comparative estimates of quality of life across countries would then be made using the PCS and the MCS. Our research using the SF-12, a subset of the SF-36, shows that the standard algorithms cannot be used globally and substantially challenges this notion for the SF-36 as a whole. As our data for the SF-12 indicate that country-specific coefficients should not be used to make cross-country comparisons, it is unlikely to be different for the SF-36.

The question of quality of life comparison between countries has been discussed for some time, and the major issue raised is how to assess equivalence between source and target populations. There is an international need for health indicators such as quality of life, which can be used to monitor health of populations for policy initiatives and cost-effectiveness studies [25–27]. Herdman et al. [27] argues it is necessary to show equivalence between translated versions of the same questionnaire and there is a growing literature on frameworks and guidelines for achieving this.

This literature deals with two major issues of equivalence. The first relates to conceptual differences, and the second is how to assess and make comparisons (measurement differences). A major problem exists in ensuring that adapted measures place the same individuals at the same point along a continuum of measurement across cultures. Beaton et al. [28] have also pointed out that subtle cultural differences do not only change the psychometric properties of an instrument but also change the statistical properties. Our research indicates we can have little confidence in the proposal that US coefficients be used to achieve this goal. We believe that further discussion and research is required on this issue of measurement and particularly whether or not we have confidence in the underlying continuum.

The US-derived coefficients for the summary components do not work for the USA because, as we have shown previously, the scores produced by coefficients obtained in orthogonal modeling are at variance with the underlying subscales [3, 4], so as summary measures of physical and mental health (as comprise the SF-12) they are inaccurate. Other researchers endorse this conclusion [6–11]. If the summary measures do not produce accurate estimates even for the USA, why then should they be promoted as a global solution? Essentially, the US-derived coefficients are flawed because the model does not include a correlation between mental and physical health.

A second question raised by the research, and also by Hawthorne et al. [5], is whether it is actually possible to make whole of country comparisons of quality of life with either US algorithms or the method promoted in our research? As discussed above, the US algorithms fail because they ignore the correlation between mental and physical health and are at variance with subscales. In our method, which acknowledges the correlation, each country's data are recalibrated to the same distribution, to have a mean of 50 and a standard deviation of 10. This effectively removes the ability to compare whole of country quality of life. What our method provides are PCS and MCS estimates that are based on country-specific coefficients, which more accurately capture the health and non-health factors driving the PCS and MCS estimates. They also provide the ability to compare population subgroups within and across countries, because the country-specific scoring coefficients yield summary scores with the same distribution in each country.

Even if the US estimates included a correlation between mental and physical health, promotion of the algorithms for use in other countries is still fraught with difficulty. In measuring quality of life, these estimates are driven by underlying health and non-health factors that are endemic to each country. We try to minimize the effect of these influences by customizing the questionnaire, as was done for Australia with some language changes [29]; however, we can never be sure that the underlying factors have been eliminated and we also know they will vary across countries through powerful influences such as economics, environment, education levels and others. All of the countries included in this research project are "Western" countries, and the argument for country (or "culture")-specific coefficients is likely to apply even more strongly in the case of, e.g., Asian, African, or Latin American countries. For example, the work of Liu et al. [30] also supports a country-specific approach for scoring SF-12. We can call these factors a country-specific bias. Given this country-specific bias, it makes more methodological sense to use country-specific coefficients, which are most likely to capture the best possible PCS and MCS scores for each country.

The research justifying the use of US-derived coefficients in other countries produced by the SF-36 development team addressed the international comparability question by comparing country-specific versus standard (US derived) scoring algorithms for the SF-36 physical and mental health summary measures [14]. The research concluded that because of a high degree of equivalence (i.e., correlation) observed within each country between country-specific and standard algorithms it was possible to use the standard (US derived) scoring algorithms in each country. However, this conclusion is disputed on the basis of the data shown in that report and we argue that a high correlation between country-specific and standard scoring is not sufficient to draw their conclusions, because differences between physical and mental health summary scales across the ten countries produced differences between three and six tenths of a standard deviation and are therefore clearly not equivalent. In addition, we seriously question the scoring methods used in the inter-country study, which were based on the production of scoring coefficients derived from a principle components extraction and an orthogonal rotation. The research conducted to justify the use of the US scoring algorithms showed a relationship existed between country-specific and US estimates in that the two second-order factors (mental and physical health) explained 82 % of the variance in eight underlying health scales in the USA and between 76 and 85 % in the scale data from the nine European countries. Given the orthogonal methods used, we conclude that these comparisons deal with accurate approximations of a poor-quality scale across countries.

Although we have used the best available techniques for analysis of these datasets, the result of comparing the 1995 Australian NHS data to the 2006 Australian ALLS data is unexpected. Although we concluded from Table 2 that common coefficients cannot be used across countries, we expected the comparison of two Australian datasets, using the same study methods, to be comparable across time.

Two major possible explanations emerge. First, this result may be explained by the analysis methods employed to adjudicate the significance of the differences between datasets. In particular, in a personal communication from Gerhard Mels of Scientific Software International, the vendors of LISREL, it was pointed out that although the properties of the test for the difference between models are well understood under maximum likelihood estimation, the behavior of the test under unweighted least squares estimation has not been studied. It is therefore possible that the results of the significance tests were pre-ordained by the method of analysis chosen, although the fact that the unrestricted model provides an adequate fit and the restricted model is inadequate supports the significance test result. Our analysis of the two randomly divided halves of the Australian ALLS dataset provides further support for the validity of the analysis procedures employed. Secondly, the differences between the two datasets may be explained by time lapse in the collection of the data which comprise more than a decade in which there could be significant changes in population quality of life. The ALLS data were collected in the second half of 2006, and the NHS data in 1995. Australian's life expectancy at birth changed between 1995 and 2006 from 75.6 to 78.7 years for males and from 81.3 to 83.5 years for females [31] and could be indicative of possible changes in quality of life over the period. If this argument is accepted, it has further implications and infers that researchers need to calculate their own scoring coefficients each time when SF-36 or SF-12 data are collected. This would require much more work in each research project and would also no doubt be opposed by the developers of the scales. We believe that this is an open question that requires more prospective or retrospective longitudinal research before an informed decision can be made.

In this research, we used the SF-12 version 1 which was available for the twelve country/language groups. No other multi-national high-quality dataset was available to us. Demonstrating the superiority of country-specific scoring coefficients in the version 1 data is a conclusion that is portable to any version of the SF-12 or SF-36 because the basic theoretical structure of the instrument is consistent in both version 1 and version 2. Further, the same scoring coefficients used for scoring the version 1 instrument are recommended by the developers for scoring version 2, and they were derived from orthogonal rotation.

Conclusion

We conclude from this study that researchers should derive their own country-specific scoring coefficients for physical and mental health summary scores if they require accurate data on which to base decisions regarding investment in health programs and the relative allocation of resources. Further research and discussion needs to take place on the issue of inter-country comparisons of health.

Compliance with ethical standards

Competing interests and funding This work was unfunded. The authors are unaware of any possible conflict of interest in the production of this publication.

Ethical standard This paper is based on a secondary analysis of various International and Australian survey files. As such, this analysis did not require formal ethics approval; however, all of the original data collections were conducted under ethics approval with the informed consent of the participants.

References

1. Wilson, D., Parsons, J., & Tucker, G. (2000). The SF-36 summary scales: Problems and solutions. *Sozial-und Präventivmedizin, 45*, 239–246.
2. Wilson, D., Tucker, G., & Chittleborough, C. (2002). Rethinking and rescoring the SF-12. *Sozial-und Präventivmedizin, 47*, 172–177.
3. Tucker, G., Adams, R., & Wilson, D. (2010). New Australian population scoring coefficients for the old version of the SF-36 & SF-12 health status questionnaires. *Quality of Life Research, 19*(7), 1069–1076.
4. Tucker, G. R., Adams, R. J., Wilson D.H. (2013) Observed agreement problems between sub-scales and summary components of the SF-36 Version 2—An alternative scoring method can correct the problem. *PLoS ONE*. 8(4): e61191.
5. Hawthorne, G., Osborne, R. H., Taylor, A., et al. (2007). The SF-36 Version 2: Critical analyses of weights, scoring algorithms and population norms. *Quality of Life Research, 16*(661), 73.
6. Simon, G. E., Revicki, D. A., Grothaus, L., et al. (1998). SF-36 summary scores. Are physical and mental health truly distinct. *Medical Care, 36*, 567–572.
7. Farrivar, S. S., Cunningham, W. E., & Hays, R. D. (2007). Correlated physical and mental health summary scores for the SF-36 and SF-12 health survey. *Health and Quality of Life Outcomes, 5*, 54.
8. Hann, M., & Reeves, D. (2008). The SF-36 summary scales are not accurately summarized by independent physical and mental component scores. *Quality of Life Research, 17*, 413–423.
9. Agnastopoulos, F., Niakis, D., & Tountas, Y. (2009). Comparison between exploratory factor analytic and SEM-based approaches to constructing SF-36 summary scores. *Quality of Life Research, 18*, 53–63.
10. Fleishman, J. A., Selim, A. J., & Kasiz, L. E. (2010). Deriving SF-12 v2 physical and mental health summary scores: A comparison of different scoring algorithms. *Quality of Life Research, 19*(2), 231–241.
11. Taft, C., Karlsson, J., & Sullivan, M. (2001). Do SF-36 summary scores accurately summarise subscale scores? *Quality of Life Research, 10*, 395–404.
12. Ware, J., & Kosinski, M. (2001). Interpreting SF-36 summary health measures: A response. *Quality of Life Research, 10*, 405–413.
13. Taft, C., Karlsson, J., & Sullivan, M. (2001). Reply to Drs Ware and Kosinski. *Quality of Life Research, 10*, 415–420.

14. Ware, J. E, Jr, Gandek, B., Kosinski, M., et al. (1998). The equivalence of SF-36 summary health scores estimated using standard and country-specific algorithms in 10 countries: Results from the IQOLA project. *Journal of Clinical Epidemiology, 51*(11), 1167–1170.
15. Stats Canada. (2011). The Adult Literacy and Life Skills Survey, 2003 and 2008 Public Use Microdata File User's Manual.
16. Australian Bureau of Statistics, Canberra. (2006). Adult Literacy and Life Skills Survey: User Guide, *Australian Bureau of Statistics*, Catalogue Number 4228.0.55.002.
17. Australian Bureau of Statistics. (1995). National Health Survey. SF-36 Population Norms Australia. Canberra: Australian Bureau of Statistics, Catalogue Number 4399.0.
18. Ware, J., Kosinski, M., & Keller, S. (1995). *SF-12: How to score the SF-12 physical and mental health summary scales* (2nd ed.). Boston: The Health Institute, New England MedicalCenter.
19. Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*, 625–641.
20. Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*, 548–570.
21. Hu, L., & Bentler, P. M. (1999). Cuttoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi:10.1080/10705519909540118.
22. Joreskog, K. G. (2000). *Latent variable scores and their uses*. ILScientific Software International: Lincolnwood.
23. Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference Chi square test statistic. *Psychometrika, 75*, 243–248.
24. Joreskog, K. G., & Sorbom, D. (1996). *LISREL user's reference guide*. Chicago, IL: Scientific Software International.
25. Bryant, F. B., & Satorra, A. (2012). Principles and Practice of Scaled Difference Chi Square Testing. *Structural Equation Modeling, 19*(3), 372–398.
26. Guilleman, E., Bombardier, L., & Beaton, D. (1993). Cross-cultural adaptation of health related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology, 46*(13), 1417–1432.
27. Herdman, M., Fox-Rushby, J., & Badia, X. (1997). Equivalence and the translation and adaptation of health related quality of life questionnaires. *Quality of Life Research, 6*(3), 4–237.
28. Beaton, D. E., Bombardier, L., Guilleman, F., et al. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3816–3891.
29. Sanson-Fisher, R. W., & Perkins, J. J. (1998). Adaptation and validation of the SF-36 Health Survey for use in Australia. *Journal of Clinical Epidemiology, 51*(11), 961–967.
30. Liu, C. J., Li, N. X., Ren, X. H., & Liu, D. P. (2010). Is traditional rural lifestyle a barrier for quality of life assessment? A case study using the Short Form 36 in a rural Chinese population. *Quality of Life Research, 19*(1), 31–36.
31. Life Expectancy Trends-Australia. Australian Social Trends, March (2011). *Australian Bureau of Statistics*. Catalogue 4102.0.

CHAPTER 7 - Discussion & recommendations

Chapter Content

- The importance of Quality of Life Measurement.
- Who is already using the results
- Areas for further research
- Recommendations and Fundamental Errors by the Developers.
- Research Conclusions.
- Summary

The importance of quality of life measurement.

Over the last century the major preoccupation with disease in the health services has moved from infectious disease, disease vectors, aetiologies and cure, to the impact of disease conditions on the activities of daily living as a consequence of the limitations of a range of chronic and degenerative conditions imposed on people's lives. Understanding this physical and psychological impact and burden and identifying ways to manage these conditions and their sequelae, often in the face of lifelong impact, is the focus of quality of life assessment and interpretation. Nearly three decades ago Fallowfield [66] described quality of life as the missing measurement in health, which focussed on the factors of disease a person has to live with.

As thinking about health and illness evolved from acute and infectious conditions to chronic age related and degenerative conditions, the major focus and concern of health planners moved to primary and secondary prevention of disease and disease conditions and the maintenance of functioning across a range of dimensions. These dimensions are characterised in the SF-36 and other quality of life instruments addressed in this thesis. The focus of health services, notwithstanding a continued interest in curative outcomes, is now largely on the minimisation of impairment, disability and handicap and optimum functioning. The importance of quality of life as a major disease concept extends across several research dimensions covering disease specific, generic and global aspects

of quality of life, outcomes for improved management, and economic outcomes and planned use of resources through cost utility analysis.

In the final analysis quality of life research now underpins how we function over time from the findings and translation of large-scale population studies down to clinical trials. Quality of life analysis also acknowledges the greatly increased complexity of modern health conditions, because of the nature of chronic and degenerative conditions, compared with infectious disease for which 'magic bullet' solutions are sought or are largely available.

Because of the importance of quality of life and the large scale studies occurring worldwide it is important that instruments such as the SF-36 'get it right' methodologically. The demonstration of valid summary scores for the SF-36 provided in this thesis is therefore a critical contribution at the very least allow others using the instrument to revisit their data and use the methods shown to re-analyse it. In addition the studies add caution to those planning future quality of life studies using the SF-36. The improvement in accuracy in past and future studies is important to decisions about the allocation of tax payers money to varying health priorities. The provision of valid summary scores for the SF-36 removes an impediment to its use.

Who is already using the results?

My first paper on Version 1 coefficients has been cited in 28 papers, 3 by us and 25 by other authors. Four were non-English language publications that are not cited in the bibliography [67-88].
My second paper on version 2 coefficients has been cited 5 times. Two papers are ours and one is unreadable/untraceable and not cited in the bibliography. The other two papers are cited [89,90].

Areas for further research

How right are Nye & Drasgow?

The CFA model used to produce scoring coefficients for Version2 of the SF-36 used representative population data from the 2004 SA Health Omnibus Survey. This survey had a sample size of 3014. To investigate the effects of sample size on the recommended fit cut-offs produced by Nye and Drasgow's formula, I re-ran the CFA using the same polychoric correlation matrix and its associated asymptotic covariance matrix, but with stated sample sizes of 400, 800, and 1200. The results are shown in the table below:

Table 5 – Nye & Drasgow cut off criteria based on different assumed sample sizes using the same input and fitted matrices.

					Cut off criteria			Cut off criteria		
n	Satorra Bentler ChiSq	df	q (no. of parms)	RMSEA	Nye & Drasgow	Hu & Bentler	SRMR	Nye & Drasgow	Hu & Bentler	AIC
400	615.588	551	79	0.01714	0.0273	0.06	0.07595	0.0606	0.08	773.588
800	1232.7188	551	79	0.03935	0.0213	0.06	0.07595	0.0496	0.08	1390.719
1200	1849.8496	551	79	0.04434	0.0178	0.06	0.07595	0.0409	0.08	2007.850
3014	4648.5378	551	79	0.04968	nan*	0.06	0.07595	nan	0.08	4806.538

- nan = not a number

These models all meet Hu & Bentler's criterion for adequate fit for maximum likelihood estimation [43]. The models were fit using diagonally weighted least squares, since unweighted least squares did not converge.

None of the models meet Nye & Drasgow's fit cut offs [40], although the model with sample size stated to be 400 meets the RMSEA criteria. Note that manipulating the sample size does not affect the SRMR. We are manipulating the sample size but using the same polychoric correlation matrices and fitted covariance matrices. Since by definition none of the residuals change, this explains why the SRMR does not change.

Clearly, the root mean square residual is not a function of sample size, although the Nye & Drasgow cut off for SRMR decreases as the sample size increases.

Note also that as the (notional) sample size increases the RMSEA increases and the Nye & Drasgow cut off for RMSEA decreases.

In this situation, where the same data matrices are analysed and the sample size is artificially manipulated, the RMSEA increases with sample size and one would expect the cut-off for adequate fit to do likewise. It appears that the cut-off for SRMR should not change at all since SRMR is not a function of sample size.

How important is SRMR when constructing scale measures?

The SRMR of the version 1 paper [46] was unacceptably high (0.2455), however the scores produced by the coefficients from that model work very effectively. This also has implications for our paper on international comparisons [49], given that the models with constant coefficients across countries were rejected based on SRMR, not RMSEA. It is, however, also noteworthy that these models were found to be significantly different to the models with unrestricted parameter estimates using the latest Chi-squared tests proposed in the literature.

Recommendations and Fundamental errors by the Developers

In Chapter 1 I set out what I believe are the problems with the scoring of SF-36 and SF-12. Here I re-iterate the problems and expand on the solutions proposed.

The developers have used an orthogonal decomposition of physical and mental health, and an orthogonal rotation of the solution, so that physical and mental health measures are not correlated. Whilst this approach may be mathematically attractive, it ignores the real life correlation that exists between physical and mental health. More accurate scores are obtained for physical and mental health components when these scores are correlated. This fact was recognized at the very early stages of the development of the SF-36 by Professor Ron Hayes, who promoted an oblique (correlated) solution in the RAND36 instrument [5], which as previously stated is identical to the SF-36 apart from the scoring algorithm.

In my publications I have shown that the use of orthogonal scoring algorithms produces component summary scores that conflict with the sub-scale scores [46,47]. This is supported by a number of other authors [5, 26, 28, 30, 33-37]. Scoring of the summary scores has been shown to work best when a correlated model for physical and mental health is employed.

The developers used an EFA of the eight sub-scale scores to produce factor score weights. An EFA of all 35 data items of the SF-36 would be expected to produce a superior result, but there is no guarantee that the solution of the EFA would have the same factor structure as the theoretical structure of the SF-36.

The sub-scale scores are adequate scales, but sub-optimal. They would be more accurate if the relevant items were combined using weights derived from a CFA or Item Response Theory (IRT) model rather than unit weighted scales. It is sensible therefore to expect a superior scale to be derived from the individual items of the SF-36 than the sub-scale scores. An EFA of the sub-scale scores or the relevant items would be improved by allowing for the real life correlation between physical and mental health in deriving scoring coefficients. An EFA is not guaranteed to produce the same factor structure as the theoretical structure of the SF-36, however a CFA uses the theoretical structure of the SF-36 and tests whether the data fits this structure. On this basis alone it can be seen that a CFA of the data items of the SF-36 provides the best method on which to base scoring algorithms.

The developers use unit weighted sub-scale scores rather than a weighted sub-scale score based on a CFA, which is inherently more accurate. The sub-scale scores are not continuous despite their appearance after they have been manipulated as specified in the scoring manual. They are still ordinal variables with a finite number of possible values. The developers have used Pearson correlations in the analysis of these data (in an EFA), whereas the nature of the data being ordinal infers that polychoric correlations are likely to produce a superior result.

As previously stated, the sub-scale scores are sub-optimal. They would be more accurate if the relevant items were combined using weights derived from a CFA or IRT model rather than unit weighted scales. Despite the manipulations called for

in the scoring manual, these sub-scale scores are not continuous or normally distributed variables, they have a finite number of possible values they can take, and are ordinal in nature. It is well established in the literature that ordinal items are best analysed using polychoric correlations (see e.g. Mindrilla [39]).

The developers have produced sub-scales with 3 items (role emotional) and 2 items (bodily pain, social functioning). To generate a weighted sub-scale score based on one factor, however, a congeneric CFA model requires four items, so there is a difficulty producing weighted scores for these sub-scales.

One possible improvement on the proprietary scoring of the sub-scales of the SF-36 would be achieved through the use of congeneric CFA models to produce weights for the combination of the relevant items to calculate sub-scale scores. Because congeneric CFA models require a minimum of four items to be identified, the structure of the SF-36 sub-scales precludes their use. The only other means of producing improved sub-scale scores flows from the fitting of the second order CFA for the full 35 items. Factor score coefficients are produced from this model that can be used to score the sub-scales, as well as the coefficients produced for scoring the summary scores. This approach does however have a drawback. Whilst the coefficients so produced would provide an accurate score for each sub-scale, there would be 35 coefficients for each sub-scale, and all 35 items would be used in the calculation of each sub-scale score. This is a conceptually unattractive method to calculate sub-scale scores.

Version 2 of the SF-36 included changes to some question wording, and the wording of some answer choices. These changes were motivated by the desire to minimise bias in completion of the instrument by subjects. There has been much work done on translation of the instrument into different languages, and adaptation of the questionnaire to different cultures, including adaptations to the Australian context [13]. These efforts to avoid bias have all been based on contextual issues and factors, but statistical bias has largely been ignored or rebutted as in Ware et al's response to the Taft criticism of the SF-36 [31].

Research Conclusions

The following conclusions are drawn from the research presented in this thesis.

- That the recommended scoring methods for both versions of the SF-36 and SF-12 are problematical.
- That statistical errors emerge in the recommended scoring methods
- That there is biological/psychological failure in the recommended method failing to allow for a correlation between mental and physical health.
- Given the extensive use of the SF-36 and SF-12 worldwide, spurious research conclusions may have been drawn.
- Those researchers worldwide can use structural equation modelling to produce scoring coefficients to re-analyse their data.
- That there is a danger in investing in decisions made on the basis of the recommended scoring methods.
- That international comparisons of SF-36 are not possible using US scoring coefficients.

Summary

In summary, I contend that scoring coefficients for the SF-36 component summary scores, and the SF-12 PCS and MCS scores, be they data from version 1 or version 2 of the questionnaires, are best based on confirmatory factor analyses rather than the original proprietary solution of an exploratory factor analysis approach involving an orthogonal rotation of a principal components extraction of factors. The failure of the developers to allow for a real world correlation between physical and mental health leads to summary scores that conflict with the sub-scales of the SF-36. I have demonstrated that my approach produces summary scores with superior measurement properties to the developers approach, and outlined the circumstances under which international comparisons are legitimate, whilst also demonstrating why the recommended scoring of international comparisons using the published US weights of the developers in not valid.

BIBLIOGRAPHY

1. Stewart A, Ware JE, (Eds). (1992) Measuring Functioning and Well-Being. The Medical Outcomes Study Approach. Durham and London. Duke University Press.
2. Geigle R, Jones SB. (1990) Outcomes measurement: a report from the front. Inquiry 27:7
3. Ware JE, Sherbourne CD. (1992) The MOS 36-Item Short Form Health Survey (SF-36); 1 Conceptual framework and item selection. Medical Care; 30: 473-83.
4. Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). The SF-36 health survey manual and interpretation guide. Boston, MA: The Health Institute, New England Medical Centre
5. Hays RD, SherbourneCD, Mazel RM. (1993) The Rand 36-Item Health Survey 1.0. Health Econ; 2(3): 217-27.
6. Ware JE, Kosinski M, & Keller SD. (1994) SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: The Health Institute
7. Quality Metric Incorporated. (2008). SF-36 v2TM and SF-12 v2 TM Health Surveys Offer Substantial Improvements. www.SF-36.org/commnity/SF36V2andSF12V2.shtml. Accessed June 20,2008.
8. Jenkinson C, Wright L, Coulter A. (1994) Criterion validity and reliability of the SF-36 in a population sample. Qual Life Res; 3(1): 7-12.
9. Sullivan M, Karlsson J, Ware JE Jnr. (1995) The Swedish SF-36 Health Survey. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. Social Science and Medicine; 41(10): 1349-58.
10. Hawthorne, G., Osborne, R. H., Taylor, A., & Sansoni, J. (2007). The SF-36 Version 2: Critical analysis of population weighting, scoring algorithms and population norms. Quality of Life Res, 16, 661–673.
11. McCallum, J. (1995). The new SF-36 health status measure: Australian validity tests. In a Paper presented to the Health Outcomes and Quality of Life Measurement Conference. Canberra: National Centre for Epidemiology and Population Health.

12. McCallum, J. (1995). The SF-36 in an Australian sample: Validating a new, generic health status measure. Australian Journal of Public Health, 19, 160–166.
13. Sanson-Fisher, R. W., & Perkins, J. J. (1998). Adaptation and validation of the SF-36 Health Survey for Use in Australia. Journal of Clinical Epidemiology, 51(11), 961–967.
14. Ware JE. (2000) SF-36 Health Survey update. Spine;24: 3130-39.
15. Aaronson NK, Acquadro c, Alonso J, Apolone G, et al. (1992) International Quality of Life Assessment (IQOLA) Project. Qual Life Res; 1(5): 349-51.
16. Goodwin RD (2003) Association between physical activity and mental disorders among adults in the United States. Prev Med 36(6): 698–703.
17. Strohle A, Hofler M, PfisterH, Muller AG, Hoyer J, et al (2007) Physical activity and prevalence and incidence of mental disorders in adolescents and young adults. Psychol Med (11): 1657–66.
18. Collingwood J.(2010) The relationship between mental and physical health. Psych Central.Available: http://psychcentral.com/lib/2010/the-relationship-betweenmental-and-physical-health..
19. Chapman DP, Perry GS, Strine TW (2005) The vital link between chronic disease and depressive disorders. Prev Chron Dis 2(1): 1–10.
20. Katon WJ (2003) Clinical and health services relationships between major depression, depressive symptoms, and general medical illness. Biol Psychiatry 54: 216–226.
21. Alonso J, Lepine J-P (2007) European Study of the Epidemiology of Mental Disorders/Mental Health Disability: A European Assessment in the Year 2000 Scientific Committee. J Clin Psychiatry (Suppl 2): 3–9
22. http://www.rcpsych.ac.uk/healthadvice/viewpoint/mentalphysicalhealth.aspx. Accessed 4/3/16
23. https://campaign.optum.com/optum-outcomes/what-we-do/health-surveys.html?gclid=CMrZxPPHrtACFdF9vQod3qIOlg Accessed 17/11/2016
24. Ware JE Jnr, Gandek B, et al. (1998) The Equivalence of SF-36 Summary Health Scores Estimated Using Standard and Country-Specific Algorithms in 10 Countries: Results from the IQOLA Project J of Clin Epid; 51: 1167-70

25. Alonso J, Ferrer M, Gandek B et al (2004) Health-related quality of life associated with chronic conditions in eight countries: Results from the International Quality of Life Assessment (IQOLA) Project Qual of Life Res; 132:283-98
26. Simon GE Revicki DA, Grothaus L, Vonkor M (1998) SF-36 summary scores. Are physical and mental health truly distinct. Med Care 36: 567–72.
27. Ware JE, Kosinski M, Bayliss MS, et al (1995) Comparison of Methods for the Scoring and Statistical Analysis of SF-36 Health Profile and Summary Measures: Summary of Results from the Medical Outcomes Study Med Care; 33:AS264
28. Wilson, D., Parsons, J., & Tucker, G. (2000). The SF-36 summary scales: Problems and solutions. Sozial- und Praventivmedizin, 45, 239–246.
29. Australian Bureau of Statistics. (1995). National Health Survey. SF-36 Population Norms Australia. Canberra: Australian Bureau of Statistics, Catalogue Number 4399.0.
30. Taft C, Karlsson J, Sullivan M (2001) Do SF-36 summary scores accurately summarise sub-scale scores? Qual Life, Res 10: 395–404.
31. Ware JE, Kosinski M (2001) Interpreting SF-36 summary health measures: A response. Qual Life Res 10: 405–413.
32. Taft C, Karlsson J, Sullivan M (2001) Reply to Drs Ware and Kosinski. Qual Life, Res 10: 415–420.
33. Nordvedt, M. W., Riise, T., Myhr, K. M., & Nyland, H. I. (2000). Performance of the SF-36, SF-12 and RAND SF36 summary scales in a multiple sclerosis population. Medical Care, 38, 1022–1028.
34. Farivar, S. S., Cunningham, W. E., & Hays, R. D. (2007). Correlated physical and mental health summary scores for the SF-36 and SF-12 health survey, V. 1. Health and Quality of Life Outcomes, 5: 54.
35. Hann M, Reeves D (2008) The SF-36 summary scales are not accurately summarized by independent physical and mental component scores. Qual Life Res 17: 413–23.
36. Agnastopoulos F, Niakis D, Tountas Y (2009) Comparison between exploratory factor analytic and SEM-based approaches to constructing SF-36 summary scores. Qual Life Res 18: 53–63.

37. Fleishman JA, Selim AJ, Kasiz LE (2010) Deriving SF-12 v2 physical and mental health summary scores: a comparison of different scoring algorithms. Qual Life Res 19(2): 231–41.
38. Rigdon EE, Ferguson CE (1991) The performance of the Polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. J Marketing Res 28: 491–97
39. Mindrilla D (2010) Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: a comparison of estimation bias with ordinal and multivariate non-normal data. International Journal of Digital Society (IJDS) 1: 60–66.
40. Nye CD, Drasgow F (2011) Assessing Goodness of Fit: Simple Rules of Thumb Simply Do Not Work. Org Res Methods 14: 548–570.
41. Flora DB, Curran PJ (2004) An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychol Methods 9: 466–91.
42. Forero CG, Maydeu-Olivares A, Gallardo-Pujol D (2009) Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. Struct Equ Modeling 16: 625–641.
43. Hu L, Bentler PM (1999) Cuttoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modelling 3: 424–53.
44. Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference Chi square test statistic. Psychometrika, 75, 243–248.
45. Bryant, F. B., & Satorra, A. (2012). Principles and Practice of Scaled Difference Chi Square Testing. Structural Equation Modeling, 19(3), 372–398.
46. Tucker, G., Adams, R., & Wilson, D. (2010). New Australian population scoring coefficients for the old version of the SF-36 and SF-12 health status questionnaires. Quality of Life Research, 19(7), 1069–1076.
47. Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2nd ed.). Hillsdale, NJ: Eribaum.
48. Tucker G, Adams R, Wilson D. (2014) Results from Several Population Studies Show That Recommended Scoring Methods of the SF-36 and the SF-12 May Lead to Incorrect Conclusions and Subsequent Health Decisions. Quality of Life Research 23:2195-2203
DOI: 10.1007/s11136-014-0669-9

49. Tucker G. R., Adams, R. J., & Wilson, D. H. (2013). Observed agreement problems between Sub-scales and summary components of the SF-36 version 2-an alternative scoring method can correct the problem. Plos One 8(4):e61191. doi: 10.1371/journal. pone.0061191. http://www.plosone.org/article/info%3Adoi%2F10. 1371%2Fjournal.pone.0061191.
50. Tucker G, Adams R, Wilson D (2016) The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36 Quality of Life Research 25(2), 267-274. DOI: 10.1007/s11136-015-1083-7
51. Sarah Louise Appleton (2010) AN EPIDEMIOLOGICAL INVESTIGATION OF THE ROLE OF PHENOTYPE IN THE ASSOCIATION OF OBESITY AND ASTHMA PhD thesis.
52. Statistical course notes from North Carolina State University – Structural Equation Modelling http://faculty.chass.ncsu.edu/garson/PA765/structur.htm
53. Moriarty D. G., Zach M. M., Kobau R. (2003) The Centers for Disease Control and Prevention Health Days Measures – Population Tracking and Perceived Physical and Mental Health Over Time. Health and Quality of Life Outcomes; 1-37
54. Erickson P. (1998) Evaluation of a population based measure of quality of life: the health and activity limitation index (HALex). Qual Life Res 7:101-114
55. Bonomi AE, Patrick DL, Bshnell DM, Martin M. (2000) Validation of the United States version of the World Health Organisation Quality of Life (WHQOL) instrument. Journal of Clinical Epidemiology 53(1): 1-12.
56. Burckhardt CS, Anderson KL. (2003) The Quality of Life Scale (QOLS): reliability, validity and utilization. Health and Quality of Life Outcomes 23:1-60. DOI: 10.1186/1477-7525-1-60 http://hqlo.biomedcentral.com/articles/10.1186/1477-7525-1-60
57. Kaplan RM, Anderson JP. (1988) A general health policy model: update and applications. Health Services Research 23(2): 203-35
58. Seiber WJ, Groessl EJ, David KM, Ganiatsw TG, Kaplan RM. (2008) Quality of Well-being Self-Administered (QWB-SA) Scale. Users Manual. Sand Diego. University of California.

59. Horsman J, Furlong W, Feeny D, Torraqnce G. (2003) The Health Utilities Index (HUI) concepts, measurement properties and applications. Health & Qual Life Outcomes PMC293474
60. Hawthorne G, Richardson J, Osborne R. (1999) The assessment of quality of life (AQoL) instrument: a psychometric measure of health-related quality of life. Qual Life Res 8(3): 209-24.
61. Whynes, David K.; TOMBOLA Group (2008-01-01). "Correspondence between EQ-5D health state classifications and EQ VAS scores". Health and Quality of Life Outcomes 6: 94.doi:10.1186/1477-7525-6-94. ISSN 1477-7525. PMC 2588564. PMID 18992139.
62. Shaw, James W.; Johnson, Jeffrey A.; Coons, Stephen Joel (2005-03-01). "US valuation of the EQ-5D health states: development and testing of the D1 valuation model". Medical Care 43(3): 203–220. ISSN 0025-7079. PMID 15725977.
63. Reenen, Mandy van (April 2015). "EQ-5D-5L User Guide" (PDF). EQ-5D. EuroQol Research Foundation.
64. Balestroni G, Bertolotti G. (2012) EuroQol 5D (EQ-5D): an instrument for measuring quality of life. Monaldi Arch Chest Dis 78(3): 155-9.
65. Wilson D, Tucker G, Chittleborough C. (2002) Rethinking and rescoring the SF-12. Sozial- und Praventivmedizin, 47, 172-177
66. Fallowfield C (1990) Quality of life: The Missing Measurement in HealthCare. Souvenir Press
67. Tavella, Rosanna, et al. "Using the Short Form-36 mental summary score as an indicator of depressive symptoms in patients with coronary heart disease."Quality of Life Research 19.8 (2010): 1105-1113.
68. EG Eakin, MM Reeves, AL Marshall (2010) Living Well with Diabetes: a randomized controlled trial of a telephone-delivered intervention for maintenance of weight loss, physical activity and glycaemic … BMC public - bmcpublichealth.biomedcentral.com
69. DJ Beales, AJ Smith, PB O'Sullivan (2012) Low back pain and comorbidity clusters at 17 years of age: a cross-sectional examination of health-related quality of life and specific low back pain impacts Journal of Adolescent …, - Elsevier
70. PB O'Sullivan, DJ Beales, AJ Smith…Low back pain in 17 year olds has substantial impact and represents an important public health disorder: a cross-sectional study. BMC public - biomedcentral.com

71. LJ Bartsch, P Butterworth, JE Byles, P Mitchell…(2011) Examining the SF-36 in an older population: analysis of data and presentation of Australian adult reference scores from the Dynamic Analyses to Optimise Ageing (… Quality of Life … - Springer
72. J Buckley, G Tucker, G Hugo, G Wittert…(2013) The Australian baby boomer population—factors influencing changes to health-related quality of life over time. Journal of aging and Health…jah.sagepub.com
73. S Coulson, P Vecchio, H Gramotnev, L Vitetta (2012) Green-lipped mussel (Perna canaliculus) extract efficacy in knee osteoarthritis and improvement in gastrointestinal dysfunction: a pilot study. Inflammopharmacology, Springer
74. TV McCann, DI Lubman, SM Cotton, B Murphy (2012) A randomized controlled trial of bibliotherapy for carers of young people with first-episode psychosis. Schizophrenia … - MPRC
75. S Coulson, H Butt, P Vecchio, H Gramotnev…(2013) Green-lipped mussel extract (Perna canaliculus) and glucosamine sulphate in patients with knee osteoarthritis: therapeutic efficacy and effects on gastrointestinal … Inflammopharmacol – Springer DOI 10.1007/s10787-012-0146-4
76. KR Holt, PL Noone, K Short, CR Elley…(2011) Fall risk profile and quality-of-life status of older chiropractic patients- Journal of manipulative …, Elsevier
77. A Wheeler, G Schrader, G Tucker, R Adams…(2013) Prevalence of depression in patients with chest pain and non-obstructive coronary artery disease. The American journal of …,- Elsevier
78. KAH Aljurany (2013) Personality characteristics, trauma and symptoms of PTSD: a population study in Iraq. ros.hw.ac.uk
79. LN Christensen (2010) The Effect of Physical Activity on Health. Masters Thesis - projekter.aau.dk
80. A Wheeler, L Denson, C Neil, G Tucker…(2014) Investigating the Effect of Mindfulness Training on Heart Rate Variability in Mental Health Outpatients: A Pilot Study. Behaviour …, - Cambridge Univ Press
81. A Talwar, S Sahni, EJ Kim, S Verma…(2015) Dyspnea, depression and health related quality of life in pulmonary arterial hypertension patients. of exercise … - ncbi.nlm.nih.gov

82. N Chehelamirani, R Sahaf…(2016) Validity and Reliability of WHOQOL-DIS Questionnaire in Iranian Older People with Disability. Journal of Rehabilitation… - rehabilitationj.uswr.ac.ir
83. R Tavella, N Cutri, G Tucker, R Adams, J Spertus, JF Beltrame (2016) Natural history of patients with insignificant coronary artery disease (ICAD). European Heart Journal:-Quality of Care and Clinical Outcomes… - ehjqcco.oxfordjournals.org (In Press)
84. A Peisker, GF Raschke, A Guentsch (2016) Longterm quality of life after oncologic surgery and microvascular free flap reconstruction in patients with oral squamous cell carcinoma. patologia oral y … - researchgate.net
85. GC Rhys (2015) Spiritual discussion, offering successful communication of the patient's ideas to the clinician is not a routine component of medical consultations in the United Kingdom. Primary Health Care: Open Access - omicsgroup.org
86. AN Roy, S Madhavan (2014) Patient Reported Health-related Quality of Life in Co-morbid Insomnia: Results from a Survey of Primary Care Patients in the United States . Primary Health Care: Open Access - omicsgroup.org
87. J Jagnoor, A De Wolf, M Nicholas, CG Maher…(2015) Restriction in functioning and quality of life is common in people 2 months after compensable motor vehicle crashes: prospective cohort study. Injury … - biomedcentral.com
88. AA Jordbru, LM Smedstad, O Klungsøyr…(2014) Psychogenic gait disorder: a randomized controlled trial of physical rehabilitation with one-year follow-up. Rehabil … - medicaljournals.se
89. J Twiss, S McKenna, L Ganderton…(2013) Psychometric performance of the CAMPHOR and SF-36 in pulmonary hypertension. BMC pulmonary … - biomedcentral.com
90. A Kelly, J Rush, E Shafonsky…(2015) Detecting short-term change and variation in health-related quality of life: within-and between-person factor structure of the SF-36 health survey. Health and quality of life outcomes…, - hqlo.biomedcentral.com

APPENDIX

Soz.- Präventivmed. 45 (2000) 239–246
0303-8408/00/060239-08 $ 1.50 + 0.20/0

David Wilson, Jacqueline Parsons, Graeme Tucker

Department of Human Services, Adelaide

The SF-36 summary scales: Problems and solutions

Summary

To determine the accuracy of the SF-36 summary mental and physical health scales in reflecting their underlying subscales using the traditional method of scoring based on factor coefficients derived through principle components analysis and orthogonal rotation. A representative Australian population survey containing the SF-36 was used to obtain factor coefficients from principle components analysis and orthogonal rotation for scoring the physical component summary (PCS) and the mental component summary (MCS) of the SF-36 in the traditional way. In addition two other methods were used to produce coefficients. The first method used maximum likelihood extraction and oblique rotation. The second method fit a structural equation model to the data in a confirmatory factor analysis. The coefficients derived by each of the methods were applied to the data of a second representative population survey. This survey also provided data on physical and mental health status which allowed comparison of the summary scores and underlying subscales according to various health states. Neither of the scoring methods based on the exploratory factor analyses methods (orthogonal and oblique) produced summary scale scores, by age group, that adequately reflected the underlying subscales. When coefficients derived using structural equation modeling were fit to the data in a confirmatory factor analysis the MCS and PCS accurately reflected their underlying subscale scores. They also produced MCS and PCS scores for the various health states as would be expected from the underlying subscales. The traditional methods of scoring the SF-36 summary scales produce results that would not be expected from the underlying subscales. The problem was only corrected by fitting a structural equation model to the data in a confirmatory factor analysis. The results advise caution in the use of the SF-36 summary scales and suggests that alternative methods of developing factor coefficients need to be employed in studies using the SF-36 summary scales.

The short form SF-36 health related quality-of-life questionnaire has been widely accepted as a generic summary of health status[1–3] and as an investigative tool in health assessment or monitoring[4–5]. Despite this a study by Simon et al. in 1998 recommended caution in the interpretation of the mental component summary (MCS) and the physical component summary (PCS) when the condition or treatment of interest had strong effects on scales with negative scoring coefficients[7]. The original subscales of the SF-36 were aggregated into the physical and mental summary components using factor analysis of subscale scores from a United States population sample[8]. Principal components extraction and orthogonal rotation produced factor score coefficients, which were then used to compute the summary scores. The findings of Simon's study showed that the baseline physical component summary, in a prospective study of patients initiating antidepressant treatment, indicated no impairment based on a population norm of 50, although patients had reported modest impairment on the physical functioning, role-physical, bodily pain and general health perceptions, all of which contribute to the physical component summary in the

hypothesized structure. In the three-month follow up stage of this study the four subscales showed moderate but statistically significant improvements, however the physical component summary was unchanged. The reason for this anomaly was attributed to the assumptions and methods used in computing the summary scores based on orthogonal factor rotation and the generation of negative scoring coefficients. In the algorithms used to score the summary components the mental health and role-emotional scales make modest negative scoring contributions to the PCS. In addition, the physical functioning, role physical and bodily pain subscales also make modest negative contributions to the MCS. The authors concluded that these negative scoring coefficients could produce unexpected results on either summary component scale.

In this study we looked again at the effects of scoring coefficients on the summary scales, how they compared with their underlying subscales and verified Simon's findings in the 1998 representative population South Australian Health Omnibus Survey (SAHOS). We then investigated other approaches of deriving scoring coefficients from the 1995 Australian National Health Survey, using i) maximum likelihood extraction and oblique rotation, and ii) structural equation modeling rather than Ware's[8] original method. These coefficients were then used to score the 1998 South Australian Health Omnibus Survey (SAHOS) and assess face validity of summary scores in comparison to subscale scores for a number of age groups (15–29 years, 30–49 years. 50–59 years and 70+ years). The SAHOS also provided information on health status, which allowed us to compare the subscales and summary scores according to physical, and mental health states reported. These summary scores were again calculated using i) principle components analysis and orthogonal rotation, ii) maximum likelihood extraction and oblique rotation, and, iii) structural equation modeling. These analyses provided an assessment of the external validity of the subscales and summary scales using the different methods of obtaining coefficients[9].

Method

Two population data sources were used for this study. The first was the 1995 Australian National Health Survey (NHS)[10] which was the second in a series of five yearly population surveys designed to obtain national benchmark information on a range of health issues. The sampling method of the NHS is a self-weighting multistage clustered area sample based on Australian Bureau of Statistics (ABS) collector districts in which households are selected with equal probability. In this survey n = 23800 households were selected and all adults aged 15 years or older were interviewed. A subset of n = 19785 were asked to complete the SF-36 quality-of-life questionnaire. Of those interviewed, n = 18492 provided some data on the SF-36 and were included in these analyses. These data were used by the ABS to calculate subscale scores using principle components analysis and orthogonal factor rotation after the method of Ware et al.[11] and produce coefficients that could be used to calculate summary scores for Australia[10]. Using the NHS data set two additional methods were used in the present study to compute factor coefficients for scoring the PCS and MCS which were then checked using the 1998 SAHOS. The factor coefficients were developed on the basis of the following logic.

First, exploratory factor analysis using maximum likelihood extraction and oblique (oblimin) rotation. Maximum likelihood extraction estimates the hypothetical factor structure, which assumes that one or more latent factors account for the correlations observed between manifest variables. The oblique rotation allows for the real world fact that physical and mental health is correlated, by allowing for this to be reflected in the rotation.

An obvious alternative to an exploratory factor analysis is to use a structural equation modelling approach to fit a confirmatory factor analysis to the implied structural model. This was undertaken using the eight subscale scores as manifest variables. Measurement error was incorporated into the model based on the reliability of the subscale measures (using Cronbach's alpha for the NHS data set as the reliability measure), by constricting the error variances of the manifest variables representing the subscale scores. This model provided a poor fit to the NHS data (RSMEA = 0.55) and the approach of calculating scores based on the subscales was therefore rejected.

A hypothetical factor structure has already been documented for the SF-36 (see Figure 1). It was therefore possible to fit a structural equation model to the data in a confirmatory factor analysis (as above). The model fit was the full measurement model, using items re-coded as detailed in the SF-36 scoring manual. As this model is based on the hypothetical factor structure, coefficients are only produced for the relevant variables related to each summary score. The other paths do not appear in the model and are therefore not estimated.

The above model was fit on the covariance matrix of 18141 observations with no missing data for all eight SF-36 subscales following imputation of data items by mean substitution, where more than half the data items in a subscale were not missing, as set out in the SF-36 scoring manual.

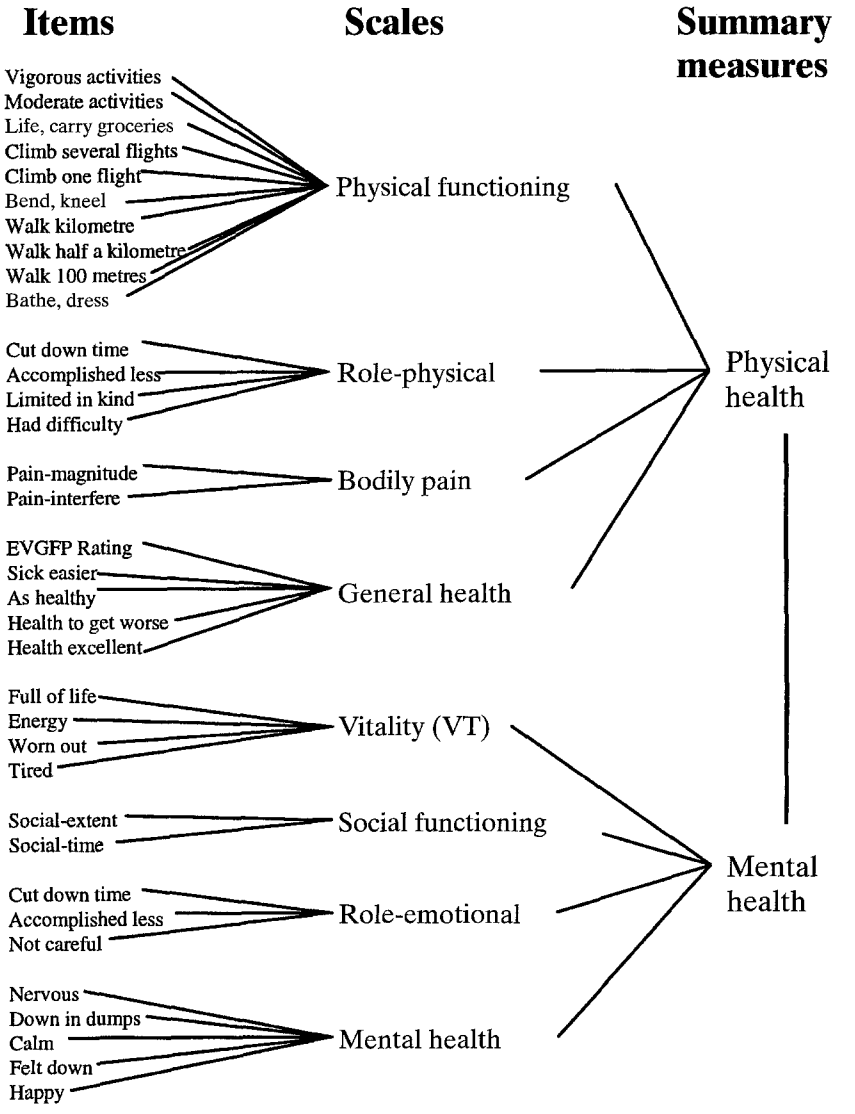


Figure 1. *Hypothesized structure of SF-36 health dimensions and the summary mental (MCS) and physical (PCS) health measures.*

Amos includes an option for handling missing data without imputation known as full information maximum likelihood (FIML) estimation. The full measurement model was also fit to the complete NHS data file of 18492 complete or partial returns, with no imputation of missing values but using FIML estimation. This model does not provide fit measures, but does provide parameters. When these parameters were compared to the model above (using imputation by mean substitution), the parameters were very similar with minor differences appearing generally at the third decimal place of the parameter. It was therefore concluded that this analysis was not biased by any problems relating to missing data and it was therefore accepted as the preferred approach.

All coefficients derived from the NHS data set were then used to score the summary scales of the SF-36 data obtained in the independent 1998 SAHOS (n = 3001, 70% response rate). The SAHOS is a statewide South Australian survey designed to obtain state benchmark information on a range of health issues. The SAHOS is also a self-weighting multistage clustered area sample selected from ABS collector districts of people aged 15 years or more who live in metropolitan Adelaide or country towns with populations over 1000. The survey is conducted annually and the method has been extensively published[3]. Households are selected with equal probability of selection within each collector district and then one adult in each household, aged 15 years or older is selected for interview according to the most recent birthday.

In addition to the SF-36, a range of other health and demographic questions were asked in the 1998 SAHOS. The survey has been running since 1990, and response rates average 73% with approximately 3000 people interviewed. In the 1998 survey the response rate was 70%. Data from the survey were weighted so they accurately represented the age, sex, household size, and geographic area of the South Australian population. The SF-36 data collected in the survey of 1998 were used to test the results of the different extraction and rotation methods.

SAHOS respondents were also asked whether or not in the previous twelve months they had used any medication for a chronic physical condition: that is one that has lasted for, or is likely to last for, six months or more. They were also asked the same question for depression (medication such as tranquilisers or anti-depressants) or a diagnosed mental illness (such as schizophrenia). From these data the study population were divided into four groups according to no medication, physical health medication only, mental health medication only or both physical and mental health medication. These implied health groups were then used to assess the external validity of the subscales and the PCS and

MCS. Mean scores for the SF-36 subscales and the PCS and MCS summary scales were calculated for each of the four health groups and compared with the overall subscale means to assess whether or not the scoring of the SF-36 subscale and summary dimension scores were consistent with the different health states as would be predicted. The external validity of the PCS and MCS subscale scores of the SAHOS data were assessed using the three methods identified above.

Data were analysed using SPSS version 9 and AMOS version 3.6.2.[12]

Results

The data in Table 1 illustrates the problem of using orthogonal or oblique rotation methods to calculate the PCS and MCS subscales. It can be seen from Table 1 that the orthogonal method produces some age-specific MCS scores that vary around the overall MCS score. This in itself is not a problem. It does, however, become a problem when the age specific MCS is higher than would be expected given the underlying subscale scores. Scores for three subscales that make up the orthogonal MCS are significantly lower for the over 70 age group (vitality, social functioning and role emotional) than for the 15–29 and 30–49 age group, yet the orthogonal MCS score for the over 70 age group is significantly higher (Tukey's honestly significant difference (HSD) <0.001) than that for the younger age groups. In the same way the oblique method produced MCS summary scores that are not significantly different across all age groups despite significantly lower subscales (vitality, social functioning

Dimension	Age group (years)				
	15–29	30–49	50–69	70+	Overall population
Physical functioning	91.2	87.5	77.6	61.4	83.0
Role physical	87.6	82.7	74.6	63.8	79.8
Bodily pain	81.0	77.7	72.2	71.0	76.5
General health	76.4	76.8	71.2	64.7	73.9
Vitality	67.3	64.5	63.7	58.9	64.3
Social functioning	88.6	88.5	88.2	84.0	87.9
Role emotional	88.3	87.1	89.0	86.5	87.8
Mental health	80.1	79.0	80.7	81.6	80.0
Summary PCS (ORTHOGONAL extraction and rotation: using ABS weights based on NHS)	52.5	51.0	46.4	41.1	49.1
Summary MCS (ORTHOGONAL extraction and rotation: using ABS weights based on NHS)	51.4	51.3	53.3	54.2	52.1
Summary PCS (maximum likelihood extraction, OBLIQUE rotation)	52.4	51.2	48.5	44.7	50.1
Summary MCS (maximum likelihood extraction, OBLIQUE rotation)	47.7	48.4	48.0	48.4	48.1
Summary PCS (using regression coefficients from STRUCTURAL EQUATION MODELING)	52.8	51.8	48.3	44.4	50.3
Summary MCS (using regression coefficients from STRUCTURAL EQUATION MODELING)	52.1	51.1	51.3	49.9	51.2

Table 1. *SF-36 dimension and summary scores for adult age groups using various summary approaches.*

and role emotional) for the over 70 age group.

These results are due to the mathematics that converts the subscales to summary scores. Each subscale score is converted to a Z score before being multiplied by the factor score coefficient. The summary score is computed by adding the Z score, multiplied by the coefficient, for each subscale (not only the four that make up the dimension). This creates a problem when a negative Z score is multiplied by a negative coefficient, as it results in a positive number inflating the opposing scale. In this case the scores on the physical subscales for people over 70 years were lower than the population norm, resulting in a negative Z score. This was then multiplied by the negative coefficient and summed to give an artificially high MCS using the orthogonal method. Table 1 also shows the MCS and PCS scores calculated using structural equation modeling. The model used to derive the coefficients for the PCS and MCS scores was found to provide an adequate fit to the data (RMSEA 95% CI = 0.068, 0.069) based on 18141 input records. Most fit indices were about 0.82. It can be seen from Table 1 that both the PCS and MCS summary scales, using structural equation modeling, produce scores that are consistent with their underlying subscales as hypothesised. The over 70

Dimension	MEDICATION TAKEN				
	No medication	Physical medication only	Mental medication only	Both physical & mental	Overall population
Physical functioning	88.7	68.4	79.0	56.6	83.0
Role physical	86.9	63.3	69.0	39.1	79.8
Bodily pain	81.2	65.2	71.6	49.6	76.5
General health	79.2	61.8	66.1	41.7	73.9
Vitality	67.6	58.3	49.2	41.0	64.3
Social functioning	90.9	83.0	77.1	58.8	87.9
Role emotional	90.6	85.5	63.0	56.5	87.8
Mental health	81.9	78.6	61.3	59.2	80.0
Summary PCS (ORTHOGONAL extraction and rotation: using ABS weights based on NHS)	51.8	41.9	50.0	37.5	49.1
Summary MCS (ORTHOGONAL extraction and rotation: using ABS weights based on NHS)	52.8	52.5	41.5	40.9	52.1
Summary PCS (maximum likelihood extraction, OBLIQUE rotation)	52.5	44.7	46.4	36.0	50.1
Summary MCS (maximum likelihood extraction, OBLIQUE rotation)	46.7	49.9	58.2	61.6	48.1
Summary PCS (using regression coefficients from STRUCTURAL EQUATION MODELING)	53.2	43.6	47.1	34.6	50.3
Summary MCS (using regression coefficients from STRUCTURAL EQUATION MODELING)	52.8	48.9	41.6	37.3	51.2

Table 2. *Subscale and summary dimension scores for four health conditions.*

age group produced an MCS score that was significantly lower than the youngest age group (Tukey HSD = 0.003) and as would be expected given the underlying subscales. The 30–49 and 50–69 age groups produced scores differing in the right direction according to their subscales, but did not achieve statistical significance. The PCS using structural equation modeling produced PCS scores as would be expected given the underlying subscales.

Table 2 shows the subscale and summary scores for the four groups with varying medical conditions according to the medication used. The PCS and MCS summary scores have again been computed using coefficients obtained by all three methods reported above. Using the logic of "known groups" validity according to Ware et al.[11] it can be seen from Table 2 that there are external validity problems with the exploratory factor analysis methods. The orthogonal model produced a higher PCS score for people taking mental health medication (Tukey HSD = 0.05) and a higher MCS score for those taking physical health medication (Tukey HSD = 0.05) when compared to the overall scores, despite lower subscale scores than the overall scores. The oblique method produced the lowest MCS score for those taking no medication and the highest MCS scores those taking both physical and mental medications. Thus the ranking was reversed from that expected. The structural equation coefficients again behaved appropriately producing higher and lower summary scores than overall where expected.

In assessing external validity, the PCS and MCS means are ordered as would be expected using the structural equation coefficients. The group reporting using no physical or mental health medication had higher MCS and PCS scores than overall. The group taking physical health medication had a lower PCS score than the overall score and also lower than the score for the group reporting mental health medication only. This latter group, in turn, had a lower MCS score than overall and those reporting physical medication only.

Discussion

The first question in this study was to identify whether or not the findings of Simon et al.[7] were verified in producing summary scores for the SF-36 that did not reflect their underlying health subscales. We would conclude from the population survey data used in this study that Simon was correct in identifying problems with the summary scales. While Simon's problem related to the PCS in assessing change in health status over time this study found problems with the MCS when comparing summary scores across age groups. We would, however, express the caution that given the large sample size in this study statistically significant differences are relatively easy to achieve. This means that we must also be careful not to over interpret small but significant differences between groups.

The second question asked was whether an alternative approach to deriving the scoring coefficients would produce summary scores with improved face validity. Structural equation modeling to derive coefficients used in producing the summary PCS and MCS scores has produced scores that are consistent with underlying subscale scores across age groups. On face value this has corrected the problems that were identified by Simon et al.[7]. The fact that structural equation modeling uses only those dimensions from the hypothesized structure that make up the PCS and the MCS also makes logical sense. When a person responds to a question related to any subscale they have the opportunity to respond to both physical and mental health issues in a way they judge to reflect their current health status. If they have a physical health problem, which also affects their mental health, or vice versa, they have the opportunity to identify the level of this effect when they answer each of the mental health questions for each subscale. The impact of mental health issues on physical health, is therefore already accounted for in the individuals direct answer to each physical health question, without accounting for it again by virtue of the statistical analysis method used (factor analysis and orthogonal rotation) which includes a weight for the MCS subscales. Similarly, the impact of physical health is reflected in the physical health questions without adding a weight for the MCS subscales. It does not make logical sense that there is a weight from subscales making up the PCS included in the MCS, or vice versa. This would appear to be double counting if the respondent has already accurately answered questions that contribute to the subscales, as occurs in the exploratory factor analysis methods. As a result, and as Simon et al.[7] point out, it produces incorrect results that are based on the factor methods used.

McCallum[13,14] in his earlier validation of the SF-36 for Australia, has suggested that structural equation modeling to calculate summary scores is inappropriate because the SF-36 items are highly inter-correlated. In his analyses the raw data did not produce a positive definite matrix but rather a singular variance/covariance matrix. In the present study, where a larger and more representative data set was used the matrix was non-singular. It should also be pointed out that McCallum did not have access to the NHS data when he conducted his validation studies.

The data shown in this study would also lead us to agree with Ware's

conclusion that oblique rotation is not an alternative scoring strategy[11]. It would appear initially that Ware did not consider the further alternative of structural equation modeling, even though he has created a hypothesized structure that would suggest such an approach. In more recent work, however, structural equation modeling was used in assessing the construct validity of the SF-36 across ten countries[15]. In this study the authors concluded that the study results confirmed the hypothesized relationships between the SF-36 items and scales.

The ten-country study mentioned above also raised the possibility that summary scale scores may vary across different cultural groups. In this study non-significant differences were observed in the PCS and MCS with greatest variation occurring in the MCS[15]. In another study conducted in Japan by some of the same investigators caution was expressed regarding some of the scales that comprise the MCS[16].

The third question asked in the present study was whether or not the PCS and MCS summary scores truly measured health in varying and understandable medical health states. Four mutually exclusive groups known to differ in the type and severity of self-reported physical and mental conditions (no conditions, physical health conditions, mental health conditions and both physical and mental conditions) were associated with the PCS and MCS as predicted according to type and severity of condition[11]. The summary scores which best represented the underlying subscales were produced from structural equation modeling.

The conclusion we would draw is that structural equation modeling corrects the effects of negative scoring coefficients produced in scoring methods based on exploratory factor analyses. It should be stated, however, that this is the first Australian assessment using structural equation modeling and further population studies should be conducted to assess reliability of the method over time.

Zusammenfassung

Die Summenskalen des SF-36: Probleme und Lösungen

Die zur Berechnung der Summenskalen der physischen (PCS) und psychischen Dimension (MCS) des SF-36 nötigen Faktorwerte wurden in einer repräsentativen australischen Bevölkerungsbefragung nach dem Standardverfahren durch eine Hauptkomponentenanalyse mit orthogonaler Rotation ermittelt. Zusätzlich wurden zwei weitere Verfahren zur Berechnung der Koeffizienten angewendet: eine Faktorenextraktion nach Maximum-Likelihood mit anschliessender schiefwinkliger Rotation und die Anpassung eines Strukturgleichungsmodells an die Daten in einer konfirmatorischen Faktoranalyse. Die so berechneten Faktorwerte wurden in einer zweiten repräsentativen Bevölkerungsbefragung verwendet. In dieser Erhebung wurden zusätzlich verschiedene Masse zur physischen und psychischen Gesundheit erhoben, die einen Vergleich der Summenskalen und der zugrunde liegenden Subskalen in Gruppen mit unterschiedlichem Gesundheitsstatus erlaubt. Keine der auf Basis explorativer Faktoranalysen (orthogonale oder schiefwinklige Rotation) berechneten Summenskalen bildet die zugrunde liegenden Subskalen in verschiedenen Altersgruppen adäquat ab. Werden die Faktorwerte in einer konfirmatorischen Faktoranalyse mit einem Strukturgleichungsmodell ermittelt, entsprechen die Summenskalen MCS und PCS den zugrunde liegenden Subskalen besser. Auch die aufgrund der Subskalen erwarteten Unterschiede in Gruppen mit unterschiedlichem Gesundheitsstatus konnten reproduziert werden. Die Standardverfahren zur Berechnung der Summenskalen des SF-36 zeigen Ergebnisse, die aufgrund der zugrunde liegenden Subskalen nicht zu erwarten sind. Eine bessere Entsprechung konnte in einer konfirmatorischen Faktoranalyse durch die Anpassung eines Strukturgleichungsmodell an die Daten erzielt werden. Die Ergebnisse weisen darauf hin, dass die Interpretation der Summenskalen des SF-36 mit Vorsicht zu erfolgen hat und dass alternative Verfahren zur Berechnung der Faktorwerte angewendet werden sollten.

References

1 *Anderson J ST C, Sullivan F, Usherwood TP.* The Medical Outcomes Study instrument – use of a new health status measure in Britain. Fam Pract 1990; *7*: 205–18.

2 *Brazier JE, Harper R, Jones NMB,* et al. Validating the SF-36 health survey questionnaire: a new outcome measure for primary care. BMJ 1992; *305*: 162–4.

3 *Wilson D, Parsons J, Wakefield M.* The health related quality-of-life of never smokers, ex-smokers and light, moderate, and heavy smokers. Prev Med 1999; *29*: 139–44.

4 *Lyons RA, Lo SV, Littlepage NC.* Perception of health amongst ever-smokers and never-smokers: a comparison using the SF-36 health survey questionnaire. Tob Control 1994; *3*: 213–5.

5 *Ziebland S.* The short form 36 health status questionnaire: clues from the Oxford region's normative

Résumé

Les scores synthétiques du SF-36: problèmes et solutions

Pour déterminer si les échelles synthétiques de santé mentale et physique du SF-36 reflètent correctement les huit échelles sous-jacentes lorsqu'on utilise la méthode traditionnelle de scorage basée sur des coefficients dérivés de l'analyse des composantes principales suivie d'une rotation orthogonale. Une enquête représentative de la population australienne utilisant le SF-36 a été mise à profit pour calculer les scores synthétiques physique (PCS) et mental (MCS) du SF-36 selon la méthode traditionnelle, à partir des coefficients issus d'une analyse factorielle exploratoire de composantes principales suivie de rotation orthogonale des huit scores SF-36 initiaux. De plus, deux autres méthodes furent utilisées pour générer les coefficients. La première méthode utilisait la maximisation de la vraisemblance et la rotation oblique. La seconde méthode appliquait un modèle d'équations structurale aux données dans une analyse factorielle confirmatoire. Les coefficients dérivés dans chacune des méthodes furent appliqués aux données d'une deuxième enquête représentative de population. Cette enquête fournit également des données sur la santé physique et mentale qui permettent de comparer les scores synthétiques aux échelles sous-jacentes selon différents états de santé. Aucune des méthodes de scorage basée sur les analyses factorielles exploratoires (orthogonale et oblique) n'a produit des scores synthétiques, par groupes d'âge qui reflétait de façon adéquate ces échelles sous-jacentes. Lorsque les coefficients dérivés des équations structurales furent appliqués aux données dans une analyse factorielle confirmatoire, le MCS et PCS reflétaient correctement les échelles sous-jacentes. Ils produisirent aussi des scores de MCS et PCS pour les différents états de santé que l'on aurait pu attendre avec les échelles sous-jacentes. Les méthodes traditionnelles de scorage des échelles synthétiques du SF-36 produisent des résultats qui n'auraient pas été attendus avec les échelles sous-jacentes. Ce problème peut être corrigé par l'application d'un modèle d'équations structurale aux données dans une analyse factorielle confirmatoire. Ces résultats suggèrent que les échelles synthétiques du SF-36 devraient utilisés avec prudence et que les méthodes alternatives pour développer les coefficients factoriels devraient utilisés dans ce genre d'étude.

data about its usefulness in measuring health gain in population surveys. J Epidemiol Community Health 1995; *49*: 102–5.

6 *Garrat AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT.* The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? BMJ 1993; *306*: 1440–4.

7 *Simon GE, Revicki DA, Grothaus L, Von Korff M.* SF-36 summary scores. Are physical and mental health truly distinct. Med Care 1998; *36*: 567–72.

8 *Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A.* Comparison of methods for the scoring and statistical analysis of the SF-36 Health Profile and Summary Measures: summary of results from the Medical Outcomes Study. Med Care 1995; *33*: AS264–AS272.

9 *McHorney CA, Ware JE, Raczek AE.* The MOS 36-item short form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1981; *19*: 247–63.

10 Australian Bureau of Statistics. National Health Survey. SF-36 Population Norms Australia. Canberra: Australian Bureau of Statistics, 1995. (Catalogue Number 4399.0).

11 *Ware JE, Kosinski M, Keller SD.* SF-36 Physical and Mental Health Summary Scales: a users manual. Boston, MA: The Health Institute, New England Medical Centre, 1994.

12 *Arbuckle JL.* Amos Users Guide Version 3.6. Small Waters Corporation. Chicago: SPSS Inc. 1997.

13 *McCallum J.* The SF-36 Physical and Mental Health Summary Scales: Australian validation. Proceedings of Health Outcomes and Quality-of-Life Measurement Conference 1995. Canberra: Australian Institute of Health and Welfare 1995.

14 *McCallum J.* The SF-36 in an Australian sample: validating a new, generic health status measure. Aust J Public Health 1995; *19*: 160–6.

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

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

Address for correspondence

David Wilson
Department of Human Services
PO Box 6, Rundle Mall PO
Adelaide
South Australia 5001

David Wilson[1], Graeme Tucker[2], Catherine Chittleborough[2]

[1] *University of Adelaide*
[2] *Department of Human Services, Adelaide*

Rethinking and rescoring the SF-12

Summary

Objectives: To derive and assess the validity of an Australian version of the SF-12 quality-of-life questionnaire.
Methods: Using regression methods and structural equation modelling to obtain item weights, an Australian version of the SF-12 was derived from Australian population survey data and compared to the existing United States (US) SF-12 variable set.
Results: The Australian version of the SF-12 explained 94% of the variation for physical components summary (PCS) and the mental components summary (MCS) of the SF-36 questionnaire. There was high level of agreement on the MCS and PCS summary scores between both versions of the SF-12 and the SF-36.
Conclusions: Although it is possible to derive a valid Australian version of the SF-12 it is concluded the US version of the SF-12 be used for reasons of international comparability, but using item weights derived from structural equation modelling.

Keywords: Quality-of-life – SF-12 – Validity.

The SF-12 is a summary quality-of-life questionnaire measuring physical and mental health with the physical component summary (PCS) and the mental component summary (MCS). It is useful for measuring these health dimensions in population or subpopulation groups. The SF-36 MCS and PCS formed the basis on which the shorter version SF-12 question items were derived in the United States (Ware et al. 1996). The 12 items chosen as the US SF-12 achieved R^2 values of 0.91 for the PCS and 0.92 for the MCS (Ware et al. 1996). Explanation of 90% of the variance in the SF-36 PCS and MCS measures in the US population were deemed to be adequate decision-making criteria in accepting a 12-item measure of quality-of-life (Ware et al. 1995). This was an important development for researchers conducting surveys and other studies in which time and cost are important in measuring health status. The SF-12 is an instrument that can be administered in three minutes with a small trade off between brevity and precision.

A recent study has, however, questioned the scoring of the SF-36 PCS and MCS (Wilson et al. 2000). This showed that when coefficients are derived using structural equation modelling (rather than by principal components analysis and orthogonal rotation) and fit to South Australian population survey data in a confirmatory factor analysis, the summary PCS and MCS more accurately reflect the underlying subscales (general health, physical health, role physical, bodily pain, role emotional, vitality, social health, and mental health) from which they are derived. The structural equation method produced SF-36 summary scale scores which had good validity across age and implied health groups when compared with scores produced by orthogonal rotation (Wilson et al. 2000). The work of rescoring the SF-36 using structural equation modelling followed on from the previous work of Simon et al. who recommended caution in the interpretation of the PCS and MCS when the condition or treatment of interest had strong effects on scales with negative scoring coefficients (1998). This means that the derivation of the SF-12 question items and the construction of this shorter quality-of-life instrument was based on problematic SF-36 summary scores. Given this previous work this study used Australian population data to derive the SF-12 on the basis of an SF-36 that was scored using structural equation modelling. As a result of this and other factors a version of the SF-12 was produced that differed to some extent in the questions that formed the US SF-12.

Soz.- Präventivmed. 47 (2002) 172–177
0303-8408/02/030172-06 $ 1.50 + 0.20/0

Historically the US version of the SF-12, using US regression weights, has been used in Australian population studies. This is inappropriate given that the weights were derived from US survey data and are now of a considerable age. An earlier attempt to develop an Australian SF-12 failed on the basis that an instrument could not be derived that adequately explained the variance of the SF-36 summary scales (McCallum 1996). Since then, however, higher quality Australian population data has become available and was used in this study to revisit the issue. The research questions asked were: 1) can a valid SF-12 be derived that adequately explains the SF-36 summary scales based on structural equation modelling, and 2) how does this Australian version of the SF-12 compare with the US version in both question content and in assessing the health of various age and health groups.

Method

Two population data sources were used for this study. The first was the 1995 Australian National Health Survey (NHS) (1995) which was the second in a series of five-yearly population surveys designed to obtain national benchmark information on a range of health issues. The sampling method of the NHS is a self-weighting multistage clustered area sample based on Australian Bureau of Statistics (ABS) collector districts in which households are selected with equal probability. In this survey n = 23800 households were selected and all adults aged 15 years or older were interviewed. A subset of n = 19785 were asked to complete the SF-36 quality-of-life questionnaire. Of those interviewed, n = 18141 provided sufficient data to be included in the analyses of the SF-36 subscale and summary scores (Ware et al. 1993). These data were used in the paper by Wilson et al. to rescore the SF-36 summary scales using structural equation modelling (2000). The NHS data were again used in this study to derive an Australian SF-12 that best explained the SF-36 PCS and MCS and also to produce regression coefficients that could be used to score the US SF-12. Having derived the Australian SF-12 items from the NHS data set the South Australian Health Omnibus (SAHOS) data set was used to confirm the validity of both the Australian and US versions of the SF-12.

Derivation of the Australian SF-12 from the NHS data set occurred after imputation of data items by mean substitution, where not more than half the data items in a subscale were missing, as set out in the SF-36 scoring manual (Ware et al. 1993). Regression methods (involving backward elimination of the least significant term) were used to select from among the SF-36 items. Adjusted R^2 values were used to determine the amount of variance explained in this Australian SF-12. The NHS data set was also used to compare the variance explained in the US version of the SF-12. Because PCS and MCS are now linear combinations of contributing variables (for each dimension) (Wilson et al. 2000), a full backward elimination model could not be fitted (one term needs to be eliminated to avoid a perfect fit). By first removing one term at random, then identifying the least significant term remaining, eliminating that term and replacing it with the first term removed, backward elimination methods can then proceed as normal.

The Australian SF-12 variables newly derived from the NHS data and the US SF-12 were then fitted to data obtained in the independent 1998 SAHOS using a sample of n = 3007 (70% response rate) in a confirmatory factor analysis. The SAHOS is a statewide survey designed to obtain benchmark information on a range of health issues. The SAHOS like the NHS Survey is also a self-weighting multistage clustered area sample selected from ABS collector districts of people aged 15 years or more who live in metropolitan Adelaide or country towns with populations over 1000. The survey is conducted annually and the method has been extensively published (Wilson et al. 1992). Households are selected with equal probability of selection within each collector district and then one adult in each household, aged 15 years or older is selected for interview according to the most recent birthday. Data from the survey are weighted to accurately represent the age, sex, household size, and geographic area of the South Australian population. Having confirmed the factor structure on the SAHOS data the same models were fitted to the NHS data to produce regression coefficients representative for Australia. Wilson et al. (2000) provide full rationale for the use of structural equation modelling and details of the methods used.

In addition to the SF-36, a range of other health and demographic questions were asked in the 1998 SAHOS and from these data it was possible to validate both versions of the SF-12, with Australian regression coefficients, using a range of age and implied health groups. In the 1998 SAHOS respondents were asked whether or not in the previous twelve months they had used any medication for a chronic physical condition: that is, one that has lasted for, or is likely to last for, six months or more. They were also asked the same question for depression (medication such as tranquillisers or anti-depressants) or a diagnosed mental illness (such as schizophrenia). From these data the study population were divided into four implied health groups according to no medication, physical health medication only, mental health medication only or both physical and mental health medication. In the previous paper by Wilson et al. (2000) these

Soz.- Präventivmed. 47 (2002) 172–177

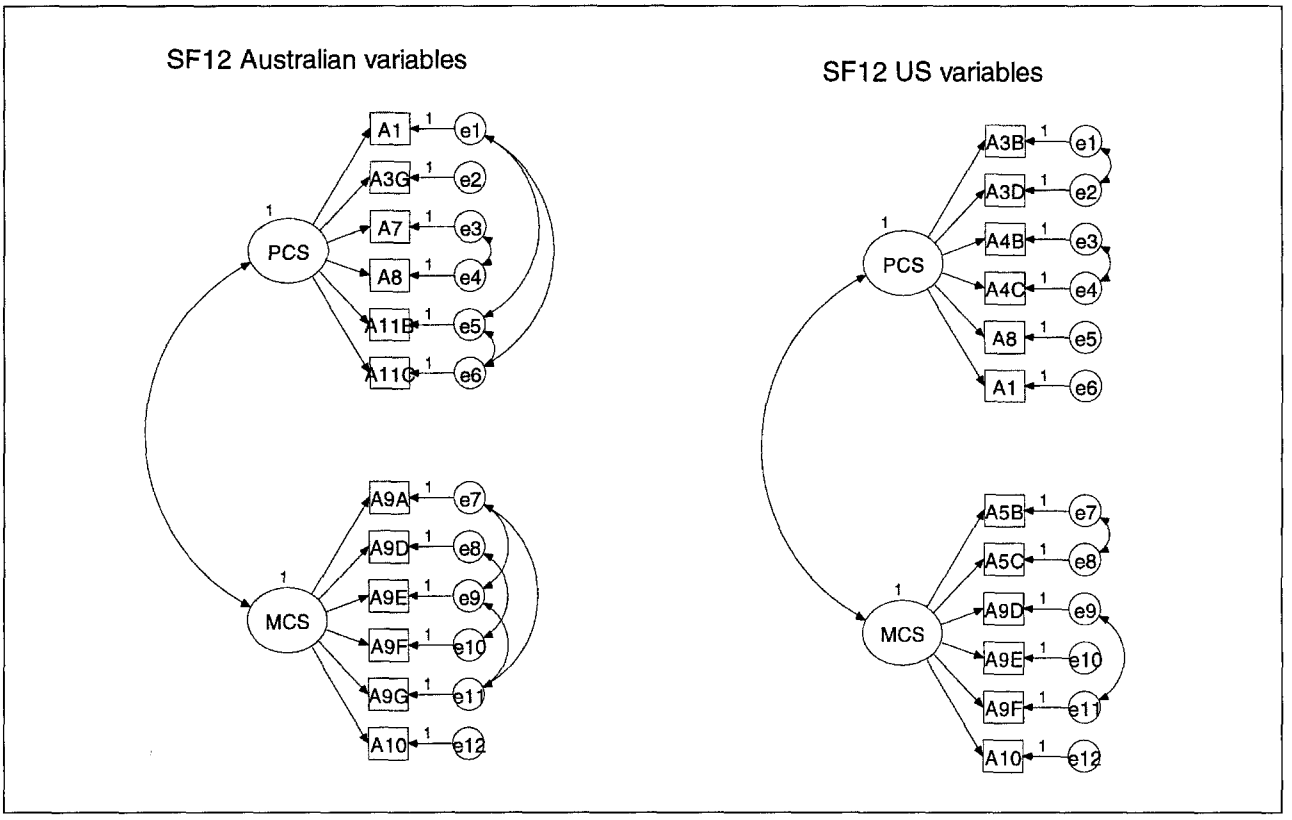


Figure 1 SF-12 model fitted

implied health groups were used to assess the external validity of two versions of the SF-36 PCS and MCS by comparing the scores with their underlying subscale scores. These versions of the PCS and MCS were derived from both structural equation modelling and principle components analyses. Scoring the SF-36 using structural equation modelling had greater validity than scoring by principal components analysis. The present study compares the Australian and US versions of the SF-12 PCS and MCS scores with the summary SF-36 PCS and MCS, using structural equation modelling, for age groups and implied health groups as an assessment of SF-12 validity.

The models fitted to the data using structural equation modelling are shown in Figure 1. The figure also shows that the covariance terms between errors were applied to items from the same subscales on the basis that these items would be expected to be more closely correlated with each other than with items from other subscales.

Table 1 SF-36 question items derived for the Australian and United States SF-12 PCS and MCs

	Australian SF-12	United States SF-12
PCS	A1 Evaluation of general health	US1 Evaluation of general health
	A8 Pain interferes with normal work	US8 Pain interferes with normal work
	A3g Walking more than a mile	US3b Moderate activities
	A7 Pain magnitude	US3d Climbing several flights
	A11b As healthy as anybody I know	US4b Accomplished less
	A11c Health to get worse	US4c Limited in kind of work
MCS	A9d Calm and peaceful	US9d Calm and peaceful
	A9e Lot of energy	US9e Lot of energy
	A9f Downhearted and blue	US9f Downhearted and blue
	A10 Social-time	US10 Social-time
	A9a Feel full of life	US5b Accomplished less
	A9g Feel worn out	US5c Not careful

Results

The twelve Australian items selected for the PCS and MCS together with their US counterparts are shown in Table 1. It can be seen that, in total, six questions from the SF-36 are

Table 2 Comparisons of Australian and United States versions of the SF-12 PCS and MCS scores with the SF-36 summary scores by age groups

	Age group				
	< 30 years	30–49 years	50–69 years	70+ years	Total
SF-36 PCS	52.8	51.8	48.2	44.4	50.3
SF-36 MCS	52.1	51.1	51.3	49.9	51.2
SF-12 PCS (Aus)	52.7	51.5[a]	48.3	45.9[a]	50.4
SF-12 MCS (Aus)	52.3[a]	50.6[a]	50.4[a]	48.5[a]	50.7[a]
SF-12 PCS (US)	52.8	51.8	48.4	44.8[a]	50.4
SF-12 MCS (US)	52.5[a]	51.0	50.9[a]	49.0[a]	51.1

[a] Statistically significantly different from the SF-36 comparable scale.

Table 3 Comparisons of Australian and United States versions of the SF-12 PCS and MCS scores with the SF-36 summary scores for implied health groups

	Medication taken				
	None	Physical only	Mental only	Both	Total
SF-36 PCS	53.1	43.6	47.1	34.6	50.3
SF-36 MCS	52.8	48.9	41.6	37.3	51.2
SF-12 PCS (Aus)	53.0[a]	44.2[a]	47.7	35.6[a]	50.4
SF-12 MCS (Aus)	52.4[a]	48.0[a]	41.9	37.5	50.7[a]
SF-12 PCS (US)	53.1	44.2[a]	47.2	35.6[a]	50.4
SF-12 MCS (US)	52.8	48.5[a]	42.4	37.7	51.1

[a] Statistically significantly different from the SF-36 comparable scale.

common to both the US and Australian version of the SF-12. For the Australian SF-12 PCS and MCS the SF-36 variation explained, as determined by the adjusted R^2, was 94% for both scales. The root mean square error (RMSEA) for the Australian model was 0.07. For the US version of the SF-12 the variance explained was 91% for the PCS and 90% for the MCS. The RMSEA, which assesses the fit of the model in relation to the degrees of freedom, was 0.08 for this model. We would not employ a model with a RMSEA equal to or greater than 0.1 (Arbuckle & Wothke 1999).

Table 2 shows the SF-12 PCS and MCS age group scores calculated in this study and compared to the SF-36 summary scores calculated in the previous SF-36 study (Wilson et al. 2000). It can be seen from Table 2 that there is a high level of agreement between both versions of the SF-12 and also a high level of agreement between these and the SF-36 summary scores. Some statistically significant differences were found for some age groups for the Australian and US SF-12 PCS and MCS when compared with the SF-36 PCS and MCS However, these differences were relatively small and statistical significance reflects the large sample sizes used. Overall only the Australian SF-12 MCS was significantly different from the SF-36 MCS. The maximum *relative* differences amounted to 3% between paired scores for the Australian SF-12 MCS and SF-36 MCS in the 70 years and older age group. This was calculated by subtracting the SF-12 score from the SF-36 score and dividing by the SF-36 score. The majority of other paired differences were less than 2% and some were less than 1%.

When compared with the theoretical health groups, based on medication taken, there was again a high level of agreement between the scores of the SF-12 versions and the SF-36 summary scales. Significant differences were observed for some implied health groups, but again the maximum relative difference did not exceed 3%. The conclusion that should be reached is that despite some statistically significant differences in these SF-12 scores when compared with the SF-36 scores, these differences are not significant in practical or clinical terms.

Discussion

In this study the SF-12 reproduced the average summary scores of the SF-36 with over 90% of the variance explained using Australian regression weights. This addresses Ware's criteria for acceptance of the instrument. In selecting the best set of twelve explanatory variables for the SF-36 PCS and MCS it should be borne in mind that all question items are part of a summary scale constructed to take advantage of the high correlation between each of the variables. Thus the variables are collinear and all strongly related to the dependent variable. Selection of question items for the Australian SF-12 are therefore based on very small differences between each item and a number of different combinations explained more than 90% of the variation in the SF-36 summary scales in the Australian data. It is also highly probable that selection of an SF-12 will be data set specific. Collinearity diagnostics conducted in this study all produced conditioning index values in excess of 20 indicating collinearity between the variables in the SF-36 subsets (Gebski et al. 1992).

The MCS and PCS summary scores for both versions of the SF-12 compared well with the SF-36 summary scores produced in the previous study by Wilson et al. (2000) and had higher agreement with the underlying scale scores (validity) than did the MCS and PCS scores obtained from orthogonal rotation. We cannot assert that the validity of results obtained in this Australian study apply to other countries. It is reasonable, however, to suggest that users of the SF-36 and SF-12 in countries other than the United States use weights that are derived in those countries as has been done in this study.

Until this study the US SF-12 has been scored using regression coefficients derived in the United States. This study has corrected the situation and shown that the US SF-12 is a valid instrument in the Australian context. For other countries using the US SF-12 question items, regression weights would also best be obtained from endogenous population data.

The Australian version of the SF-12 explained more of the variance for both the PCS and the MCS than did the US version and the goodness of fit diagnostics (RMSEA) may suggest that the US version is not quite as good a model as the Australian version. Given that it does explain 90% of the variation in the SF-36 PCS and MCS with a RMSEA of 0.08 it is, however, a very adequate model. With this in mind we can conclude that the preferred version of the SF-12 for the Australian context should be the US SF-12. This is based on both the model adequacy and the fact that this version has international comparability in quality-of-life scores. In addition, despite better statistical diagnostics for the Australian variable set in these analyses, the US variable set actually performed better when compared to the SF-36 and provided smaller comparative differences in the validity study. Further support for using the US SF-12 comes from the fact that the Australian SF-12 did not cover all of the subscales in the questions derived. For the PCS role-physical was omitted and for the MCS role-emotional was omitted. Overall therefore the US SF-12 is a better instrument for both the Australian and international research context.

Zusammenfassung

Überdenken und Neuauswertung des SF-12-Fragebogens

Fragestellung: Eine australische Version des SF-12-Fragebogens zur Erfassung der Lebensqualität ableiten und dessen Validität beurteilen.

Methoden: Unter Verwendung von Regressions- und Strukturgleichungsmodellen zur Bestimmung von Gewichtungsfaktoren wurde eine australische Version des SF-12-Fragebogens abgeleitet. Hierzu wurden Daten einer Befragung der australischen Bevölkerung herangezogen. Der australische Fragebogen wurde anschliessend mit dem bereits bekannten U.S. SF-12-Variablensatz verglichen.

Ergebnisse: Die australische Version des SF-12 erklärte 94% der Variation für die physische (PCS) und psychische Summenskala (MCS) des SF-36-Fragebogens. Zwischen den beiden Versionen des SF-36 und SF-12 stimmten die MCS- und PCS-Summenwerte sehr gut überein.

Schlussfolgerungen: Es ist möglich, eine valide australische Version des SF-12 zu erhalten. Aus Gründen der internationalen Vergleichbarkeit ist es jedoch besser, die U.S. Version des SF-12 einzusetzen, aber unter Verwendung von Gewichtungsfaktoren, die anhand struktureller Vergleichsmodelle abgeleitet wurden.

Résumé

Repenser et redéfinir le score du SF-12

Objectifs: Etablir une version australienne du questionnaire SF-12 sur la qualité de vie et évaluer sa validité.

Méthode: A partir de méthodes de régression et de pondération par des modèles d'équations structurelles, une version australienne du SF-12 a été établie à partir d'une enquête de population et comparée avec l'ensemble des variables composant la version nord-américaine existante du SF-12.

Résultats: La version australienne du SF-12 expliquait 94% de la variation du score synthétique physique (PCS) et mental (MCS) du questionnaire SF-36. Il y avait un haut niveau de concordance pour ces deux scores entre les deux versions du SF-12 et du SF-36.

Conclusions: Bien qu'il soit possible d'établir une version australienne valide du SF-12, l'utilisation de la version américaine du SF-12 peut être utilisée pour des raisons de comparabilité internationale, mais en utilisant un système de pondération spécifique basé sur des modèles d'équation structurelles.

References

Arbuckle JL, Wothke W (1999). Amos 4.0 user's guide. Chicago: SmallWaters Corporation.

National Health Survey (1995). SF-36 population norms Australia. Canberra: Australian Bureau of Statistics.

Gebski V, Leung O, McNeil D, Lunn D (1992). Spida users manual. New South Wales: Statistical Computing Laboratory.

McCallum J (1996). The shorter form: analysis of SF12 items in Australian data: proceedings. Canberra: Integrating Health Outcomes Measurement in Routine Health Care Conference.

Simon GE, Revicki DA, Grothaus L, Von Korff M (1998). SF-36 summary scores: are physical and mental health truly distinct? Med Care *36*: 567–72.

Ware J, Kosinski M, Keller SD (1996). A 12-Item Short Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care *34*: 220–33.

Ware JE, Kosinski M, Keller SD (1995). SF-12: how to score the SF-12 Physical and Mental Health Summary Scales. Boston, Mass: The Health Institute, New England Medical Center.

Ware JE, Snow KK, Kosinski MA, Gandek B (1993). SF-36 Health Survey manual and interpretation guide. Boston, Mass: The Health Institute, New England Medical Center.

Wilson D, Parsons J, Tucker G (2000). The SF-36 summary scales: problems and solutions. Soz Praventivmed *45*: 239–46.

Wilson D, Wakefield M, Taylor A (1992). The South Australian Health Omnibus Survey. Health Promot J Aust *2*: 47–9.

Address for correspondence

A/Professor David Wilson
Department of Medicine
University of Adelaide
Queen Elizabeth Hospital
Woodville, South Australia 5011

Tel.: +1 61 8 82226047
Fax: +1 61 8 82226042

e-mail: David.Wilson@dhs.sa.gov.au