

Blome, Christine; Augustin, Matthias

**Working Paper**

## Measuring change in subjective wellbeing: Methods to quantify recall bias and recalibration response shift

HCHE Research Paper, No. 12

**Provided in Cooperation with:**

Hamburg Center for Health Economics (hche), University of Hamburg

*Suggested Citation:* Blome, Christine; Augustin, Matthias (2016) : Measuring change in subjective wellbeing: Methods to quantify recall bias and recalibration response shift, HCHE Research Paper, No. 12, University of Hamburg, Hamburg Center for Health Economics (HCHE), Hamburg

This Version is available at:

<https://hdl.handle.net/10419/145973>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



**Universität Hamburg**  
DER FORSCHUNG | DER LEHRE | DER BILDUNG

**hche**

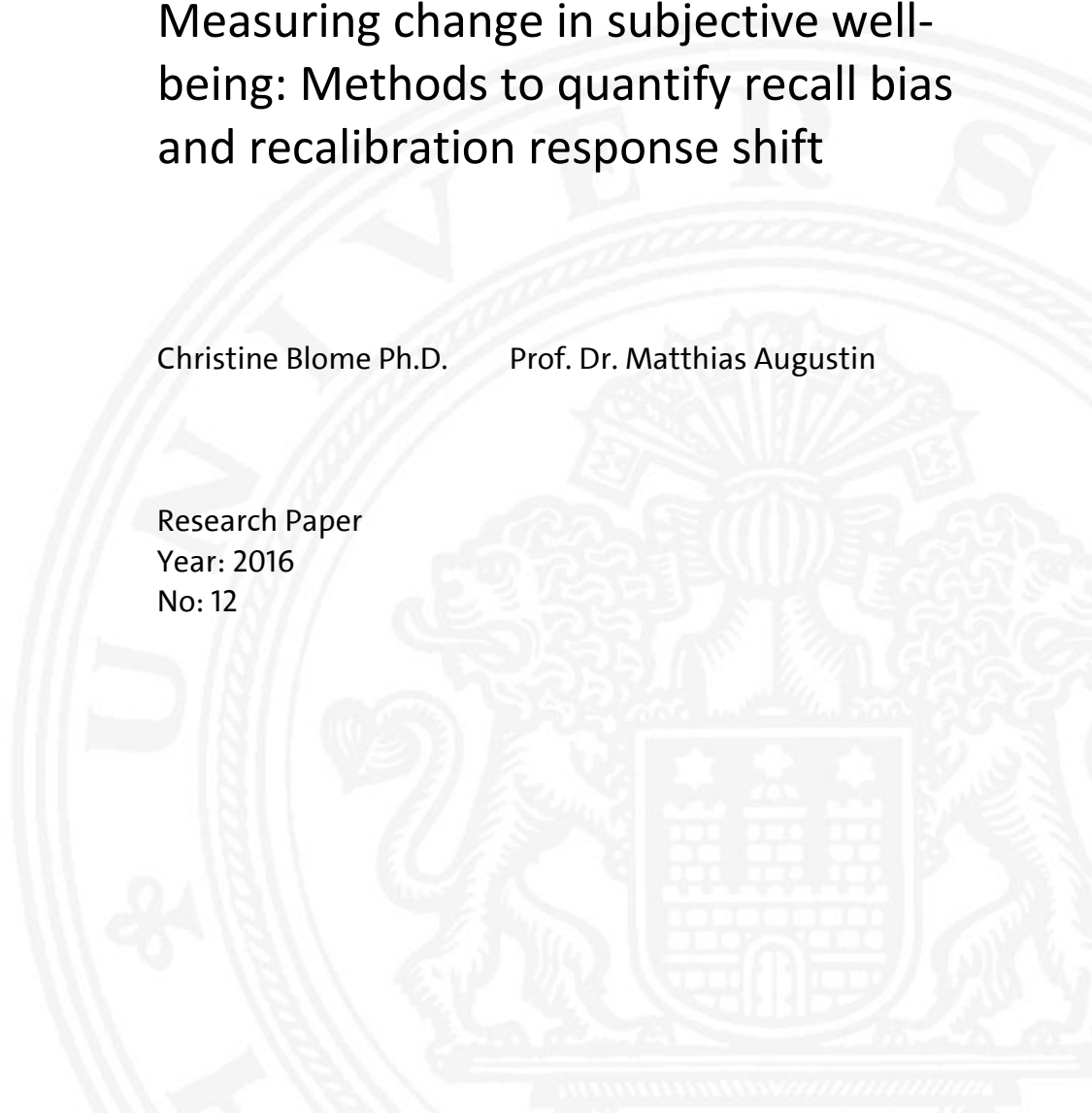
Hamburg Center  
for Health Economics

## Measuring change in subjective well-being: Methods to quantify recall bias and recalibration response shift

Christine Blome Ph.D.

Prof. Dr. Matthias Augustin

Research Paper  
Year: 2016  
No: 12





# Measuring change in subjective well-being: Methods to quantify recall bias and recalibration response shift

Christine Blome Ph.D.

Prof. Dr. Matthias Augustin

hche Research Paper No. 12

<http://www.hche.de>

## Abstract

We propose to use subjective well-being (SWB) measures to determine patient-relevant treatment benefit. Benefit can be measured either prospectively (pre-post) or retrospectively, but both approaches can be biased: Prospective evaluation may be subject to response shift; retrospective evaluation may be subject to recall bias. As prospective and retrospective evaluations often differ in effect size and since there is no gold standard to compare against, the extent of the two biases needs to be determined. Response shift includes reprioritization, reconceptualization, and recalibration. We argue that in SWB measures only recalibration, but not reprioritization and reconceptualization are validity threats.

We review approaches to quantify recall bias, response shift, or both in the measurement of health-related quality of life. We discuss which of these approaches are most suitable for application to SWB measurement, where only recall bias and recalibration are to be quantified, ignoring the other two response shift types.

Some approaches of bias detection will not be applicable to SWB measurement, because they do not distinguish between recalibration and other types of response shift, or quantify reprioritization and/or reconceptualization alone. For other approaches, it is unclear whether underlying assumptions apply to SWB measurement. Anchor recalibration, structural equation modelling, and ROSALI are most suitable, the latter two with some limitations. Anchor recalibration was considered by its developers to be too difficult for participants to understand in its current form.

Refining the anchor recalibration method may provide the most promising way to quantify both scale recalibration and recall bias.

Keywords: health-related quality of life; thentest; response shift; recall bias; scale recalibration; subjective well-being

Christine Blome, Ph.D.  
Institute for Health Services Research  
in Dermatology and Nursing

University Medical  
Center Hamburg-Eppendorf  
Martinistraße 52  
20246 Hamburg

Prof. Dr. Matthias Augustin  
Institute for Health Services Research  
in Dermatology and Nursing

University Medical  
Center Hamburg-Eppendorf  
Martinistraße 52  
20246 Hamburg

Disclosure:

The authors of this article have no conflict of interest.

## Introduction

### *Subjective well-being as an indicator of patient benefit*

It is important to determine patient-relevant treatment benefit, as medical treatment should not only improve objective indicators of health but should also be beneficial from the patients' perspective. One of the most relevant patient-reported outcomes (PRO) is health-related quality of life (HRQoL). It is usually understood as a rather broad construct including both subjective components and (subjective assessments of) objective components, as for example in [1]: "HRQOL (symptoms, functioning, perceptions of physical, mental, and social health)." Accordingly, items on these different dimensions of life are included in the typical HRQoL questionnaire.

As an alternative, we propose to determine patient-relevant treatment benefit with measures of subjective well-being (SWB), independent of objective functioning [2]. We thereby follow Dolan & Kahneman [3, 4] who proposed to measure "the Q in the QALY" [3] via the affect balance of patients who experience the health state in question – an approach they call "experienced utility". Accordingly, we propose using (the emotional component of) SWB [5] as an outcomes measure in clinical trials. One of the reasons we took this "radically subjective" [6] view is that focusing on SWB alone should make benefit assessment both strictly patient-centered and non-redundant to other indicators of patient benefit, such as morbidity or subjective health status, as argued before [2, 7]. Symptoms, for example, can certainly have a negative influence on SWB. This association however is far from perfect, both intra- and inter-individually, which might make symptoms a possible predictor of well-being but argues against them as a constituent part of it. In addition, symptoms are assessed in both HRQoL and morbidity questionnaires, which leads to difficulties in differentiating the two dimensions of benefit [8] as well as to potential double-counting of benefits. A further reason for our proposition is the finding that not all impairments have the same importance to all patients and that importance changes over time [9].

The affective component of SWB can be measured with retrospective questionnaires that assess positive and negative affect within a given time period (e.g., the Positive and Negative Affect Schedule (PANAS) [10]) or by asking patients to rate their momentary SWB with the help of mobile experience sampling or daily diaries [11-13].

*Bias in the measurement of change: response shift and recall bias*

Patient benefit is usually measured prospectively, i.e., by assessing PRO both before and after treatment. Alternatively, it can be measured only retrospectively (after treatment), for example by asking patients to both rate their current HRQoL and recall their former HRQoL, or to directly rate the change they have experienced. Prospective and retrospective assessments often differ in effect size or even in direction [14], raising the question of which approach measures change more accurately.

This question cannot be answered by convergent validation, because due to the subjective nature of PROs, there is no gold standard against which to compare prospective and retrospective assessments [15]; one can only test for plausible associations between the PRO and convergent criteria. These criteria are also measured either prospectively or retrospectively, making them susceptible to the same potential biases. We therefore need to determine the extent of bias in prospective and retrospective PRO assessment in order to estimate their validity. The two approaches are subject to different potential biases: Prospective evaluation may be biased by response shift, whereas retrospective evaluation may be biased by recall bias.

Recall bias occurs if patients completing a retrospective questionnaire cannot accurately recall their former state because of memory effects, including both directional and nondirectional effects [2, 16, 17]. Nondirectional recall bias means that, due to chance, prior states are sometimes underestimated and sometimes overestimated, whereas directional recall bias means that a patient mainly underestimates or mainly overestimates their former state [2].

Response shift is a concept that has been developed for HRQoL measurement, but is also applicable to other PROs. According to a seminal and widely adopted classification by Sprangers and Schwartz [14, 18], response shift includes three different processes: reprioritization (a change in the value the patient assigns to different areas of life), reconceptualization (a change in the patient's definition of what HRQoL is), and recalibration (a change in the patient's internal standards of measurement, leading to a different understanding of the response scale). Response shift can be viewed from either a measurement or conceptual perspective [1]. In this article, we take a measurement perspective, that is, we are interested in response shift processes as potential biases – factors that lead to a situation where observed change does not represent true change in the construct of interest [1].

Some of the literature describes all three types of response shift (also) as potential measurement biases to HRQoL assessment (e.g., as "validity threats" [19], "bias" [20], "attenuating or inflating ... effects" [21] or "confounder" [22]). Others do not see reprioritization and reconceptualization as potential sources of measurement error but define them as adaptation processes leading to changes in true HRQoL values [23-25, 26]. These adaptations have also been regarded as a wanted outcome [22, 27], for example in palliative care [20].

### *Bias in SWB measurement due to response shift*

From the measurement perspective, the question arises which of the three response shift phenomena can also introduce bias in the assessment of SWB.

Reprioritization can bias HRQoL assessments in the following way: HRQoL is often determined with a sum score over impairment items (such as pain or mobility restrictions). These sum scores are usually calculated without weighting, thereby implicitly weighting all items with 1. This equal importance of all HRQoL areas probably does not match the preferences of each patient in the first place – in other words, the potential influence on their quality of life will not actually be the same for all areas of life assessed in the questionnaire. If priorities change (i.e., reprioritization response



shift occurs, for example due to adaptation or coping), the sum scores will not be comparable between time points in terms of indicating the patient's HRQoL – the measurement of change will be biased.

The same is true if reconceptualization occurs: Items that matched a patient's definition of HRQoL at the first assessment may not do so anymore after this definition has changed, and vice versa.

In SWB measurement, in contrast, we argue that these two response shift types will not be a validity threat. Reprioritization implies that patients consider (and experience) other aspects of life to be important to their well-being than before. Their levels of SWB, however, will still be comparable when assessed directly with SWB measures – for example with the PANAS asking, among others, how "irritable" or "enthusiastic" patients were in a given period. Feeling good is still feeling good, irrespective of how the emotion has been achieved. The same applies to reconceptualization, where completely different aspects of life have become important for the patient's well-being. (This is not to say that adaptation processes leading to reprioritization and reconceptualization are not worth investigating. Quite the opposite; it is extremely important from a conceptual perspective to understand how adaptation mediates the association between objective health and SWB.)

Recalibration occurs when the same degree of the construct of interest will be labeled with a different response option than before. For example, a patient who needs a walking aid rates her mobility as "very much" impaired at age 50, but only as "somewhat" impaired twenty years later at age 70, comparing herself with other people her age at both assessments. When the construct of interest is objective mobility, the two responses will be incomparable due to recalibration. This can also happen to – and bias – SWB measures: The same emotional experience might be rated "moderately enthusiastic" at one point in time but "quite a bit enthusiastic" later, because the patient's understanding of these response options has changed. This will introduce bias, as two subjectively equal experiences of SWB are given different values.

Thus, we believe that recalibration, but not reprioritization and reconceptualization are validity threats to SWB measures.

### *Quantifying recall bias and response shift*

As argued above, the magnitude of recall bias and response shift needs to be quantified in order to find out if prospective or retrospective SWB assessment is less biased and thus measures change more validly. A range of different methods have been developed for quantifying recall bias and response shift in HRQoL assessment, where all three types of response shift are potential sources of bias to be accounted for [28-42].

In contrast, for SWB measurement, we have argued that reprioritization and reconceptualization represent true changes in well-being, and the two biases that must be quantified are recall bias and *scale recalibration*. Therefore, this article aims to answer the following question: What methods have been developed to quantify recall bias and/or scale recalibration, and which of these methods are most suitable to SWB assessment, where only recall bias and scale recalibration are to be quantified, ignoring the other two response shift types?

### **Approaches to quantifying recall bias and scale recalibration**

There are a range of different approaches to quantifying recall bias and/or scale recalibration, which we will outline and discuss below. Some aim to detect recall bias, some scale recalibration; others aim to do both. Further approaches aim to detect response shift types other than recalibration (i.e., reconceptualization and reprioritization).

### *Thentest*

The most common approach for determining either scale recalibration or recall bias in HRQoL is the thentest method [28, 29]. Patients rate their HRQoL at time 1 (pretest) and time 2 (posttest). At time 2, they also provide a retrospective rating of their HRQoL at time 1 (thentest). The thentest has been very useful for showing that prospective and retrospective effects can differ markedly and in both directions: sometimes, the effect is larger if determined prospectively, sometimes if determined retrospectively [14].

If pretest and thentest ratings differ, this can be interpreted as evidence of recall bias: Patients do not precisely remember their former HRQoL and thus make a biased thentest assessment. However, the pretest-thentest-difference can also be interpreted as evidence of scale recalibration: The patients' understanding of the response scale has changed so that they assign a different response category to their (correctly remembered) former HRQoL than they did in the pretest. Thus, differences between pretest and thentest can be explained by both recalibration and recall bias. The thentest alone is therefore not sufficient to distinguish between recall bias and scale recalibration and to determine true change values [23, 27, 30, 33, 43-45]; a thentest can only show that at least one of them is present. Critical discussion of the method further challenges the assumption underlying the thentest that cognitive processes are more similar between posttest and thentest than between pretest and posttest [46], and argues that thentest results are contaminated by social desirability responding [28, 47]. These limitations will apply to the use of the thentest method in SWB measurement, too.

In contrast, many studies using a thentest approach mainly interpreted the effect as scale recalibration (or, more generally, as response shift) and assumed the posttest-thentest-difference to be the true effect, while only briefly mentioning recall bias as an alternative explanation (e.g., as a study limitation) [13, 26, 27, 35]. Some discussed the possibility of recall bias but consider it negligible because (a) the patients were asked to re-evaluate their former health instead of recalling their former response and (b) the time between pretest and thentest was only three months [35]. Still others "assume

that in the case of a deeply felt change in health [...] recall will not cause memory difficulties for most respondents. Therefore, in our study we expect that recall bias did not have a major influence on the results" [27].

Other studies mainly (or only) discuss recall bias, assuming the pretest-posttest-difference to be the true effect [36-38].

### *Recall test*

An approach that explicitly aims to distinguish between recall bias and recalibration response shift in HRQoL involves performing a "recall test" in addition to the thentest [30, 31]. While the thentest asks patients to recall their former HRQoL, the recall test asks them to recall their pretest *response*. The difference between recall test and pretest is interpreted as pure memory effect (i.e., recall bias). It is subtracted from the total pretest-thentest-difference, which includes both scale recalibration and recall bias effects, leaving only recalibration effects. Unlike the thentest, this method thus has the advantage of differentiating between recall bias and scale recalibration. It is also easily applicable.

This approach assumes that any difference between pretest and recall test is due to incorrect recall, but does not take into account that the difference may also be caused by scale recalibration. When patients are asked to remember which answer they chose in the pretest, they can proceed as follows: They can either try to directly recall which answer they gave or try to recall their former state and reconstruct what they probably answered based on this memory. In the latter case, they could well base their answer on their current, and possibly recalibrated, understanding of the scale. There is only no room for scale recalibration in the recall test if patients are aware that their scale understanding has changed since the pretest and if they can recall their former understanding. We assume that people will likely more accurately recall their former SWB than their former understanding of a response scale, as their former SWB is of much higher importance to them personally. For these reasons, we assume that the

approach of using a recall test might not sufficiently distinguish recall bias from scale recalibration to quantify their relative extent in SWB measurement.

#### *Comparison with variables unaffected by scale recalibration*

A method by Schwartz et al. [32] also aims to distinguish recall bias from scale recalibration in HRQoL. Patients with multiple sclerosis completed a thentest on the use of assistive devices for ambulation – an area assumed not to be subject to scale recalibration because it is "a fact". Accordingly, any pretest-thentest-difference was interpreted as recall bias. This difference, expressed as percentage of overall variance, was then subtracted from the pretest-thentest-difference in other areas that were assumed to (possibly) be subject to scale recalibration on account of being "internal, subjective experiences," such as fatigue. The remaining variance was interpreted as due to scale recalibration, so the respective percentages of variance due to recall bias and scale recalibration could be compared. This is an elegant approach that may allow bias comparison in items on objectively verifiable content with items where this is not (or not easily) possible.

However, we think this method is limited when used for bias detection in SWB assessment, where one would have to assume that recall bias has the same magnitude in observable facts and the completely subjective experience of well-being. Findings from cognitive psychology show that recall can highly depend on the content to be remembered – for instance, pleasant events are remembered better than unpleasant ones, and classmate names are remembered better than street names [48]. It seems likely that patients also have a harder time remembering a prior degree of SWB than remembering whether they used a walking aid or not (although, to our knowledge, this has not yet been tested empirically). We therefore think this approach is not sufficient to determine the relative extent of recall bias and scale recalibration in SWB assessment.

#### *Anchor recalibration*

The anchor recalibration method aims to quantify scale recalibration in HRQoL measurement [33]. At time 1, patients are asked to describe how they understand the end points of the HRQoL response scale in their own words by stating what best and worst imaginable HRQoL mean to them, for example regarding physical capability. At time 2, patients again describe their current understanding of best and worst imaginable HRQoL for each item. They are then asked to compare their time 1 and time 2 descriptions, and to locate their time 1 descriptions on the scale as they understand it at time 2. The time 1 HRQoL data are transformed accordingly in order to statistically control for scale recalibration.

This method might also help to quantify recall bias by calculating the difference between pretest data adjusted for scale recalibration and an additional thentest.

However, according to the authors themselves, participants in the anchor recalibration exercise found it hard to determine "whether anchors over time coincided or not, and if not in which direction they differed" [33]. We therefore conclude that anchor recalibration is a promising approach for quantifying recalibration in SWB, but requires some further refinement to make it feasible for patients.

### *Structural equation modelling*

Structural equation modelling (SEM) has been proposed by Oort [34] to quantify response shift, including recalibration, in HRQoL data, and has since been widely applied for this purpose [49-53] as well as for quantifying recall bias by combining it with a thentest [29]. The method is based on the comparison of confirmatory factor analyses on HRQoL items assessed at two (or more) points in time, with non-invariance of specific SEM parameters between the time points considered evidence of different response shift processes. This statistically sophisticated approach has important advantages: No additional data assessments are needed (such as thentest, recall test, or qualitative descriptions of response options); it is designed to both detect response shift and determine true effect sizes, the latter by comparing common factor means, without being susceptible to recall bias; and it determines the three types of response

shift separately. In addition, Oort introduces an important distinction between two types of recalibration: *uniform recalibration*, where "all response options change in the same direction and to the same extent", which will affect the observed variables' means, and *non-uniform recalibration*, where the scale will "stretch or shrink", which will also affect the observed variables' variance and the covariance between them [34].

Uniform recalibration is assumed if the intercepts in the model differ between the two time points; in this case, changes in the mean of a variable cannot be explained by a corresponding change in the common factor's mean. If there is uniform recalibration in a variable, this will lead to a change in the variable mean but not in factor means. However, the converse is not necessarily true – as an alternative explanation, the variable-specific true value may have changed without any recalibration (i.e. a changed scale understanding without a corresponding true change in the construct to be measured), but not the true value of the common factor. A possible cause of such an effect is that an intervention has improved only and specifically one aspect of the construct that is measured by only this item (in the PANAS, for example, this would be one specific emotion). As pointed out by Donaldson (2005) [54], non-invariances found in the SEM approach are necessary, but not sufficient, conditions for concluding that response shift has occurred.

Non-uniform recalibration is assumed if the error variances differ between time points; then, changes in the variance of a variable cannot be explained by a corresponding change in the common factor's variance. Oort [34] states that "there may be other reasons, other than recalibration response shift, that may cause changes in intercepts and residual variances." As examples, he mentions coping and pain management, which may improve functioning and pain (measured by single variables) without improving physical health (the common factor), but argues that coping and pain management can be considered causes of recalibration. As a second alternative explanation we would add that the variance of the true, variable-specific value may have changed, for example because an intervention improved only one specific aspect of the construct measured by only one item (e.g., actual pain), but differently so in different patients, and without fully corresponding changes in the common factor (e.g.,

physical health). These true change in variable-specific aspects may even be *fully* accountable for the observed changes in means and variances, so that changes in the residual variances are indicative (but not proof) that scale recalibration has occurred.

This approach may be helpful to quantify recalibration effects in SWB instruments as well; however, alternative explanations to non-invariant intercepts and residual variances should also be taken into consideration.

#### *RespOnse Shift ALgorithm in Item response theory (ROSALI)*

Recently, a new IRT-based method has been proposed by Guilleux et al. called RespOnse Shift ALgorithm in Item response theory (ROSALI) [35]. The procedure allows to determine of both uniform and non-uniform recalibration, which is assumed if item difficulties change significantly. It also allows to determine reprioritization and can be used to estimate true change by controlling for different types of response shift. This approach has the big advantage of being robust to missing data; it can be used to control for recalibration only and may thus be applicable to SWB data, too. In this approach, item difficulty is assumed to indicate recalibration response shift – an assumption that seems plausible for the application to multi-item SWB measures as well. As in SEM, a limitation is that alternative explanations are possible, for example a change in the variable-specific true value – e.g., a change in a single emotion without a corresponding change in overall SWB.

#### *Approaches not designed to quantify the effect of recalibration alone*

Several approaches to determine response shift in HRQoL will not be applicable for SWB instruments because they were not designed to quantify the specific effect of recalibration, which is the only response shift process we assume to bias SWB measurement. These methods will be outlined briefly.

Latent trajectory analysis is a specific type of structural equation modeling that examines patterns in discrepancies between predicted and observed HRQoL values [36,



37, 38]. It does not distinguish between different types of response shift and thus, it is not meant to detect the specific effect of recalibration.

Recursive partitioning tree analysis uses a nonparametric statistical index to iteratively divide respondents into increasingly homogenous subgroups [38, 39]. In contrast to latent trajectory analysis, it does distinguish scale recalibration from the other two types of response shift. However, it *quantifies* only the overall size of response shift. Scale recalibration is not quantified; instead, qualitative indicators determine whether recalibration is present or not. This method would therefore also not be applicable if one wished to *quantify* recalibration response shift *only*.

A related method is growth curve analysis [18]. It does not measure the different types of response shift. The authors state that it is "useful as a first step in primary or secondary analysis to determine whether response shift is likely to have occurred" [18], but not to quantify scale recalibration.

Lix et al. [40] use different statistical measures of relative importance (i.e., measures that "discriminate between groups or predict group membership", to detect reprioritization response shift. These methods also do not aim to quantify scale recalibration.

A very straightforward approach to assessing response shift is direct assessment, that is, to ask patients if they have experienced any. Hinz et al. [41] asked urologic cancer patients: "In the last three months, has your opinion about what health is changed?", with response options ranging from "not at all" to "completely." This approach only addresses reconceptualization response shift, but is not designed to quantify scale recalibration.

Vignettes – short descriptions of health states – have been used by Korfage et al. to detect response shift in patients with prostate cancer [42]; health states vignettes were based on the EQ-5D plus information on urinary, bowel and erectile dysfunction. Patients evaluated the vignettes on visual analogue scales anchored 0 = "very bad" to 10 = "very good," and a change in evaluations over time was mainly interpreted as evidence of reprioritization response shift. (Scale recalibration may also influence

patients' responses in this method, as their understanding of the magnitude of "very bad" may have changed in addition to a change in priorities.)

## **Conclusion**

We proposed to use SWB as an indicator of patient-relevant treatment benefit, and argued that in SWB, only scale recalibration, but not reprioritization and reconceptualization, is a source of bias. Many approaches to quantify the relative extent of response shift and/or recall bias in HRQoL have some limitations when used for SWB measurement. Some of them have not been designed to quantify the specific effect of scale recalibration in the first place. Others may not reliably distinguish between scale recalibration and recall bias in SWB (thentest, recall test) or are less suitable for SWB instruments because they are based on the assumption that recall bias has the same magnitude in variables on objective and subjective aspects (approach of comparing with variables unaffected by scale recalibration). To quantify scale recalibration in SWB, the anchor recalibration method [33], structural equation modelling [34], and ROSALI [35] are most suitable.

In the anchor recalibration approach, patients found it difficult to compare the two sets of scale anchors they had provided at different points in time. One possible way to enhance the method's feasibility may be to have the patients describe each response option of each item in their own words (instead of the highest and lowest option only) at time 1, and at time 2, have them first complete the questionnaire. Afterwards, their scale descriptions of time 1 should be presented to them, and they should then complete the questionnaire again according to this former understanding. Differences between the two questionnaire assessments at time 2 could then be assigned to scale recalibration. In this version of anchor recalibration, the challenging task of comparing different anchors would no longer be necessary. Adding a thentest to this procedure might also help quantify recall bias. This approach would, however, only be feasible for short questionnaires where the overall number of response options to be described is

limited; otherwise, the procedure would probably be too lengthy and thus burden participants.

We conclude that for the measurement of SWB, only scale recalibration is a source of bias, but not reprioritization and reconceptualization. For SWB, refining the anchor recalibration method may provide the most promising way to quantify both scale recalibration *and* recall bias – and thereby tell if prospective or retrospective measurement of change in SWB is more valid.

### **Funding**

This study was funded by the German Federal Ministry of Education and Research (BMBF) within the context of the Hamburg Centre for Health Economics (HCHE), grant no. 01EH1101B.

### **Ethical approval**

This article does not contain any studies with human participants or animals performed by any of the authors.

## References

1. Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, 62(11), 1126-1137.
2. Blome, C., & Augustin, M. (2015). Measuring Change in Quality of Life: Bias in Prospective and Retrospective Evaluation. *Value in Health*, 18(1), 110–115.
3. Dolan, P. (2008). Developing methods that really do value the 'Q' in the QALY. *Health Economics, Policy and Law*, Jan; 3(Pt 1), 69-77.
4. Dolan, P., & Kahneman, D. (2008). Interpretations of utility and their implications for the valuation of health. *The Economic Journal*, 118(525), 215–234.
5. Pavot, W., & Diener, E. (2013). Happiness experienced: The science of subjective well-being. In S. David, I. Boniwell, & A.C. Ayers (Eds.), *The Oxford handbook of happiness* (pp. 134-151). Oxford, UK: Oxford University Press.
6. Wilm, S., Leve, V., & Santos, S. (2014). Ist Lebensqualität das, was Patienten wirklich wollen? Einschätzungen aus einer hausärztlichen Perspektive. [Is it quality of life that patients really want? Assessment from a general practitioner's perspective.] *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 108(2-3), 126-129.
7. Blome, C. (2016). Lebensqualität als radikal subjektives Wohlbefinden: methodische und praktische Implikationen [Quality of life as radically subjective well-being: methodological and practical implications]. In R. Kipke, L. Kovács L & R. Lutz (Eds.), *Lebensqualität in der Medizin. Messung, Konzept, Konsequenzen* (pp. 223-236). Wiesbaden: Springer.
8. Lohrberg, D., Augustin, M., & Blome, C. (2016). The Definition and Role of Quality of Life in Germany's Early Assessment of Drug Benefit: A Qualitative Approach. *Quality of Life Research*, 25(2), 447-455.
9. Blome, C., Gosau, R., Radtke, M. A., Reich, K., Rustenbach, S. J., Spehr, C., et al. (2015). Patient-relevant treatment goals in psoriasis. *Archives of Dermatological Research*, 2015 Dec 19. [Epub ahead of print]
10. Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070.
11. Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks, CA: Sage.
12. Scollon, C. N., & Kim-Prieto, C. (2003). Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses. *Journal of Happiness Studies*, 4, 5-34.
13. Diener, E., & Tay, L. (2013). Review of the Day Reconstruction Method (DRM). *Social Indicators Research*, 116(1), 255-267.
14. Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A. G., & Fayers, P. M. (2006). The clinical significance of adaptation to changing health: a meta-analysis of response shift. *Quality of Life Research*, 15(9), 1533–1550.
15. Mookink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010). Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Medical Research Methodology*, 10, 82.
16. Mancuso, C.A., & Charlson, M.E. (1995). Does recollection error threaten the validity of cross-sectional studies of effectiveness? *Medical Care*, 33(Suppl), A577–88.
17. Coughlin, S. S. (1990). Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology*, 43(1), 87-91.

18. Schwartz, C. E., & Sprangers, M. A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine*, 48(11), 1531–1548.
19. Sprangers, M. A. & Schwartz, C. E. (2010). Do not throw out the baby with the bath water: build on current approaches to realize conceptual clarity. Response to Ubel, Peeters, and Smith. *Quality of Life Research*, 19(4), 477–479.
20. Preston, N. J., Fayers, P., Walters, S. J., Pilling, M., Grande, G. E., Short, V., et al. (2013). Recommendations for managing missing data, attrition and response shift in palliative and end-of-life care research: part of the MORECare research method guidance on statistical issues. *Palliative Medicine*, 27(10), 899-907.
21. Dabakuyo, T. S., Guillemin, F., Conroy, T., Velten, M., Jolly, D., Mercier, M., et al. (2013). Response shift effects on measuring post-operative quality of life among breast cancer patients: a multicenter cohort study. *Quality of Life Research*, 22(1), 1–11.
22. Barclay-Goddard, R., Epstein, J. D., & Mayo, N. E. (2009). Response shift: a brief overview and proposed research priorities. *Quality of Life Research*, 18(3), 335-346.
23. Ubel, P. A., Peeters, Y., & Smith, D. (2010). Abandoning the language of "response shift": a plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life. *Quality of Life Research*, 19(4), 465–471.
24. Ubel, P. A., & Smith, D. M. (2010). Why should changing the bathwater have to harm the baby? *Quality of Life Research*, 19(4), 481–482.
25. McClimans, L., Bickenbach, J., Westerman, M., Carlson, L., Wasserman, D., & Schwartz, C. (2013). Philosophical perspectives on response shift. *Quality of Life Research*, 22(7), 1871-1878.
26. Schwartz, C. E., & Rapkin, B. D. (2004). Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health and Quality of Life Outcomes*, 2:16.
27. Nolte, S., Elsworth, G. R., Sinclair, A. J., Osborne, R. H. (2012). The inclusion of 'then-test' questions in post-test questionnaires alters post-test responses: a randomized study of bias in health program evaluation. *Quality of Life Research*, 21(3), 487-494.
28. Schwartz, C. E., & Sprangers, M. A. G. (2010). Guidelines for improving the stringency of response shift research using the then-test. *Quality of Life Research*, 19(4), 455–464.
29. Verdam, M. G. E., Oort, F. J., Visser, M. R. M., & Sprangers, M.A.G. (2012). Response shift detection through then-test and structural equation modelling: Decomposing observed change and testing tacit assumptions. *Netherlands Journal of Psychology*, 67(3), 58-67
30. Schwartz, C. E., Rapkin, B. D., & Rapkin, B. A. (2012). Understanding appraisal processes underlying the then-test: a mixed methods investigation. *Quality of Life Research*, 21(3), 381–388.
31. McPhail, S., & Haines, T. (2010). Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health and Quality of Life Outcomes*, 8, 65.
32. Schwartz, C. E., Sprangers, M. A., & Carey, A. et al. (2004). Exploring response shift in longitudinal data. *Psychology & Health*, 19(1), 51–69.
33. Visser, M. R. M., Oort, F. J., & Sprangers, M. A. G. (2005). Methods to detect response shift in quality of life data: a convergent validity study. *Quality of Life Research*, 14(3), 629–639.
34. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, 14(3), 587-598.
35. Guilleux, A., Blanchin, M., Vanier, A., Guillemin, F., Falissard, B., Schwartz, C. E., Hardouin, J. B., & Sébille, V. (2015). RespOnse Shift ALgorithm in Item response theory

- (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcome studies. *Quality of Life Research*, 24(3), 553-564.
36. Ahmed, S., Mayo, N., Scott, S., Kuspinar, A., & Schwartz, C. (2011). Using latent trajectory analysis of residuals to detect response shift in general health among patients with multiple sclerosis article. *Quality of Life Research*, 20(10), 1555–1560.
  37. Mayo, N. E., Scott, S. C., Bernstein, C. N., & Lix, L. M. (2015). How are you? Do people with inflammatory bowel disease experience response shift on this question? *Health and Quality of Life Outcomes*, 13:52.
  38. Schwartz, C. E., Sprangers, M. A., Oort, F. J., Ahmed, S., Bode, R., Li, Y., et al. (2011). Response shift in patients with multiple sclerosis: an application of three statistical techniques. *Quality of Life Research*, 20(10), 1561–1572.
  39. Li, Y., & Schwartz, C. E. (2011). Data mining for response shift patterns in multiple sclerosis patients using recursive partitioning tree analysis. *Quality of Life Research*, 20(10), 1543–1553.
  40. Lix, L. M., Sajobi, T. T., Sawatzky, R., Liu, J., Mayo, N. E., Huang, Y., et al. (2013). Relative importance measures for reprioritization response shift. *Quality of Life Research*, 22(4), 695-703.
  41. Hinz, A., Finck Barboza, C., Zenger, M., Singer, S., Schwalenberg, T., & Stolzenburg, J.-U. (2011). Response shift in the assessment of anxiety, depression and perceived health in urologic cancer patients: an individual perspective. *European Journal of Cancer Care*, 20(5), 601–609.
  42. Korfage, I. J., de Koning, Harry J, & Essink-Bot, M.-L. (2007). Response shift due to diagnosis and primary treatment of localized prostate cancer: a then-test and a vignette study. *Quality of Life Research*, 16(10), 1627–1634.
  43. Hill, L. G. (2005). Revisiting the Retrospective Pretest. *American Journal of Evaluation*, 26(4), 501–517.
  44. Norman, G. (2003). Hi! How are you? Response shift, implicit theories and differing epistemologies. *Quality of Life Research*, 12(3), 239–249.
  45. Williamson, O. D., Gabbe, B. J., Sutherland, A. M., & Hart, M. J. (2013). Does recall of preinjury disability change over time? *Injury Prevention*, 19(4), 238–243.
  46. Taminiau-Bloem, E. F., Schwartz, C. E., van Zuuren, F. J., Koeneman, M. A., Visser, M. R., Tishelman, C., et al. (2015). Using a retrospective pretest instead of a conventional pretest is replacing biases: a qualitative study of cognitive processes underlying responses to thentest items. *Quality of Life Research*, 2015 Nov 16. [Epub ahead of print].
  47. Sprangers, M. (1989). Subject bias and the retrospective pretest in retrospect. *Bulletin of the Psychonomic Society*, 27(1), 11-14.
  48. Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*: Cambridge: Cambridge University Press.
  49. Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, 14(3), 599–609.
  50. King-Kallimanis, B. L., Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Structural equation modeling of health-related quality-of-life data illustrates the measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, 62(11), 1157-1164.
  51. Gandhi, P. K., Ried, L. D., Huang, I. C., Kimberlin, C. L., & Kauf, T. L. (2013). Assessment of response shift using two structural equation modeling techniques. *Quality of Life Research*, 22(3), 461-471.

52. Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., & Cohen, S. R. (2005). The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *Journal of Clinical Epidemiology*, 58(11), 1125-1133.
53. Ahmed, S., Bourbeau, J., Maltais, F., & Mansour, A. (2009). The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique. *Journal of Clinical Epidemiology*, 62(11), 1165-1172.
54. Donaldson, G. W. (2005). Structural equation models for quality of life response shifts: promises and pitfalls. *Quality of Life Research*, 14(10), 2345-2351.

**hche Research Paper Series**, ISSN 2191-6233 (Print), ISSN 2192-2519 (Internet)

---

- 2011/1 Mathias Kifmann and Kerstin Roeder, Premium Subsidies and Social Insurance: Substitutes or Complements? March 2011
- 2011/2 Oliver Tiemann and Jonas Schreyögg, Changes in Hospital Efficiency after Privatization, June 2011
- 2011/3 Kathrin Roll, Tom Stargardt and Jonas Schreyögg, Effect of Type of Insurance and Income on Waiting Time for Outpatient Care, July 2011
- 2012/4 Tom Stargardt, Jonas Schreyögg and Ivan Kondofersky, Measuring the Relationship between Costs and Outcomes: the Example of Acute Myocardial Infarction in German Hospitals, August 2012
- 2012/5 Vera Hinz, Florian Drevs, Jürgen Wehner, Electronic Word of Mouth about Medical Services, September 2012
- 2013/6 Mathias Kifmann, Martin Nell, Fairer Systemwettbewerb zwischen gesetzlicher und privater Krankenversicherung, July 2013
- 2013/7 Mareike Heimeshoff, Jonas Schreyögg, Estimation of a physician practise cost function, August 2013
- 2014/8 Mathias Kifmann, Luigi Siciliani, Average-cost Pricing and Dynamic Selection Incentives in the Hospital Sector, October 2014
- 2015/9 Ricarda Milstein, Jonas Schreyögg, A review of pay-for-performance programs in the inpatient sector in OECD countries, December 2015
- 2016/10 Florian Bleibler, Hans-Helmut König, Cost-effectiveness of intravenous 5 mg zoledronic acid to prevent subsequent clinical fractures in postmenopausal women after hip fracture: a model-based analysis, January 2016
- 2016/11 Yauheniya Varabyova, Rudolf Blankart, Jonas Schreyögg, Using Nonparametric Conditional Approach to Integrate Quality into Efficiency Analysis: Empirical Evidence from Cardiology Departments, May 2016
- 2016/12 Christine Blome Ph.D., Prof. Dr. Matthias Augustin, Measuring change in subjective well-being: Methods to quantify recall bias and recalibration response shift



The Hamburg Center for Health Economics is a joint center of Universität Hamburg and the University Medical Center Hamburg-Eppendorf (UKE).



# hche | Hamburg Center for Health Economics

Esplanade 36  
20354 Hamburg  
Germany  
Tel: +49 (0) 42838-9515/9516  
Fax: +49 (0) 42838-8043  
Email: [info@hche.de](mailto:info@hche.de)  
<http://www.hche.de>  
ISSN 2191-6233 (Print)  
ISSN 2192-2519 (Internet)

HCHE Research Papers are indexed in RePEc and SSRN.  
Papers can be downloaded free of charge from <http://www.hche.de>.