
CHERE
CENTRE FOR HEALTH ECONOMICS
RESEARCH AND EVALUATION


UNIVERSITY OF
TECHNOLOGY SYDNEY

UTS:CHERE

The Centre for Health Economics Research and Evaluation (CHERE) was established in 1991. CHERE is a centre of excellence in health economics and health services research. It is a joint Centre of the Faculties of Business and Nursing, Midwifery and Health at the University of Technology, Sydney, in collaboration with Central Sydney Area Health Service. It was established as a UTS Centre in February, 2002. The Centre aims to contribute to the development and application of health economics and health services research through research, teaching and policy support. CHERE's research program encompasses both the theory and application of health economics. The main theoretical research theme pursues valuing benefits, including understanding what individuals value from health and health care, how such values should be measured, and exploring the social values attached to these benefits. The applied research focuses on economic and the appraisal of new programs or new ways of delivering and/or funding services. CHERE's teaching includes introducing clinicians, health services managers, public health professionals and others to health economic principles. Training programs aim to develop practical skills in health economics and health services research. Policy support is provided at all levels of the health care system by undertaking commissioned projects, through the provision of formal and informal advice as well as participation in working parties and committees.

University of Technology, Sydney
City campus, Haymarket
PO Box 123 Broadway NSW 2007
Tel: +61 2 9514 4720
Fax: + 61 2 9514 4730
Email: mail@chere.uts.edu.au
www.chere.uts.edu.au

Validation and calibration of the SF-36 health transition question in the Household, Income and Labour Dynamics in Australia (HILDA) survey

Stephanie Knox[1], Madeleine King[1]

CHERE WORKING PAPER 2007/15

1. Centre for Health Economics Research and Evaluation
Faculty of Business
University of Technology, Sydney

First Version: December 2007
Current Version: December 2007

ABSTRACT

Background:

Cross-sectional population surveys depend on retrospective self-report if they are to estimate changes in health status over time. For example the health transition question (HTQ) from the SF-36 health survey asks the respondent to rate his/her health compared to one year ago, with response categories “much better”, “somewhat better”, “about the same”, “somewhat worse”, “much worse”. Transition questions are often used to anchor the interpretation of the magnitude of prospectively measured change in health status. However little has been done to calibrate the categories of health transition questions against other measures of clinically important change in health. Such a calibration of the health transition question would increase its usefulness as a stand alone item in cross-sectional surveys.

Methods:

The sample included 9,649 adults from the longitudinal Household Income and Labour Dynamics of Australia (HILDA) study, who had completed the SF-36 in wave 1 (2001) and wave 2(2002).

The anchor of important clinical change was having developed a new long-term condition(s) between wave 1 and wave 2, serious enough to affect daily activities.

Prospective changes were calculated as within-person differences in scores between interviews for each of the 8 SF-36 domain scales. Mean change scores adjusted for age and sex, and standardised response means were calculated for the categories of the HTQ and compared with the mean changes for the group with a known clinical change.

Results:

The adjusted mean change in scale scores and the standardised response means for the group who described their health as “somewhat worse” than one year ago were similar in magnitude to those who had developed a long-term health condition in the last year.

Conclusion:

In the context of a population survey, the group who described their health as “somewhat worse” than one year ago had on average experienced a decline in clinical health status of a magnitude equivalent to having developed one or more new long-term health conditions serious enough to affect daily activities. This calibration is useful for describing the average change in clinical health status for groups categorised by the HTQ. However, large variations in individual change scores within HTQ categories indicate that the HTQ alone does not have good predictive validity at the individual level.

Validation and calibration of the SF-36 health transition question in the Household, Income and Labour Dynamics in Australia (HILDA) survey

BACKGROUND

In cross-sectional surveys the only way to identify respondents who may have experienced an important recent change in health status is through retrospective self-report. One example is the Australian National Health Survey (NHS) a cross-sectional study of population health undertaken every 4 years that has used the global health transition question from the Short form 36 item (SF-36) general health survey as a retrospective measure of self-reported change in health status 'in the past year' (ABS 2006).

The SF-36 was designed as a generic health status instrument to be used in the context of both clinical and general population studies (Ware, Snow et al. 1993). The SF-36 includes 36 questions, 35 of which are aggregated into 8 scales measuring the domains of Physical Functioning, Physical Role, Bodily Pain, General Health, Vitality, Social Functioning, Emotional Role, and Mental Health. The additional global health transition question (HTQ) asks respondents to rate their general health compared with one year ago, with five categories of response, "Much better", "Somewhat better", "About the same", "Somewhat worse", "Much worse".

The SF-36 instrument has been used in major population studies in Australia, both longitudinal and cross-sectional (ABS 1997; Watson and Wooden 2004) and the Australian norms for the SF-36 version 1 were established using the 1995 NHS (ABS 1997). While the full SF-36 was dropped from the NHS in 2001, the HTQ was retained then dropped in 2005 (ABS 2006).

The use of brief items that validly measure health status can reduce the burden of questions on respondents and help minimise rates of non-response (Ware, Snow et al. 1993). However such items need to be valid and interpretable as measures of change in health status. Issues around the use of global health transition questions include the difficulty of establishing reliability for single items, the global or general nature of the question and the potential for recall bias (Norman, Sridhar et al. 2001).

There is evidence that retrospective self-report is biased towards the respondent's present health state (Norman, Stratford et al, 1997). If retrospective recall were unbiased it should be positively correlated with follow-up scores and negatively correlated with the baseline scores (Norman, Stratford et al, 1997). However studies have shown that retrospective recall of change is positively correlated with follow-up scores and either un-correlated or positively correlated with baseline scores (Norman, Stratford et al, 1997, Cella, Hahn et al. 2002). This indicates that respondents with good health at follow-up are more likely to assume that their health has recently improved, and respondents with poor health at follow-up are more likely to assume that it has worsened.

Despite this problem, retrospective transition questions have often been used as the external criterion or anchor of change against which to establish the size of the minimal important clinical change (MCID) for the domain scales of various instruments (Perneger et al, Garratt et al, Cella et al, Fitzpatrick et al). The MCID was initially defined by Jaeschke and colleagues as 'the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side-effects and cost, a change in the patient's management'. (Jaeschke, Singer et al. 1989). Other authors have extended the MCID concept to include both improvement and worsening in health status (Norman, Stratford & Regehr, 1997, Cella, Hahn & Kelly 2002). This extension was recently expressed in an update of Jaeschke et al's MCID definition by Guyatt and colleagues: 'the smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and which would lead the clinician to consider a change in the patient's management' (Guyatt, Osoba et al. 2002). The MCID is a key concept in planning sample size, since studies must be powered to detect the MCID. It is also a key concept in interpreting the results of studies, particularly those where sample size has been based on another outcome. The quantification of this important concept is difficult because it is not clear which specific external measures to use to calibrate the scales of specific instruments.

There has been some debate in the literature around the use of retrospective transition questions to determine the MCID. Some authors argue that this approach implies that the minimal detectable change is both clinical and important (Garratt, Ruta et al. 1994; Wright 1996; Beaton 2003; Middel, Goudriaan et al. 2006). However not all change that is discernible to the respondent need necessarily be important either subjectively or clinically. Responses to transition questions are not actual measures of clinical change, but are measures of the magnitude of change in health as perceived by the respondent (Garratt et al 1994, Middell et al 2006). Furthermore the threshold of discernible difference that can be expressed by the respondent depends on the categories provided for response (Wright 1996; McHorney 1999). Therefore this paper will use the term minimal detectable difference (MDD) for the change in health status described by the HTQ (Wright 1996; Beaton 2003) to avoid confusion with measures of known important clinical change. The minimal detectable differences that can be expressed by the SF-36 HTQ are the categories of "somewhat worse" and "somewhat better" health compared with a year ago.

One of the important issues around the use of retrospective self-report as a measure against which to calibrate the responsiveness of health status scales is a question of direction of validation. Generally the reliability and validity of global retrospective measures of change is less well established than the reliability and validity of scales designed for prospective measurement of health status (Norman et al 1997). Perhaps therefore it is more appropriate to establish the validity of the retrospective question via serial changes in the scale scores, not the reverse.

Another question around the use of the health transition question that remains unanswered and that forms the central focus of this paper is that of interpreting the clinical meaning of the minimal discernible difference in the HTQ. How big are the changes in health status reported by respondents as "somewhat better" or "somewhat worse" health than a year ago, and what are the clinical implications of changes of this size?

Although many other studies have focussed on concordance between the HTQ and prospective changes in the SF-36 scales (Perneger, Etter et al. 1997) to our knowledge little has been done to calibrate the HTQ against an external measure of known clinical change and thus quantify the magnitude of change represented by the minimum discernible difference measured by the HTQ. Establishing the size of the MDD in clinical terms would extend the usefulness of the HTQ, especially when it is used as a benchmark for change in health status.

This question of the clinical meaning of the MDD can be addressed by looking for external clinical measures of change against which to calibrate the HTQ. Triangulation with a third external clinical anchor also serves to break the potentially circular argument of the direction of validation for prospective and retrospective measures, by using an external criterion of known clinical change in health.

The Household Income and Labour Dynamics in Australia (HILDA) study provides an opportunity to explore the clinical meaning of the HTQ, since it is a longitudinal population survey which includes the full SF-36 (version 1) in every 12 month interview wave. Prospective changes in health status can be calculated from changes in the respondent's score on each SF-36 scale between interviews. The change in scale scores cover the same 12 month interval as the respondent is asked to recall on the HTQ so allows the prospective and retrospective measures to be directly compared.

The HILDA study also asks respondents whether they have developed any long-term health problem or disability since last interview. This question provides an anchor of *clinically important change* of known magnitude that occurred during the period covered by the retrospective health transition question (Deyo and Inui 1984; Jaeschke, Singer et al. 1989; Crosby, Kolotkin et al. 2003). The categories of the HTQ can thus be calibrated against this clinical anchor by comparing the size of prospective changes in SF-36 scale scores for both questions.

This approach takes the position that prospective measures are the "gold-standard" for measuring change since they are based on serial measures of current health status and are therefore free from recall bias. Prospective change scores can therefore be used as an intermediary measure to calibrate

the retrospective HTQ against the known change in clinical health. Measures of prospective change have their own validity problems, including the presence of ceiling and floor effects on scale scores that reduce a scales sensitivity to future changes in actual health (McHorney 1999). However by using prospective change in scores to calibrate the HTQ to the clinical anchor any shortcomings in the prospective measures will apply equally to both sides, since the method of calibration we propose involves comparability of change rather than absolute magnitude of change on the scale scores.

Our aim in this paper was to calibrate the response categories of the HTQ to a known clinically important change in health by calculating the magnitude of prospective change scores on the SF-36 scales for the group who had developed a new long-term condition since last interview with the change scores for each category of the HTQ. By using a large population study, we anticipate that the interpretation of the meaning of the health transition question can be used in other population studies, in particular cross-sectional surveys that rely on the retrospective HTQ as a measure of change in health status.

METHOD

Design

The Household Income and Labour Dynamics of Australia (HILDA) study is a longitudinal population survey commenced in 2001. The HILDA sample of respondents was drawn from a representative sample of Australian households. Follow-up waves of interviews are completed every 12 months. Details of the HILDA method and sample are published elsewhere (Watson and Wooden 2004). At each interview wave respondents complete the Australian SF-36 Version 1 health survey (Sanson-Fisher and Perkins 1998).

The sample

We included respondents aged 18 years and over who completed the SF-36 health survey in HILDA wave 1 (2001) and wave 2 (2002), the same age range sampled in the National Health Survey 1995 for comparison with Australian population norms (ABS 1997).

Measures

Changes in health status were measured in three ways.

1) Retrospective assessment of change was measured by the responses to wave 2 (2002) of the SF-36 health transition question (HTQ).

The HTQ asks respondents “compared to one year ago how do you rate your health in general now?” The response categories “somewhat worse” or “somewhat better” define the *minimally detectable changes* of the health transition question.

2) The external anchor of *clinically important change* was defined as at least one new long-term health condition lasting at least 6 months, first developed since last interview, that affects everyday activities. This category was based on two questions in HILDA. Respondents are shown a list of conditions (see Appendix) and asked.

i) “Do you have any long-term health condition, impairment or disability (such as these) that restricts you in your everyday activities, and has lasted or is likely to last, for 6 months or more? Response: Yes/no

ii) “Did you first develop [this condition / any of these conditions] after date of last interview?” Response: Yes/no

3) *Prospective change scores* in health status were calculated as within-person changes in scores between baseline (wave 1, 2001) and follow up (wave 2, 2002) for each of the SF-36 scales.

Analysis

We calculated the prospective change in scores between baseline and follow-up for each individual on each SF-36 scale.

We then calculated mean prospective changes for:

1) the group which had developed any new long-term condition and the group which had not;

2) the five groups defined by the five categories of the HTQ.

To account for differences between individual in change scores within groups we calculated the standardised response mean (SRM) for each group (Fischer, Stewart et al. 1999; Husted, Cook et al. 2000).

SRM = Unadjusted mean change in score/standard deviation of change scores.

The SRM expresses the mean change in terms of variability among individuals. Because the mean and the standard deviation are both in the same units (the domain scale), the SRM is a unitless measure of effect that can be validly compared across different scales (Fischer, Stewart et al. 1999; Husted, Cook et al. 2000).

To adjust mean change scores for potential confounding with age and sex we ran a series of linear regression models for each SF-36 scale (Husted et al 2000). Change in the scale score was the dependent variable, the levels of the HTQ question was the group variable and age and sex were included as covariates. From the regression coefficients we calculated the adjusted mean change scores for each category of the HTQ substituting the sample mean values for age and sex into each equation.

We repeated the analysis with new long-term condition as the group variable to calculate the adjusted mean change in scale scores for the group with a known clinically important change in health.

The adjusted mean changes and SRMs for the categories of the HTQ were compared with those of the group with new long-term conditions, to calibrate the magnitude of change captured by responses to the HTQ in terms of a known *clinically important change*.

We also examined the distribution of scores of each of the domains at baseline for ceiling and floor effects. We defined these as percentage of scores within less than 5 points of the maximum or minimum possible score. We examined present state bias in the HTQ by through correlations between the HTQ and baseline and follow-up scores.

RESULTS

There were 13,191 respondents 18 years and over in wave 1 of whom 9,649 (73.0%) completed the SF-36 health questionnaire in waves 1 and 2. At baseline, the mean age of the final sample was 45.7 years, 53.4% were female, 50.1% reported being in very good or excellent health, and nearly one quarter reported an existing long-term condition or disability (Table 1).. At follow-up, 3.7% (n=350) reported developing a long-term health condition or disability since the last interview. These respondents were older than average (mean 55.4 years), and had on average poorer self-assessed health at baseline (Table 1).

T*able 1: Respondent characteristics at baseline: total sample compared respondents with a new long-term health condition*

Respondent characteristics	**Total sample N = 9,649**	**New long-term health condition n = 350**
Female gender (%)	53.4	52.3
Mean age (years) at baseline	45.7	55.4
Self-assessed health at baseline (missing = 98)		
% Excellent	15.1	5.5
% Very Good	35.6	21.2
% Good	33.0	35.4
% Fair	13.2	29.0
% Poor	3.2	9.0
Existing long-term health condition at baseline (%)	23.8	48.0
New long-term health condition at follow-up (%)	3.7	

The distribution of the retrospective health transition question is shown in Table 2. More than 70% of the total sample reported their health as "about the same" as a year ago and 13% reported their health was "much worse" or "somewhat worse" than a year ago. However among respondents who had developed a new long-term health condition since last interview, 55% reported that their health was "somewhat worse" or "much worse" than a year ago.

Table 2: Distribution of retrospective Health Transition Question wave 2 by new long-term health condition.

Health Transition question wave 2	**No new long-term condition N=9,299**	**New long-term condition N=350**
Much better now than one year ago (%)	4.8	2.0
Somewhat better now than one year ago	11.1	8.3
About the same as one year ago	72.8	34.0
Somewhat worse now than one year ago	10.2	44.0
Much worse now than one year ago	1.0	11.7

Respondents who at follow-up reported "somewhat worse" or "much worse" health compared with one year ago were more likely to report "fair" or "poor" health at baseline (see Table 3)

Table 3: Distribution of general health status wave 1 by Health Transition Question wave 2

	Health Transition question wave 2				
General health wave 1	**Much better n = 457**	**Somewhat better n = 1065**	**About the same n = 6891**	**Somewhat worse n = 1103**	**Much worse n=133**
Excellent (%)	13.62	12.56	17.07	6.87	3.82
Very Good (%)	33.04	36.44	38.52	20.16	11.45
Good (%)	36.38	36.82	32.58	31.81	22.14
Fair (%)	14.06	12.08	10.19	29.79	36.64
Poor (%)	2.9	2.09	1.64	11.37	25.95

Table 4a: The percentage of respondents who were near the ceiling of each SF-36 scale at baseline: Total sample and HTQ categories "somewhat better" and "much better"*

SF-36 scale	Total sample	Somewhat better	Much better
Physical functioning	30.4	32.7	25.6
Role physical	66.4	64.1	56.5
Bodily pain	31.9	31.6	30.6
General health	8.5	7.8	8.1
Vitality	1.0	0.5	2.2
Social functioning	50.5	42.3	37.6
Role emotional	72.0	67.6	61.3
Mental health	7.5	6.2	8.3

*near ceiling defined as scores > 95 points on the scale

Table 4b : The percent of respondents who were near the floor of each SF-36 scale at baseline: Total sample and HTQ categories "somewhat worse" and "much worse"*

SF-36 scale	Total sample	Somewhat worse	Much worse
Physical functioning	2.3	2.9	5.3
Role physical	14.7	32.5	51.9
Bodily pain	1.4	3.2	15.0
General health	2.6	5.4	6.0
Vitality	1.2	2.1	3.0
Social functioning	0.8	1.8	10.5
Role emotional	11.7	23.7	35.3
Mental health	0.8	1.4	0.0

*near floor defined as scores < 5 on the scale

Tables 4a and 4b shows the ceiling and floor effects of the scales at baseline for the total sample and for respective categories of the HTQ at follow-up. Scores near or at the ceiling of the SF-36 scales at baseline can register very little to no improvement in health status, and scores near or at the floor can register little or no worsening. Therefore ceiling effects attenuate the magnitude of mean prospective change scores for the groups who report better health at follow relative to the group who reported their health status as "about the same" as a year ago. Similarly floor effects attenuate the magnitude of mean change scores for those who reported worse health at follow-up

Five of the SF-36 scales had pronounced ceiling effects in the total sample at baseline, particularly the role-physical, role-emotional and social functioning domains (Table 4a).. The subset of respondents who reported "much better" health than a year ago at follow-up had a similar pattern of ceiling effects as the total sample.

There were few floor effects in the total sample at baseline (Table 4b). However, floor effects were apparent in the subset of respondents who reported at follow up that their health was "much worse", particularly for the role-physical and role-emotional domains.

Table 5: Correlation between retrospective HTQ with baseline SF-36 scale scores, follow up scores and change in scale scores*

SF-36 scale	Correlation of HTQ± with wave 1 scores*	Correlation of HTQ± with wave 2 scores*	correlation of HTQ± with prospective SF-36 change scores
bodily pain	0.14	0.26	0.13
general health	0.15	0.30	0.21
mental health	0.05	0.15	0.12
physical functioning	0.16	0.26	0.16
role-emotional	0.05	0.16	0.10
role-physical	0.12	0.28	0.17
social functioning	0.08	0.22	0.16
vitality	0.11	0.25	0.17

*Spearman rank correlation coefficients (all coefficients significant p < .0001)
±Health Transition question coding has been reversed (ie. coded 5 for "much better" health than a year ago and 1 for "much worse" health than a year ago), consistent with the direction of the SF-36 scales, so that correlations are more intuitively interpretable.

Table 5 shows the correlations between the Health Transition Question and the SF scale scores at baseline and follow-up and the correlation with prospectively measured change scores. The HTQ was moderately correlated with follow-up scores; thus lower follow-up scores were associated with self-reported worsening of health and higher follow-up scores were associated with self-reported improvements in health. The HTQ was less strongly correlated with baseline scores and the direction of the relationship was the same as for the follow-up scores. This lack of symmetry of the correlation patterns between the HTQ and baseline and follow up scores indicates that the HTQ was appreciably biased towards present health status. Correlations were weakest for mental health and role emotional scales. The HTQ was only weakly correlated with the change in scale scores.

Table 6: Mean changes over the 12 months in SF-36 scores for respondents who did and did not develop a new long-term health problem since last interview.

SF-36 scale	Long-term condition	Unadjusted mean change score (SD)	Mean change score, adjusted for age and sex (95%CL) ±
general health	no new condition	-0.5 (14.6)	-0.5 (-0.9,-0.2)
	new condition(s)	-8.0 (19.0)	-7.9 (-9.5,-6.3)
bodily pain	no new condition	0.1 (22.1)	0.1 (-0.4,0.5)
	new condition(s)	-7.2 (29.8)	-7.0 (-9.4,-4.6)
mental health	no new condition	0.1 (15.0)	0.1 (-0.2,0.4)
	new condition(s)	-2.5 (17.5)	-2.4 (-4.0,-0.8)
physical functioning	no new condition	-0.1 (17.4)	-0.1 (-0.5,0.2)
	new condition(s)	-9.1 (25.7)	-8.9 (-10.8,-7.0)
role emotional	no new condition	1.1 (34.1)	1.1 (0.4,1.8)
	new condition(s)	-7.0 (45.0)	-6.7 (-10.4,-2.9)
role physical	no new condition	0.3 (33.9)	0.3 (-0.5,1.0)
	new condition(s)	-19.1 (45.3)	-19.0 (-22.7,-15,3)
social functioning	no new condition	0.2 (22.7)	0.2 (-0.3,0.6)
	new condition(s)	-8.9 (28.9)	-8.6 (-11.0,-6.2)
vitality	no new condition	-0.1 (16.6)	-0.1 (-0.4,0.2)
	new condition(s)	-5.3 (20.2)	-5.1 (-6.9,-3.3)
pcs	no new condition	-0.1(7.7)	-0.1 (-0.3,0.1)
	new condition(s)	-5.4(11.3)	-5.4 (-6.3,-4.5)
mcs	no new condition	0.2(9.2)	0.1 (-0.1,0.3)
	new condition(s)	-1.1(10.9)	-1.0 (-2.1,0.0)

± Adjusted means calculated from regression coefficients using the sample mean age of 46 years and sample proportion of 53% female.

Table 6 compares the adjusted and unadjusted mean changes in SF-36 scale scores over the 12 months for those with and without a new long-term health condition. The mean change in scale scores do not change substantially after adjusting for age and sex. For those with a new chronic condition the largest decreases in SF-36 scores are for the role-physical and physical functioning domains. The group without a new chronic health condition had negligible average change on all scales.

Table 7 compares the unadjusted mean change in scale scores with the mean changes adjusted for age and sex across the categories of the HTQ. There was little difference in the adjusted means compared with the unadjusted means.

The mean change in scale scores for the group with a new long-term health condition (Table 6) were generally similar in magnitude to the mean change in scores for the HTQ category “somewhat worse” than a year ago (Table 7).

Table 8 compares the standardised response means (SRM) for each of the SF-36 scales across the categories of the HTQ and for the group with a new long-term health condition. The SRMs for the group with a new long-term health condition were small to moderate, indicating large variability among individuals in change scores relative to the mean change. A comparison with the categories of the HTQ again indicated that the SRMs for those with a new long-term condition were similar in magnitude to the SRMs for the category “somewhat worse” health than a year ago.

The effect sizes were not symmetrical for the categories of improved and worsened health status. Those who reported “much worse” health than a year ago had moderate to large negative SRMs, while the groups who reported “much better” health than a year ago showed small positive effect sizes on all scales.

Table 7: Mean change in SF-36 scores over 12 months for respondents by retrospective self-report on the health transition question: Unadjusted and adjusted means

SF-36 scale	Health transition category	Mean unadjusted change (SD)	Mean change adjusted for age and sex (95% CL)
general health	much better	6.3 (17.7)	6.4 (5.1,7.7)
	somewhat better	2.5 (15.1)	2.6 (1.7,3.5)
	about the same	-0.2 (13.4)	-0.2 (-0.5-0.2)
	somewhat worse	-9.1 (16.7)	-9.1 (-9.9,-8.2)
	much worse	-17.0 (20.9)	-17.4 (-19.9,-14.8)
bodily pain	much better (ref)	8.0 (27.1)	8.0 (6.0,10.1)
	somewhat better	3.6 (22.1)	3.6 (2.3,5.0
	about the same	0.1 (21.1)	0.1 (-0.4,0.7)
	somewhat worse	-7.2 (25.0)	-7.3 (-8.6,-6.0)
	much worse	-15.4 (31.5)	-15.6 (-19.4,-11.7)
mental health	much better (ref)	3.4 (18)	3.6 (2.2,4.9)
	somewhat better	2.8 (15.7)	3.0 (2.0,3.9)
	about the same	0.2 (14.0)	0.2 (-0.2,0.5)
	somewhat worse	-3.7 (17.2)	-3.8 (-4.7,-2.9)
	much worse	-10.2 (22.4)	-10.4 (-12.9,-7.8)
physical functioning	much better (ref)	3.2 (19.7)	3.2 (1.5,4.9)
	somewhat better	1.8 (18.8)	1.8 (0.7,2.9)
	about the same	0.2 (16.6)	0.2 (-0.2,0.6)
	somewhat worse	-6.3 (20.3)	-6.3 (-7.3,-5.2)
	much worse	-15.6 (27.7)	-15.6 (-18.7,-12.5)
role emotional	much better (ref)	7.9 (42.2)	7.7 (4.5,11.0)
	somewhat better	5.6 (36.7)	5.5 (3.4,7.6)
	about the same	1.2 (30.7)	1.2 (0.4,2.0)
	somewhat worse	-6.4 (45.0)	-6.4 (-8.5,-4.3)
	much worse	-21.6 (55.9)	-21.6 (-27.7,-15.5)
role physical	much better (ref)	12.3 (38.5)	12.4 (9.2,15.6)
	somewhat better	5.1 (35.5)	5.3 (3.2,7.4)
	about the same	0.7 (31.6)	0.7 (-0.1,1.5)
	somewhat worse	-15.4 (42.6)	-15.7 (-17.7,-13.6)
	much worse	-25.0 (43.3)	-25.4 (-31.5,-19.4)
social functioning	much better (ref)	7.4 (28.0)	7.5 (5.4,9.6)
	somewhat better	3.5 (23.7)	3.5 (2.2,4.9)
	about the same	0.5 (21.0)	0.5 (-0.0,1.0)
	somewhat worse	-8.4 (26.8)	-8.5 (-9.8,-7.1),
	much worse	-18.7 (33.7)	-18.8 (-22.7,-14.9)
vitality	much better (ref)	6.4 (19.0)	6.5 (5.0,8.0)
	somewhat better	3.7 (17.0)	3.8 (2.8,4.8)
	about the same	-0.1 (15.6)	-0.1 (-0.4,0.3)
	somewhat worse	-6.8 (18.8)	-6.9 (-7.9,-5.9)
	much worse	-11.7 (24.2)	-11.9 (-14.7,-9.0)
pcs	much better (ref)	2.6(9.0)	2.7 (1.9,3.4)
	somewhat better	0.9(8.2)	0.9 (0.4,1.4)
	about the same	0.0(7.1)	0.1 (-0.1,0.3)
	somewhat worse	-4.0(9.6)	-4.0 (-4.5,-3.5)
	much worse	-7.0(11.7)	-7.1 (-8.6,-5.6)
mcs	much better (ref)	2.6(11.1)	2.7 (1.8,3.5)
	somewhat better	1.9(10.0)	1.9 (1.3,2.5)
	about the same	0.2(8.4)	0.2 (-0.0,0.4)
	somewhat worse	-2.6(11.1)	-2.6 (-3.2,-2.0)
	much worse	-6.6(13.1)	-6.6 (-8.3,-4.9)

± Adjusted means calculated from regression coefficients using the sample mean age of 46 years and 53% female.

Table 8: Standardised response means (SRM) for change in each SF-36 scale across Health Transition question responses

SF-36 scale	Much better	Somewhat better	The same	Somewhat worse	Much worse	New long-term condition
	457	**1065**	**6891**	**1103**	**133**	**360**
bodily pain	0.30	0.16	0.01	-0.29	-0.49	-0.24
general health	0.36	0.16	-0.02	-0.53	-0.81	-0.42
Mental health	0.19	0.18	0.01	-0.22	-0.45	-0.14
Physical functioning	0.16	0.09	0.01	-0.31	-0.56	-0.35
role emotional	0.19	0.15	0.04	-0.14	-0.39	-0.16
role physical	0.32	0.14	0.02	-0.36	-0.58	-0.42
social functioning	0.27	0.15	0.02	-0.32	-0.55	-0.31
Vitality	0.34	0.22	0.00	-0.36	-0.49	-0.26
PCS	0.29	0.10	0.01	-0.41	-0.60	-0.48
MCS	0.24	0.19	0.02	-0.23	-0.50	-0.10

SRM effect sizes: small = 0.2 moderate = 0.5 large = 0.8 (Cohen 1988)

Figures 1 to 10 show the box plots of the distribution of the change in scale scores and physical and mental component summary scores for each category of the HTQ.

All scales show an appreciable proportion of respondents with change scores in the opposite direction to the stated change on the HTQ. For example for those who reported “much better” health on the HTQ, the median change on the general health scale was an increase of 5 percentage points. However 25% of the “much better” group recorded a decrease on the general health scale of greater than 5 percentage points. For those who reported “somewhat worse” health on the HTQ the median change on the general health scale was a decrease of 8 percentage points, however 10% recorded an increase of greater than 10 percentage points.

The mode change in score was zero or close to zero across all categories of the HTQ for all scales including the pcs and mcs.

Figure 1

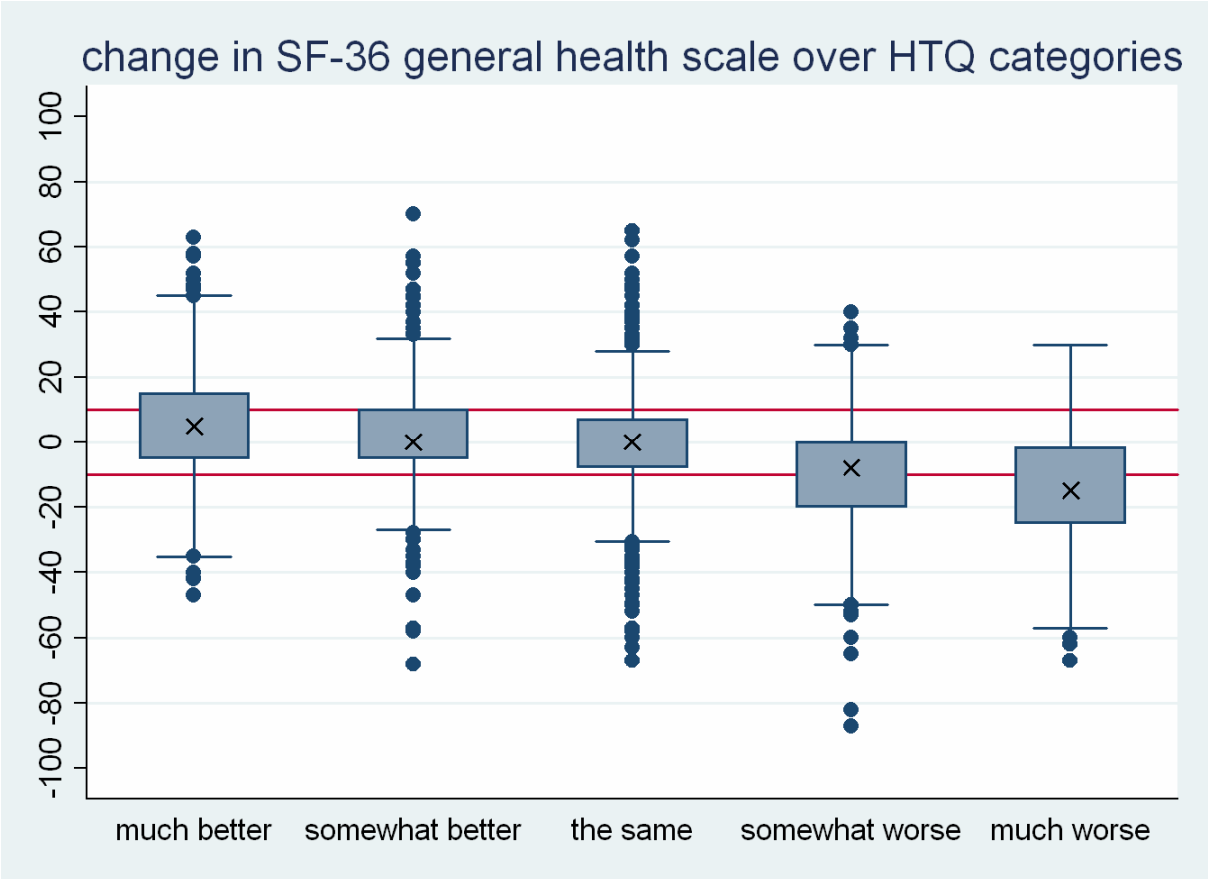


Figure 2

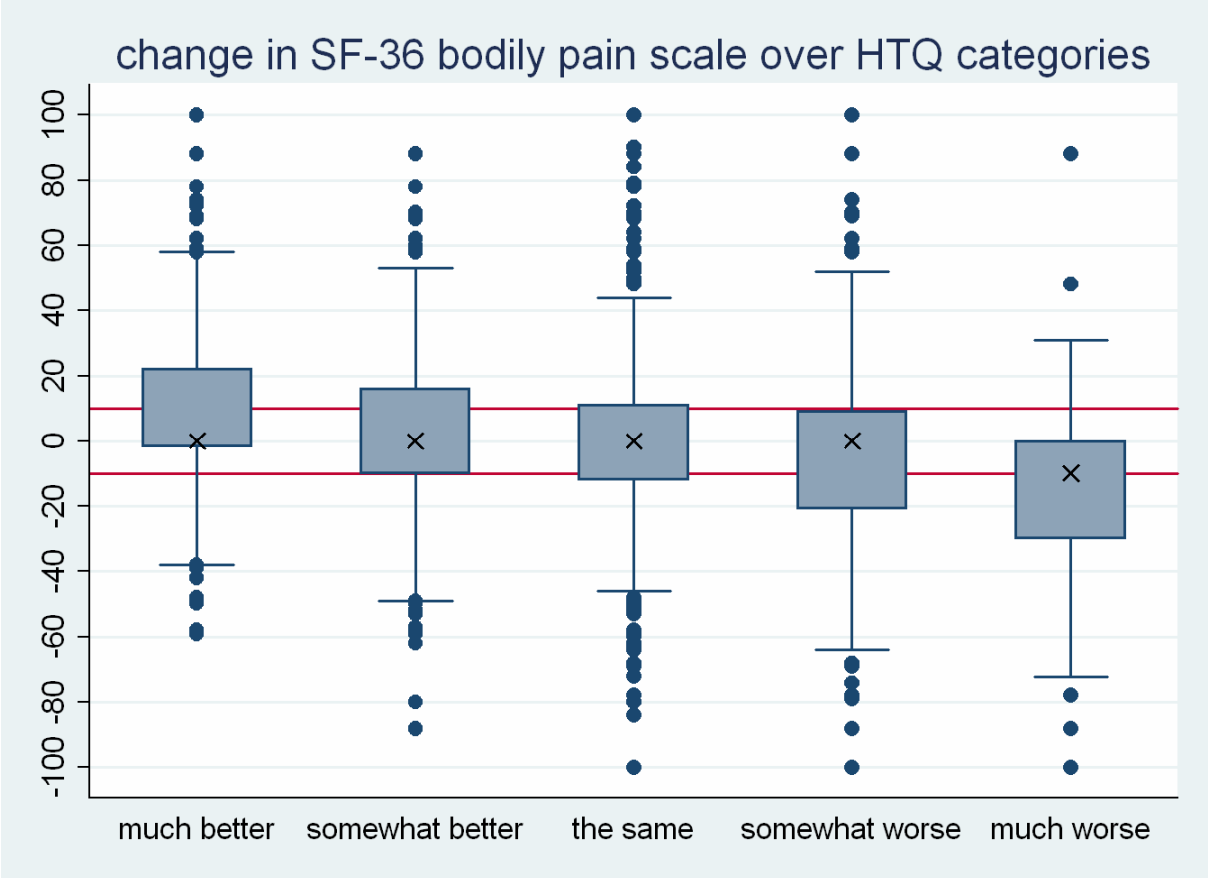


Figure 3

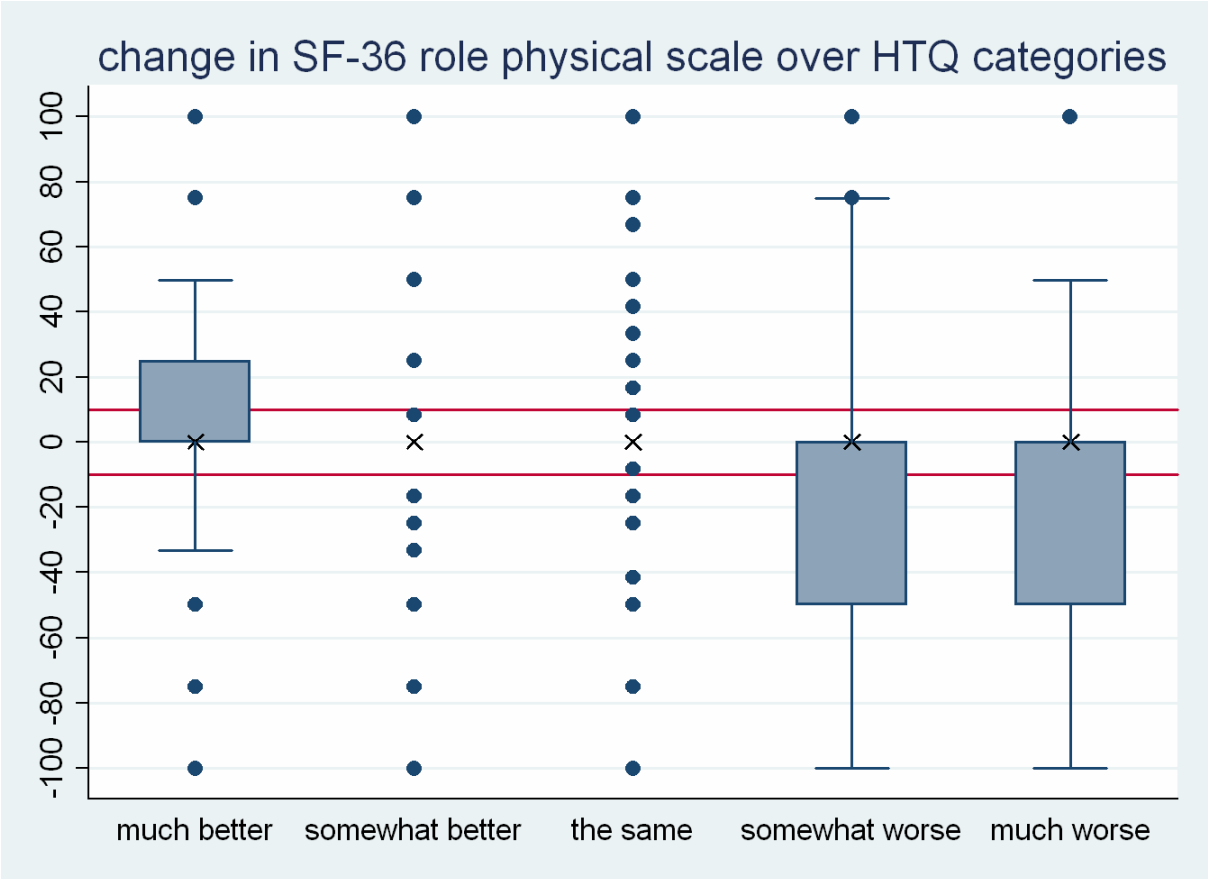


Figure 4

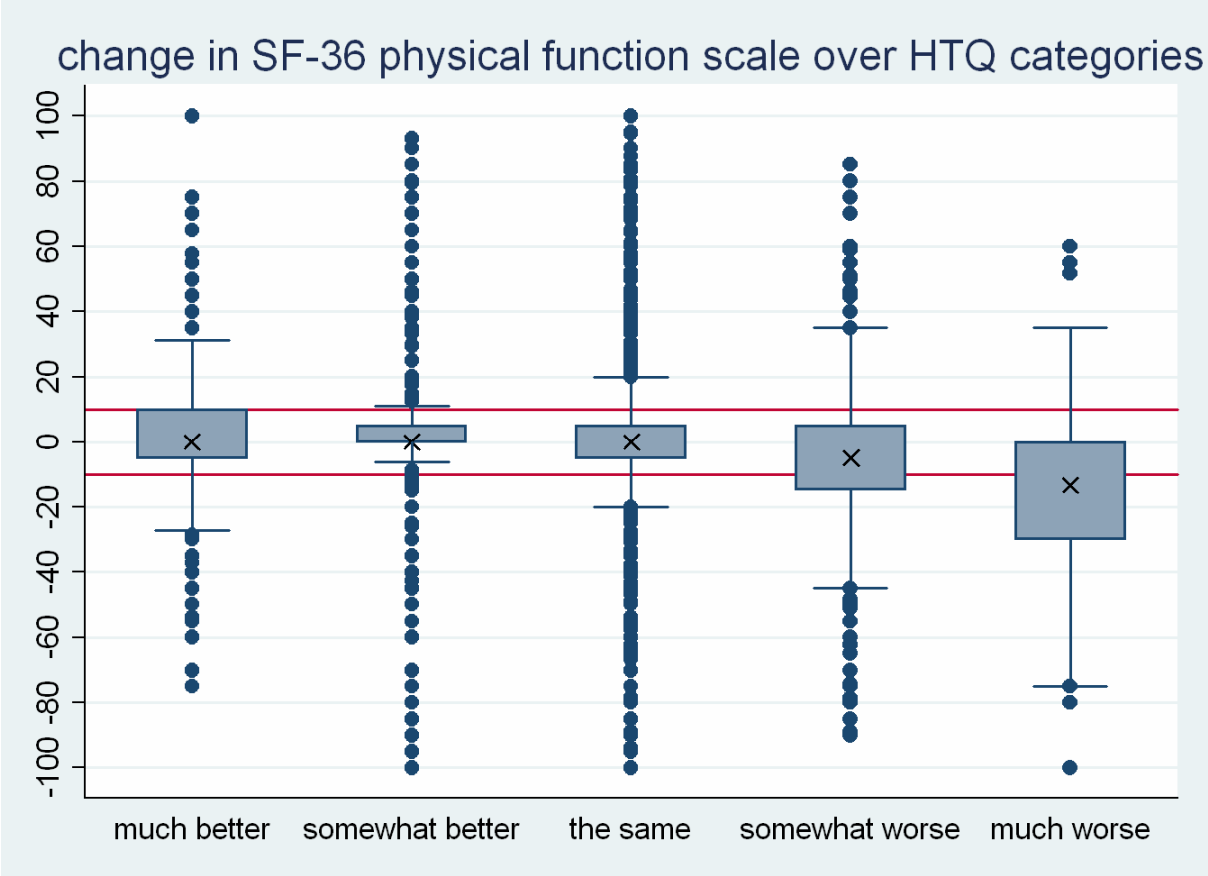


Figure 5

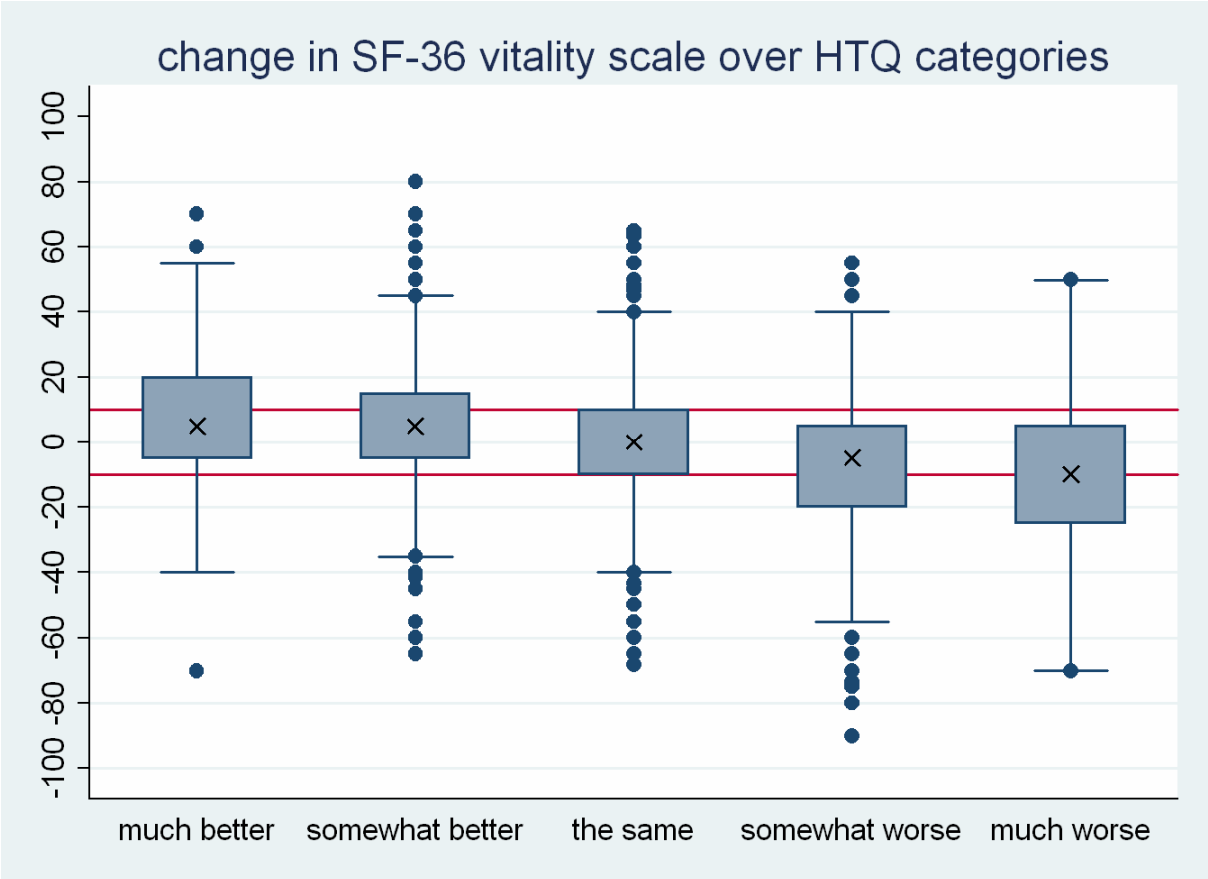


Figure 6

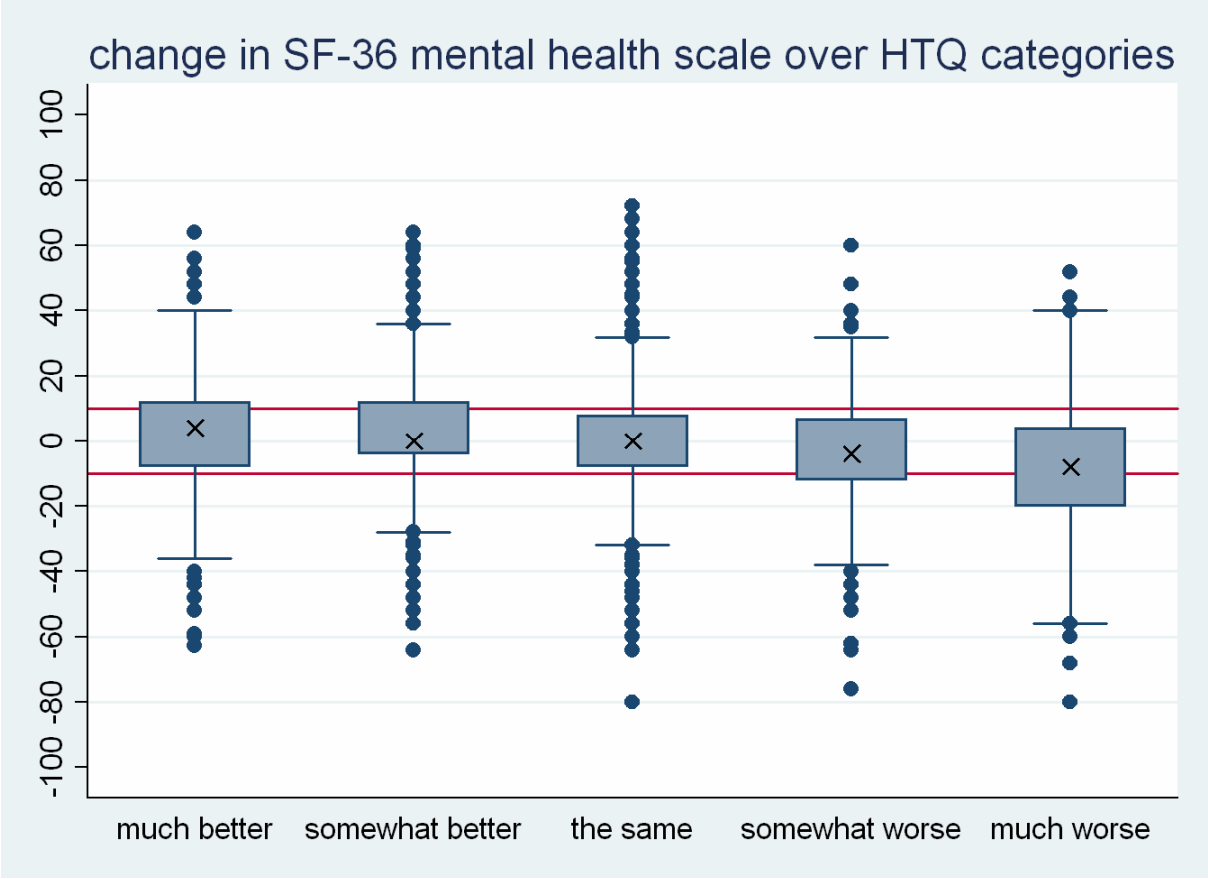


Figure 7

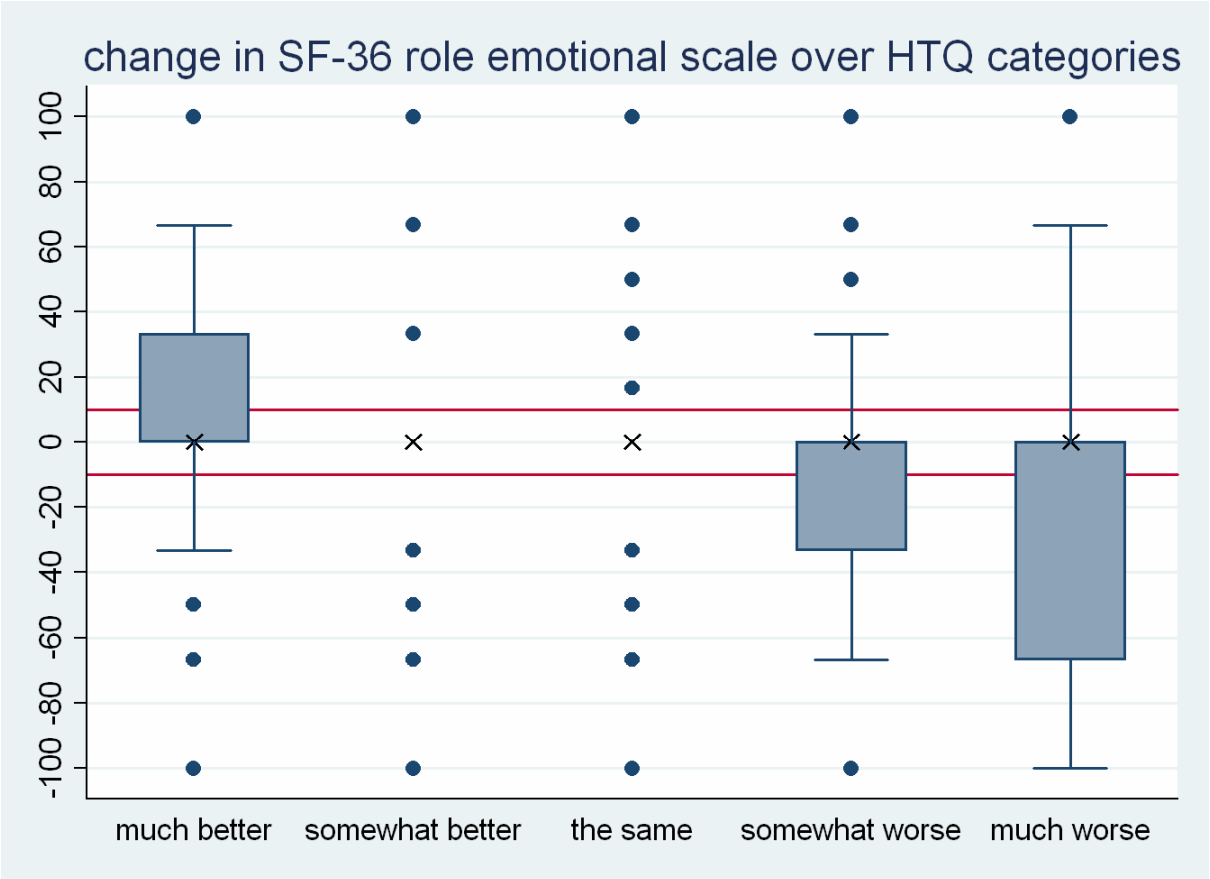


Figure 8

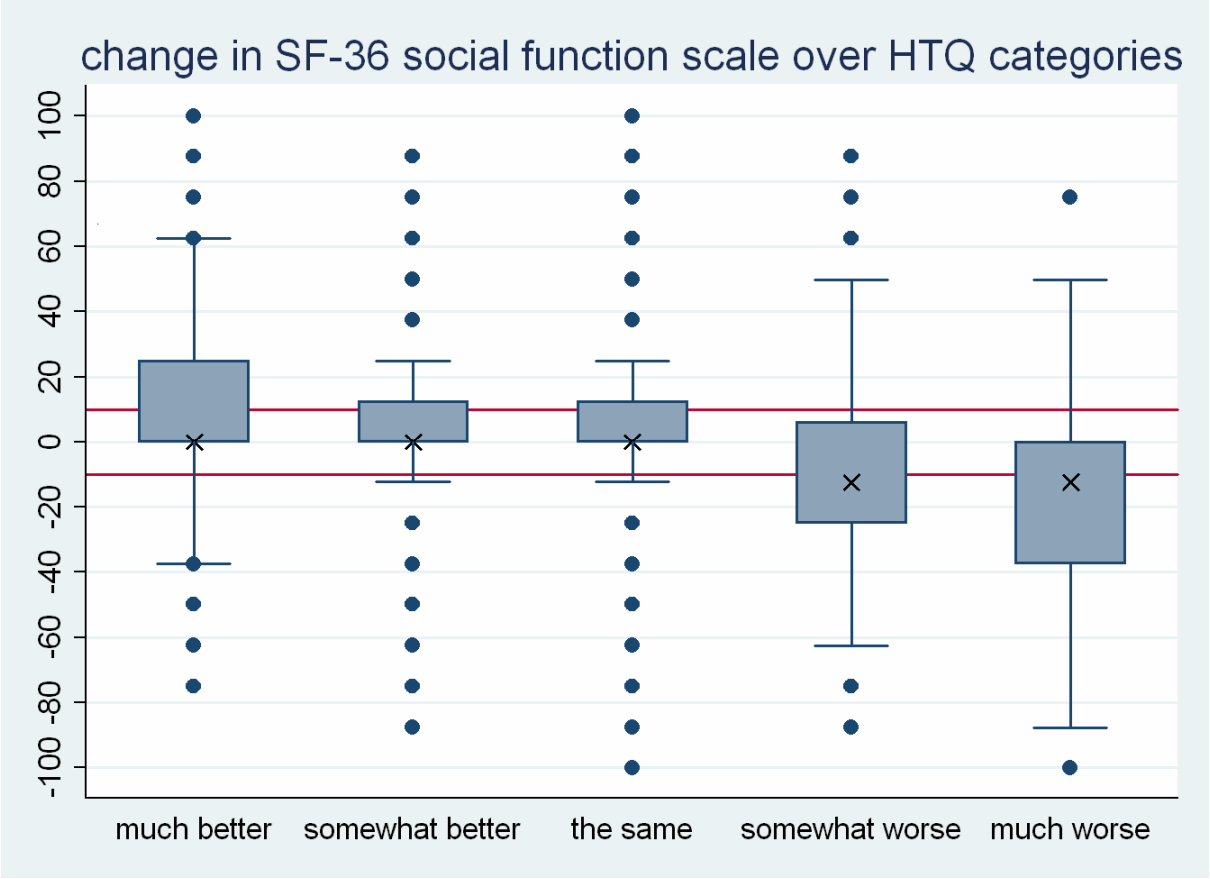


Figure 9

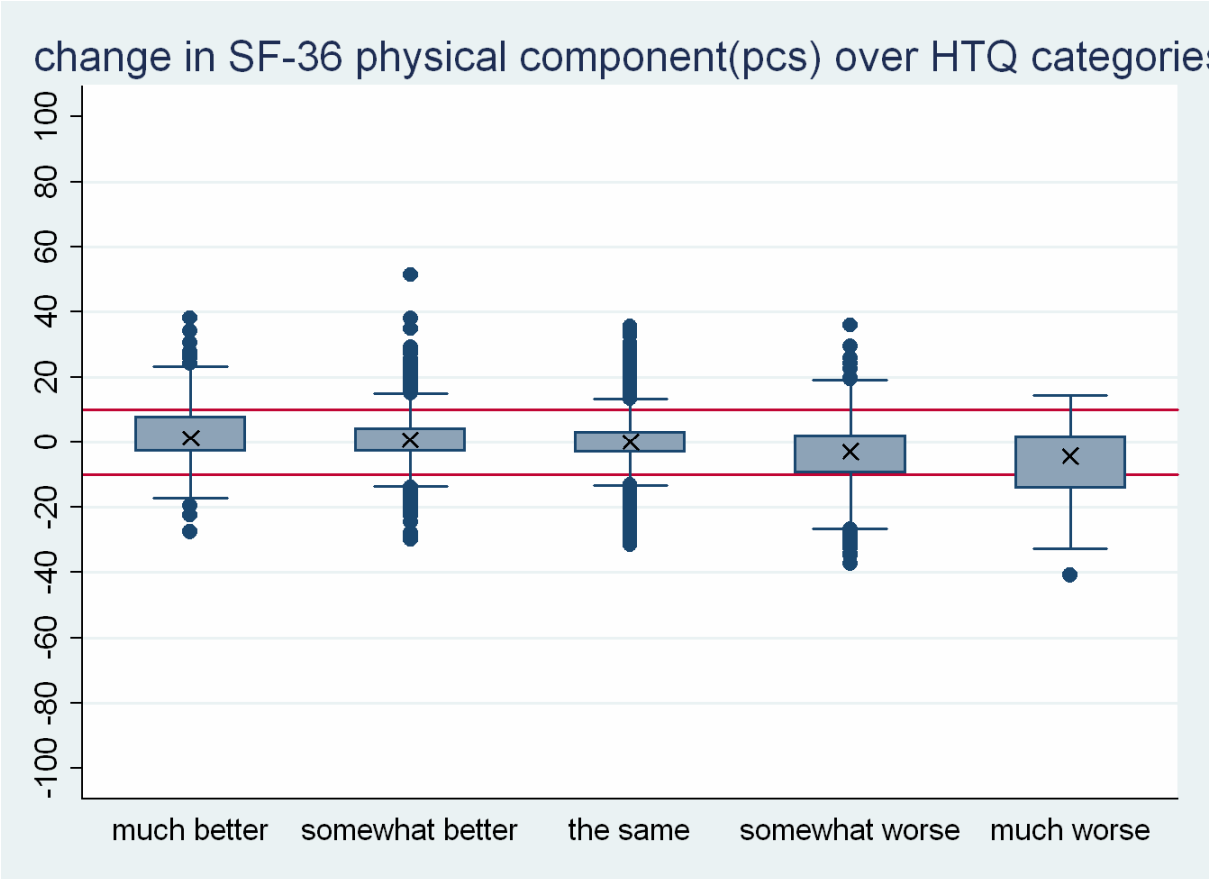
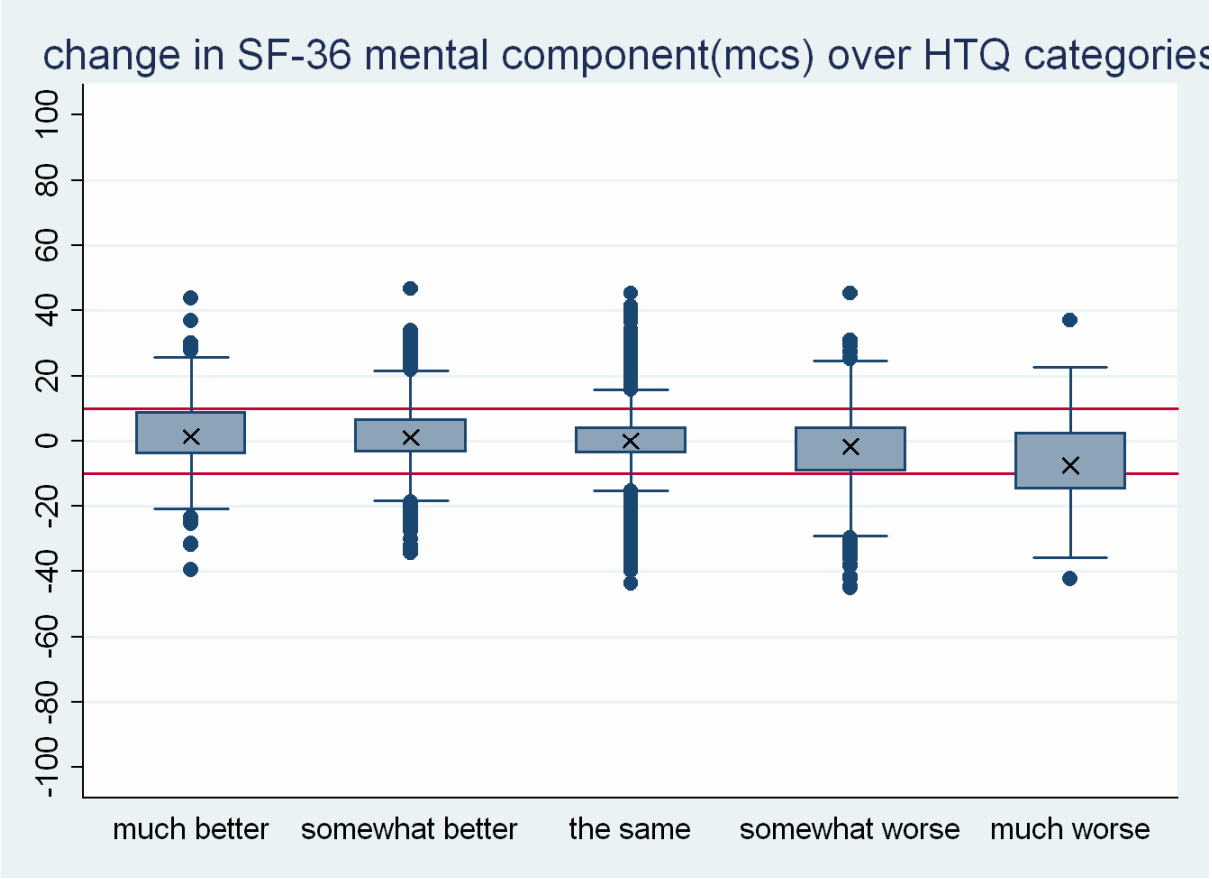


Figure 10



DISCUSSION

Using data from the longitudinal HILDA study we calibrated the categories of the SF-36 health transition question against an external anchor of clinically important change in health by comparing the magnitude of prospective within-person changes in the SF-36 scales. We found that the group who described their health as "somewhat worse" than one year ago had experienced changes in health status similar in magnitude to those who had developed a long-term health condition in the last year. These results provide a clinical interpretation for the minimum discernible difference measured by the category "somewhat worse" on the HTQ.

Since we used the onset of a new long-term health problem as the external anchor, or criterion for calibrating the HTQ, it was important to demonstrate that this anchor had the expected relationship with the prospectively measured change scores. This was the case: the mean change in SF-36 scale scores of the group who did not report a new long-term health condition was close to zero for all domains, while there was an appreciable negative mean change in scale scores for those who had recently developed a new long-term health problem.

We also found that at the aggregate level, the retrospective HTQ scores had the expected relationship with the prospective change on SF-36 scales. The mean change in SF-36 scores for the group who reported no change in their health status on the HTQ was negligible for all domains with a significant positive gradient in mean change in scores for the categories of "somewhat better" and "much better" and a significant negative gradient in mean change for the categories "somewhat worse" and "much worse".

Population studies, cross-sectional short questions

Our findings indicate that the retrospective HTQ is a valid and useful question that can identify groups of people who have experienced a clinically important change in health status in population surveys, particularly in cross-sectional surveys where other measures of change are not possible. Our results are based on a large and representative Australian population sample and add to what is known on the application of the SF-36 in population surveys (Hemingway, Stafford et al. 1997; Perneger, Etter et al. 1997; Mozes, Maor et al. 1999) . Our recommendation for using the retrospective HTQ as a proxy for clinical change in health is however moderated by some important caveats.

Individual versus group changes

The mean changes in scale scores for each HTQ category were in the expected direction, however the effect sizes for change were small due to large variability among individuals in their change scores. This included many respondents who had changes in scores that were in the opposite direction to their stated change in health status on the HTQ. These discrepancies may arise in part due to the global nature of the HTQ in the SF-36 compared to the scales which are domain-specific ((Fitzpatrick, Ziebland et al. 1993). A respondent may experience both improvement and worsening concurrently in different domains of health but may value particular domains above others when making a retrospective assessment of change in health status.

This variability of change scores within HTQ categories indicates that although we have calibrated the HTQ in terms of average clinical change in health at the group level, the HTQ cannot be used to reliably predict the magnitude of clinical change experienced by a particular individual. Thus if the HTQ is used to group respondents in terms of average clinical change, some individuals will be misclassified. For example, some of those people who perceived their health was somewhat worse a year down the track did not actually experience a decline their health status (as measured prospectively at the two time points), indeed some may indeed have experienced an improvement prospectively. These results indicate that those who perceived their health to be poor at follow-up are most likely to have a biased perception of deterioration at follow-up.

Present state bias

We found that the HTQ correlated more strongly with follow-up scale scores than with baseline scores or with prospective measured change scores. This means that respondents with poor health status at follow up were inclined to over-estimate any worsening of health on the HTQ, while respondents with good health status at follow-up over-estimated improvements in health. These findings support other research indicating that retrospective recall of recent changes in health status is subject to present state bias, ie recall is more influenced by the respondent's present health status than by actual change in health status from baseline (Norman, Stratford et al. 1997; Middel, Goudriaan et al. 2006),

Global clinical anchor

The clinical anchor was also a global measure that was based on two brief questions that did not identify which specific long-term conditions nor how many conditions the respondent had recently developed. Therefore some of the variability in change scores around the clinical anchor may be explained by the particular chronic condition(s) developed by each respondent.

Worsening vs Improving

As with previous studies, we found that those who reported improvements in health had smaller prospective changes in scores than those who reported worsening (Perneger, Etter et al. 1997; Cella, Hahn et al. 2002). This could be due in part to the substantial ceiling effects in baseline scale scores among the HILDA respondents, as might be expected from a mostly healthy population sample. This renders the SF-36 scales less sensitive to improvements in health status. Cella, Hahn et al. (2002) argue that the lack of symmetry between improving and worsening health may also arise because respondents value or notice small improvements in health status more than comparable declines (Cella, Hahn et al. 2002). Our analysis only calibrated the HTQ against a clinical anchor of deterioration in health, and floor effects which might affect this calibration were less evident. It would be informative to find an equivalent clinical anchor for improvement in health against which to calibrate the minimal discernible improvement in health status on the HTQ.

Future directions

The clinical meaning of "somewhat worse health than a year ago" would be expected to be context specific and may have different clinical meanings in clinical and general populations. However if the clinical meaning is stable within a particular context then the interpretation of the minimal discernible difference in the HTQ defined in HILDA should be generalisable to other population studies. HILDA provides an opportunity to re-run this analysis in subsequent interview waves to test whether the clinical meaning of the HTQ is a stable effect over time. Changes to the HILDA questionnaire in later interview waves also allow specific chronic conditions to be identified, rather than collapsing all conditions into a single global question. Repeating the analysis on later waves of HILDA would therefore allow us to analyse new mental health conditions and new physical health conditions separately.

CONCLUSIONS

This analysis revealed that in the HILDA survey, as in previous studies:

- retrospective self-report of change in health status is subject to present state bias;
- prospective measures of improvement were subject to ceiling effects;
- individuals within a HTQ category vary considerably in their prospectively measured change scores;
- the HTQ is more likely to reflect changes on physical domains of the SF-36 than in mental and emotional domains.

Given these caveats, we conclude that the HTQ is a valid measure of prospective change in health status at the group level, and the minimally detectable decline in health measured by the health transition question category 'somewhat worse' represents a decline in health equivalent in magnitude to developing one or more new long-term health conditions.

The meaning and magnitude in clinical terms of the categories of the HTQ may be context specific, differing for example between clinical and non-clinical populations (McHorney, 1999). Nevertheless, it is reasonable to assume that the size of the minimum detectable difference on the SF-36 HTQ relative to a clinically important change calculated in the HILDA study will be generalisable to other population studies, in particular cross-sectional studies that depend on respondents' retrospective self-report of health changes and are looking for valid brief measures to reduce the response burden.

REFERENCES

ABS (1997). National Health Survey: SF36 Population Norms, Australia. Cat No 43990.0. Canberra, Australian Bureau of Statistics.

ABS (2006). 2004-05 National Health Survey: Users' Guide. Cat No 4363.0.55.001, Australian Bureau of Statistics.

Beaton, D. E. (2003). "Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation." Medical Care **41**(5): 593-596.

Cella, D., E. A. Hahn, et al. (2002). "Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening." Quality of Life Research **11**(3): 207-21.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ, Lawrence Earlbaum Associates.

Crosby, R. D., R. L. Kolotkin, et al. (2003). "Defining clinically meaningful change in health-related quality of life." Journal of Clinical Epidemiology **56**(5): 395-407.

Deyo, R. A. and T. S. Inui (1984). "Toward clinical applications of health status measures: Sensitivity of scales to clinically important changes." Health Services Research **19**(3): 275-289.

Fischer, D., A. L. Stewart, et al. (1999). "Capturing the patient's view of change as a clinical outcome measure.[see comment]." JAMA **282**(12): 1157-62.

Fitzpatrick, R., S. Ziebland, et al. (1993). "Transition questions to assess outcomes in Rheumatoid Arthritis." British Journal of Rheumatology **32**: 807-881.

Garratt, A. M., D. A. Ruta, et al. (1994). "SF 36 health survey questionnaire: II. Responsiveness to changes in health status in four common clinical conditions." Quality in Health Care **3**(4): 186-92.

Guyatt, G. H., D. Osoba, et al. (2002). "Methods to explain the clinical significance of health status measures." Mayo Clinic Proceedings **77**: 371-383.

Hemingway, H., M. Stafford, et al. (1997). "Is the SF-36 a valid measure of change in population health? Results from the Whitehall II Study." BMJ **315**(7118): 1273-9.

Husted, J. A., R. J. Cook, et al. (2000). "Methods for assessing responsiveness: a critical review and recommendations." Journal of Clinical Epidemiology **53**(5): 459-468.

Jaeschke, R., J. Singer, et al. (1989). "Measurement of health status: Ascertaining the minimal clinically important difference." Controlled Clinical Trials **10**: 407-415.

McHorney, C. A. (1999). "Health status assessment methods for adults: Past accomplishments and future challenges." Annual Review of Public Health **20**: 309-335.

Middel, B., H. Goudriaan, et al. (2006). "Recall bias did not affect perceived magnitude of change in health-related functional status." Journal of Clinical Epidemiology **59**(5): 503-11.

Mozes, B., Y. Maor, et al. (1999). "Do we know what global ratings of health-related quality of life measure?" Quality of Life Research **8**(3): 269-73.

Norman, G. R., F. G. Sridhar, et al. (2001). "Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life." Medical Care **39**(10): 1039-47.

Norman, G. R., P. Stratford, et al. (1997). "Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach." Journal of Clinical Epidemiology **50**(8): 869-79.

Perneger, T. V., J. F. Etter, et al. (1997). "Prospective versus retrospective measurement of change in health status: a community based study in Geneva, Switzerland." Journal of Epidemiology & Community Health **51**(3): 320-5.

Sanson-Fisher, R. and J. Perkins (1998). "Adaptation and validation of the SF-36 Health Survey for use in Australia." Journal of Clinical Epidemiology **51**(11): 961-967.

Ware, J. E., Jr. , K. Snow, et al. (1993). SF-36 Health Survey: Manual and Interpretation Guide. Boston, The Health Institute.

Watson, N. and M. Wooden (2004). "The HILDA Survey Four Years On." The Australian Economic Review **37**(3): 343-349.

Wright, J. G. (1996). "The Minimal Important Difference: Who's to Say What Is Important?" Journal of Clinical Epidemiology **49**(11): 1221-1222.

APPENDIX

List of chronic conditions on HILDA showcard

DISABILITIES/ HEALTH CONDITIONS WHICH:
• Have lasted 6 months or more,
• Restrict everyday activity, and
• Can not be corrected by medication or medical aids

• Sight problems not corrected by glasses or contact lenses
• Hearing problems
• Speech problems
• Blackouts, fits or loss of consciousness
• Slow at learning or understanding things
• Limited use of arms or fingers
• Difficulty gripping things
• Limited use of feet or legs
• Nerves or emotional conditions which require treatment
• Any restriction on physical activity or physical work
• Any disfiguration or deformity
• Any mental illness which requires help or supervision
• Long term effects as a result of a head injury, stroke or other brain damage
• A long-term condition or ailment which is still restrictive even though it is being treated or medication is being taken for it
• Any restriction caused by Arthritis, Asthma, Heart Disease, Alzheimer's Disease, Dementia or any other long-term condition

Source: http://www.melbourneinstitute.com/hilda/qaires/ShowcardsW2.pdf