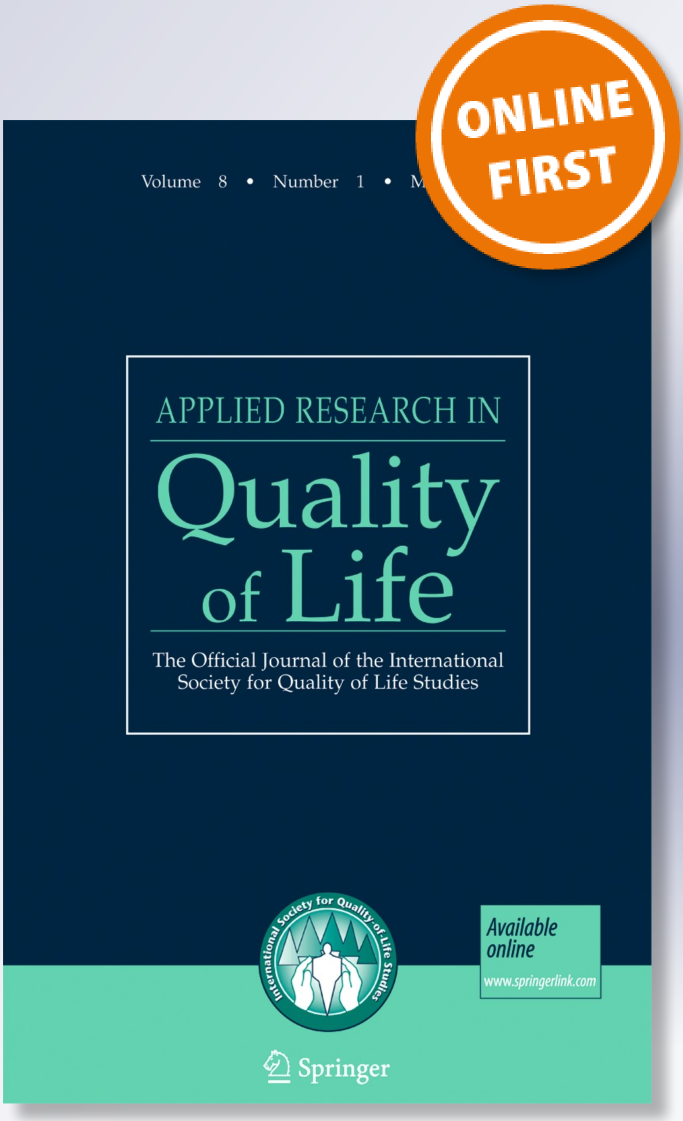
# *Threats to the Internal Validity of Spinal Surgery Outcome Assessment: Recalibration Response Shift or Implicit Theories of Change?*

**Joel A. Finkelstein, Brian R. Quaranto & Carolyn E. Schwartz**

**Applied Research in Quality of Life**
The Official Journal of the International
Society for Quality-of-Life Studies

ISSN 1871-2584

Applied Research Quality Life
DOI 10.1007/s11482-013-9221-2



ONLINE FIRST

Volume 8 • Number 1 • M

APPLIED RESEARCH IN
Quality
of Life

The Official Journal of the International
Society for Quality of Life Studies

International Society for Quality-of-Life Studies

Available online
www.springerlink.com

Springer

Applied Research Quality Life
DOI 10.1007/s11482-013-9221-2

# Threats to the Internal Validity of Spinal Surgery Outcome Assessment: Recalibration Response Shift or Implicit Theories of Change?

Joel A. Finkelstein · Brian R. Quaranto · Carolyn E. Schwartz

Received: 14 August 2012 / Accepted: 17 February 2013

**Abstract** A recalibration response shift will cause the patient to think about a self-report measure's response options differently after a health state change. Commonly assessed using the retrospective-pretest design ("then-test"), recent guidelines suggest adjusting then-test estimates for competing explanations. This prospective longitudinal study investigated recalibration response shift after adjusting for implicit theories of change in patients undergoing spinal surgery. The Oswestry Disability Index (ODI) and Short Form-36 (SF-36) were collected before surgery and at 6 weeks and 3 months after spinal decompression surgery. Then-tests of the measures were also collected at all post-tests. Recalibration response shift was operationalized as the then-minus-pre difference score on the evaluative SF-36. Implicit theories of change were operationalized as the then-minus-pre difference score on the perception-based ODI. Improved vs. No-Effect patient groups were compared using the Minimally Important Difference (±15 points) as a cut-off on the Visual Analogue Scale (VAS) items for back and leg pain. Logistic regression analyses investigated whether recalibration response shift had an independent effect distinguishing patient groups,

**Electronic supplementary material** The online version of this article (doi:10.1007/s11482-013-9221-2) contains supplementary material, which is available to authorized users.

J. A. Finkelstein
Division of Orthopaedics, Sunnybrook Health Sciences Center, Sunnybrook Center for Spinal Trauma, University of Toronto, Toronto, ON, Canada

B. R. Quaranto · C. E. Schwartz (✉)
DeltaQuest Foundation, Inc, 31 Mitchell Road, Concord, MA 01742, USA
e-mail: carolyn.schwartz@deltaquest.org

C. E. Schwartz
Department of Medicine and Orthopaedic Surgery, Tufts University Medical School, Boston, MA, USA

Published online: 16 March 2013

after adjusting for implicit theories of change. The sample (baseline $n=169$, mean age 52, 39 % female) was well-educated, and 1/3 were working. All then-minus-pre difference scores were non-zero at each time point and were stable over time. In the adjusted models distinguishing Improved versus No Effect groups, then-minus-pre ODI difference scores were significant in the majority of the adjusted models at all timepoints, but only one then-minus-pre SF-36 difference score—for physical functioning recalibration—was significant and only at 6-weeks post-surgery. This suggests that implicit theories of change bias the estimation of post-surgical outcomes, but that recalibration response shift biased only the estimation of physical functioning and only at 6 weeks post-surgery. Recalibration response shift and implicit theories of change can both be sources of bias in patient-reported outcome assessment. Our findings suggest that implicit theories of change are a greater threat to validity in this patient sample. Future research using the then-test should control for implicit theories of change to minimize misspecification of effects.

**Keywords** Response shift · Implicit theories of change · Response bias · Minimal clinically important difference · Then-test · Spine surgery · Functional outcomes

## Introduction

Response shift is a psychological process whereby over time, a patient will alter his/her self-evaluation of health-related constructs such as health-related quality of life (QOL), disability or pain. This can be due to: 1) changes in internal standards (i.e., recalibration); 2) changes in values (i.e., reprioritization); or, 3) redefinition of the concept (reconceptualization)(Sprangers and Schwartz 1999). A response shift will cause the patient to rate the same self-report measure differently in retrospection. Interest in response shift began in the 1990's when clinicians began to recognize that this phenomenon could obfuscate important treatment-related changes. Response shift has been studied and recognized in patients with multiple sclerosis (Schwartz et al. 2004), cancer (Jansen et al. 2000; Bernhard et al. 1999; Chapman et al. 1998; Hagedoorn et al. 2002), stroke (Ahmed et al. 2004; Ahmed et al. 2005), diabetes(Wikby et al. 1993; Postulart and Adang 2000), dental disorders (Ring et al. 2005), and in the fields of geriatric medicine (Daltroy et al. 1999; Heidrich and Ryff 1993; Rijken et al. 1995), palliative care(Rees et al. 2004; Schwartz et al. 2004; Schwartz et al. 2002; Schwartz et al. 2005), and orthopaedics (Razmjou et al. 2006; Finkelstein et al. 2009). Oort et al. demonstrated that for clinical interventions in cancer patients, adjusting for response shift increased effect sizes from small to moderate (Oort and Sprangers 2005), rendering them clinically significant (Norman et al. 2003).

Most clinicians anecdotally report observing response shift. For example, it is not uncommon for spine surgeons to observe that, although their patients express satisfaction with the surgery and would do it again with 20:20 hindsight, their back pain, leg pain or function is not scored as well as would be expected on patient-reported outcomes. This type of inconsistency may reflect changes in internal standards, values or conceptualization of QOL (Sprangers and Schwartz 1999). This discrepancy is particularly striking in the context of a health care system that rations access to expensive interventions (e.g., spinal surgery) such that only patients who are strongly expected to benefit from surgery would be allowed to undergo surgery (Pearson et al. 2012). Despite the stringent screening by spine surgeons and substantial wait for surgery (e.g., 1 year or more), there is a notable subgroup of patients who exhibit less gain than expected (Kurd et al. 2012). It would be useful to understand the causes of this noted discrepancy between expected and observed scores.

Outcome research has burgeoned in the past two decades, with increasing standardization and sophistication in standards for measuring patient-reported outcomes. In the context of spine outcome research, Deyo et al.'s (1998a) seminal call for uniform standards in measuring patient-reported outcomes has led to a standard core of outcome tools for clinical spine research, including the Short Form-36 (SF-36)(Ware and Sherbourne 1992), numeric rating scales for back and leg pain using the 10-point Likert-scale Visual Analogue Scale (VAS), and the Oswestry Disability Index (ODI) (Fairbank and Pynsent 2000; Toyone 2005), a disease-specific functional outcome measure. Interpretation of a clinically meaningful change has been facilitated by empirical evidence documenting what a minimally important difference (MID) is on various measures in the spine surgery patient population (Ostelo et al. 2008). Using the MID as a cut-off in defining successful outcome post-surgery may be valuable for reducing noise in comparing patient groups.

In parallel with this growth in outcome research, the field of response shift research has grown substantially. The methodological tools for detecting response shift have increased in number, sophistication, and sensitivity (Schwartz and Sprangers 1999; Schwartz et al. 2011; Li and Rapkin 2009; Li and Schwartz 2011; Sajobi et al. 2012). Whereas most of these newer methods rely on increasingly complex statistical models and require large sample sizes (e.g., 200–500), the retrospective-pretest design (i.e., "then-test") (Howard et al. 1979; Sprangers et al. 1999) is relatively simple and feasible with small to moderate sample sizes. The most commonly used method for detecting recalibration response shifts since early in the development of the field, the then-test requires a minor modification to questionnaire items such that the respondent is asked to retrospectively re-evaluate his/her level on the items at baseline from his/her current perspective. Recalibration response shift is operationalized as the then-minus-pre difference score (Fig. 1). The standard post-minus-pre difference score operationalizes the reported treatment effect; the full treatment effect is operationalized as the then-minus-post difference score since it is assumed that both of these scores, collected at the same time, share the same internal standards (Sprangers et al. 1999)(Fig. 1).

Critiques of the method have focused on its confounding with recall bias (Schwartz et al. 2004; Ahmed et al. 2005) and with implicit theories of change (Norman 2003). Implicit theories of change refer to the idea that patients do not remember their initial state and instead extrapolate backwards from their present state. Implicit theory presumes that memory or recall of the pre-treatment state is poor so that the retrospective judgment of the initial state is reconstructed and the prospective judgment is more valid. According to the implicit theory, people begin their recollection by asking themselves how they are currently, followed by asking themselves how they think things have changed and then infer what their initial state must have been like (Schwartz, Sprangers et al. 2004). The cognitive processes underlying recalibration response shift as compared to implicit theories are different. For

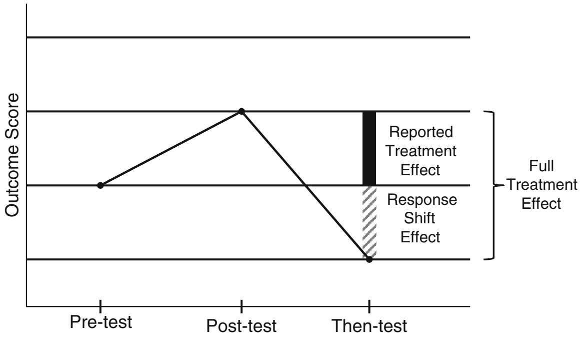


**Fig. 1** The retrospective-pretest design (i.e., then-test) asks the respondent to retrospectively re-evaluate his/her level on the items at baseline from his/her current perspective. Recalibration response shift is operationalized as the then-minus-pre difference score. The standard post-minus-pre difference score operationalizes the reported treatment effect. The full treatment effect is operationalized as the then-minus-post difference score since it is assumed that both of these scores, collected at the same time, share the same internal standards

example, if one is interested in learning about changes in a patient's level of bodily pain, recalibration response shift would be reflected by a change in the meaning of "severe" pain. For example, "severe" would refer to more debilitating pain after an extended period of back pain. Consequently, patients might be less likely to consider their pain "severe" post-surgically even if it were worse than a non-back patient's level of pain. In contrast, implicit theories of change would be reflected by a patient thinking about how she is doing today post-surgery (e.g.,. relatively pain-free), thinking "but I know I had surgery 3-months ago and this surgery was intended to improve my pain; therefore my pain must have been worse 3 months ago". Although social desirability is a relevant cognitive process for patient-reported outcome research, it is distinct from both recalibration response shift and implicit theories of change (Schwartz and Sprangers 2010a).

Schwartz and Rapkin (2004) discuss how discrepancies between expected and observed scores for different types of items reflect different measurement concepts (Table 1). For example, perception-based items reflect a process that involves judgment where a knowledgeable Other's rating, such as those given by a spouse or close friend, would be expected to converge with the patient's own rating. Discrepancy in perception would reflect a response bias. In the context of longitudinal research, the specific response bias would be implicit theories of change. In contrast, evaluation-based items reflect a judgment process based on idiosyncratic and subjective criteria; one would not expect knowledgeable Others to give the same answer as the patient because both parties would have unique evaluative criteria. Discrepancy in evaluation for these types of items would reflect response shift.

Recent guidelines on the use of the then-test have thus suggested that research on the then-test should adjust then-minus-pre difference scores for recall bias and implicit theories of change (Schwartz and Sprangers 2010a), Schwartz and Sprangers have proposed that this adjustment could be accomplished by adjusting then-minus-pre difference scores on an evaluative QOL measure (e.g., the generic Short-Form-36 (SF-36)) by then-minus-pre difference scores on performance-based measures (e.g., timed 20′ walk) to adjust for recall bias, and by then-minus-pre difference scores on perception-based measures (e.g., the Oswestry Disability Index

**Table 1** Correspondence between type of measurement and bias reflected by expected-observed discrepancy

| Item type | Item example | Measurement assumptions | Discrepancy in judgement reflects | Use to adjust for |
|---|---|---|---|---|
| Performance-based | Timed walk of 1/4 of a mile | Measurement process is independent of a judgment | Error of Measurement | Recall Bias |
| Perception-based | Pain prevents me from walking more than 1/4 of a mile[a] | Measurement involves judgement, but raters expected to converge | Response Bias | Implicit Theories of Change |
| Evaluation-based | How much does your health limit you in walking 1/4 of a mile? (A lot? A little? Not at all?)[b] | Measurement involves judgment using idiosyncratic criteria | Response Shift | Response Shift |

[a] This exemplary item was taken from the Oswestry Disability Index

[b] This exemplary item was taken from the SF-36ä. It is one of the items used to compute the Physical Functioning domain score

(ODI)) to adjust for implicit theories of change (Schwartz and Sprangers 2010a)(Table 1). The present study sought to empirically test the hypotheses proposed by Schwartz and Sprangers (2010a) by investigating recalibration response shift effects in spine patients, after adjusting for implicit theories of change using the perception-based ODI. The analysis sought to examine differences between patient groups defined as Improved vs. No Effect using the MID of the VAS items for back and leg pain. Thus, our purpose is to compare the independent explanatory power of recalibration response shift and implicit theories of change in predicting patient groupings distinguished by surgical outcome.

## Materials and Methods

*Design and Eligibility Criteria* The study is a longitudinal observational cohort study. All patients were over 18 years of age. There was no upper age exclusion criterion. Patients undergoing elective posterior lumbar spinal decompression surgery for either spinal stenosis at one or two levels or for disc herniation were candidates for recruitment. The clinical indication for surgery was leg dominant pain without lumbar instability. Patients requiring spinal fusion were excluded. Patients were treated by three fellowship-trained spinal surgeons in the Division of Orthopaedics at Sunnybrook Health Sciences Center, a tertiary care hospital associated with the University of Toronto. Patients were excluded from the study if they were unable to complete the questionnaires in English, if they had visual or cognitive impairment or any other disability that prevented them from completing the outcome measures independently, or if they were unable to give consent for participation.

*Procedure* Eligible patients were approached by research personnel not involved in their care and informed consent was obtained. This study was approved by the research ethics review board at Sunnybrook Health Sciences Center. Participants completed the questionnaires pre-operatively, and post-operatively at 6 weeks, and 3 months. Pre-operative questionnaires were completed in the clinic within 2 weeks prior to surgery and took approximately 20–30 min to complete. After completion, questionnaire booklets were checked for completeness prior to the patient leaving the clinic and the study assistant gathered any missing information.

*Measures* Standardized spine outcome measures were collected in this study, as per Deyo et al.'s recommendations (1998b). The 10-item disease specific ODI (Hagg et al. 2003) measured perceived pain during activities of daily living—the tool asks patients to describe the impact of pain of 10 distinct life domains, using descriptions of observable behaviors (e.g., "Pain prevents me from lifting heavy weights off the floor but I can if they are conveniently positioned, for example on a table."). In this perception-based tool, items reflect a process that involves judgment where a knowledgeable Other's rating, such as those given by a spouse or close friend, would be expected to converge with the patient's own rating. The generic SF-36 (Hagg et al. 2003) comprises eight subscales assessing evaluative functional health—the tool asks patients to assess the level of difficulty or limitation across different health domains, and to attribute this limitation to physical health or emotional problems. In this evaluation-based tool, items reflect a judgment process based on idiosyncratic and subjective criteria; one would not expect knowledgeable Others to give the same answer as the patient because both parties would have unique evaluative criteria. Two Likert-scaled VAS items were included to measure back and leg pain (Gudex et al. 1996). Other baseline data collected were patient demographics (age and sex), duration of symptoms, employment status (working at present, retired, student, homemaker), and compensation status (currently on disability or compensation), as well as associated co-morbid health conditions and other musculoskeletal conditions.

*Response Shift Detection Method* The then- test was the basis for measurement of recalibration response shift (Fairbank and Pynsent 2000). This method was implemented on both the ODI and the SF-36 items so that respondents were asked to retrospectively rate their function on those items thinking back to pre-operative experience from their current post-operative perspective. Thus, patients completed the ODI and SF-36 twice: first with standard instructions, and then with then-test instructions. As shown in Fig. 1, recalibration response shift is operationalized as the difference between the initial pre-test values and the retrospective pre-test values (i.e., then-minus-pre scores).

## Defining Patient Groupings

Patients were grouped for hypothesis testing using a cut-off based on the published Minimally Important Difference (MID) (i.e., ±15 points) on the VAS in the spine patient population (Ostelo et al. 2008) to define patients who were improved by the spine surgery (i.e., change on both leg and back VAS greater than the 15-point MID) as compared to those with no effect (i.e., change on neither leg nor back VAS greater than the MID).

## Statistical Methods

The standard analysis for working with then-test data is to compute difference scores using the then-test (e.g., then-minus-pre reflects recalibration response shift), and testing the null hypothesis that the difference score is equal to zero (Sprangers et al. 1999). One sample t-tests within each time point were used to evaluate whether then-minus-pre difference scores were statistically different from zero. Paired t-tests were used to compare mean then-minus-pre difference scores between time points to evaluate whether the differences were stable over time. If they were stable over time, this would justify taking a mean of the follow-up time points for subsequent analyses. This approach minimized the number of comparisons and thereby reduced the chance of false findings. We therefore used a mean then-test score computed across all three follow-up time points, and subtracted from this score the baseline score on the specific subscale or scale score. Logistic regression analysis was used to evaluate whether the above-mentioned grouping (i.e., Cure versus No Effect based on MID as dependent variable) were associated with then-minus-pre difference scores on ODI summary scores and individual SF-36 subscales or composite scores, testing the main effects of implicit theories of change and recalibration response shift separately. Multivariable logistic regression analysis was then used to adjust the evaluation-based then-minus-pre difference scores (i.e., recalibration response shift based on the SF-36) for the perception-based then-minus-pre difference scores (i.e., implicit theories of change-based on the ODI).

Another way of thinking about the research question is in terms of testing misclassification of outcome. We tested a hierarchical series of multivariable logistic regression models, beginning with standard change scores on the SF-36 and ODI, and adding in the then-minus-pre scores on the ODI (implicit theories of change) and SF-36 (recalibration response shift). Thus we ran the suggested analyses as follows:

**Model 1:** *STANDARD SF-36 CHANGE SCORE MODEL*

Predictor: Observed SF-36 subscale change (post minus pre);

**Model 2:** *STANDARD SF-36 CHANGE SCORE MODEL WITH STANDARD ODI CHANGE SCORE MODEL*

Predictors: Observed SF-36 subscale change (post minus pre)+observed ODI change (post minus pre);

**Model 3:** *INDEPENDENT EFFECT OF IMPLICIT THEORIES OF CHANGE*

Predictors: Observed SF-36 subscale change (post minus pre)+observed ODI change (post minus pre)+ODI then-minus-pre;

**Model 4:** *INDEPENDENT EFFECT OF RECALIBRATION RESPONSE SHIFT*

Predictors: Observed SF-36 subscale change (post minus pre)+observed ODI change (post minus pre)+ODI then-minus-pre+SF-36 subscale post-minus-pre;

The c-statistic is a measure of how well the model predicts the outcome, and we would be looking at the model with the highest c-statistic.

## Results

The sample included 169 patients with baseline data, 102 patients with 6-week follow-up, 106 patients with 3-month follow-up data, and 68 patients with 6-month follow-up data. Table 2 shows how the frequency distributions of the above-mentioned grouping variable and baseline values on demographic and outcome scores. Supplemental Table S1 provides the mean values on the ODI and SF-36 domain scores for 6-weeks, and 3-months. The mean age of the sample was 51.96 (SD=16.46, range 20–84) and 39 % of the patients were female. The sample was well-educated, with over half having completed college or a graduate degree. Approximately one-third of the sample was currently working. Most patients reported using pain medication at baseline. Table 2 provides further information on patient diagnoses, co-morbidities, demographic characteristics, and baseline scores on outcome measures.

### Patient Groupings

Figure 2 shows a scatter plot of patient grouping used in our analyses. The *Improved* versus-*no effect hypothesis* compared patients in the bottom left quadrant to patients whose back and leg pain differences were less than the MID (i.e., fell within the red-lined areas on the x- and y-axes). Excluded from this analysis were six patients whose scores fell outside of the red-lined area *and* who were not in the bottom left quadrant. Including these six outliers in the analysis would have been problematic because they were not numerous enough to be analyzable using inferential statistics.

### Then-Minus-Pre Difference Scores at Each Time Point

All of the then-minus-pre difference scores were significantly different from zero at 6 weeks ($n$=82), and 3 months ($n$=85) (Table 3). The only exception was a trend difference for the SF-36 Mental Health subscale ($p$=0.07).

### Stability of Then-Minus-Pre Difference Scores Over Time

Paired t-tests comparing then-minus-pre difference scores at 6 weeks and 3 months suggested that the difference scores were generally consistent over time (Table 4). The mean scores shown represent then-minus-pre difference scores, where the then-test score being used is specific to a given follow-up time point. Thus, it was empirically justified to use the mean of post-test scores for the final hypothesis-testing. If the then-minus-pre difference scores (i.e., recalibration) were not stable over time, one would not be empirically justified to use the mean post-test score.

**Table 2** Sample demographics

| Variable | N (%) |
|---|---|
| Time points | |
| Baseline | 169 |
| 6 week follow-up | 102 |
| 3 month follow-up | 106 |
| 6 month follow-up | 68 |
| Grouping variable | |
| No effect | 58 (34.2) |
| Improved | 64 (37.9) |
| Missing | 47 (27.8) |
| Gender | |
| Gender (% Male) | 94 (55.6) |
| Surgical diagnosis | |
| Disc herniation | 64 (61.5) |
| Spinal stenosis | 30 (28.8) |
| Spondylolithesis | 1 (1.0) |
| Other | 9 (9.0) |
| Co-morbidities | |
| Depression | 17 (10.1) |
| Cardiac conditions | 12 (7.1) |
| Diabetes | 11 (6.5) |
| Thyroid conditions | 6 (3.5) |
| Cancer | 8 (4.7) |
| Pulmonary conditions | 5 (2.3) |
| Stroke | 1 (0.5) |
| Peripheral neuropathy | 1 (0.5) |
| Other | 23 (13.6) |
| Education | |
| Less than high school | 13 (8.3) |
| Graduated from high school or GED | 27 (17.2) |
| Some college or technical school | 33 (21.0) |
| Graduated from college | 37 (23.6) |
| Postgraduate school or degree | 47 (29.9) |
| Employment status at baseline | |
| Working | 55 (35.8) |
| On leave of absence | 14 (11.8) |
| Unemployed | 7 (4.1) |
| Retired | 39 (20.7) |
| Disabled | 14 (11.8) |
| Homemaker | 7 (4.1) |
| Pain medication use at baseline | |
| Narcotic | 37 (35.6)[a] |
| NSAIDS or other antiinflammatory | 42 (40.4)[a] |
| Other pain medications | 18 (17.3)[a] |
| Smoking status: (%) current smoker | 51 (30.6) |

**Table 2** (continued)

| Variable | Mean (SD) |
|---|---|
| Age | 51.96 (16.46) |
| Range: | [20–84] |
| Baseline outcome scores | |
| Physical functioning | 35.1 (22.8) |
| Role physical | 22.6 (30.4) |
| Bodily pain | 29.35 (18.5) |
| General health | 67.82 (20.3) |
| Vitality | 36.74 (20.7) |
| Social functioning | 49.17 (30.1) |
| Role emotional | 52.64 (41.6) |
| Mental health | 61.78 (21.9) |
| ODI | 23.00 (9.05) |

[a] Sample size for these variables is 104

## MID-Based Comparisons

Logistic regression models revealed significant recalibration response shift effects in most domains in distinguishing Improved versus No Effect groups based on the ±15-point MID on the VAS (see Online Supplemental Table S1 for the unadjusted recalibration response shift effects). These effects were, however, completely explained by implicit theories of change with only one exception. Whereas the then-minus-pre difference score on the physical functioning subscale was not significant in the unadjusted model, after adjusting for implicit theories of change using the ODI then-minus-pre difference score, its parameter estimate became statistically significant at 6 weeks post-surgery (Table 5). This finding suggests a recalibration response shift in physical functioning at 6 weeks post-surgery in distinguishing Improved patients, after adjusting for implicit theories of change. This effect was not maintained at 3-months post-surgery (Table 6).

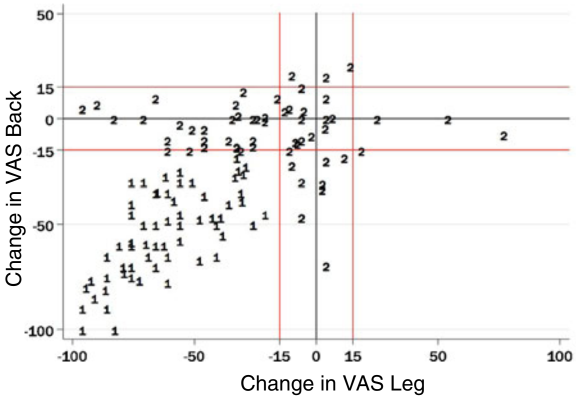


**Fig. 2** This scatter plot denotes the patient groupings used in our analyses. The *Improved*-versus-*No Effect hypothesis* compared patients in group "1" ("Improved" group) shown in the *bottom left quadrant* to patients to patients in group "2" ("No effect" group) whose back and leg pain differences were less than the MID (i.e., fell within the red-lined areas on the x- and y-axes). Excluded from this analysis were those patients whose scores fell outside of the red-lined area *and* who were not in the *bottom left quadrant*

**Table 3** One sample T-tests comparing then-minus-pre difference scores to zero at each time point

| 6 weeks | | | | | |
|---|---|---|---|---|---|
| | N | Mean | SD | T | p< |
| Physical functioning | 82 | −35.72 | 27.36 | −11.83 | 0.00 |
| Role physical | 82 | −23.17 | 36.38 | −5.76 | 0.00 |
| Bodily pain | 82 | −34.07 | 26.12 | −11.82 | 0.00 |
| General health | 82 | −13.86 | 20.54 | −6.11 | 0.00 |
| Vitality | 82 | −20.61 | 21.85 | −8.54 | 0.00 |
| Social functioning | 82 | −29.27 | 29.93 | −8.85 | 0.00 |
| Role emotional | 82 | −27.24 | 52.15 | −4.73 | 0.00 |
| Mental health | 82 | −15.39 | 19.34 | −7.21 | 0.00 |
| ODI | 60 | 14.22 | 11.70 | 9.41 | 0.00 |
| 3 months | | | | | |
| | N | Mean | SD | T | p |
| Physical functioning | 84 | −33.77 | 28.99 | −10.68 | 0.00 |
| Role physical | 84 | −32.14 | 39.49 | −7.46 | 0.00 |
| Bodily pain | 84 | −31.38 | 24.86 | −11.57 | 0.00 |
| General health | 84 | −13.08 | 21.33 | −5.62 | 0.00 |
| Vitality | 84 | −20.93 | 21.33 | −9.00 | 0.00 |
| Social functioning | 84 | −25.89 | 31.69 | −7.49 | 0.00 |
| Role emotional | 84 | −37.30 | 48.10 | −7.11 | 0.00 |
| Mental health | 84 | −18.95 | 22.33 | −7.78 | 0.00 |
| ODI | 67 | 16.12 | 10.23 | 12.90 | 0.00 |

### Predictive Value

The c-statistics from the hierarchical series of logistic regression analyses are presented in Fig. 3, and Supplementary Table 1 presents the c-statistics. These findings suggest that only recalibration in bodily pain had predictive value in identifying Improved versus No-Effect group membership. The proxy for implicit theories of change has independent predictive value only for the domain of general health. An examination of the regression output supported a significant recalibration response shift in physical functioning, as found in the analyses presented in Table 5.

## Conclusions

Our findings suggest that implicit theories of change are prevalent across domains and follow-up time points in the assessment of patient outcome after spinal surgery. Our results also suggest that recalibration response shift in physical functioning occurred in our sample only at 6 weeks post-surgery but not at 3-months post-surgery. Our findings support the stability of then-minus-pre scores on all of the tested measures over time. Although both recalibration response shift and implicit theories of change can be sources of bias in patient-

**Table 4** Paired T-tests investigating stability of difference scores over time

| | 6 weeks | | | 3 months | | | t | p |
|---|---|---|---|---|---|---|---|---|
| | N | Mean | SD | N | Mean | SD | | |
| Physical functioning | 82 | −35.72 | 27.36 | 84 | −33.77 | 28.99 | −0.45 | 0.66 |
| Role physical | 82 | −23.17 | 36.38 | 84 | −32.14 | 39.49 | 1.52 | 0.13 |
| Bodily pain | 82 | −34.07 | 26.12 | 84 | −31.38 | 24.86 | −0.68 | 0.50 |
| General health | 82 | −13.86 | 20.54 | 84 | −13.08 | 21.33 | −0.24 | 0.81 |
| Vitality | 82 | −20.61 | 21.85 | 84 | −20.93 | 21.33 | 0.10 | 0.92 |
| Social functioning | 82 | −29.27 | 29.93 | 84 | −25.89 | 31.69 | −0.71 | 0.48 |
| Role emotional | 82 | −27.24 | 52.15 | 84 | −37.30 | 48.10 | 1.29 | 0.20 |
| Mental health | 82 | −15.39 | 19.34 | 84 | −18.95 | 22.33 | 1.10 | 0.27 |
| ODI | 60 | 14.22 | 11.70 | 67 | 16.12 | 10.23 | −0.98 | 0.33 |

reported outcome assessment, our findings suggest that implicit theories of change are the greater threat to validity in this patient sample.

Our findings also support that considering implicit theories of change and recalibration response shift can improve the prediction of post-surgical outcome for general health and bodily pain, respectively. These latter findings are consistent with findings from other work done by our group utilizing the Lix relative importance method to detect reprioritization response shift (Schwartz et al. 2013). This approach uses logistic regression and discriminant function analysis to identify what domains are most informative in distinguishing known-group membership, and how these domains change in importance over time. In this other study utilizing the same data set, we found that bodily pain went from the least important to the most important domain in distinguishing Improved versus No Effect groups post-surgery, and physical functioning took on an increasingly important role in distinguishing these groups.

The philosophical question about these data is what constitutes the 'truth' in outcome assessment. The outcome of surgery is a patient-based perception which is affected by subjective interpretation. To our knowledge, this study breaks new ground in two ways. First, it is the first study to examine recalibration response shift effects after adjusting for implicit theories of change, Both of these processes are subjective factors that bias or obfuscate the 'truth' in outcomes assessment. Past research on a range of patient populations has utilized an unadjusted then-minus-pre difference score for estimating response shift effects. Our findings suggest that the putative recalibration effects revealed may actually not reflect recalibration response shift at all. This is consistent with Norman's critique of the then-test method(Norman 2003). Recent research using mixed methods has documented that the then-test method reflects a host of other appraisals (e.g., better current coping, better self-care, current mental health problems, comfort/contentment), and only a small proportion (15 %) of responses reflect recalibration.

The second way in which this study is significant, is in documenting the duration of response shift effects over time. We documented a transient recalibration response

**Table 5** Unadjusted SF-36 then-minus-pre difference scores to test MID-based patient groupings

| | Odds ratio | 95 % CI | | Std. error | z | p | Pseudo-R2 |
|---|---|---|---|---|---|---|---|
| 6 weeks (*N*=67) | | | | | | | |
| Physical functioning | 0.99 | 0.97 | 1.01 | 0.01 | −1.15 | 0.25 | 0.01 |
| Role physical | 0.99 | 0.98 | 1.00 | 0.01 | −1.25 | 0.21 | 0.02 |
| Bodily pain | 0.97 | 0.95 | 0.99 | 0.01 | −2.86 | 0.00 | 0.11 |
| General health | 0.98 | 0.95 | 1.00 | 0.01 | −1.96 | 0.05 | 0.05 |
| Vitality | 0.98 | 0.96 | 1.00 | 0.01 | −1.90 | 0.06 | 0.04 |
| Social functioning | 0.99 | 0.97 | 1.00 | 0.01 | −1.66 | 0.10 | 0.03 |
| Role emotional | 1.00 | 0.99 | 1.01 | 0.00 | −0.54 | 0.59 | 0.00 |
| Mental health | 0.97 | 0.95 | 1.00 | 0.01 | −2.00 | 0.05 | 0.05 |
| ODI (*n*=47) | 1.08 | 1.02 | 1.14 | 0.03 | 2.55 | 0.01 | 0.12 |
| 3 months (*N*=67) | | | | | | | |
| Physical functioning | 0.96 | 0.94 | 0.99 | 0.01 | −3.28 | 0.00 | 0.15 |
| Role physical | 0.99 | 0.98 | 1.00 | 0.01 | −1.63 | 0.10 | 0.03 |
| Bodily pain | 0.97 | 0.95 | 0.99 | 0.01 | −2.80 | 0.01 | 0.10 |
| General health | 0.97 | 0.94 | 0.99 | 0.01 | −2.44 | 0.02 | 0.07 |
| vitality | 0.96 | 0.94 | 0.99 | 0.01 | −2.77 | 0.01 | 0.10 |
| Social functioning | 0.97 | 0.95 | 0.99 | 0.01 | −2.89 | 0.00 | 0.11 |
| Role emotional | 0.98 | 0.97 | 0.99 | 0.01 | −3.05 | 0.00 | 0.12 |
| Mental health | 0.96 | 0.94 | 0.99 | 0.01 | −2.79 | 0.01 | 0.10 |
| ODI (*n*=51) | 1.12 | 1.04 | 1.20 | 0.04 | 3.01 | 0.00 | 0.17 |

shift effect in reported physical functioning after a powerful catalyst (i.e., spinal surgery): the response shift effects were detected at 6 weeks but not 3 months. This finding may suggest that when the health state change is powerful but transient (i.e., one recovers from surgery-related disability in a matter of months), then the recalibration response shift effects are also transient. It should be noted, however, that even if transient, these response shift effects would impact the comparability of outcome scores within- and between-patients over time.

The limitations of the present work should be noted. Primarily, this study did not control for recall bias, which is another important potential confounder of the then-test method (Schwartz, Sprangers et al. 2004; Ahmed 2004; Mancuso 1995; Middel 2006; Nolte et al. 2009). Future research should evaluate whether the then-minus-pre difference scores associated with evaluation-based measures have an independent effect in predicting patient groups after adjusting for both implicit theories of change (using perception-base measures) and recall bias (using performance-based measures)(Schwartz and Sprangers 2010b). This methodological approach would yield a more rigorous test of the recalibration response shift hypothesis. Second, our finding that recalibration response shift in physical functioning occurred in our sample at 6 weeks post-surgery may be spurious given the number of comparisons. This finding should serve to generate a hypothesis to be tested and replicated in future research. Third,

**Table 6** Hypothesis testing using patient groupings based on MID controlling for implicit theories of change

| | Odds ratio | 95 % CI | | Std. error | z | p | Pseudo-R2 | % correctly classified (C-statistic) |
|---|---|---|---|---|---|---|---|---|
| 6 weeks (*N*=44) | | | | | | | | |
| Physical functioning | 1.05 | 1.00 | 1.09 | 0.02 | 2.04 | 0.04 | 0.20 | 0.66 |
| ODI | 1.19 | 1.06 | 1.35 | 0.07 | 2.84 | 0.01 | | |
| Role physical | 1.00 | 0.98 | 1.02 | 0.01 | −0.33 | 0.74 | 0.13 | 0.64 |
| ODI | 1.07 | 1.01 | 1.14 | 0.03 | 2.20 | 0.03 | | |
| Bodily pain | 0.98 | 0.95 | 1.01 | 0.02 | −1.21 | 0.23 | 0.15 | 0.72 |
| ODI | 1.05 | 0.97 | 1.13 | 0.04 | 1.23 | 0.22 | | |
| General health | 0.99 | 0.95 | 1.03 | 0.02 | −0.44 | 0.66 | 0.13 | 0.62 |
| ODI | 1.07 | 1.00 | 1.15 | 0.04 | 1.84 | 0.07 | | |
| Vitality | 1.00 | 0.97 | 1.04 | 0.02 | 0.03 | 0.97 | 0.12 | 0.62 |
| ODI | 1.08 | 1.00 | 1.16 | 0.04 | 2.10 | 0.04 | | |
| Social functioning | 0.99 | 0.97 | 1.02 | 0.01 | −0.42 | 0.68 | 0.13 | 0.68 |
| ODI | 1.07 | 0.99 | 1.15 | 0.04 | 1.77 | 0.08 | | |
| Role emotional | 1.01 | 0.99 | 1.02 | 0.01 | 0.82 | 0.41 | 0.13 | 0.66 |
| ODI | 1.09 | 1.02 | 1.16 | 0.04 | 2.61 | 0.01 | | |
| Mental health | 0.99 | 0.95 | 1.03 | 0.02 | −0.73 | 0.47 | 0.13 | 0.62 |
| ODI | 1.07 | 1.00 | 1.14 | 0.04 | 1.87 | 0.06 | | |
| 3 months (*N*=48) | | | | | | | | |
| Physical functioning | 0.98 | 0.95 | 1.02 | 0.02 | −0.83 | 0.41 | 0.18 | 0.71 |
| ODI | 1.08 | 0.97 | 1.20 | 0.06 | 1.46 | 0.14 | | |
| Role physical | 1.00 | 0.98 | 1.02 | 0.01 | 0.32 | 0.75 | 0.18 | 0.73 |
| ODI | 1.12 | 1.03 | 1.22 | 0.05 | 2.75 | 0.01 | | |
| Bodily pain | 1.00 | 0.96 | 1.03 | 0.02 | −0.21 | 0.83 | 0.18 | 0.73 |
| ODI | 1.11 | 1.02 | 1.21 | 0.05 | 2.38 | 0.02 | | |
| General health | 1.00 | 0.96 | 1.04 | 0.02 | −0.02 | 0.98 | 0.17 | 0.73 |
| ODI | 1.11 | 1.03 | 1.21 | 0.05 | 2.57 | 0.01 | | |
| Vitality | 1.02 | 0.97 | 1.06 | 0.02 | 0.71 | 0.48 | 0.18 | 0.67 |
| ODI | 1.14 | 1.03 | 1.27 | 0.06 | 2.60 | 0.01 | | |
| Social functioning | 0.99 | 0.96 | 1.02 | 0.02 | −0.39 | 0.70 | 0.18 | 0.75 |
| ODI | 1.10 | 1.00 | 1.21 | 0.05 | 2.03 | 0.04 | | |
| Role emotional | 1.00 | 0.98 | 1.01 | 0.01 | −0.60 | 0.55 | 0.18 | 0.67 |
| ODI | 1.11 | 1.03 | 1.19 | 0.04 | 2.63 | 0.01 | | |
| Mental health | 1.00 | 0.96 | 1.04 | 0.02 | 0.07 | 0.94 | 0.17 | 0.73 |
| ODI | 1.12 | 1.02 | 1.22 | 0.05 | 2.40 | 0.02 | | |

characterizing a group of people with a range of −16 to −100 as the same "Improved" group, may result in a 'noisy' classification system. Further, since the VAS is an evaluative-based tool, it is quite possible that recalibration response shift effects will occur in it as well. Thus, the MID may change differentially

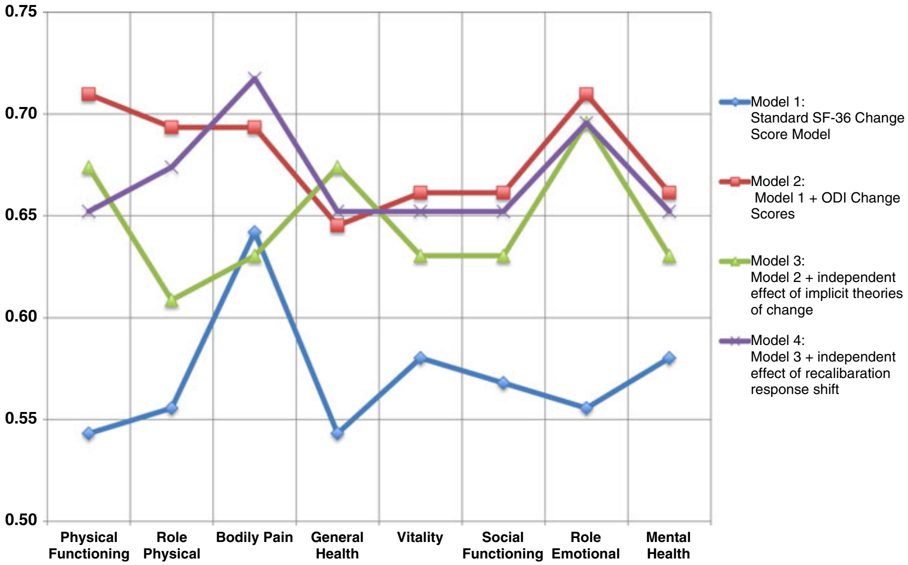


**Fig. 3** This line graph displays the classification statistics from the hierarchical series of logistic regression models testing the independent predictive value of implicit theories of change and recalibration response shift. This figure suggests that recalibration response shift in bodily pain is the only domain with independent predictive value in identifying Improved versus No-Effect groups. In contrast, implicit theories of change has independent predictive value only for the domain of general health. See Supplementary Table 1

in patients who experience response shift effects as compared to those who do not. Future research should consider using an external criterion variable to differentiate Improved vs. No Effect groups, such as a validated clinician-observed performance-based scale. Finally, our data set has problems related to missing data. Recent work on the impact of missing data patterns on response shift detection would suggest that this problem would lead to an underestimation of response shift effects(Sajobi et al. 2012).

## Summary

In summary, the current study suggests that in the measurement of outcome following spinal decompression surgery using validated self-administered questionnaires, recalibration response shift may influence patient report of physical functioning relatively soon after surgery but does not appear to have long-term effects. In contrast, implicit theories of change appear to influence patient-reported outcomes for an extended period of time. These implicit theories—patient assumptions about how their post-surgical course should go—are a possible form of bias that should be assessed in more depth. Future research might apply cognitive interviewing methodology(Willis 2005) to understanding this phenomenon in greater depth.

**Acknowledgments** We gratefully acknowledge helpful discussions with Bruce Rapkin, Ph.D., Aileen Davis, Ph.D., Gary Naglie, M.D., FRCP(C), FGSA, and Alex Kiss, Ph.D., on an earlier draft of this manuscript; and Aimee Gallant for her assistance with data collection.

## References

Ahmed. (2004). Response shift influenced estimates of change in health-related quality of life poststroke. *Clinical Epidemiology, 57*(6), 10.

Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., & Cohen, S. R. (2004). Response shift influenced estimates of change in health-related quality of life poststroke. *Journal of Clinical Epidemiology, 57*(6), 561–570.

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.

Ahmed, S., Mayo, N. E., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R. (2005). Change in quality of life in people with stroke over time: true change or response shift? *Quality of Life Research, 14*, 611–627.

Bernhard, J., Hurny, C., Maibach, R., Herrmann, R., & Laffer, U. (1999). Quality of life as subjective experience: reframing of perception in patients with colon cancer undergoing radical resection with or without adjuvant chemotherapy. Swiss Group for Clinical Cancer Research (SAKK). *Annals of Oncology, 10*(7), 775–782.

Chapman, G. B., Elstein, A. S., Kuzel, T. M., Sharifi, R., Nadler, R. B., Andrews, A., et al. (1998). Prostate cancer patients' utilities for health states: how it looks depends on where you stand. *Medical Decision Making, 18*(3), 278–286.

Daltroy, L. H., Larson, M. G., Eaton, H. M., Phillips, C. B., & Liang, M. H. (1999). Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. *Social Science & Medicine, 48*(11), 1549–1561.

Deyo, R. A., Battie, M., Beurskens, A. J., Bombardier, C., Croft, P., Koes, B., et al. (1998a). Outcome measures for low back pain research. A proposal for standardized use. *Spine (Phila Pa 1976), 23*(18), 2003–2013.

Deyo, R. A., Battie, M., Beurskens, A. J., Bombardier, C., Croft, P., Koes, B., et al. (1998b). Outcome measures for low back pain research. A proposal for standardized use. *Spine, 23*(18), 2003–2013.

Fairbank, J. C., & Pynsent, P. B. (2000). The oswestry disability index. *Spine, 25*(22), 2940–2952.

Finkelstein, J. A., Razmjou, H., & Schwartz, C. E. (2009). Response shift and outcome assessment in orthopedic surgery: is there is a difference between complete vs. partial treatment? *Journal of Clinical Epidemiology, 82*, 1189–1190.

Gudex, C., Dolan, P., Kind, P., & Williams, A. (1996). Health state valuations from the general public using the visual analogue scale. *Quality of Life Research, 5*(6), 521–531.

Hagedoorn, M., Sneeuw, K. C., & Aaronson, N. K. (2002). Changes in physical functioning and quality of life in patients with cancer: response shift and relative evaluation of one's condition. *Journal of Clinical Epidemiology, 55*(2), 176–183.

Hagg, O., Fritzell, P., & Nordwall, A. (2003). The clinical importance of changes in outcome scores after treatment for chronic low back pain. *European Spine Journal, 12*(1), 12–20.

Heidrich, S. M., & Ryff, C. D. (1993). The role of social comparisons processes in the psychological adaptation of elderly adults. *Journal of Gerontology, 48*(3), 127–136.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pre-tests. *Applied Psychology Measurement, 3*(1), 1–23.

Jansen, S. J., Stiggelbout, A. M., Nooij, M. A., Noordijk, E. M., & Kievit, J. (2000). Response shift in quality of life measurement in early-stage breast cancer patients undergoing radiotherapy. *Quality of Life Research, 9*(6), 603–615.

Kurd, M. F., Lurie, J. D., Zhao, W., Tosteson, T., Hilibrand, A. S., Rihn, J., et al. (2012). Predictors of treatment choice in lumbar spinal stenosis: a spine patient outcomes research trial study. *Spine, 37*(19), 1702–1707.

Li, Y., & Rapkin, B. (2009). Classification and regression tree analysis to identify complex cognitive paths underlying quality of life response shifts: a study of individuals living with HIV/AIDS. *Journal of Clinical Epidemiology, 62*, 1138–1147.

Li, Y., & Schwartz, C. E. (2011). Data mining for response shift patterns using recursive partitioning tree analysis. *Quality of Life Research, 20*(10), 1543–1553.

Mancuso. (1995). Does recollection error threaten the validity of cross-sectional studies of effectiveness? *Medical Care, 33*(4 Suppl), 12.

Middel. (2006). Recall bias did not affect perceived magnitude of change in health-related functional status. *Journal of Clinical Epidemiology, 59*(5), 9.

Nolte, S., Elsworth, G. R., Sinclair, A. J., & Osborne, R. H. (2009). Tests of measurement invariance failed to support the application of the "then-test". *Journal of Clinical Epidemiology, 62*(11), 1173–1180.

Norman, G. (2003). Hi! How are you? Response shift, implicit theories and differing epistemologies. *Quality of Life Research, 12*(3), 239–249.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care, 41*, 582–592.

Oort, F. J. V. M., & Sprangers, M. A. G. (2005). An application of structural equation modeling to dtect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research, 14*, 599–609.

Ostelo, R. W., Deyo, R. A., Stratford, P., Waddell, G., Croft, P., Von Korff, M., et al. (2008). Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine (Phila Pa 1976), 33*(1), 90–94.

Pearson, A., Lurie, J., Tosteson, T., Zhao, W., Abdu, W., Mirza, S., et al. (2012). Who should have surgery for an intervertebral disc herniation? Comparative effectiveness evidence from the spine patient outcomes research trial. *Spine, 37*(2), 140–149.

Postulart, D., & Adang, E. M. (2000). Response shift and adaptation in chronically ill patients. *Medical Decision Making, 20*(2), 186–193.

Razmjou, H., Yee, A., Ford, M., & Finkelstein, J. A. (2006). Response shift in outcome assessment in patients undergoing total knee arthroplasty. *The Journal of Bone and Joint Surgery. American Volume, 88*(12), 2590–2595.

Rees, J., MacDonagh, R., Waldron, D., & O'Boyle, C. (2004). Measuring quality of life in patients with advanced cancer. *European Journal of Palliative Care, 11*(3), 104–106.

Rijken, M., Komproe, I. H., Ros, W. J. G., Winnubst, J. A. M., & van Heesch, N. C. A. (1995). Subjective well-being of elderly women: conceptual differences between cancer patients, women suffering from chronic ailments and healthy women. *British Journal of Clinical Psychology, 34*, 289–300.

Ring, L. H., Höfer, S., Heuston, F., Harris, D., & O'Boyle, C. A. (2005). Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. *Health and Quality of Life Outcomes, 3*, 55.

Sajobi, T. T., Lix, L. M., Clara, I., Walker, J., Graff, L. A., Rawsthorne, P., et al. (2012). Measures of relative importance for health-related quality of life. *Quality of Life Research, 21*, 1–11.

Sajobi, T. T., Lix, L. M., Schwartz, C. E., Quaranto, B. R., & Finkelstein, J. A. (2012). Effect of missing data on relative importance measures for response shift detection: an example from an orthopedic population. *Quality of Life Research, 21*(Supplement 1), 2–3.

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.

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.

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*, 455–464.

Schwartz, C. E., Wheeler, H. B., Hammes, B., Basque, N., Edmunds, J., Reed, G., et al. (2002). Early intervention in planning end-of-life care with ambulatory geriatric patients: results of a pilot trial. *Archives of Internal Medicine, 162*(14), 1611–1618.

Schwartz, C. E., Merriman, M., Reed, G., & Hammes, B. (2004). Measuring patient treatment preferences in end-of-life care research: applications for advance care planning interventions and response shift research. *Journal of Palliative Medicine, 7*(2), 233–245.

Schwartz, C. E., Sprangers, M. A. G., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology and Health, 19*(1), 51–69.

Schwartz, C. E., Merriman, M. P., Reed, G., & Byock, I. (2005). Evaluation of the missoula-VITAS quality of life index—revised: research tool or clinical tool? *Journal of Palliative Medicine, 8*(1), 121–135.

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.

Schwartz, C. E., Sajobi, T., Lix, L., Quaranto, B. R., Finkelstein, J. A. (2013). Changing values, changing outcomes: The influence of reprioritization response shift on outcome assessment after spine surgery. Quality of Life Research. doi:10.1007/s11136-013-0377-x.

Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science & Medicine, 48*(11), 1507–1515.

Sprangers, M. A., Van Dam, F. S., Broersen, J., Lodder, L., Wever, L., Visser, M. R., et al. (1999). Revealing response shift in longitudinal research on fatigue–the use of the thentest approach. *Acta Oncologica, 38*(6), 709–718.

Toyone. (2005). Patients' expectations and satisfaction in lumbar spine surgery. *Spine, 30*(23), 5.

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*(6), 473–483.

Wikby, A., Stenstrom, U., Hornquist, J. O., & Andersson, P. O. (1993). Coping behaviour and degree of discrepancy between retrospective and prospective self-ratings of change in quality of life in type 1 diabetes mellitus. *Diabetic Medicine, 10*(9), 851–854.

Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks: Sage Publications.