Validation and calibration of the SF-36 health transition question against an external criterion of clinical change in health status

Abstract

Purpose:

Cross-sectional surveys depend on retrospective health transition questions (HTQ) to estimate recent changes in health status. This paper assesses the validity of the SF-36 HTQ and calibrates its categories against change assessed prospectively on the SF-36 domain scales in a subgroup known to have experienced clinically important change in health status.

Methods:

Adults (n =9,649) from a longitudinal population survey completed the SF-36 in 2001 and 2002. Prospective measures were calculated as mean changes in SF-36 scale scores adjusted for age and gender, and also expressed as standardised response means.

Comparison groups were those who had developed a long-term health condition since last interview and the HTQ response categories for those who had not developed any new condition.

Results:

Those with a new condition and those without a new condition but who described their health as 'somewhat worse' than a year ago had comparable declines in health status on all domain scales except role physical, where those with a new condition experienced a greater decline.

Conclusions

This analysis demonstrates the validity and limitations of the HTQ as a measure of change in population studies The calibration is useful for interpreting the meaning of the HTQ categories at the group level but not the individual level.

Keywords

Health transition question, anchor-based methods, population studies, prospective change, retrospective recall

Validation and calibration of the SF-36 health transition question against an external criterion of clinical change in health status

Stephanie A Knox,

Centre for Health Economics Research and Evaluation (CHERE), University of Technology (UTS), Sydney, NSW, Australia.

+61 2 9514 4758

+61 2 9514 4730

Stephanie.knox@chere.uts.edu.au

Madeleine T King,

Centre for Health Economics Research and Evaluation (CHERE), University of Technology (UTS), Sydney, NSW, Australia.
and
School of Psychology, University of Technology (UTS), Sydney, NSW, Australia.

Background

Cross sectional studies rely on retrospective recall to identify and quantify recent changes in respondents' health status. The short form 36 item (SF-36) survey is a health status instrument that includes a global health transition question (HTQ) that 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' [1]. For example, the Australian National Health Survey, a cross-sectional population survey, included the HTQ as part of the full SF-36 in 1996, then as a stand alone item in 2001 [2]. In principle, such a brief retrospective item can add value to cross-sectional surveys where there is no opportunity for follow-up by identifying groups of subjects who have experienced a recent change in health status. However, such an item is useful only if it is a valid measure of prospective change in health, and if its magnitude is interpretable in clinically meaningful terms.

There is a body of work establishing the validity of the SF-36 multi-item scales in population studies [1, 3-5]. However the validity of the SF-36 retrospective HTQ has received relatively little attention. The issues of reliability and validity of a global single item retrospective measure of change have been described by Norman et al (1997)[6] who point out that inter-item consistency, the usual test of reliability for multi-item scales, is not possible for single items. They also describe the problem of 'present state bias' where subjects are inclined to judge their change in health status in relation to their present health state; 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. A number of studies have demonstrated present state bias in global retrospective transition questions [6, 7]. A more recent study has found that present state bias may be most pronounced when using a single global transition item and is ameliorated by using multi-item retrospective scales [8]. In terms of validity, it has been argued that the HTQ has face validity as a measure of subjectively meaningful change from the patient's perspective, as seen in the routine use of retrospective report in clinical decision making [7].

Although a HTQ is arguably less well-validated than multi-item scales, it has been common practice to use a HTQ as an external 'anchor' against which to calibrate the responsiveness of the scales of health status instruments [7, 9, 10]. In particular retrospective questions such as the HTQ have been used to ascertain the magnitude of the minimal clinically important change (MCID) measured by prospective changes in scale scores of health status instruments [11]. The MCID is defined by the smallest category of the retrospective HTQ (somewhat better/worse) and assumes that the HTQ represents true change [6]. The term 'minimal clinically important difference' however includes three problematic assumptions in its terminology that need clarification. Firstly the term MCID assumes that the respondent would consider any detectable change to be an important change. Yet a subject may detect and report changes that he/she does not necessarily deem to be important[12]. A second and related problem is the dependence of the MCID on the scale of available response categories on the HTQ. Retrospective questions with different numbers of response categories may

produce different MCIDs [12]. Thirdly the term MCID assumes that the HTQ is a proxy for an actual clinical change when in fact the HTQ itself is a subjective measure of change. The magnitude of actual clinical change measured by the HTQ remains unclear and is the focus of this paper.

To avoid problems of terminology, we use the more accurate term 'minimal detectable difference' (MDD) to describe the change in health status reported by the HTQ [13, 14]. The MDD that respondents can report with the HTQ corresponds to the categories 'somewhat worse' and 'somewhat better' health than a year ago. Applying the term MDD to the HTQ will then allow us to define our measure of actual clinical change without confusion in the use of terms.

To our knowledge little has been done to ascertain the clinical meaning of the categories of the HTQ. Yet understanding the magnitude of clinical change associated with the MDD would make the HTQ a much more useful benchmark of change in health status. This question of the clinical meaning of the HTQ can be approached by using a measure of known clinical change which is external to the health status instrument and which occurred during the period covered by the HTQ. The HTQ response categories can then be calibrated against the anchor by comparing the relative size of prospective change in scale scores.

The Household Income and Labour Dynamics in Australia (HILDA) study provides an opportunity for exploring the clinical meaning of the HTQ, since it is a longitudinal population survey which includes the full SF-36 in every 12 month interview wave. In the HILDA study, respondents are also asked whether they have developed any long-term health problem or disability since last interview, responses to which provides an anchor of clinically important change [11, 15, 16].

Our aim in this paper was to calibrate the response categories of the HTQ to this anchor of clinically important change in health. We did this by comparing prospective change scores on the SF-36 scales for the group who had developed a new long-term condition since last interview with the prospective change scores for the MDD group of the HTQ. This approach takes the position that prospective change scores 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. We use them as an intermediary measure to calibrate the retrospective HTQ against the known change in clinical health. By calibrating the HTQ within a large Australian population study we anticipate that the clinical interpretation of the HTQ can be generalised to 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 respondents were 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 [17]. At each interview wave respondents complete the Australian SF-36 Version 1 health survey [3].

We included respondents aged 18 years and over who completed the SF-36 health survey in HILDA Wave 1 (2001) and Wave 2 (2002).

Measures

Changes in health status were measured in three ways.

1) Prospective changes 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 eight SF-36 scales.

2) Retrospective change in health status was measured by responses to the SF-36 HTQ in Wave 2.

3) The external anchor of clinically important change was defined as the development at follow-up (Wave 2) of at least one new long-term health condition since the previous interview (baseline, Wave 1). The clinical anchor was derived from two questions in HILDA. Respondents were shown a list of conditions (detailed in the 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.

Comparison groups

Respondents who answered yes to both questions on new conditions were classified as having experienced a clinically important change since the last interview and were included in the New Conditions group.

Respondents who had not developed a new long term condition since last interview were then divided into five HTQ sub-groups according to their responses to the HTQ, ie 'Much Better' HTQ subgroup, 'Somewhat Better' subgroup, 'The Same' sub-group, 'Somewhat Worse' subgroup and 'Much Worse' subgroup. The New Conditions group formed the anchor against which we compared the HTQ subgroups. Excluding those with a new condition from the HTQ subgroups ensured that the comparison groups were mutually exclusive.

We compared the magnitude of prospective change scores on SF-36 scales for the New Conditions group with the change scores for each of the HTQ sub-groups (with no new condition).

Analysis

Raw change scores were adjusted for comparability across groups using methods recommended by Husted and colleagues [10]. Firstly we used linear regression to calculate mean change scores for each group adjusted for age and gender.

To calculate the mean change scores for the HTQ sub-groups we used the sub-sample of respondents without a new long term condition.

We fitted a separate model for each SF-36 scale with change in SF-36 scale score as the dependent variable and each level of the HTQ question as the group variable, adjusted for age and gender.

We repeated the analysis using the whole sample and fitted new long-term condition (yes or no) as the group variable to calculate the adjusted mean change in scale scores New Conditions group.

The second measure we used to estimate change in scale scores was the standardised response mean (SRM) [10].

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

The SRM takes into account variability of change between persons by adjusting the mean change for within group variance.

The adjusted mean changes and SRMs for the HTQ sub-groups were compared with those of the New Conditions group to calibrate the magnitude of change captured by responses to the HTQ relative to a a known clinically important change.

To assess present state bias in the HTQ we examined the correlations between the HTQ and the scale scores at baseline and follow-up.

Results

There were 13,191 respondents 18 years and over in Wave 1, of whom 9,649 (73%) completed the SF-36 health questionnaire in Waves 1 and 2. These individuals formed the sample for this paper. At baseline, their mean age was 45.7 years, 53.4% were female, 50.7% reported being in very good or excellent health at baseline (Table 1). Respondents who dropped out after Wave 1 but who had completed the SF-36 in Wave 1 (n = 1329) were somewhat younger (mean age 41.5 years) than the included sample, and a similar proportion had very good or excellent health at baseline (53.2%). The included sample and dropouts had the same rate of existing conditions at baseline (24.3% and 23.9%).

At Wave 2, 3.6% (n=350) reported developing a new long-term health condition or disability since the last interview. These respondents (the New Conditions group) were older (mean age 55.4 years), and had poorer self-assessed health at baseline than did the group with no new long term condition (Table 1). Table 2 shows the distribution of the HTQ for the whole sample, for the New Conditions group, and for the those without a new long-term health condition (collapsed

across HTQ categories). In the New Conditions group, 55.7% reported their health as worse than a year ago compared with 11.2% of those without a new condition.

****

TABLEs 1 and 2 about here

****

Table 3 shows the correlation between the HTQ and baseline, follow-up and change scores. The correlations between the HTQ and change scores were positive but weak. Correlations with follow-up scores were weak to moderate. With baseline scores the correlations were very weakly positive.

****

TABLE 3 about here

****

Table 4 compares the prospectively calculated SF-36 mean change scores of those with and without a new long-term health condition. The New Conditions group had a mean decline of 5 points or more and small to moderate negative SRMs for all scales except mental health[18]. In contrast, the average change in scale scores and SRMS were close to zero for the group without a new chronic health condition. When those without a new condition were analysed by HTQ category (Table 5), it can be seen that the mean change scores and SRMs follow the expected gradient, with the largest improvements and deteriorations being seen in the 'much better' and 'much worse' groups, respectively. The magnitude of the SRMs for the New Conditions group was similar in sign and magnitude to those of the 'somewhat worse' HTQ subgroup.

Figure 1 shows the profile of mean changes in scale scores adjusted for age and gender. The profiles of change scores highlight the similarity between the New Conditions group and the 'somewhat worse' HTQ subgroup. These two groups only differed significantly in change scores for the physical function ($p=0.004$) and the role physical scales ($p=0.03$). Figure 1 also demonstrates the asymmetry in the magnitude of mean change scores on the HTQ. The negative mean changes of the 'somewhat worse' and 'much worse' subgroups were larger than the positive mean change of the corresponding 'somewhat' and 'much better' subgroups.

(Note that the mean change scores in Tables 4 and 5 are unadjusted, while the change scores in Figure 1 are adjusted for age and gender. Adjustment did not change the results substantially.)

****

TABLEs 4 and 5 and Figure 1 about here

****

In Tables 4 and 5, the standard deviations in change scores are large relative to the mean change for all the SF-36 scales. This means that a proportion of respondents who reported retrospective change in health status on the HTQ had prospective change scores that were in the opposite direction to their HTQ category. This is illustrated in Figure 2, which shows the distributions of the individual (within-

person) change scores for the general health scale for each of the HTQ sub-groups and for the New Conditions group. For example, of the 92 people who did not have a new health condition at Wave 2 and reported feeling much worse at Wave 2, 75% had a negative change score (reflecting deterioration in their prospectively measured health consistent with their HTQ response), but the remaining 25% had a positive change score (reflecting improvement in their prospectively measured health, which was not consistent with their HTQ response). The same was true for the 949 people who reported feeling somewhat worse at Wave 2, as reflected in the location of the third quartile (upper edge of the box) of their distribution at the zero change line. Somewhat larger proportions of inconsistent observations were seen in the two HTQ improved groups (since the third quartiles of their distributions were below the zero change line). For the 'no change' HTQ subgroup, approximately 50% of the distribution was above or below zero. About 50% of the individuals in this group had prospective change scores within about 10 points of zero, but some individuals had quite large change scores, both positive and negative.

****

FIGURE 2 about here

****

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 mean group differences in prospective within-person changes in the SF-36 scales. The mean change in scale scores of the group who did not report a new long-term health condition was close to zero for all SF-36 scales, while there was an appreciable negative mean change in scale scores for those who had recently developed a new long-term health problem. This demonstrated that the prospective measures of change were responsive to our selected anchor of known clinical change, a necessary pre-requisite for our proposed calibration method. For the sub-group who described their health as 'somewhat worse' than one year ago, the mean changes in scores of the majority SF-36 scales, were in the vicinity of those of the New Conditions group. We interpret these findings as indicating that those who retrospectively reported their health as 'somewhat worse' than a year ago experienced a decline in health similar in magnitude to those who had developed a long-term health condition in the last year. These results therefore provide a clinical interpretation for the 'minimal detectable difference' measured by the category 'somewhat worse' on the HTQ.

Excluding respondents with a new long-term condition from the HTQ subgroups allowed a comparison of mutually exclusive groups. However the 'somewhat worse' HTQ sub-group lost the greatest proportion of respondents to the New Conditions group, since 44% of the New Conditions group reported their health as 'somewhat worse' than a year ago. We note that the definition of the New Conditions group required that the condition was serious enough to affect daily activities, and that this definition has some overlap with items of both the physical function and role physical scales of the SF-36. Thus the larger declines on the physical functioning and role physical scales for those with a new condition relative to those who report their health as 'somewhat worse' may be an artefact of the removal of respondents with predominantly physical conditions from the HTQ sub-groups to the New Conditions group. Even within the 'somewhat worse health' subgroup the change profile still showed greatest declines on the physical scales which may indicate that when physical roles and functioning decline that is when respondents notice that their health has become worse

Twenty three percent of the HTQ subgroups had an existing long-term condition at baseline. However existing conditions did not explain decline on the domain scales for the HTQ subgroups (results not shown) and the causes of decline in health status for those without a new long-term condition remain unexplained.

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 [4, 9, 19] . Our findings indicate that, at least at the group level, the retrospective HTQ is a valid and useful question that can distinguish groups of respondents who improve from those who deteriorate. This makes the SF-36 HTQ a reasonable candidate for cross-sectional population surveys which require a measure of change in health status, and are looking for brief measures to reduce the response burden. However, our recommendation for using the retrospective SF-36 HTQ as

a valid proxy for prospectively measured change in health is moderated by some important caveats.

The variability in individual change scores indicates some level of misclassification using the HTQ. One potential source of misclassification is present state bias which our results demonstrated. Responses to the HTQ were positively correlated with follow up scale scores, but were NOT similarly negatively correlated with baseline scores, indicating that respondents were strongly influenced by their present health status when assessing recent changes in health status [6, 7]. The SF-36 HTQ may be prone to present state bias because it
is based on a one-year recall period, whereas other HTQ are based on much shorter periods. The global nature of the HTQ may also increase the risk of bias, where multi-item domain-based transition questions may elicit more accurate recall [8].

However not all variability in change scores necessarily indicates misclassification. There was a similar amount of variability in the change scores for the New Conditions group. Less misclassification error would be expected in this more 'objective' clinical change group. Variability may be explained by the global question on new conditions that did not specify which new conditions the respondent had developed. Therefore change in health status for the New Conditions group would be expected to be heterogenous across domains which would contribute to the observed variability in change scores. New diagnosis may also result in improvement for some of this group as they respond to the commencement of treatment.

Our results demonstrate that the HTQ is valid for describing the average clinical change in health status at the group level, and within group variability does not neccessarily disqualify the HTQ for this purpose. However the variability of change scores does indicate that the HTQ cannot reliably predict the magnitude of clinical change experienced by any particular individual, and if used to categorise individuals into groups be extent of change, a sizable proportion (but not a majority) will be misclassified. However, decline (or improvement) in health status in a single domain may prompt a respondent to consider his/her health has worsened (or improved) even if other domains change in the other direction and this could explain some of the variability in change scores on the other domains These limitation of the HTQ as a measure of change need to be acknowledged and taken into account, not only in the design of future health outcome surveys, but also in the use of the HTQ as an anchor for assessing minimum clinically important difference

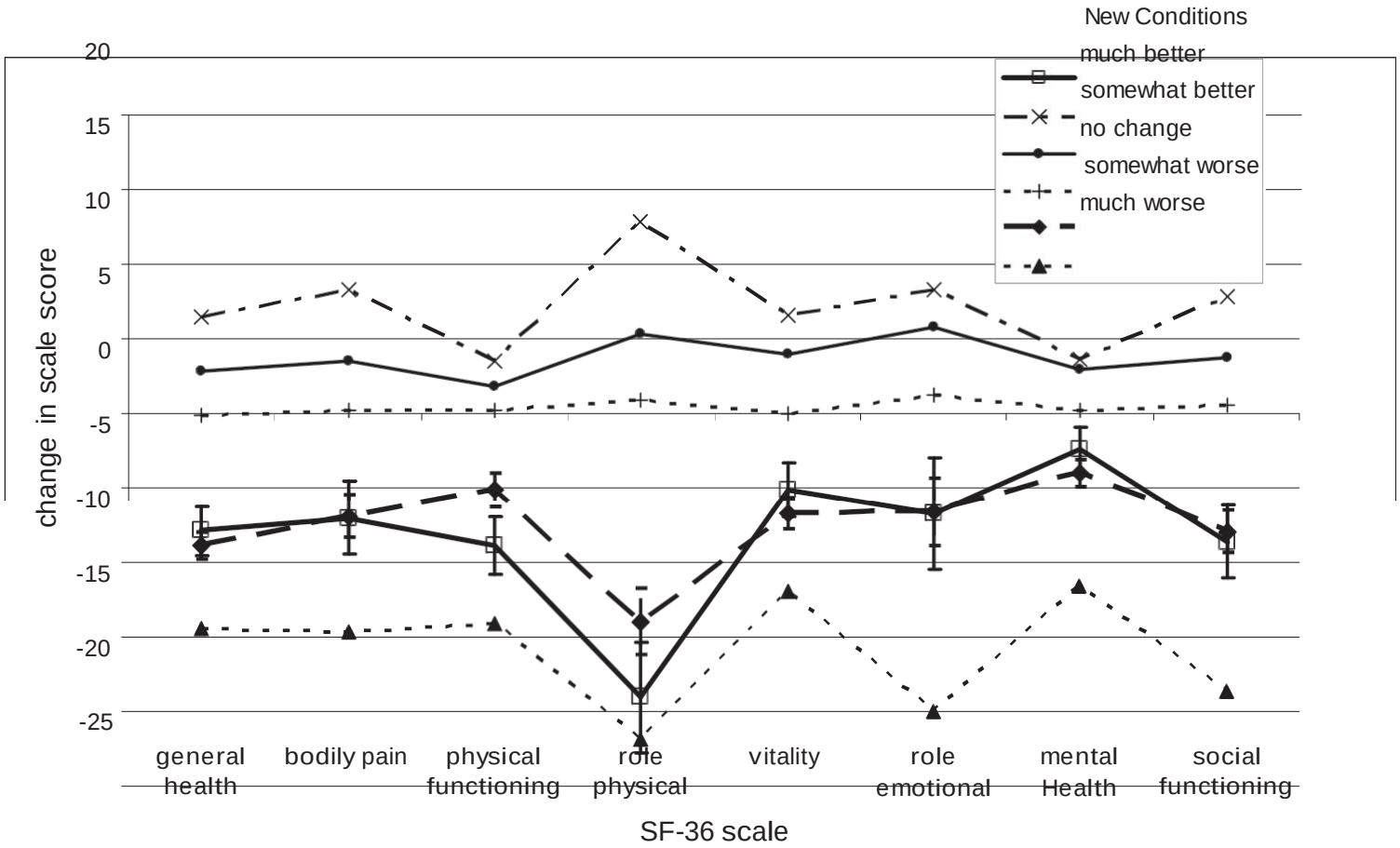
The clinical meaning of 'somewhat worse health than a year ago' may be context specific and may have different clinical meanings between clinical and general populations [20]. Nevertheless, the results presented in this paper contribute to our understanding and interpretation of the minimal detectable difference for the SF-36 HTQ.

References

1. Ware, J.E., Jr. , et al., SF-36 Health Survey: Manual and Interpretation Guide. 1993, Boston: The Health Institute.
2. ABS, 2004-05 National Health Survey: Users' Guide. Cat No 4363.0.55.001. 2006, Australian Bureau of Statistics.
3. Sanson-Fisher, R. and J. Perkins, Adaptation and validation of the SF-36 Health Survey for use in Australia. Journal of Clinical Epidemiology 1998. 51(11): p. 961-967.
4. Hemingway, H., et al., Is the SF-36 a valid measure of change in population health? Results from the Whitehall II Study. BMJ, 1997. 315(7118): p. 1273-9.
5. ABS, National Health Survey: SF36 Population Norms, Australia. Cat No 43990.0. 1997, Australian Bureau of Statistics: Canberra.
6. Norman, G.R., P. Stratford, and G. Regehr, Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. Journal of Clinical Epidemiology, 1997. 50(8): p. 869-79.
7. Cella, D., E.A. Hahn, and K. Dineen, Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Quality of Life Research, 2002. 11(3): p. 207-21.
8. Middel, B., et al., Recall bias did not affect perceived magnitude of change in health-related functional status. Journal of Clinical Epidemiology, 2006. 59(5): p. 503-11.
9. Perneger, T.V., J.F. Etter, and A. Rougemont, Prospective versus retrospective measurement of change in health status: a community based study in Geneva, Switzerland. Journal of Epidemiology & Community Health, 1997. 51(3): p. 320-5.
10. Husted, J.A., et al., Methods for assessing responsiveness: a critical review and recommendations. Journal of Clinical Epidemiology, 2000. 53(5): p. 459-468.
11. Jaeschke, R., J. Singer, and G.H. Guyatt, Measurement of health status: Ascertaining the minimal clinically important difference. Controlled Clinical Trials, 1989. 10: p. 407-415.
12. Wright, J.G., The Minimal Important Difference: Who's to Say What Is Important? Journal of Clinical Epidemiology, 1996. 49(11): p. 1221-1222.
13. Norman, G.R., J.A. Sloan, and K.W. Wyrwich, Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation.[see comment]. Medical Care, 2003. 41(5): p. 582-92.
14. Beaton, D.E., Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. Medical Care, 2003. 41(5): p. 593-596.
15. Deyo, R.A. and T.S. Inui, Toward clinical applications of health status measures: Sensitivity of scales to clinically important changes. Health Services Research, 1984. 19(3): p. 275-289.
16. Crosby, R.D., R.L. Kolotkin, and G.R. Williams, Defining clinically meaningful change in health-related quality of life. Journal of Clinical Epidemiology, 2003. 56(5): p. 395-407.
17. Watson, N. and M. Wooden, The HILDA Survey Four Years On. The Australian Economic Review, 2004. 37(3): p. 343-349.

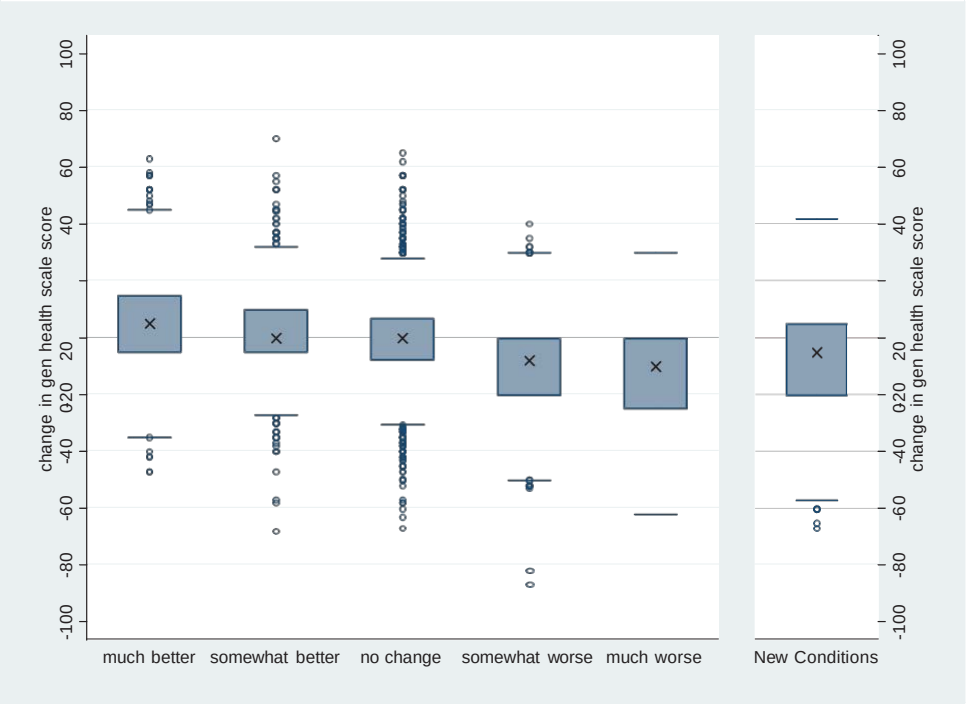
18. Cohen, J., Statistical Power Analysis for the Behavioral Sciences. 2 ed. 1988, Hillsdale, NJ: Lawrence Earlbaum Associates.
19. Mozes, B., Y. Maor, and A. Shmueli, Do we know what global ratings of health-related quality of life measure? Quality of Life Research, 1999. 8(3): p. 269-73.
20. McHorney, C.A., Health status assessment methods for adults: Past accomplishments and future challenges. Annual Review of Public Health, 1999. 20: p. 309-335.

Figure 1 Prospectively measured mean change scores[1] for those with a new long term health condition (New Conditions) at Wave 2, and by HTQ subgroup at Wave 2 for those without a new long term condition.



1. Calculated from Wave 1 to Wave 2, adjusted for age and gender; 95% confidence intervals for the New Conditions group (n=350) and those without a new long term condition who reported ‘somewhat worse’ health at Wave 2 (n=949)

Figure 2 Box-and-whisker plot of distribution of change in general health scale scores for each HTQ sub-group (without a new long-term health condition) and for the group with a new long-term health condition (New Conditions)



The box = interquartile range, from the lower quartile (Q1) to the upper quartile (Q3), thus contains 50% of the observations.
The lower whisker = Q1-1.5 x interquartile range
The upper whisker = Q3+1.5 x interquartile range
'x' indicates the mean
'o' indicates a value more extreme than the upper and lower whisker

Tables

Table 1: Respondent characteristics at baseline: total sample compared with respondents with those who developed a new long-term health condition at follow-up.

Respondent characteristics	New long-term health condition at follow-up (New Conditions group) n = 350 %	No New long-term health condition n = 9299 %	Total sample N = 9,649 %
Baseline (Wave 1)			
Gender (% female)	52.3	53.5	53.4
Mean age (years) at baseline	55.4	45.4	45.7
Self-assessed health at baseline (missing = 98)			
Excellent	5.5	15.4	15.1
Very Good	21.2	36.1	35.6
Good	35.4	32.9	33.0
Fair	29.0	12.6	13.2
Poor	9.0	3.0	3.2
Existing long-term health condition at baseline	48.0	22.9	23.8

Table 2: Distribution of HTQ responses at follow up for those who reported a new long-term condition at follow up (New Conditions group) compared with those who did not develop a new long-term condition at follow-up.

Respondent characteristics	New long-term health condition at follow-up (New Conditions group) % (n)	No New long-term health condition (HTQ sub-groups) % (n)	Total sample N = 9,649 %
Follow up (Wave 2)			
Much better now than one year ago	2.0 (7)	4.8 (450)	4.7
Somewhat better now than one year ago	8.3 (29)	11.1(1036)	11.0
About the same as one year ago	34.0 (119)	72.8 (6772)	71.4
Somewhat worse now than one year ago	44.0 (154)	10.2 (949)	11.4
Much worse now than one year ago	11.7 (41)	1.0(92)	1.4

Table 3: Correlation between retrospective HTQ with baseline SF-36 scale scores, follow up scores and change in scale scores.

SF-36 domains	Correlation of HTQ± with Wave 1 SF-36 scores*	Correlation of HTQ± with Wave 2 SF-36 scores*	correlation of HTQ± with prospective change in SF-36 scale 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) to more meaningfully align the HTQ the SF-36 scales in terms of positive correlation.

Table 4: Unadjusted mean change (standard deviation, SD) in SF-36 scores over the 12 months and standardised response mean (SRM) for respondents who had and had not developed a new long-term condition at follow-up.

	New Conditions group (n=350)		No new condition (n = 9299)	
SF-36 domain	Mean change (SD)	SRM	Mean change (SD)	SRM
general health	-8.0 (19.0)	-0.42	-0.5 (14.6)	-0.03
bodily pain	-7.2 (29.8)	-0.24	0.1 (22.1)	0.00
mental health	-2.5 (17.5)	-0.14	0.1 (15.0)	0.01
physical functioning	-9.1 (25.7)	-0.35	-0.1 (17.4)	-0.01
role emotional	-7.0 (45.0)	-0.16	1.1 (34.1)	0.03
role physical	-19.1 (45.3)	-0.42	0.3 (33.9)	0.01
social functioning	-8.9 (28.9)	-0.31	0.2 (22.7)	0.01
vitality	-5.3 (20.2)	-0.26	-0.1 (16.6)	-0.01
pcs	-5.4(11.3)	-0.48	-0.1 (7.7)	-0.01
mcs	-1.1(10.9)	-0.10	0.2 (9.2)	0.02

SRM effect sizes: small = 0.2 moderate = 0.5 large = 0.8

Table 5: Prospectively measured mean change (standard deviation) in SF-36 domain scores over the 12 months and standardised response means for respondents without a new long-term condition by retrospective self-report on the health transition question

Health transition category at Wave 2	SF-36 domain	Mean unadjusted change (SD) from Wave 1 to Wave 2	standardised response means
much better (n = 450)	general health	6.4(17.7)	0.36
	bodily pain	8.4(26.9)	0.31
	mental health	3.5(18.1)	0.19
	physical functioning	3.5(19.5)	0.18
	role emotional	8.4(41.7)	0.2
	role physical	12.7(38.4)	0.33
	social functioning	7.8(27.9)	0.28
	vitality	6.5(19.0)	0.34
	pcs	2.8(9.0)	0.31
	mcs	2.7(11.1)	0.24
somewhat better (n = 1036)	general health	2.7(15.1)	0.18
	bodily pain	3.5(21.9)	0.16
	mental health	2.9(15.8)	0.18
	physical functioning	1.8(18.8)	0.09
	role emotional	5.7(36.7)	0.16
	role physical	5(35.1)	0.14
	social functioning	3.7(23.7)	0.16
	vitality	3.9(16.9)	0.23
	pcs	0.8(8.1)	0.10
	mcs	1.9(10)	0.19
about the same (n = 6772)	general health	-.19(13.4)	-0.01
	bodily pain	0.2(21.1)	0.01
	mental health	0.2(14.0)	0.01
	physical functioning	0.3(16.5)	0.02
	role emotional	1.2(30.5)	0.04
	role physical	0.9(31.4)	0.03
	social functioning	0.6(21.0)	0.03
	vitality	0(15.5)	0.00
	pcs	0.1(7.1)	0.01
	mcs	0.2(8.4)	0.02
somewhat worse (n = 949)	general health	-8.8(16.1)	-0.54
	bodily pain	-6.8(24.0)	-0.29
	mental health	-3.9(17.0)	-0.23
	physical functioning	-5.1(19.0)	-0.27
	role emotional	-6.5(45.5)	-0.14
	role physical	-13.8(41.6)	-0.33
	social functioning	-7.9(26.0)	-0.30
	vitality	-6.7(18.4)	-0.36
	pcs	-3.3(9.2)	-0.36
	mcs	-2.8(11.0)	-0.25
much worse (n = 92)	general health	-14.1(20.8)	-0.68
	bodily pain	-14.5(28.1)	-0.52
	mental health	-11.3(21.4)	-0.53
	physical functioning	-14.1(25.7)	-0.55
	role emotional	-19.8(53.0)	-0.37
	role physical	-21.4(44.0)	-0.49
	social functioning	-18.5(31.9)	-0.58

	vitality	-11.7(24.1)	-0.49
	pes	-5.5(10.4)	-0.53
	mcs	-7.6(12.4)	-0.61

SRM effect sizes: small= 0.2 moderate= 0.5 large= 0.8

APPENDIX

SHOWCARD: List of chronic conditions shown to respondents
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