Published in final edited form as:

Med Decis Making. 2011 ; 31(2): 260–269. doi:10.1177/0272989X10380925.

Standard error of measurement of five health utility indexes across the range of health for use in estimating reliability and responsiveness

Mari Palta, Ph.D.¹, Han-Yang Chen, M.S.², Robert M. Kaplan, Ph.D.³, David Feeny, Ph.D.⁴, Dasha Cherepanov, Ph.D.³, and Dennis Fryback, Ph.D.¹

¹ Department of Population Health Sciences, University of Wisconsin, Madison, WI

² Center for Urban Population Health, Milwaukee, WI

³ Department of Health Services, University of California, Los Angeles, CA

⁴ Center for Health Research, Kaiser Permanente Northwest Region, Portland, OR

Abstract

Background—Standard errors of measurement (SEMs) of health related quality of life (HRQoL) indexes are not well characterized. SEM is needed to estimate responsiveness statistics and provides guidance on using indexes on the individual and group level. SEM is also a component of reliability.

Purpose—To estimate SEM of five HRQoL indexes.

Design—The National Health Measurement Study (NHMS) was a population based telephone survey. The Clinical Outcomes and Measurement of Health Study (COMHS) provided repeated measures 1 and 6 months post cataract surgery.

Subjects—3844 randomly selected adults from the non-institutionalized population 35 to 89 years old in the contiguous United States and 265 cataract patients.

Measurements—The SF6-36v2™, QWB-SA, EQ-5D, HUI2 and HUI3 were included. An item-response theory (IRT) approach captured joint variation in indexes into a composite construct of health (theta). We estimated: (1) the test-retest standard deviation (SEM-TR) from COMHS, (2) the structural standard deviation (SEM-S) around the composite construct from NHMS and (3) corresponding reliability coefficients.

Results—SEM-TR was 0.068 (SF-6D), 0.087 (QWB-SA), 0.093 (EQ-5D), 0.100 (HUI2) and 0.134 (HUI3), while SEM-S was 0.071, 0.094, 0.084, 0.074 and 0.117, respectively. These translate into reliability coefficients for SF-6D: 0.66 (COMHS) and 0.71 (NHMS), for QWB: 0.59 and 0.64, for EQ-5D: 0.61 and 0.70 for HUI2: 0.64 and 0.80, and for HUI3: 0.75 and 0.77, respectively. The SEM varied considerably across levels of health, especially for HUI2, HUI3 and EQ-5D, and was strongly influenced by ceiling effects.

Limitations—Repeated measures were five months apart and estimated theta contain measurement error.

Contact information: Mari Palta, Department of Population Health Sciences, University of Wisconsin, 610 Walnut Street, Madison, WI 53726.

Presented at: Annual meeting of the International Society for Quality of Life Research (ISOQOL), October 2008. Society of Medical Decision Making, October 2008.

Conclusions—The two types of SEM are similar and substantial for all the indexes, and vary across the range of health.

Introduction

Preference-based indexes of health related quality of life (HRQoL) have been widely used to evaluate the utility of interventions and policies impacting health outcomes. They transform answers to questions describing health states into scores interpretable in absolute terms as anchored by 0, a level of health equivalent to dead and 1, full health. It is important that the questions defining a utility score reliably capture clinically important differences in health states. Lack of reliability greatly interferes with assessment of whether individual patients change, and increases the sample size needed to accurately determine the average impact of health conditions and interventions (1,2,3). Reliability is most commonly assessed by the intraclass correlation coefficient (ICC) (4), which strongly depends on the standard error of measurement (SEM). An important distinction between the two is that SEM is relatively sample independent, while the ICC also depends on the total variation of an index in the population under consideration (4, 5). Terwee et al. (1) recommend that investigators report the SEM of outcomes used in their research.

The SEM plays a direct role in estimating Guyatt's responsiveness statistic (6) a standardized measure often used to assess sensitivity of indexes to health interventions (7,8), or more informally the “signal to noise ratio”. Guyatt's measure reflects the error of measurement under “stable” conditions, obtained as $SEM \sqrt{2}$, and is equivalent to the responsiveness parameter arising from item response theory (IRT) (see e.g. Baker, 9). Norman, Wyrwich and Patrick (10) comprehensively discussed and compared this and other choices of standard deviation for computing change in standardized units. Others have redefined SEM, as also including variation under non-stable conditions (1). The latter definition includes a potential component of variation that reflects varying response to the health condition or treatment inducing the change. Other options use standard deviations in control groups or normative populations, which are influenced by the range of health present in the particular group or population. In this paper we will focus on the use of the standard error of measurement (SEM) under stable conditions, as it is an inherent property of an index, and relatively independent of the intervention or population under study.

Added interest in the SEM arises from application of HRQoL indexes in clinical practice. Hays, Harivar and Liu (11) recommended that a minimally important difference (MID) be estimated via anchor based methods, where the criterion for adequate responsiveness of a measure is whether its change is at least as large on the original raw scale as that produced by a difference in health that is small, yet perceived by an individual. A recent paper (1) demonstrated that MID must be compared to SEM to provide guidance on how useful an index could be when used to monitor changes in individual patients. For example, achieving 95% specificity and 80% sensitivity for a given MID requires that SEM be of magnitude only one fourth of the MID (1). Others (12, 13) have also recommended that both MID and standard scores be used, especially when comparing different instruments.

We will address the SEM of 5 preference scored indexes: SF6D_36v2, QWB-SA, EQ5D, HUI2 and HUI3 across the range of health in the general population. We compute two conceptually different forms of SEM. One is the usual test-retest standard deviation (SEM-TR), estimated from repeated measures of patients from 3 cataract surgery clinics participating in the Clinical Outcomes and Measurement of Health Study (COMHS) (14). The 5 indexes were obtained at 1 and 6 months post surgery, thereby providing repeated measures. Various HRQoL and clinical measures used in COMHS indicate that the period from 1 to 6 months after cataract surgery was one of stability for the overwhelming majoring

of patients. The other is a structural standard deviation (SEM-S) of each index around an item-response theory (IRT)-derived composite construct (“theta”) of overall underlying health status captured by the joint variation of indexes in the National Health Measurement Survey (NHMS) (15, 16). In the NHMS, 5 preference scored instruments were administered via telephone to a national sample of 35-89 olds. The SEM-S is similar in principle to the total error reported for the HUI2 by Torrance et al. (17). We find that SEM-TR and SEM-S are of similar magnitude overall, and take advantage of the large sample size of the NHMS to estimate SEM-S separately at different levels of underlying health. For comparison with other studies, we also report overall reliability coefficients computed from the estimated standard deviations in NHMS.

Methods

Study design

The methodology of the National Health Measurement Survey (NHMS) has been previously described (15). Briefly, the NHMS employed a random digit dialed (RDD) telephone interview of a nationally representative sample of non-institutionalized adults in the U.S. aged 35 to 89 years. U.S. telephone exchanges were divided into strata with very high, high, medium, and low percentages of blacks. The sample was differentially drawn from these strata under a pre-allocated sampling design that increased the yield of black households in the sample that was called, yet allowed later statistical adjustment back to the U.S. population. The sampling also over-represented older adults. Of eligible respondents, 3,853 completed the interview, corresponding to an estimated response rate of 56%. During the initial data cleaning process, self-reported age could not be determined or was outside the specified sampling frame (i.e., age 35-89) for nine respondents, and these cases were eliminated from the analytic dataset, leaving a final sample size of 3,844. Trained interviewers at the University of Wisconsin Survey Center conducted the interviews from June 2005 through August 2006, using computer assisted telephone interview (CATI) software.

Distributions of the demographic characteristics of the NHMS survey sample, and population norms by gender for non-institutionalized U.S. adults aged 35 to 89 have previously been published for each of the 5 HRQoL indexes (15).

Patients age 35 and above undergoing cataract surgery were participants in the Clinical Outcomes and Measurement of Health Study (COMHS). These patients were recruited at 3 clinics (UCSD, UCLA, and UW-Madison) and self-administered mailed questionnaires prior to surgery and at 1 and 6 months post surgery. Of 378 patients entering the study, after deleting 3 patients above age 89 and one with a change of over 1.0 on the HUI3, 265 had repeated measures at 1 and 6 months on at least one of the preference scored indexes and were available for use in our analysis.

The SF-36v2™, QWB-SA, EQ-5D and HUI questionnaires (18-22, 17, 23, 24) were administered in randomized order across respondents in the NHMS and collated in randomized order for COMHS. In both studies, each measure was scored using the algorithm appropriate to or distributed with the measure. The algorithms yield summary scores for five indexes SF6D_36v2, EQ-5D (US scoring system, 22), QWB-SA, HUI2 and HUI3 that represent overall HRQoL anchored by 0.0 (dead) and 1.0 (full health). The HUI2, HUI3, and EQ-5D allow for scores less than zero, representing “health states worse than dead”. SF6D_36v2 scoring ranges from 0.30 to 1.0, QWB-SA from 0.09 to 1.0, EQ-5D from -0.11 to 1.0, HUI2 from -0.03 to 1.0, and HUI3 from -0.36 to 1.0. Previous analyses (15) showed the EQ-5D, HUI2 and HUI3 to have skewed distributions with ceiling effects, while the SF6D_36v2 and QWB-SA had population distributions closer to the normal distribution.

Statistical analysis

Except when otherwise specified, analyses were generated by SAS/STAT software, Version 9.1 of the SAS System for Unix, Copyright 2002-2003, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

SEM assumed constant across health—We first assumed the SEM to be constant across health and obtained two different estimators, the overall test-retest SEM-TR based on repeated measures, and overall structural SEM-S based on the variation of the indexes around the construct they have in common. Reliability coefficients for each index were also estimated from the two samples.

Data from the cataract surgery group were used to directly estimate SEM-TR as the standard deviation of the difference between time points divided by the square root of 2. Here, and in the computation of reliability, we assume that both overall variance and error variance is equal at the two time points, an assumption that is likely to hold. The approach adjusts for any systematic trend in the index between the time points, but can include variation created by non-systematic individual changes in health during the 1-6 months post cataract surgery. The mean trend between time points was statistically non-significant, except for the EQ-5D, which demonstrated a downward shift of 0.028 between time points ($p=0.0011$ by t- and signed rank tests).

Estimation of the structural standard deviation SEM-S from NHMS utilized an item response theoretic (IRT) approach previously applied to the data set to capture the joint variation of the 5 indexes (16). The approach models the common entity captured by the indexes (referred to as “theta”) on a standardized (mean=0, SD=1) nearly normally distributed scale using SCORIGHT (25). The indexes were analyzed as categorized into intervals defined by <0 , $0 - <0.25$, $0.25 - <0.5$, $0.5 - <0.75$, $0.75 - <0.95$, $0.95 - 1$. SCORIGHT uses Bayesian estimation of Samejima's (26) ordinal response model via Markov Chain Monte Carlo (MCMC) techniques (27). SCORIGHT is one of several IRT programs which allow inter-item correlations above and beyond those induced by the theta common to all items. This latter property was needed to account for non-independence of the HUI2 and HUI3. Model fit was assessed by inspection of fit plots for the 5 indexes and chi-square tests obtained by MODFIT software (28) and was found to be adequate. Theta was re-estimated with other and more densely spaced cutpoints for the indexes, and by alternative software, and the resulting estimates correlated very highly.

R-square estimates from the IRT model cannot be directly applied to obtain estimates of SEM-S of the indexes on their preference scored near continuous scale. Instead, we regressed each index on estimates of theta. To minimize collinearity of the predictors with the index, we did not use the previously published (16) original theta estimates based on all five indexes as predictor values in these regressions, but produced four new sets of theta estimates. These were obtained based on four subsets: SF-6D, QWB-SA and EQ-5D (i.e. leaving out the HUI based indexes), and as all combinations of HUI2, HUI3 with 2 of the other 3 measures.

Subsequent analyses of NHMS data utilized post-stratified survey weights to make estimators of variation represent the underlying US population. The SEM-S of each index was obtained from the residual variation of the index around a weighted least squares fitted regression curve of the index on a 5th degree polynomial in the respective non-collinear thetas. A high degree polynomial was used to ensure model fit, assessed to be satisfactory by inspecting residual plots and plots of observed versus modeled mean scores.

Finally, as we are interested in the standard deviation of each index around the true underlying construct of health the residual standard deviation was adjusted for estimation error in theta. Although we use several indexes to estimate theta, some measurement error remains, and the residual variance of an index around a theta estimated with error will be inflated compared to the standard deviation around the true value. The R^2 will be reduced proportionally to the reliability (denoted here by λ) of the estimated predictor values (29). We apply the correction to R^2 appropriate for the linear case as a reasonable approximation also in our case of polynomial regression. The quantity λ is obtained from the respective IRT models (30). The following formula was applied to obtain the adjusted SEM:

$$\text{SEM-S} = \sigma_{\text{index}} (1 - R^2_{\text{adjusted}})^{1/2} = \sigma_{\text{index}} (1 - R^2_{\text{regression}/\lambda})^{1/2}$$

where σ_{index} is the weighted estimator of the population standard deviation of the index and $R^2_{\text{regression}}$ is obtained from the weighted regression of the index on the respective thetas.

The reliability of each index was estimated by the formula $1 - (\text{SEM} / \sigma_{\text{index}})^2$, where SEM is adjusted SEM-S or SEM-TR and σ_{index} is the standard deviation estimate from above for SEM-S and the standard deviation in an index 1 month post surgery as estimated from COMHS.

SEM allowed to vary across health—Residual plots from the NHMS and Bland-Altman plots based on COMHS (differences between two repeated measurements plotted against the mean of the same two measurements) (31, 32) demonstrated that neither version of SEM was constant across health levels for most of the indexes. We then utilized the large sample size of the NHMS to estimate SEM-S within different ranges of health.

We estimated SEM-S from NHMS within subgroups defined from original theta estimates (16) by cut points -2, -1.5, -0.5, 0.5 and 1 SD from mean theta. Original theta was used only to define these subgroups in a uniform manner and to translate the cut points into corresponding expected values of each index via a 5th degree polynomials. The SEM-S within each subgroup was estimated as the standard deviation of the residuals around the fifth degree polynomial in the new non-collinear thetas described above, adjusted for estimation error in theta via multiplication by the ratio $(1 - R^2_{\text{regression}})^{1/2} / (1 - R^2_{\text{adjusted}})^{1/2}$ arising from above.

For comparison with SEM-TR across health, interval specific estimates of SEM-S were multiplied by 2 to reflect the standard deviation of the difference in scores under stable conditions. These were further multiplied by 1.96 and superimposed on Bland-Altman plots of cataract patient data to assess the fit of the interval specific SEM-S estimates to the repeated measures. The repeated measures difference would be expected to fall within these limits approximately 95% of the time.

Results

A description of both samples is provided in Table 1, as well as of the population underlying NHMS. The cataract sample was slightly older, included a greater percentage of white race and was better educated. Descriptive statistics on the 5 indexes are in Tables 2 and 3 and show that the means of the indexes and the percentage at the ceiling of indexes are higher in the population underlying NHMS. Table 2 also shows estimates of SEM-TR and reliability coefficients for the indexes based on the COMHS cataract sample.

Table 3 shows population estimates from the NHMS of unadjusted and adjusted SEM-S, as well as the estimated reliability coefficients of the theta estimates used as the predictor in the regression for each index, and the estimated reliability of the indexes themselves. The four

sets of theta estimates used as predictors in the regression models, all correlated at 0.95 and above with the original theta estimates. The original theta estimates had estimated reliability of 0.87, and reliabilities of those based on subsets of indexes (as in Table 3) were only slightly lower. Comparison of results in Tables 2 and 3 shows that SEM-TR and adjusted SEM-S estimates are quite consistent. Reliability coefficients for the indexes are lower in the cataract sample.

Bland-Altman plots based on data from months 1 and 6 in the cataract group are shown in Figures 1-5. We see that SEM-TR as reflected in the absolute size of the differences, tends to be less for index values near 1, except for the QWB-SA. The non-constancy of the standard deviation is particularly striking for the HUI2 and HUI3 and for the EQ-5D.

The essential results of our analyses along the spectrum of underlying health are summarized in Table 4. The first 5 rows show how the cut points used for categorizing health correspond to index values as predicted from the model of each index on the original estimates of theta. It is clear that the indexes take on quite different preference scored values for similar levels of estimated overall health.

The second block of entries is the estimated SEM-S within intervals. We see that SEM-S is quite similar across all the indexes close to mean theta. Standard deviations for the EQ-5D, HUI2 and HUI3 are much larger at low values of health (theta) and very small close to their ceiling. For the EQ-5D, HUI2 and HUI3, 88%, 15% and 17% of the population was estimated to fall at the ceiling of 1 in the interval 0.5-1.0 of underlying health (theta) and 99%, 55% and 55% at the ceiling, respectively, in the interval >1.0 of overall health. In comparison 2% in the 0.5-1.0 and 16% in the >1.0 interval fell at the ceiling of the SF6D_36v2.

Intervals constructed from interval specific adjusted SEM-S estimates from NHMS to capture 95% of the differences between 6 and 1 month time points in the cataract sample are superimposed on Figures 1-5. The intervals follow the contours of differences well, except that the scarcity of observations at the lowest values of indexes makes it difficult to assess fit in this range. Close to expectation, for SF-6D 95% of differences fell inside the interval, for QWB-SA 95%, for EQ-5D 96%, for HUI2 92% and for HUI3 94%.

Discussion

Several methods of estimating SEMs of 5 commonly used preference scored HRQoL indexes showed these standard deviations to be substantial, and in most ranges of health, well above an often used value for “minimally important difference” (MID) of 0.03-0.04 (33-35), although values of MID as high as 0.07 have been suggested (36). According to previous literature, this would make the indexes investigated inappropriate for individual patient monitoring (1), although it must be recognized that HRQoL indexes and their subscales may often be used only as ancillary to other information. A recent publication provides guidance on how to apply SEM in assessing the uncertainty in clinical change scores (37).

Indexes differed in the magnitude of their SEMs with the HUI3 having the largest and the SF6D_36v2 the smallest standard deviation. This conclusion held for both SEM-TR based on test-retest, and SEM-S based on variation of each index around a joint construct of underlying health. Importantly, SEM varied considerably across the range of health, so that average SEM depends on the population composition. Our SEM estimates may be helpful in choosing the most precise index for a certain range of health. However, ceiling effects play a central role and cause SEM to be artificially small close to the maximum index value of 1. SEM in the mid range of health is quite comparable across indexes.

Reliability coefficients for health outcomes measures can be estimated using a variety of methods. The common element is the creation of a ratio of true to observed variance. Some investigators use measures of internal consistency, while others use estimates derived from repeated applications of the measures to the same populations. This analysis primarily uses a method that depends on several modeling assumptions. Nonetheless the reliabilities of the computed from the estimated SEM fell firmly within ranges of previously reported values except for the QWB-SA (38). In the latter overview, reliability coefficients were tabulated from a range of disease specific and community studies, with the middle of the range being 0.71 for SF-6D, 0.72 for EQ-5D and 0.76 for the HUI3. From the NHMS we have 0.71 for SF-6D, 0.70 for EQ-5D and 0.77 for HUI3. As noted previously, however, reliability coefficients are dependent on the range of health in the population under study, and COHMS does indeed provide lower estimates. A population based study in Canada (39), arrived at a reliability estimate of 0.77 for the HUI3, which is identical to our NHMS estimate. The reliability coefficients for the indexes estimated by us and others are adequate or almost adequate for population studies (40)

Our estimates for QWB-SA reliability of 0.59 and 0.64 are well below the reliability of 0.90 previously reported. However, previous estimates of QWB reliability used an entirely different methodology. It may be noted that QWB-SA was found the least strongly related to the construct of underlying health in the IRT analysis (16), and the reliability estimate from NHMS may therefore reflect some unique variance being included in SEM-S. The IRT analysis identifies common variance across measures. While all five indexes include items on physical and emotional health and symptoms such as pain and discomfort, the QWB-SA differs from other measures because it includes an extensive set of items on symptoms and health problems, some of which are acute. The unique symptom-problem content of the QWB-SA may explain why the QWB-SA was less strongly related to the shared construct and some of the variability between visits in the COHMS. Hence, the reliability of QWB-SA may have been underestimated in our analyses.

We further found that SEM varies across the range of health, although less for the QWB-SA and the SF6D_36v2 than for HUI2, HUI3 and EQ-5D. This non-constancy may lead to misleading estimates of responsiveness and reliability from studies of patients representing a limited range of health. For example ceiling effects may lead to underestimation of SEM and corresponding overestimation of reliability and responsiveness in healthy samples. Notably, our overall SEM is estimated as lower from the NHMS where the percentages falling at the ceiling of indexes are higher than in the cataract sample. The differences in SEM between indexes also somewhat mirror the differences in index ranges, where the minimum observed value of the HUI3 is -0.34, but of the SF6D_36v2 is as high as 0.30. Hence they are partly explained by index scaling. Our results (Table 4) provide some insight into signal to noise ratio in different ranges of health, and show that different indexes may be best in different ranges. However, we found the signal to noise ratio more sensitive to modeling choices, such as cut-points chosen for the indexes in the IRT model than were the SEM estimates themselves.

We estimated two conceptually different SEMs across two separate samples representing a general population, and post-cataract surgery patients. Given these differences, the similarity of the results is surprising and reassuring. Nonetheless, some caution is in order.

The structural SEM-S in the general population, around the underlying measure of health contains some unique variance, i.e. sensitivity of an index to health conditions not reflected in the other indexes. The unique variance would be considered measurement error if the goal is to estimate the core construct of health common to all indexes, but not if the goal is to measure the construct represented by the specific index itself. On the other hand, some

collinearity in the prediction models underlying SEM-S may have remained and have led to underestimation. Such collinearity may have arisen from correlated errors in responses to questions that are similar between indexes.

Test-retest SEM-TR from the cataract sample almost surely contains variance due to short term fluctuations in health such as due to acute illness episodes. Hence, SEM-TR is quite likely an overestimate of SEM as short term health fluctuations would be considered measurement error if the goal is to measure the impact of chronic illness only. Our study of SEM-TR has the weakness of not having access to repeated measures closer than 5 months apart, although stability of long term health is difficult to confirm in any study. Short time intervals are well known to raise the alternative problem of recall bias.

Our method to adjust for reliability of theta is not precise. First of all, the reliability coefficient used was derived from an IRT procedure that did not take sampling weights into account (16). We adopted this approach to be faithful to our previously published methodology, and also because different methods attempting to produce weighted reliability coefficients did not yield consistent results. Second, the method of adjustment is technically correct only when linear relationship is used to predict index scores and for the overall estimates of SEM. The complexity of our model precluded a more exact solution. In spite of these caveats, SEM-TR and unadjusted and adjusted are all close enough to provide a reasonably narrow range for the size of SEM for the five indexes. In addition, intervals constructed from SEM-S capture close to the expected percentage of differences between time points from the repeated measures.

In addition to generating better understanding of preference scored indexes, our analysis provides guidelines on the magnitude of SEMs of indexes, which should be useful in assessing responsiveness in studies too small to provide reliable internal error standard deviation estimates.

Acknowledgments

Grant: This research was supported by grant P01-AG020679 from the National Institute on Aging.

References

1. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HCW. Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol.* 2009; 62:62–67. [PubMed: 18619810]
2. Miller JD, Malthaner RA, Goldsmith CH, Goeree R, Higgins D, Cox PG, Tan L, Road JD. for the Canadian Lung Volume Reduction Surgery Study. A randomized clinical trial of lung volume reduction surgery versus best medical care for patients with advanced emphysema: A two-year study from Canada. *Ann Thorac Surg.* 2006; 81:314–21. [PubMed: 16368389]
3. Ploeg J, Brazil K, Hutchison B, Kaczorowski J, Dalby DM, Goldsmith CH, Furlong W. Effect of preventive primary care outreach on health related quality of life among older adults at risk of functional decline: randomised controlled trial. *BMJ.* 2010; 340:c1480. [PubMed: 20400483]
4. Nunnally, JC.; Bernstein, IH. *Psychometric Theory.* 3rd. New York: McGraw-Hill; 1994.
5. Shrout PE, Spitzer RL, Fleiss JL. Quantification of agreement in psychiatric diagnosis revisited. *Arch Gen Psychiatry.* 1987; 44:172–77. [PubMed: 3813814]
6. Guyatt G, Walter GS, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 1987; 40:171–8. [PubMed: 3818871]
7. Blanchard C, Feeny D, Mahon JL, Bourne R, Rorabeck C, Stitt L, et al. Is the Health Utilities Index valid in total hip arthroplasty patients? *Qual Life Res.* 2004; 13:339–48. [PubMed: 15085906]

8. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Qual Life Res.* 2003; 12:349–62. [PubMed: 12797708]
9. Baker, F. The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation. College Park, MD: University of Maryland; 2001.
10. Norman GR, Wyrwich KW, Patrick DL. The mathematical relationship among different forms of responsiveness coefficients. *Qual Life Res.* 2007; 16:815–22. [PubMed: 17351823]
11. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *Journal of Chronic Obstructive Pulmonary Disease.* 2005; 2:63–7. [PubMed: 17136964]
12. Guyatt GH, Osoba B, Wu AW, Wyrwich K, Norman GR. the Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings.* 2002; 77:371–83. [PubMed: 11936935]
13. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. The Clinical Significance Consensus Meeting Group. Estimating clinically significant differences in quality of life outcomes. *Qual Life Res.* 2005; 14:285–95. [PubMed: 15892420]
14. Kaplan RM, Hays RD, Feeny D, Palta M, Ganiats TG, Fryback DG. Five preference-based indexes in cataract and heart failure patients were not equally responsive to change. *J Clin Epidemiol.* 2010 in press.
15. Fryback DG, Dunham NC, Palta M, Hanmer J, Buechner J, Cherepanov D, et al. U.S. norms for six generic health-related quality-of-life indexes from the National Health Measurement Study. *Med Care.* 2007; 45:1162–70. [PubMed: 18007166]
16. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim JS. Comparison of five health-related quality – of-life indexes using item response theory analysis. *Med Decis Making.* in press.
17. Torrance GW, Feeny DH, Furlong WJ, Barr R, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system Health Utilities Index Mark 2. *Med Care.* 1006; 34:702–22. [PubMed: 8676608]
18. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics.* 2002; 21:271–92. [PubMed: 11939242]
19. Kaplan RM, Sieber WJ, Ganiats TG. The Quality of Well-being Scale: Comparison of the interviewer-administered version with a self-administered questionnaire. *Psychology & Health.* 1997; 12:783–91.
20. Andresen EM, Rothenberg BM, Kaplan RM. Performance of a self-administered mailed version of the Quality of Well-Being (QWB-SA) questionnaire among older adults. *Med Care.* 1998; 36:1349–60. [PubMed: 9749658]
21. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med.* 2001; 33:337–43. [PubMed: 11491192]
22. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: Development and testing of the D1 valuation model. *Med Care.* 2005; 43:203–29. [PubMed: 15725977]
23. Feeny, D.; Torrance, G.; Furlong, W. Health Utilities Index. In: Spilker, B., editor. *Quality of Life and Pharmacoeconomics in Clinical Trials.* Philadelphia, PA: Lippincott-Raven Press; 1996.
24. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care.* 2002; 40:113–28. [PubMed: 11802084]
25. Wang, X.; Bradlow, ET.; Wainer, H. User's Guide for SCORIGHT (Version 3.0): A Computer Program for Scoring Tests Built of Testlets Including a Module for Covariate Analysis. Princeton, NJ: Educational Testing Service; 2005.
26. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, No 17. 1969
27. Wang X, Bradlow ET, Wainer H. A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement.* 2002; 26(1):109–128.
28. [accessed 1 July 09] MODFIT (Version 1.1, 2001_ StephenStark, IRT Modeling Lab, University of Illinois at Urbana-Champaign. http://work.psych.uiuc.edu/irt/mdf_modfit.asp
29. Fuller, WA. *Measurement Error Models.* New York: John Wiley & Sons; 1987.

30. Sireci SG, Thissen D, Wainer H. On the reliability of testlet-based tests. *Journal of Educational Measurement*. 1991; 28:237–47.
31. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician*. 1983; 32:307–17.
32. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; i:307–10. [PubMed: 2868172]
33. Kaplan RM. The minimally clinically important difference in generic utility-based measures. *COPD: Journal of Chronic Obstructive Pulmonary Disease*. 2005; 2:91–7.
34. Majumdar SR, Johnson JA, Bowker SL, Booth G, Dolovich L, Ghali WA, et al. A Canadian Consensus for the Standardized Evaluation of Quality Improvement Interventions in Type 2 Diabetes. *Canadian Journal of Diabetes*. 2005; 29:220–9.
35. Feeny, D. Preference-Based Measures: Utility and Quality-Adjusted Life Years,” Chapter 6.2. In: Peter, Fayers; Ron, Hays, editors. *Assessing Quality of Life in Clinical Trials*. 2nd. Oxford; Oxford University Press; 2005.
36. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health status utility values? The case of the SF-6D. *Health and Quality of Life Outcomes*. 2003; 1 electronic journal: <http://www.hqlo.com/content/1/1/4>.
37. de Vet HCW, Terluin B, Knol DL, Roorda LD, Mookink LB, Ostelo RWJG, Hendriks EJM, Bouter LM, Terwee CB. Three ways to quantify uncertainty in individually applied “minimally important change” values. *J Clin Epidemiol*. 2010; 63:37–45. [PubMed: 19540720]
38. Sinnott, PL.; Joyce, VR.; Barnett, PG. *Guidebook*. Menlo Park CA: VA Palo Alto, Health Economics Resource Center; 2007. *Preference Measurement in Economic Analysis*.
39. Boyle MH, Furlong W, Feeny D, Torrance G, Hatcher J. Reliability of the Health Utilities Index-Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. *Qual Life Res*. 1995; 4:249–57. [PubMed: 7613535]
40. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health surveys adequate? *Qual Life Res*. 1995; 4:293–307. [PubMed: 7550178]

Figure 1. Bland-Altman plots of SF6D_36v2 values at 6 months minus 1 month post cataract surgery (open circles) versus mean of same values. Intervals expected to contain 95% of points (solid circles) based on adjusted SEM-S estimated from the NHMS are superimposed.

Figure2.

Bland-Altman plots of QWB-SA values at 6 months minus 1 month post cataract surgery (open circles) versus mean of same values. Intervals expected to contain 95% of points (solid circles) based on adjusted SEM-S estimated from the NHMS are superimposed.

Figure 3.

Bland-Altman plots of EQ-5D values at 6 months minus 1 month post cataract surgery (open circles) versus mean of same values. Intervals expected to contain 95% of points (solid circles) based on adjusted SEM-S estimated from the NHMS are superimposed.

Figure 4.

Bland-Altman plots of HUI2 values at 6 months minus 1 month post cataract surgery (open circles) versus mean of same values. Intervals expected to contain 95% of points (solid circles) based on adjusted SEM-S estimated from the NHMS are superimposed.

Figure 5.

Bland-Altman plots of HUI3 values at 6 months minus 1 month post cataract surgery (open circles) versus mean of same values. Intervals expected to contain 95% of points (solid circles) based on adjusted SEM-S estimated from the NHMS are superimposed.

Table 1

Description of samples

Item	NHMS Sample		NHMS Population *	COMHS Cataract Sample	
	N	(%)		N	(%)
Gender					
Male	1641	42.7	47.1	102	39
Female	2203	57.3	52.9	163	62
Age, yr					
35-44	642	16.7	31.6	2	1
45-54	826	21.5	23.8	20	8
55-64	684	17.8	19.9	59	22
65-74	965	25.1	14.2	72	27
75-89	727	18.9	10.6	112	42
Race					
White	2562	66.7	81.2	232	88
Black	1086	28.3	10.5	9	3
Other races	178	4.6	7.6	19	7
Missing	18	1	0.7	5	2
Education (highest level)					
< High school	464	12.1	8.3	11	4
High School	1159	30.2	28.2	40	15
Some post-high school	856	22.3	22.1	73	28
4-yr college degree or higher	1341	34.9	40.9	137	52
Missing	24	1	0.5	4	2
Total	3844	100	100	265	100

* Estimates are weighted to non-institutionalized population age 35-89 in the 48 contiguous states of the US.

Table 2
Overall standard deviations based on 1 and 6 month visits in COMHS cataract group based on formula (1)

Index	N Cataract group	Mean (SD) of index at 1 months post-surgery	% at ceiling	SEM-TR Based on 1 and 6 month visits	Reliability
SF-6D	236	0.74 (0.12)	1.7%	0.068	0.66
QWB	265	0.62 (0.13)	0%	0.087	0.59
EQ-5D	246	0.85 (0.15)	36%	0.093	0.61
HUI2	247	0.82 (0.17)	5.3%	0.100	0.64
HUI3	245	0.73 (0.26)	4.5%	0.134	0.75

Table 3
Overall standard deviations based on model from NHMS and formula (2)

Index	N NHMS	Mean (SD) of index NHMS [*]	% at ceiling [*]	Unadjusted SEM-S [*]	Reliability of theta excluding instrument	Adjusted SEM-S ^{*,†}	Reliability [*]
SF-6D	3739	0.79 (0.13)	4.2%	0.084	0.84	0.071	0.71
QWB	3758	0.66 (0.16)	3.2%	0.107	0.83	0.094	0.64
EQ-5D	3812	0.87 (0.15)	43%	0.099	0.83	0.084	0.70
HUI2	3558	0.85 (0.16)	14%	0.101	0.79 [‡]	0.074	0.80
HUI3	3567	0.81 (0.24)	14%	0.152	0.79 [‡]	0.117	0.77

^{*} Estimates are weighted to non-institutionalized population age 35-89 in the 48 contiguous states of the US.

[†] Adjusted to account for measurement error in the theta used in regressions corresponding to each index.

[‡] Excluding both HUI2 and HUI3

Table 4
Structural standard deviation SEM-S of indexes within intervals of health

Latent Health (theta)	N*	Popula- tion	%at ceiling†	SF-6D	QWB-SA	EQ-5D	HUI2	HUI3
Interval endpoints								
Index values corresponding to theta interval endpoints‡								
-2				0.53	0.38	0.46	0.36	0.12
-1.5				0.60	0.46	0.63	0.53	0.33
-0.5				0.73	0.60	0.82	0.81	0.74
0.5				0.86	0.73	0.93	0.93	0.95
1				0.90	0.77	0.98	0.96	0.97
Interval								
SEM-S of indexes w/i interval‡								
theta -2.0	226		3.7%	0.064	0.093	0.173	0.138	0.182
-2.0 < theta	243		4.0%	0.058	0.097	0.157	0.130	0.203
-1.5 < theta	994		24%	0.079	0.083	0.081	0.102	0.177
-0.5 < theta	1295		35%	0.077	0.084	0.085	0.062	0.095
0.5 < theta	602		18%	0.059	0.115	0.070	0.041	0.048
theta >1.0	484		16%	0.057	0.103	0.024	0.026	0.031

* In NHMS sample

† Estimates are weighted to non-institutionalized population age 35-89 in the 48 contiguous states of the US

‡ Adjusted for estimation error in theta by multiplication by the ratio of adjusted to unadjusted from Table 2.