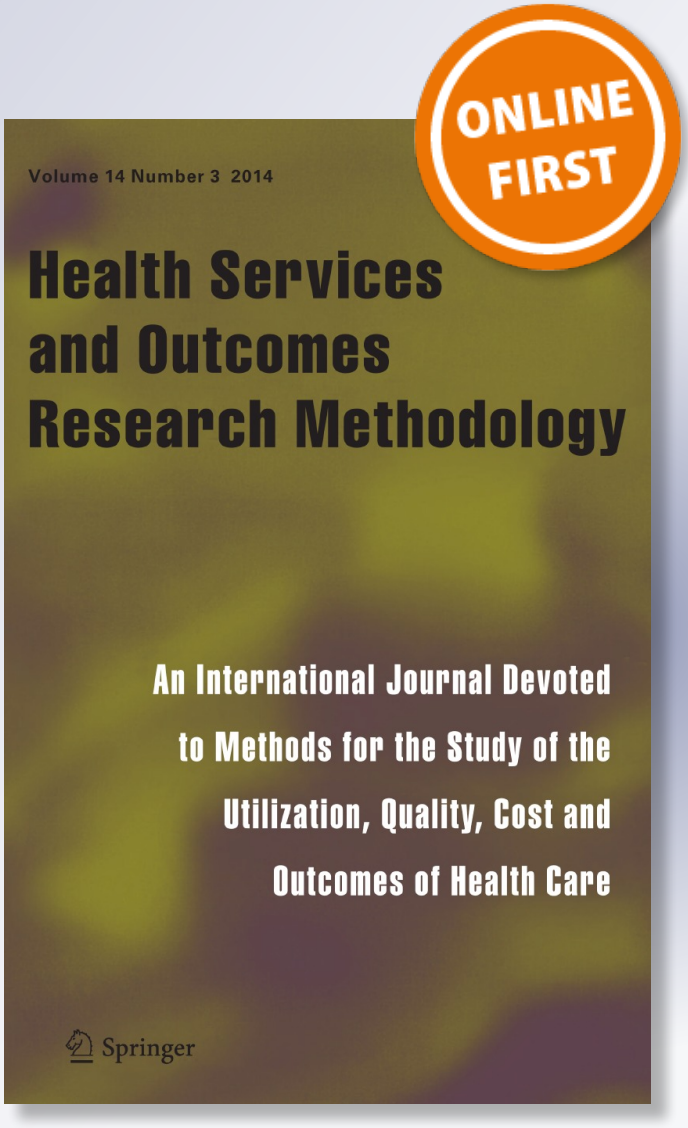
Item response theory approaches to harmonization and research synthesis

Robert D. Gibbons, Marcelo Coca Perraillon & Jong Bae Kim

Health Services and Outcomes Research Methodology
An International Journal Devoted to Methods for the Study of the Utilization, Quality, Cost and Outcomes of Health Care

ISSN 1387-3741

Health Serv Outcomes Res Method
DOI 10.1007/s10742-014-0125-x



Springer

Health Serv Outcomes Res Method
DOI 10.1007/s10742-014-0125-x

Item response theory approaches to harmonization and research synthesis

Robert D. Gibbons · Marcelo Coca Perraillon · Jong Bae Kim

Received: 10 February 2014 / Revised: 7 July 2014 / Accepted: 23 August 2014

Abstract The need to harmonize different outcome metrics is a common problem in research synthesis and economic evaluation of health interventions and technology. The purpose of this paper is to describe the use of multidimensional item response theory (IRT) to equate different scales which purport to measure the same construct at the item level. We provide an overview of multidimensional IRT in general and the bi-factor model which is particularly relevant for applications in this area. We show how both the underlying true scores of two or more scales that are intended to measure the same latent variable can be equated and how the item responses from one scale can be used to predict the item responses for a scale that was not administered but are necessary for the purpose of economic evaluations. As an example, we show that a multidimensional IRT model predicts well both the EQ-5D descriptive system and the EQ-5D preference index from SF-12 data which cannot be directly used to perform an economic evaluation. Results based on multidimensional IRT performed well compared to traditional regression methods in this area. A general framework for harmonization of research instruments based on multidimensional IRT is described.

Keywords Harmonization · Item response theory · Research synthesis · Cost effectiveness · QALY

1 Introduction

With increasing availability of large-scale databases from multiple studies and cohorts, the opportunities for research synthesis have grown tremendously. While meta-analytic methods have been available for decades, they rely on the pooling of effect size estimates

R. D. Gibbons (✉) · M. C. Perraillon · J. B. Kim
Departments of Medicine and Health Studies, University of Chicago, 5841 S. Maryland Avenue, MC 2007 Office W260, Chicago, IL 60637, USA
e-mail: rdg@uchicago.edu

Published online: 06 September 2014

from individual studies or units. The advantage of traditional meta-analysis is that the individual studies or datasets do not require that the same outcome measure be available, but rather pool data summary measures such as standardized mean differences or odds ratios. Yet, this flexibility is not without cost in that it precludes synthesis of longitudinal data and person-level specific information and potential random-effects. While statistical methods for research synthesis of discrete and continuous endpoints, such as multi-level models are available (Hedeker and Gibbons 2006 for an overview), they require a common endpoint to be measured for all subjects on all measurement occasions in all studies. Consistent endpoints, however, are often unavailable and the research synthesis is then limited to the subset of studies for which commonly measured data are available, potentially biasing results since the remaining studies may not be representative of the entire population of studies.

The need for harmonization of outcome measures goes beyond research synthesis. Harmonization of outcome measures is often needed in economic evaluations of treatments and technologies. In cost-effectiveness studies, the cost of alternative treatments and technologies are compared to their benefits. A common measure of benefits is the quality-adjusted life years (QALYs). QALYs are calculated by combining life years gained with a measure of societal preferences or "utilities" over different levels of health functioning. To obtain a measure of societal preferences over different levels of health functioning, health status needs to be measured. Over the years, the EuroQol five-dimension questionnaire EQ-5D has become the standard way of measuring health status in economic evaluations. The EQ-5D descriptive system is a generic, non-disease specific instrument that describes general health functioning. The EQ-5D descriptive system is transformed into a measure of preferences by applying country-specific scoring algorithms derived from large-scale valuation studies. Even though the EQ-5D is widely-used in economic evaluations, it is not yet routinely employed in clinical trials and other datasets that could potentially be used to conduct economic evaluations. On the other hand, these studies do have generic measures of health functioning, like the Short Form 12 (SF-12) survey, which measures physical and mental functioning. The SF-12 instrument, however, has not been extensively used in valuation studies and therefore cannot be transformed into preferences, a key component of many economic evaluations.

A growing literature has focused on mapping or predicting either the EQ-5D preference index (the resulting scale after applying country-specific scoring algorithms) or the response patterns (to which scoring algorithms can then be applied) using the SF-12 component as predictors (Longworth and Rowen 2013). Both types of prediction are challenging. Prediction of the EQ-5D preference index is difficult because of its unusual distribution. The EQ-5D preference index is bounded, it has multiple modes, and a large percent of respondents may have the same score. On the other hand, predicting the response patterns by respondent is also difficult because the EQ-5D is made of five questions with three answers each for a total of 3^5 (243) possible response patterns.

The focus of this paper is on the use of multidimensional item response theory (IRT) to equate different scales which purport to measure the same construct at the item level, so that studies using different scales of measurement for the outcome of interest can be harmonized into a common metric. We provide an overview of the general methodology that can be applied to diverse types of applied problems in research synthesis and economic evaluation. As an example, we concentrate on the problem of predicting the EQ-5D response patterns and its resulting preference index in a dataset representative of the U.S. population. The use of multidimensional IRT for this particular application is a natural starting point for the prediction of the EQ-5D using the SF-12 as both instruments are

intended to measure the same underlying construct. We show that a multidimensional IRT model is able to make predictions that are better than commonly used models for predicting both the responses to the EQ-5D descriptive system and the EQ-5D preference index.

2 The logic of item response theory

Classical and IRT methods of measurement differ dramatically in the ways in which items are administered and scored. The difference is clarified by the following analogy originally suggested by R.D. Bock. Imagine a track and field meet in which ten athletes participate in men's 110-m hurdles race and also in men's high jump. Suppose that the hurdles race is not quite conventional in that the hurdles are not all the same height and the score is determined, not only by the runner's time, but also by the number of hurdles successfully cleared, *i.e.*, not tipped over. On the other hand the high jump is conducted in the conventional way: the cross bar is raised by, say, 2 cm increments on the uprights, and the athletes try to jump over the bar without dislodging it.

The first of these two events is like a traditionally scored objective test: runners attempt to clear hurdles of varying heights analogous to questions of varying difficulty that examinees try to answer correctly in the time allowed. In either case, a specific counting operation measures ability to clear the hurdles or answer the questions. On the high jump, ability is measured by a scale in millimeters and centimeters at the highest scale position of the cross bar the athlete can clear. IRT measurement uses the same logic as the high jump. Test items are arranged on a continuum at certain fixed points of increasing difficulty. The examinee attempts to answer items until she can no longer do so correctly. Ability is measured by the location on the continuum of the last item answered correctly. In IRT, ability is measured by a scale point, not a numerical count.

These two methods of scoring the hurdles and the high jump, or their analogues in traditional and IRT scoring of objective tests, contrast sharply: if hurdles are arbitrarily added or removed, number of hurdles cleared cannot be compared with races run with different hurdles or different numbers of hurdles. Even if percent of hurdles cleared were reported, the varying difficulty of clearing hurdles of different heights would render these figures non-comparable. The same is true of traditional number-right scores of objective tests: scores lose their comparability if item composition is changed.

The same is not true, however, of the high jump or of IRT scoring. If the bar in the high jump were placed between the 2 cm positions, or if one of those positions were omitted, height cleared is unchanged and only the precision of the measurement at that point on the scale is affected. Indeed, in the standard rules for the high jump, the participants have the option of omitting lower heights they feel they can clear. Similarly, in IRT scoring of tests, a certain number of items can be arbitrarily added, deleted or replaced without losing comparability of scores on the scale. Only the precision of measurement at some points on the scale is affected. This property of scaled measurement, as opposed to counts of events, is the most salient advantage of IRT over classical methods of educational and psychological measurement.

This example should make it clear that once calibrated using items from two scales, IRT allows different subjects to be measured on only one of the two scales yet valid scores of the underlying latent variable of interest remain estimable.

Most applications of IRT are based on unidimensional models which assume that all of the association between the items is explained by a single primary latent dimension or

factor (e.g., mathematical ability). However, many constructs of interest are inherently multidimensional, for example, the previously mentioned EQ-5D and SF-12 scales contain items which investigate both physical health and mental health and would therefore violate the conditional independence assumption of a unidimensional IRT model. Alternatively, there may be methodological differences between the two (or more) scales being harmonized, for which the scales themselves become additional factors in the model. If we attempt to fit such data to a traditional unidimensional IRT model, we will typically have to discard the majority of candidate items (e.g. all of the mental health items or all of the items from one of the two scales) to achieve a reasonable fit of the model to the data. By contrast, the bi-factor IRT model (Gibbons and Hedeker 1992) permits each item to tap the primary dimension of interest (e.g. overall well-being) and one sub-domain (e.g., physical well-being), thereby accommodating the residual dependence and allowing for the retention of the majority of the items in the final model.

The bi-factor model of Gibbons and Hedeker (1992) was the first example of a confirmatory item factor analysis model, and they showed that it is computationally tractable regardless of the number of dimensions, in stark contrast to exploratory item factor analytic models. Furthermore the estimated bi-factor loadings are rotationally invariant, greatly simplifying interpretability of the model estimates. In the following, we provide a general overview of the statistical foundation of the model, and parameter trait estimation. The notation that we use denotes an item as j and a subject as i. There are N subjects, n items and m_j categories for item j. The number of dimensions is denoted as d and a particular dimension is indexed by v. A particular item response category is indexed by h.

2.1 Multidimensional item response theory

As described by Bock and Gibbons (2010), IRT-based item factor analysis makes use of all information in the original categorical responses and does not depend on pairwise indices of association such as tetrachoric or polychoric correlation coefficients. For that reason it is referred to as *full information* item factor analysis. It works directly with item response models giving the probability of the observed categorical responses as a function of latent variables descriptive of the respondents and parameters descriptive of the individual items. It differs from the classical formulation in its scaling, however, because it does not assume that the response process has unit standard deviation and zero mean; rather it assumes that the *residual* term has unit standard deviation and zero mean. The latter assumption implies that the response processes (for each item) have zero mean and standard deviation equal to

$$\sigma_{y_j} = \sqrt{1 + \Sigma_v^d \alpha_{jv}^2},$$

where α_{jv} represents the factor loading for item j on dimension v and y_j represents the unobservable response process for the jth item. Inasmuch as the scale of the model affects the relative size of the factor loadings and thresholds, we rewrite the model for dichotomous responses in a form in which the factor loadings are replaced by factor slopes, a_{jv}, and the threshold is absorbed in the intercept, c_j:

$$y_j = \sum_{v=1}^{d} a_{jv}\theta_v + c_j + \varepsilon_j.$$

To convert factor slopes into loadings we divide by the above standard deviation and similarly convert the intercepts to thresholds:

$$\alpha_{jv} = a_{jv}/\sigma_{y_j} \quad \text{and} \quad \gamma_j = -c_j/\sigma_{y_j}.$$

Conversely, to convert to factor analysis units, we change the standard deviation of the residual from 1 to

$$\sigma^*_{\varepsilon_j} = \sqrt{1 - \Sigma_v^d \alpha_{jv}^2},$$

and change the scale of the slopes and intercept accordingly:

$$a_{jv} = \alpha_{jv}/\sigma^*_{\varepsilon_j} \quad \text{and} \quad c_j = -\gamma_j/\sigma^*_{\varepsilon_j}.$$

For polytomous responses, the model generalizes as:

$$z_j = \sum_{v=1}^{d} a_{jv}\theta_v,$$

$$P_{jh}(\theta) = \Phi(z_j + c_{jh}) - \Phi(z_j + c_{j,h-1}),$$

where $\Phi(z_j + c_{j0}) = 0$ and $\Phi(z_j + c_{jm_j}) = 1 - \Phi(z_j + c_{j,m_j-1})$. In the context of item factor analysis, this is the multidimensional generalization of the *graded* model introduced by Samejima (1969).

2.2 Confirmatory item factor analysis

In confirmatory factor analysis, indeterminacy of rotation is resolved by assigning arbitrary fixed values to certain loadings of each factor during maximum likelihood estimation. An important example of confirmatory item factor analysis is the bi-factor pattern for general and content-specific factors (i.e. subdomains), which applies to tests and scales with item content drawn from several well-defined sub-areas of the domain in question. To analyze these kinds of structures for dichotomously scored item responses, Gibbons and Hedeker (1992) developed full-information item bi-factor analysis for binary item responses, and Gibbons and colleagues extended it to the polytomous case (Gibbons et al. 2007). To illustrate, consider a set of n test items for which a d-factor solution exists with one general factor and $d - 1$ subdomains. These subdomains may be content based (e.g., cognitive and somatic subdomains of the primary dimension of depressive severity) or method-related factors such as different scales that were designed to measure a common primary dimension such as the overall quality of one's health. The bi-factor solution constrains each item j to a non-zero loading α_{j1} on the primary dimension and a second loading (α_{jv}, $v = 2,\ldots,d$) on not more than one of the $d - 1$ subdomains. For four items, the bi-factor pattern matrix might be

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix}$$

This structure, which Holzinger and Swineford (1937) termed the ''bi-factor'' pattern, also appears in the inter-battery factor analysis of Tucker (1958) and is one of the confirmatory factor analysis models considered by Joreskog (1969). In the latter case, the model is restricted to test scores assumed to be continuously distributed. However, the

bi-factor pattern might also arise at the item level (Muthen 1989). Gibbons and Hedeker (1992) showed that paragraph comprehension tests, where the primary dimension represents the targeted process skill and additional factors describe content area knowledge within paragraphs, were described well by the bi-factor model. In this context, they showed that items were conditionally independent between paragraphs, but conditionally dependent within paragraphs.

The bi-factor restriction leads to a major simplification of likelihood equations that (1) permits analysis of models with large numbers of subdomains or group factors since the integration always simplifies to a two-dimensional problem, (2) permits conditional dependence among identified subsets of items, and (3) in many cases, provides more parsimonious factor solutions than an unrestricted full-information item factor analysis, at least in part due to its rotational invariance.

2.3 The Bi-factor Model

In the bi-factor case, the graded response model is

$$z_{jh}(\theta) = \sum_{v=1}^{d} a_{jv}\theta_v + c_{jh},$$

where only one of the $v = 2,\ldots,d$ values of a_{jv} is non-zero in addition to a_{j1}. Assuming independence of the θ, in the unrestricted case, the multidimensional model above would require a d-fold integral in order to compute the unconditional probability for response pattern $\mathbf{u}$, i.e.,

$$P(\mathbf{u} = \mathbf{u}_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L_i(\theta)g(\theta_1)g(\theta_2)\cdots g(\theta_d)d\theta_1 d\theta_2\cdots d\theta_d,$$

where $L_i(\theta)$ is the conditional probability of the response pattern (conditional on the latent variables $\boldsymbol{\theta}$). Numerical approximation of this marginal probability is limited as the number of dimensions increases beyond 5 or 6. Gibbons and Hedeker (1992) showed that for the binary response model, the bi-factor restriction always results in a two-dimensional integral regardless of the number of dimensions, one for θ_1 and the other for θ_v, $v > 1$.

For the graded response model, the probability of a value less than the category threshold $\gamma_{jh} = -c_{jh}/y_j$ can be obtained by substituting γ_{jh} for γ_j in the previous equation. Let $\delta_{ijh} = 1$ if person i responds positively to item j in category h and $\delta_{ijh} = 0$ otherwise. The unconditional probability of a particular response pattern $\mathbf{u}_i$ is then

$$P(\mathbf{u} = \mathbf{u}_i) = \int_{-\infty}^{\infty}\left\{\prod_{v=2}^{d}\int_{-\infty}^{\infty}\left[\prod_{j=1}^{n}\prod_{h=1}^{m_j}\left[\Phi_{jh}(\theta_1,\theta_v) - \Phi_{jh-1}(\theta_1,\theta_v)\right)^{\delta_{ijh}\cdot u_{jv}}\right] g(\theta_v)d\theta_v\right\} g(\theta_1)d\theta_1,$$

which can be approximated to any degree of practical accuracy using two-dimensional Gauss–Hermite quadrature, since for both the binary and graded bi-factor response models, the dimensionality of the integral is 2 regardless of the number of subdomains (i.e., $d - 1$) that comprised the scale.

Complete details of the likelihood equations and their solution are provided in Gibbons et al. (2007).

Trait Estimation: In practice, the ultimate objective is to estimate the trait level of person i on the primary dimension the instrument was designed to measure. For the

bi-factor model, the goal is to estimate the latent variable θ_1 for person i. A good choice for this purpose (Bock and Aitkin 1981) is the expected a posteriori (EAP) value (Bayes estimate) of θ_1, given the observed response vector $\mathbf{u}_i$ and levels of the other subdimensions $\theta_2 \ldots \theta_d$. The Bayesian estimate of θ_1 for person i is:

$$\hat{\theta}_{1i} = E(\theta_{1i}|\mathbf{u}_i) = \frac{1}{P_i}\int_{\theta_1} \theta_{1i}\left\{\prod_{v=2}^{d}\int_{\theta_v} L_{iv}(\theta_v^*)g(\theta_v)d\theta_v\right\}g(\theta_1)d\theta_1.$$

Similarly, the posterior variance of $\hat{\theta}_{1i}$, which may be used to express the precision of the EAP estimator, is given by

$$V(\theta_{1i}|\mathbf{u}_i) = \frac{1}{P_i}\int_{\theta_1} (\theta_{1i} - \hat{\theta}_{1i})^2\left\{\prod_{v=2}^{d}\int_{\theta_v} L_{iv}(\theta_v^*)g(\theta_v)d\theta_v\right\}g(\theta_1)d\theta_1.$$

These quantities can be evaluated using Gauss–Hermite quadrature as previously described.

In some applications, we are also interested in estimating a person's location on the secondary domains of interest as well. For the vth sub-domain, the EAP estimate and its variance can be written as:

$$\hat{\theta}_{vi} = E(\theta_{vi}|\mathbf{u}_i) = \frac{1}{P_i}\int_{\theta_v} \theta_{vi}\left\{\int_{\theta_1} L_{iv}(\theta_v^*)g(\theta_1)d\theta_1\right\}g(\theta_v)d\theta_v,$$

and

$$V(\theta_{vi}|\mathbf{u}_i) = \frac{1}{P_i}\int_{\theta_v} (\theta_{vi} - \hat{\theta}_{vi})^2\left\{\int_{\theta_1} L_{iv}(\theta_v^*)g(\theta_1)d\theta_1\right\}g(\theta_v)d\theta_v.$$

2.4 Harmonization

We can use the bi-factor model to provide harmonization between two or more measures. The subdomains can represent the individual scales or unique substantive subdomains such as physical and mental health items. Since estimation is based on full-information marginal maximum likelihood, the missing data assumption is missing at random (MAR) Rubin (1976). This means that we can still estimate the primary domain score, even if item responses for only a subset of the scales are available (e.g., SF-12 administered EQ-5D not administered). As such, the simplest approach to harmonization is to use the primary dimension score based on the available scales as an estimate of the primary dimension score for the missing scale. Regression methods can then be used to re-express the IRT estimated score into the original total score metric.

An alternative approach is to use the available scales to estimate the actual item responses for the missing scale(s). To do this, note that

$$z_j = \sum_{v=1}^{d} a_{jv}\theta_v, \quad p_{jh}(\theta) = \Phi(z_j + c_{jh}) - \Phi(z_j + c_{j,h-1}),$$

where $\Phi(z_j + c_{j0}) = 0$ and $\Phi(z_j + c_{jm_j}) = 1 - \Phi(z_j + c_{j,m_{j-1}})$.

We can then select the category for a missing scale item that has maximum probability based on the theta estimate(s) for the available scale(s). A somewhat more sophisticated approach is to sample n random draws from the distribution of θ_1, which is normal with

mean $\bar{\theta}_1$ and variance $V(\theta_1)$, compute the estimated category probabilities, and select the category with maximum average probability.

3 Illustration

To illustrate the general methodology, we obtained data from the 2000 Medical Expenditure Survey (MEPS). The MEPS is a nationally representative survey of the non-institutionalized U.S. population. Data are collected from families and individuals as well as their medical providers and employers. The MEPS was designed to provide accurate and representative estimates of national health care utilization and expenditure. Interviews are conducted at the household level with one respondent per family for approximately five interviews over a two-year span. The survey collects detailed information on respondents' demographics, health care utilization and expenditures, self-reported medical conditions, insurance coverage, and socioeconomic status. For a subset of respondents, the self-reported utilization data are validated and complemented with medical providers' interviews. In the year 2000, the MEPS asked respondents to complete both the EQ-5D and the SF-12 questionnaires. The questionnaires that included the EQ-5D and SF-12 instruments were part of a self-administered instrument that contained questions deemed to be unreliable if reported by a proxy. This questionnaire was distributed by mail to adults older than 17 as of July 2000 (Lawrence and Fleishman 2004; Agency for Healthcare Research and Quality 2005; Fleishman 2005). From the 15,151 respondents who were asked to complete the EQ-5D and SF-12 instruments in the year 2000, we selected 12,950 subjects who had no missing data in any of the items in both surveys. The average age of the subjects was 44.11 (SD = 16.8) and 46.55 % were male.

3.1 EQ-5D

The EQ-5D is a generic, non-disease-specific instrument for describing and valuing health-related quality of life (Brooks et al. 2003). As shown in Table 1, the EQ-5D descriptive system consists of five questions addressing five domains of health: mobility, self-care, usual activities, pain and discomfort, and anxiety and depression. Each question has three possible answers that capture a respondent's ability to perform each of the five domains of health: no problems, some or moderate problems and extreme problems or unable to perform the activity. The EQ-5D describes a total of 3^5 (243) possible response patterns, each defining a so-called "health state." Perfect health is defined as having no problem in any of the five domains, while the worst possible state is being unable to perform any of the five activities.

To transform the EQ-5D descriptive system into a measure that represents preferences, health states are scored using an algorithm derived from a valuation study. The algorithm to score the MEPS EQ-5D questionnaire was obtained from a valuation study conducted in the U.S. (Shaw et al. 2005). The study sample represented the civilian non-institutionalized U.S population, consisting of English- and Spanish-speaking adults, aged 18 and older who resided in the continental United States in 2002. The valuation study involved extensive interviews in which respondents were asked to evaluate a subset of health states using the time trade-off (TTO) approach (Torrance et al. 1972). In addition to the 243 states from the EQ-5D descriptive system, two states were added: unconsciousness and immediate death. Linear models were then employed to predict valuations for all possible health states. The

Table 1 EQ-5D and SF-12 items and response categories

Instrument	Item
EQ-5D	(1) Mobility
	1. I have no problems walking about
	2. I have some problems walking about
	3. I am confined to bed
	(2) Self-care
	1. I have no problems with self-care
	2. I have some problems washing or dressing myself
	3. I am unable to wash or dress myself
	(3) Usual activities (e.g. work, study, housework, family or leisure activities)
	1. I have no problems with performing my usual activities
	2. I have some problems with performing my usual activities
	3. I am unable to perform my usual activities
	(4) Pain/discomfort
	1. I have no pain or discomfort
	2. I have moderate pain or discomfort
	3. I have extreme pain or discomfort
	(5) Anxiety/depression
	1. I am not anxious or depressed
	2. I am moderately anxious or depressed
	3. I am extremely anxious or depressed
SF-12	(6) In general, would you say your health today is:
	1. Excellent
	2. Very good
	3. Good
	4. Fair
	5. Poor
	The following two questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?
	(7) Moderate activities such as moving a table, pushing vacuum cleaner, bowling, or playing golf:
	1. No. Not limited at all
	2. Yes. Limited a little
	3. Yes. Limited a lot
	(8) Climbing several flights of stairs:
	1. No. Not limited at all
	2. Yes. Limited a little
	3. Yes. Limited a lot
	During the past 4 weeks, have you had any of the following problems with your work or regular activities as a result of your physical health?
	(9) Accomplished less than you would like?
	1. No
	2. Yes

Table 1 continued

Instrument	Item
	(10) Were limited in the kind of work or other activities
	1. No
	2. Yes
	During the past 4 weeks, were you limited in the kind of work you do or other regular activities as a result of any emotional problems (such as feeling depressed or anxious)
	(11) Accomplished less than you would like?
	1. No
	2. Yes
	(12) Didn't do work or other activities as carefully as usual:
	1. No
	2. Yes
	(13) 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
	2. A little bit
	3. Moderately
	4. Quite a bit
	5. Extremely
	The next three questions are about how you feel and how things have been during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks:
	(14) Have you felt calm and peaceful?
	1. All of the time
	2. Most of the time
	3. A good bit of the time
	4. Some of the time
	5. A little of the time
	6. None of the time
	(15) Did you have a lot of energy?
	1. All of the time
	2. Most of the time
	3. A good bit of the time
	4. Some of the time
	5. A little of the time
	6. None of the time
	(16) Have you felt downhearted and blue?
	1. None of the time
	2. Some of the time
	3. A good bit of the time
	4. Most of the time
	5. All of the time

Table 1 continued

Instrument	Item
	(17) During the past 4 weeks, how much of the time has your physical health or your emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)? 1. None of the time 2. Some of the time 3. A good bit of the time 4. Most of the time 5. All of the time

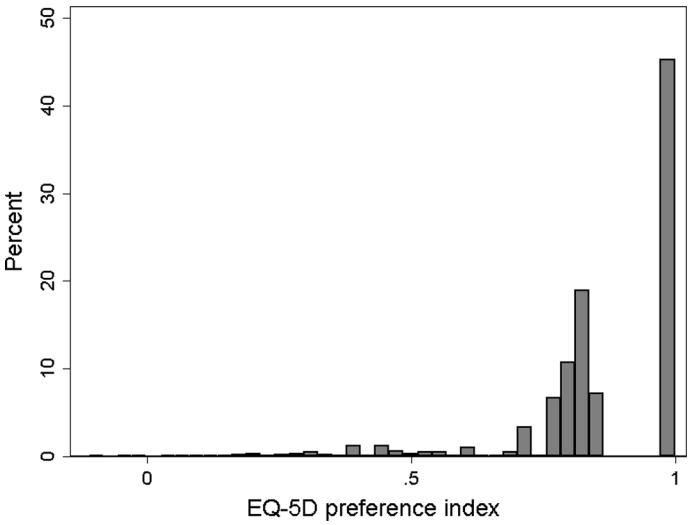


Fig. 1 Distribution of EQ-5D preference index

most valuable health state, perfect health, receives a maximum score of 1. On the other hand, immediate death was scored as 0; although some states were allowed to receive a negative score. Negative scores indicate that immediate death is preferable to the current health state (Agency for Healthcare Research and Quality 2005; Shaw et al. 2005). Figure 1 shows the distribution of the EQ-5D preference index. A large proportion of individuals (45.28 %) have an EQ-5D index of 1, the highest possible value that can be attained, indicating preferences for perfect health. The distribution exhibits three modes, at approximately 0.45, 0.83, and 1. The EQ-5D preference index is by design bounded between −0.109 and 1.

3.2 SF-12

The 12-item short-form (SF-12) health survey is an instrument derived from the longer 36-item short-form (SF-36) instrument (Ware et al. 1996). The purpose of the instrument is to measure general health functioning. The 12 questions of the SF-12 instrument measure

physical or emotional limitations, physical functioning, pain, general health, vitality, social functioning, and mental health problems. It provides two summary scores, the Physical Component Summary (PCS) and the Mental Component Summary (MCS). Higher SF-12 scores indicate better functioning in each domain. Scores are standardized; the mean score in the population is 50 with a standard deviation of 10 points. Table 1 shows all the items and possible responses for both instruments. For the analyses, we re-scaled some of the original SF-12 items so that the first answer indicates better health functioning.

3.3 Previous mapping studies

Several methods have been proposed for predicting the EQ-5D preference index from the SF-12 components using MEPS data. In general, these methods can be divided into two types depending on whether the prediction target is the EQ-5D preference index or the responses to the EQ-5D descriptive system. One of the earliest approaches using MEPS data focused on the prediction of the mean EQ-5D preference index using mean values of the physical and mental components of the SF-12 instrument (Lawrence and Fleishman 2004). Other research used individual-level data and ordinary least squares (OLS) regressions and found that the models explained approximately 63 percent of the total variance (Franks et al. 2004). Recognizing that the EQ-5D index is bounded at 1, researchers instead recommended the use of Tobit and censored least absolute deviations (CLAD) models instead of simple linear regression (Sullivan and Ghushchyan 2006). Two-part models, which are commonly used to model cost data bounded at zero (Duan et al. 1983), have also been proposed to account to the bulk of responses concentrated at EQ-5D values of 1 (Li and Fu 2009). In most regression approaches, the outcome is the observed EQ-5D preference index and the predictors are the mental and physical scores derived from the SF-12 along with a set of covariates such as demographic characteristics and comorbid conditions reported in the MEPS dataset. Most useful regression-based approaches limit the number of covariates to those variables also present in secondary data sources (e.g. sex and age). When EQ-5D preference index predictions are needed for a sample of patients with certain conditions rather than the general population, only observations with the same conditions are used from the MEPS to estimate regression coefficients.

Other studies have instead focused on predicting the EQ-5D descriptive system responses rather than the preference index. Some used a multinomial model for predicting the probability of answering a combination of item and response level (Gray et al. 2006) while others used Bayesian networks (Le and Doctor 2011).

A study comparing different methodologies (OLS, CLAD, two-part models, and multinomial models) using MEPS data concluded that OLS was the best method for predicting the overall preference index but that the accuracy of OLS deteriorated in less healthy groups (Chuang and Kind 2009). We compare EQ-5D preference index predictions derived from our IRT models with those obtained from OLS regression. Predictors included the mental and physical components of the SF-12 survey, age and its square, and an indicator for male.

To illustrate the general approach, we considered a simple bifactor model in which all 17 items (5 items from the EQ-5D and 12 from the SF-12) load on the primary well-being dimension and each scale loads on its own sub-domain. Thus, we assume that all questions are used to measure the same underlying construct. Predictions for each item of the EQ-5D were calculated using the estimates obtained from the bi-factor model assuming that only items from the SF-12 were available. For this application, we used the bi-factor generalization of the polytomous IRT model to accommodate the ordinal nature of the raw item

Table 2 Item factor loadings for primary dimension and scale-specific sub-domains

Instrument	Item	Description	Primary	EQ-5D	SF-12
EQ-5D	1	Mobility	0.853	0.336	0.000
	2	Self-care	0.805	0.382	0.000
	3	Usual activities	0.914	0.264	0.000
	4	Pain	0.813	0.120	0.000
	5	Anxiety/depression	0.667	−0.095	0.000
SF-12	6	General health	0.731	0.000	0.028
	7	Moderate activities	0.899	0.000	−0.239
	8	Climbing stairs	0.873	0.000	−0.232
	9	Accomplished less (physical)	0.926	0.000	−0.039
	10	Limited (physical)	0.945	0.000	−0.136
	11	Accomplished less (mental)	0.734	0.000	0.613
	12	Limited (mental)	0.720	0.000	0.579
	13	Pain	0.858	0.000	−0.064
	14	Felt calm	0.519	0.000	0.504
	15	Energy	0.691	0.000	0.299
	16	Felt blue	0.461	0.000	0.478
	17	Interference social activities	0.753	0.000	0.303

responses and variability in number of categories between scales. We assumed a normal ogive item response function with underlying bivariate normal trait distribution as described by Gibbons et al. (2007). Improvement in fit of the bi-factor model over a unidimensional alternative was determined using a likelihood ratio Chi square statistic (Bock and Aitkin, 1981). Absolute fit was evaluated by comparing observed and expected marginal category response rates.

4 Results

The bi-factor model significantly improved fit over a simple unidimensional alternative ($p < 0.0001$). Table 2 displays the estimated item factor loadings and Table 3 displays the estimated item thresholds. Table 2 reveals that all 17 items had strong loadings on the primary well-being dimension (from 0.46 to 0.95). In general the mental health items (EQ-5D item 5 and SF-12 items 14 and 16) had the lowest loadings. This is likely due to the majority of items from both scales measuring physical well-being. The EQ-5D has only one question related to mental health (item 5). The SF-12 subdomain was uniquely characterized by limitations at work produced by emotional problems (items 11 and 12) and emotional problems in general (items 14 and 16), which had lower loading on the primary dimension. The EQ-5D subdomain was uniquely characterized by mobility and self-care problems (items 1 and 2); however, these two items also loaded highly on the primary dimension.

Table 3 presents the estimated category thresholds for each of the 17 items. Items for which the lower category thresholds are smaller (e.g. SF-12 item 6 involving general health, where the highest category represents poor health), indicate that respondents more

Table 3 Item category thresholds

Instrument	Item	Description	1	2	3	4	5
EQ-5D	1	Mobility	0.754	2.403			
	2	Self-care	1.632	2.442			
	3	Usual activities	0.734	1.972			
	4	Pain	−0.040	1.587			
	5	Anxiety/depression	0.484	1.836			
SF-12	6	General health	−1.030	−0.036	0.909	1.790	
	7	Moderate activities	0.614	1.267			
	8	Climbing stairs	0.471	1.157			
	9	Accomplished less (physical)	0.637				
	10	Limited (physical)	0.696				
	11	Accomplished less (mental)	0.800				
	12	Limited (mental)	0.942				
	13	Pain	0.008	0.649	1.146	1.759	
	14	Felt calm	−1.168	0.059	0.579	1.243	1.868
	15	Energy	−1.318	−0.211	0.329	1.011	1.598
	16	Felt blue	−0.421	0.504	1.158	1.479	1.857
	17	Interference social activities	0.328	0.815	1.396	1.811	

commonly reported poor health relative to items like EQ-5D item 2, where very few respondents reported being unable to take care of themselves.

Table 4 presents the observed and estimated category proportions. In general, there appears to be close agreement between the observed and model-based estimates of the response proportions, confirming absolute fit of the model to the observed data.

Table 5 presents the summary statistics and correlations among the various primary well-being estimates for all 17 items, the 5 EQ-5D items and the 12 SF-12 items. For all three estimates the mean scores are close to zero and standard deviations are close to 1.0 as expected. The lower range of estimated primary trait scores appears to be underestimated by the EQ-5D, which could lead to disagreement between the observed and estimated EQ-5D scores when the estimate is based on the SF-12 items. This is born out in the correlations where the correlation between the SF-12 and the combined SF-12 and EQ-5D score is $r = 0.98$. By contrast, the correlation between the EQ-5D and the combined SF-12 and EQ-5D correlation is reduced to $r = 0.85$ and the correlation of the primary well-being scores between the estimates based on the EQ-5D items only and the estimate based on the SF-12 only is $r = 0.77$ (see Fig. 2).

Table 6 presents the results for prediction of the individual EQ-5D item responses from the primary well-being estimate based on the SF-12 items only. The overall percent agreement between the predicted and observed EQ-5D categories were 89, 96, 90, 74 and 77 % for the 5 EQ-5D items respectively. The lower levels of agreement were for pain and mental health. Sampling from the estimated distribution of the well-being estimates provide similar results and did not improve agreement (88, 97, 89, 71, and 76 %). It is interesting to note that these results were not dramatically different from results using the estimated score based on the EQ-5D to predict the actual EQ-5D response categories (89, 97, 96, 78, and 78 %). This suggests that it may be lack of model fit that contributes to the lower predictive accuracy for the pain and mental health items.

Table 4 Observed and estimated proportions

Instrument	Item	Observed proportions						Estimated proportions					
EQ-5D	1	0.823	0.173	0.004				0.775	0.217	0.008			
	2	0.963	0.033	0.004				0.949	0.044	0.007			
	3	0.823	0.162	0.015				0.769	0.207	0.024			
	4	0.538	0.421	0.040				0.484	0.460	0.056			
	5	0.725	0.250	0.024				0.686	0.281	0.033			
SF-12	6	0.170	0.370	0.315	0.115	0.030		0.152	0.334	0.333	0.145	0.037	
	7	0.784	0.142	0.074				0.730	0.167	0.103			
	8	0.739	0.170	0.092				0.681	0.195	0.124			
	9	0.798	0.202					0.738	0.262				
	10	0.815	0.185					0.757	0.243				
	11	0.833	0.167					0.788	0.212				
	12	0.866	0.134					0.827	0.173				
	13	0.562	0.233	0.109	0.067	0.028		0.503	0.238	0.132	0.087	0.039	
	14	0.135	0.426	0.188	0.161	0.068	0.022	0.121	0.402	0.195	0.174	0.076	0.031
	15	0.105	0.360	0.211	0.196	0.087	0.041	0.094	0.323	0.213	0.215	0.101	0.055
	16	0.364	0.354	0.174	0.051	0.036	0.021	0.337	0.356	0.184	0.054	0.038	0.032
	17	0.681	0.149	0.108	0.040	0.022		0.629	0.164	0.126	0.046	0.035	

Table 5 Estimates of the general well-being dimension

Summary statistics	*N*	Mean	SD	Minimum	Maximum
All items (1–17)	12,950	−0.00001	0.938	−1.911	3.674
EQ-5D items (1–5)	12,950	−0.00007	0.824	−0.741	3.232
SF-12 items (6–17)	12,950	−0.00007	0.930	−2.033	3.329

Pearson correlation coefficients ($N = 12{,}950$)	All items	EQ-5D items	SF-12 items
All items	1	–	–
EQ-5D items	0.853	1	–
	<0.001		
SF-12 items	0.983	0.771	1
	<0.001	< .001	

Table 6 Prediction of EQ-5D total scores from estimated item responses

(1) Observed and expected total score (summed for items 1–5)

Primary factor from:	Observed total score (0) ($N = 12{,}950$)		Expected total score (E) ($N = 12{,}950$)		O–E ($N = 12{,}950$)		Correlation of O & E (SE)
	Av	SD	Av	SD	Av	SD	
Items 6–17	6.215	1.564	5.905	1.479	0.309	0.947	0.808 (<0.0001)
Items 6–17[a]	6.215	1.564	5.965	1.425	0.249	0.882	0.830 (<0.0001)
Items 1–5	6.215	1.564	5.789	1.440	0.428	0.573	0.931 (<0.0001)
Items 1–17	6.215	1.564	5.917	1.476	0.298	0.809	0.860 (<0.0001)

[a] 100 of a primary theta drawn from the normal distribution (mean of est. theta_i, var of est. theta_i) for each i are used to calculate probability

Table 7 Summary statistics observed versus predicted EQ-5D preference index

	N	Mean	SD	Minimum	Maximum
Observed	12,950	0.861	0.164	−0.109	1.000
Predicted	12,950	0.925	0.124	−0.038	1.000

When the estimated category scores for the EQ-5D are used to compute the observed EQ-5D total score, we find that the estimated total scores are underestimated by approximately 5 %. Using the SF-12 based predictions of the EQ-5D item responses, correlations of $r = 0.81$ and $r = 0.83$ were observed for the maximum predicted category and for the average of 100 draws from the distribution of the SF-12 estimated well-being score. The upper bound on attainable correlation was $r = 0.93$ which is what was found for predicting the EQ-5D item responses from the estimated primary well-being score based on the observed EQ-5D responses.

For economic evaluations, the scale of interest is the EQ-5D preference index. We scored both the observed and expected EQ-5D response patterns for each of the 12,950

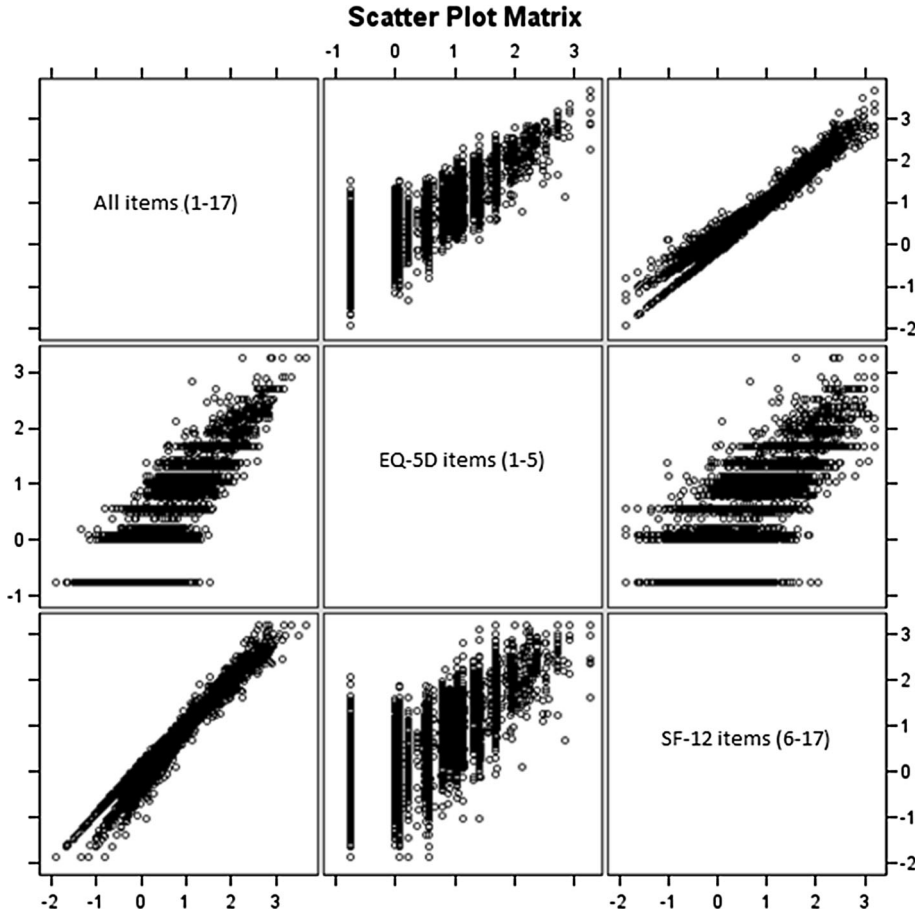


Fig. 2 Joint distributions of well-being scores based on complete and partial information

respondents using the standard scoring algorithm for the U. S population (Agency for Healthcare Research and Quality 2005). Table 7 presents summary statistics for the observed and predicted EQ-5D preference index. The IRT model is able to predict well the observed characteristics of the preference index. The estimated mean square error (MSE) of the prediction is 0.012 and the correlation between the observed and predicted preference index is 0.82. Our IRT model performed better than OLS regression. Using the same sample, the MSE of predictions from OLS regression is 0.014, with a correlation between observed and predicted preference index of 0.725. These results are similar to those obtained by Chuang and Kind 2009, who found that OLS was the best regression-based method.

5 Discussion

There are several areas of research and practice which require the imputation of an unmeasured instrument score from an available instrument score. Traditionally, this has

been done using regression methods for the observed scores of one instrument to predict the unobserved score, sometimes supplemented with auxiliary information such as demographic characteristics of the subject. This is important in research synthesis where different measures may have been used in different studies and a re-analysis of the data rather than a meta-analysis is desired. This is also important in health-based quality of life assessments where we seek a preference-based measurement (e.g. the EQ-5D) but the only available data are for a non-preference based assessment (e.g. the SF-12).

In this paper, we have taken a different approach to this problem based on equating true scores rather than observed scores using multidimensional item response theory. Using an IRT model that has been jointly calibrated for two or more instruments, it is possible to estimate the true score of interest when only a subset of the instruments has been administered. In our example, we jointly calibrated a bifactor IRT model to the EQ-5D and SF-12 in a sample that administered both, and then examined the accuracy of predictions assuming that one of the two scales was missing. A byproduct of the IRT approach is that we can also use the estimated true score based on one test (e.g. SF-12), to obtain estimated individual item responses for the other test (e.g. EQ-5D) that have maximum probability. We have suggested two different approaches to do this, one based on the point estimate of the true score and the other based on sampling from the distribution of the true score.

A key advantage of the IRT approach to predict the EQ-5D instrument is that this method predicts the response patterns rather than the preference index. After applying country-specific scoring algorithms to the response patterns, the resulting predicted preference index preserves the observed characteristics of the index: scores are bounded, have multiple modes, and a large proportion of observed scores are equal to 1. Further work on this example should consider a bifactor model based on the substantive subdomains of mental and physical health. Applications related to research synthesis should be studied in greater detail as well.

Acknowledgments This work was supported by NIMH grant MH66302 (RDG) and AHRQ Grants 1U18HS016973 (RDG) and T32HS000084 (MCP).

References

Agency for Healthcare Research and Quality: Calculating the U.S. Population-based EQ-5D™ Index Score. http://www.ahrq.gov/professionals/clinicians-providers/resources/rice/EQ5Dscore.html (2005). Accessed 1 June 2013

Bock, R.D., Aitkin, M.: Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika **46**, 443–459 (1981)

Bock, R.D., Gibbons, R.D.: Factor analysis of categorical item responses. In: Nering, M., Ostini, R. (eds.) Handbook of Polytomous Item Response Theory Models: Development and Applications. Lawrence Erlbaum, Florence (2010)

Brooks, R., Rabin, R., Charro, F.D.: The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective: Evidence from the EuroQol BIO MED Research Programme. Springer, New York (2003)

Chuang, L.-H., Kind, P.: Converting the SF-12 into the EQ-5D. Pharmacoeconomics **27**(6), 491–505 (2009). doi:10.2165/00019053-200927060-00005

Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P.: A comparison of alternative models for the demand for medical care. J. Bus. Econ. Stat. **1**(2), 115–126 (1983). doi:10.2307/1391852

Fleishman, J.A.: Demographic and Clinical Variations in Health Status. http://meps.ahrq.gov/data_files/publications/mr15/mr15.pdf (2005). Accessed 1 Nov 2013

Franks, P., Lubetkin, E.I., Gold, M.R., Tancredi, D.J., Jia, H.: Mapping the SF-12 to the EuroQol EQ-5D Index in a national US sample. Med. Decis. Mak. **24**(3), 247–254 (2004)

Gibbons, R.D., Hedeker, D.: Full-information item bifactor analysis. Psychometrika **57**, 423–436 (1992)

Gibbons, R.D., Bock, R.D., Hedeker, D., Weiss, D., Bhaumik, D.K., Kupfer, D., Frank, E., Grochocinski, V., Stover, A.: Full-information item bifactor analysis of graded response data. Appl. Psychol. Meas. **31**, 4–19 (2007)

Gray, A.M., Rivero-Arias, O., Clarke, P.M.: Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. Med. Decis. Mak. **26**(1), 18–29 (2006). doi:10.1177/0272989X05284108

Hedeker, D., Gibbons, R.D.: Longitudinal Data Analysis. Wiley, New York (2006)

Holzinger, K.J., Swineford, F.: The bifactor method. Psychometrika **2**, 41–54 (1937)

Joreskog, K.G.: A general approach to confirmatory maximum likelihood factor analysis. Psychometrika **34**, 183–202 (1969)

Lawrence, W.F., Fleishman, J.A.: Predicting EuroQoL EQ-5D preference scores from the SF-12 Health Survey in a nationally representative sample. Med. Decis. Mak. **24**(2), 160–169 (2004). doi:10.1177/0272989X04264015

Le, Q.A., Doctor, J.N.: Probabilistic mapping of descriptive health status responses onto health state utilities using Bayesian networks: an empirical analysis converting SF-12 into EQ-5D utility index in a national US sample. Med. Care **49**(5), 451–460 (2011). doi:10.1097/MLR.0b013e318207e9a8

Li, L., Fu, A.Z.: Some methodological issues with the analysis of preference-based EQ-5D index score. Health Serv. Outcomes Res. Methods **9**(3), 162–176 (2009). doi:10.1007/s10742-009-0053-3

Longworth, L., Rowen, D.: Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. Value Health **16**(1), 202–210 (2013). doi:10.1016/j.jval.2012.10.010

Muthen, B.O.: Latent variable modeling in heterogeneous populations. Psychometrika **54**, 557–585 (1989)

Rubin, D.B.: Inference and missing data. Biometrika **63**, 581–592 (1976)

Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychom. Monogr. Suppl. **35**, 139–139 (1969)

Shaw, J.W., Johnson, J.A., Coons, S.J.: US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med. Care **43**(3), 203–220 (2005)

Sullivan, P.W., Ghushchyan, V.: Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. Med. Decis. Mak. **26**(4), 401–409 (2006). doi:10.1177/0272989X06290496

Torrance, G.W., Thomas, W.H., Sackett, D.L.: A utility maximization model for evaluation of health care programs. Health Serv. Res. **7**(2), 118–133 (1972)

Tucker, L.R.: An inter-battery method of factor analysis. Psychometrika **23**, 111–136 (1958)

Ware Jr, J., Kosinski, M., Keller, S.D.: A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med. Care **34**(3), 220–233 (1996)