

The Response Shift Phenomenon in Clinical Trials

By Steven McPhail and Terry Haines

As clinical research has embraced the patient's perspective, use of self-reported measures targeting more subjective constructs have become commonplace during primary outcome selection.¹⁻³ Making sense of self-reported health outcomes amongst clinical trial participants is not always as easy as one might first anticipate. While objective clinical measures may lend themselves to an occasional inconvenience during evaluation or analysis, it is perhaps these self-reported outcomes of a more subjective nature that are most likely to bestow additional complexity on an unsuspecting investigator. Nonetheless, when trial participants experience changes in their health, it is often these types of outcomes that provide meaning to the changes.^{1,3,4} Additionally, outcomes like health-related quality of life may not only reveal whether changes in health have had a meaningful impact on patients, but may also allow for comparisons of benefit (or detriment) to be made across interventions and patient conditions, as well as inform economic evaluation in clinical trials through cost-utility analysis.⁵

There are various methods of assessing self-reported aspects of patient health amongst trial participants to attain a score or outcome amenable to conventional statistical analysis (as opposed to qualitative analytical approaches).^{1,2,5-7} One of the simplest approaches is to use a direct rating scale, such as a visual analogue scale.^{5,8} Other indirect approaches can be quite complex, such as the calculation of multi-attribute utility (i.e., desirability) scores.^{5,9} Multi-attribute utility is a summary score representing the desirability of a certain health state on a scale, where death and perfect health are represented by 0 and 1, respectively.^{5,9} It is commonly calculated by applying weighting systems derived from population-based investigations to participants' discrete survey responses.^{5,9} Regardless of the approach, the point of evaluating constructs of this nature is to capture the patient's point of view of his or her own health state (or components of it that are of interest to the investigation at hand).

Despite the substantial effort that frequently accompanies development and validation of instruments used to evaluate self-reported outcomes, a participant's understanding of a target construct may change between assessments.¹⁰⁻¹⁴ Consider the example below of participants taking part in a fictitious randomized trial investigating the effect of a low back pain education and self-management group on pain and health-related quality of life (compared to conventional medical management control).

Low Back Pain Education and Self-management Group Example

Aim: To investigate the effect of a four-week low back pain education and self-management group program in addition to conventional medical management on pain and health-related quality of life over a six-month period.

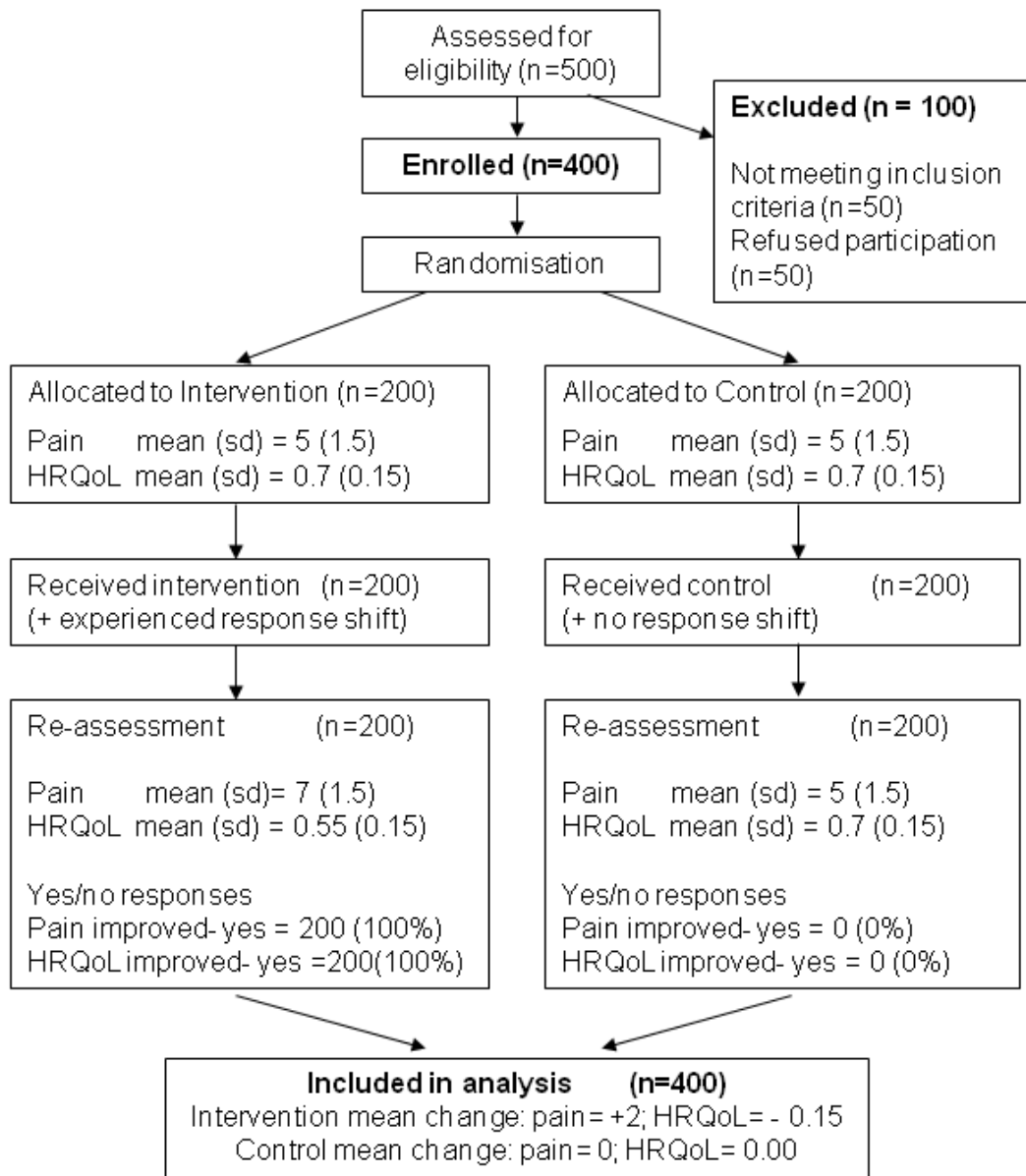
Design: Randomized controlled trial (see Figure 1 for study design).

Participants: Community-dwelling adults with chronic low back pain.

Intervention: Participation in a recently developed four-week low back pain education and self-management group program held at community recreation facilities, in addition to usual medical management (analgesics, etc.).

Control: Usual medical management (analgesics, etc.) only.

Figure 1. Study Design: Fictitious Randomized Controlled Trial



Outcome measure 1a: Participants self-report their pain on a zero (no pain) to ten (worst imaginable pain) scale at baseline and at a six-month follow up.

Outcome measure 1b: At the six-month follow up, participants are also asked to give a dichotomous answer (yes/no) as to whether their pain has improved since the baseline assessment.

Outcome measure 2a: Participants are also asked to fill out a generic health-related quality of life questionnaire at baseline and at six-month follow up. At each assessment, completion of this questionnaire results in an overall score on a continuous scale, on which 0.00 represents death and 1.00 represents perfect health.

Outcome measure 2b: At the six-month follow up, participants are also asked to give a dichotomous answer (yes/no) as to whether their health-related quality of life has improved since the baseline assessment.

In this example, the intervention group experiences a change in their understanding of the constructs under investigation. Prior to group attendance, the mean score from both the intervention and control groups is five out of ten for pain and 0.70 for health-related quality of life. However, while attending the back pain group intervention at community recreation facilities, intervention-group participants see many adults of their own age engaging in recreational activities (tennis, basketball, etc.). Many remember the enjoyable physical activities (e.g., hiking or kicking a ball with their children) they can no longer do because of their back pain. As a result, many of the intervention group participants realize they are greatly limited by their back pain. During the group program, they are also reminded of the importance of staying active in preventing lifestyle diseases (e.g., heart disease). They realize that perhaps their pain and health-related quality of life is worse than they had previously rated at the baseline assessment.

At the six-month follow up assessment, intervention group participants remember their time attending the back pain group. They consider the many useful things they learned and the positive lifestyle changes they have put into place since undertaking the program. However, they also remember how much worse off they were in relation to other adults at the recreation facility. With all factors considered, their mean pain rating now increases to seven out of ten and mean health-related quality of life score now declines to 0.55.

In contrast, control group members have gone about their life as usual. They have not learned anything useful, but neither have they compared themselves to physically active adults. Their understanding of the constructs under investigation has not changed in any systematic way; neither have their priorities or the internal scale on which they consider their pain and health-related quality of life. They again rate their pain as five out of ten and health-related quality of life as 0.70.

All intervention group participants report "yes," when asked if their pain and health-related quality of life has improved since baseline, even though their quantitative measures have worsened. As expected, the control participants report "no" to this same question (Figure 1).

Has attending the back pain education and self-management group increased participants' pain and reduced their health-related quality of life? While logic suggests not, statistical analysis of outcomes 1a and 2a would support this conclusion, particularly if there were no dichotomous responses (outcomes 1b and 2b) included at the follow up assessment.

The paradoxical finding from this simplistic illustration highlights an important phenomenon, termed "response shift," that has the potential to invalidate comparisons of longitudinal measures (over time) in clinical trials.^{10,11,14} Response shift is a change in one's internal perception or understanding of a construct and is thought to be made up of three components: reconceptualization, reprioritization and recalibration:^{10,11,14}

- **Reconceptualization** is a change in one's understanding of which elements or components are included in a target construct. In the example above, for some participants, these components changed to include the ability to take part in social physical recreation activities they had not previously considered when reporting their health-related quality of life.
- **Reprioritization** is a change in one's preferences regarding the relative importance of certain components within the construct. In the example above, for some participants, being able to play with their children became a higher priority component in their assessment of health-related quality of life.

- **Recalibration** is a change in the value scaling of certain health states (or aspects of health states) in relation to others. In the example above, a participant may have considered a certain health state (perhaps his own) to be near the top of the scale at 8/10 but not as good as another health state (perhaps his wife's), which he considered to be 9/10. However, after seeing other healthy adults playing tennis, he adjusts his scaling approach so that the health states he previously valued at 8/10 and 9/10, respectively, now represent 6/10 and 7/10, respectively, on this scale after the recalibration has occurred.

Such changes in a participant's understanding of a construct can invalidate comparisons of longitudinal self-reported outcomes in clinical trials, whether or not he or she is aware of these changes in understanding.^{10,11,14-16} While a false negative conclusion in the example above is not the end of the world, the intervention could have been pharmaceutical or surgical. Another trial might compare an experimental back surgery to a "watch and wait" approach. The surgery group may experience a very painful immediate post-operative recovery. Compared to the pain they experienced immediately after the surgery, their ongoing pain in the weeks that follow seems inconsequential. As a result, at the post surgery assessment, they report a reduction in pain of 3 points, compared to their pre-surgery assessment (due to recalibration), despite actually feeling more pain than they had prior to the surgery. The study's positive results lead to more studies, with similar results, and widespread adoption of a deleterious surgical intervention.

These hypothetical cases illustrate common real-life occurrences. Response shifts can confound results in a broad range of clinical trials. Some examples of response shifts previously reported include investigations amongst patients with hearing impairments (response shift occurred after their hearing aids were fitted),¹⁷ amongst patients who received dental implants,¹⁸ and amongst stroke patients reporting health-related quality of life.¹⁹ Even if the response shift is equal for both treatment and placebo/no-treatment groups of participants, it can mask the actual effect.

Other interesting empirical evidence of the response shift phenomenon has been reported. Perhaps the most widely reported and discussed finding is that people with severe illness, disability or chronic disease often report similar levels of health-related quality of life to people considered to be in good health.^{10,11,14-16} To an objective observer, their health-related quality of life may appear significantly degraded. Another interesting example is pancreas and kidney transplant patients who appeared to experience response shifts in opposite directions when reporting their health-related quality of life (dependent on whether transplantation was successful),²⁰ Also interesting were sub-groups of cancer patients receiving radiation therapy who experienced response shift differently in the self-reported construct of fatigue (depending on if they were in the early stages of adapting to increasing levels of fatigue or experiencing decreasing levels of fatigue).²¹

Considerable effort has gone into the development of methods to measure the size and direction of response shift (and thus take its effect into account with appropriate analysis procedures). Previously described methods to detect response shift include individualized,^{22,23} preference-based^{24,25} and qualitative²⁶ methods, as well as successive comparison,²⁷ design^{17,21} and statistical approaches.^{12,13,28} Although imperfect, these methods help elucidate the problem of response shift. However, they are often time consuming or burdensome on trial participants. In-depth discussion of methods to detect response shift have been reported previously.^{10,13} Perhaps the two approaches most amenable to use in clinical trials to detect the presence of response shift are the "then test" approach and "structural equation modeling," due to their relatively minimal burden on participants and research staff.

Avoiding and Mitigating Response Shift

Current methods for avoiding and mitigating response shift are imperfect, but methodological research continues. Validation of a particular method for a particular trial is problematic.

The “then test” is the simplest and perhaps most widely reported method of detecting response shift amongst trial participants.^{10,13,18} A “then test” requires the respondent to complete a retrospective report of their previous health state at a certain point in time (for example, their pre-intervention health state) from their current perspective (i.e., “how were you back then”). If participants from the fictitious trial described above were to complete a “then test” at the post-intervention follow up, they would also have been required to rate (from their current perspective) how they now believe their pain and health related quality of life were at the initial baseline assessment. If their look back shows better or worse assessments than reported originally, a response shift has presumably occurred.

Despite its popularity, simplicity and amenability to trial contexts, the “then test” approach is not without flaws. The greatest risk when using this approach is that respondents will not be able to accurately recall what their health state was actually like at the previous assessment (recall bias). Additionally, recent evidence has emerged amongst patients undertaking chronic disease self-management interventions that the “then test” approach may contain psychometric flaws resulting from implicit theory of change, social desirability, and recall biases.²⁹ In the implicit theory of change bias, the patient may feel obliged to report improvement (by indicating a lower “then test” score than their current score) because time has passed and they have received treatment (and therefore should be better off than they were previously). In the social desirability bias, the patient may, for example, want to please the physician. In the recall bias, the patient simply cannot accurately remember his or her prior state.

While the “then test” relies on participant recall, structural equation modeling measures underlying components that contribute to a self-reported assessment and attempts to mathematically evaluate whether response shift has occurred.^{12,13,29} For example, participants would assess five different aspects of their health-related quality of life individually, as well as give an overall health-related quality of life rating. By collecting this information (before and after an intervention), structural equation modeling may be able to detect that, say, “undertaking recreation activities” gained importance after the intervention, thereby creating a response shift.

The statistical complexity of structural equation modeling adds additional intricacy to data analysis procedures. Validation is also problematic. However, the method generally does not impose extensive additional time costs on participants and staff involved in data collection.

Despite its potential, only a few investigations have provided empirical evidence supporting the use of structural equation modeling.^{12,13,28} Additionally, structural equation modeling has been applied primarily to the construct of self-reported health-related quality of life and has not yet been applied broadly to other important self-reported constructs.^{12,13,26} It is also noteworthy that initial empirical evidence reports some similarity between structural equation modeling and “then test” results, but not all methods to detect response shift are in agreement.¹³

The trial experience itself may be a particularly important cause of response shift. During clinical trials, participants undertake activities that are likely to elicit self-reflection and potentially internal re-evaluation of health-related constructs. It is therefore important that all groups have the same (or as similar as possible) trial-related experiences. Even differences that seem to be innocuous may have unforeseen effects. For example, in the

back pain education and self-management study above, the researchers did not appreciate the impact of using community recreation facilities.

Given the difficulties of measuring and analyzing response shift, the best option may be to minimize potential causes of response shift occurring differentially between groups. Differentials are likely to depend on the nature of the intervention under investigation. Placebos in drug trials may help equilibrate response shifts across treatment arms. Sham surgeries serve the same purpose, but may be unethical. Depending on the circumstances, a sham educational program may be ethical. Consider the example of vertebroplasty (injecting a cement-like substance into osteoporotic spinal fractures).^{30,31} This often painful procedure has become routine practice based on findings from a number of small investigations with less-than-ideal study designs.³²⁻³³ These investigations (which did not contain a sham intervention comparator or blinded assessment) indicated that the procedure resulted in a substantial and lasting reduction in symptoms.³²⁻³³ However, two recent well-designed investigations (double-blind, sham intervention) revealed there were no additional benefits in the group that had the cement substance injected, in comparison to the sham intervention group.^{30,31} It is entirely likely that the earlier investigations were flawed by response shift and/or the placebo effect.

Finally, self-assessment questions themselves may create a response shift if not asked in the same way or if participants do not give the same consideration to the question at each assessment. Hence, it is important to use outcome measures with sound test-retest reliability and standardized outcome administration procedures.⁵⁻⁷ It is also important that these measures have high levels of construct and criterion-related validity. Participants should also be encouraged to give adequate consideration to standardized rating scales, including scale anchors (such as “best imaginable” or “worst imaginable”). For example, consider a participant who unconsciously replaces a top anchor of “best imaginable health” with “previous health” on a rating scale at a baseline assessment. At a follow-up assessment in which the participant correctly understands the anchor, he or she may give more consideration to how good his or her health could potentially be, and may recalibrate his or her response further away from the anchor (resulting in what may appear to be a decline in health), despite no actual change in health occurring.

Summary

Response shift may cause paradoxical, illogical or incorrect findings during clinical trials. Response shift is a naturally occurring process believed to be part of natural coping and adaptive mechanisms. However, it has the potential to invalidate study findings. To improve the validity of comparisons from longitudinal datasets in clinical trials, it is important to use research designs that are likely to reduce response shift occurring differently between groups or potentially measure and adjust for response shift, if necessary. This is particularly important if there is potential for this phenomenon to occur differentially between groups, dependent on the nature of the intervention being delivered. Methodological research on response shift has not found any panaceas, but progress is being made. To advance the science, it is very important for reports on clinical research trials to discuss response shift.

References

1. Sullivan M. The new subjective medicine: taking the patient's point of view on health care and health. *Soc Sci Med* 2003;56: 1595-1604
2. Kind P. Measuring quality of life in evaluating clinical interventions: an overview. *Ann Med* 2001; 33: 323-327

3. Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. *Copd* 2005;2:57-62
4. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;56:395-407
5. Hickey A, Barker M, McGee H, et al. Measuring health-related quality of life in older patient populations: a review of current approaches. *Pharmacoeconomics* 2005;23:971-993
6. Mishoe SC, Maclean JR. Assessment of health-related quality of life. *Respir Care* 2001;46:1236-1257
7. Bullinger M. Assessing health related quality of life in medicine. An overview over concepts, methods and applications in international research. *Restor Neurol Neurosci* 2002;20:93-101
8. Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? *Soc Sci Med* 1997;45:1289-1297
9. Dolan P, Roberts J. Modelling valuations for Eq-5d health states: an alternative model using differences in valuations. *Med Care* 2002;40:442-446
10. Schwartz C, Sprangers M. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science and Medicine* 1999;48:1531-1548
11. Sprangers M, Schwartz C. Integrating response shift into health-related quality of life research: a theoretical model. *Social Science and Medicine* 1999;48:1507-1515
12. Oort FJ, Visser MR, Sprangers MA. An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Qual Life Res* 2005;14:599-609
13. Visser MR, Oort FJ, Sprangers MA. Methods to detect response shift in quality of life data: a convergent validity study. *Qual Life Res* 2005;14:629-639
14. Rapkin BD, Schwartz CE. Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health Qual Life Outcomes* 2004;2:14
15. Schwartz CE, Andresen EM, Nosek MA, et al. Response shift theory: important implications for measuring quality of life in people with disability. *Arch Phys Med Rehabil* 2007;88:529-536
16. Schwartz CE, Rapkin BD. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health Qual Life Outcomes* 2004;2:16
17. Joore MA, Potjewijd J, Timmerman AA, et al. Response shift in the measurement of quality of life in hearing impaired adults after hearing aid fitting. *Qual Life Res* 2002;11:299-307
18. Ring L, Hofer S, Heuston F, Harris D, O'Boyle CA. Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. *Health Qual Life Outcomes* 2005;3:55
19. Ahmed S, Mayo NE, Wood-Dauphinee S, Hanley JA, Cohen R. The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *J Clin Epidemiol* 2005;58:1125e33.
20. Adang EM, Kootstra G, Engel GL, et al. Do retrospective and prospective quality of life assessments differ for pancreas-kidney transplant recipients? *Transpl Int* 1998;11:11-15
21. Sprangers MA, Van Dam FS, Broersen J, et al. Revealing response shift in longitudinal research on fatigue--the use of the thentest approach. *Acta Oncol* 1999;38:709-718

22. Browne JP, O'Boyle CA, McGee HM, et al. Individual quality of life in the healthy elderly. *Qual Life Res* 1994; 3: 235-244
23. O'Boyle CA, McGee H, Hickey A, et al. Individual quality of life in patients undergoing hip replacement. *Lancet* 1992; 339: 1088-1091
24. Schwartz CE, Cole BF, Gelber RD. Measuring patient-centered outcomes in neurologic disease. Extending the Q-TWiST method. *Arch Neurol* 1995; 52: 754-762
25. Schwartz CE, Cole BF, Vickrey BG, et al. The Q-TWiST approach to assessing health-related quality of life in epilepsy. *Qual Life Res* 1995; 4: 135-141
26. Rapkin BD, Fischer K. Personal goals of older adults: issues in assessment and prediction. *Psychol Aging* 1992; 7: 127-137
27. Schwartz CE, Peng CK, Lester N, et al. Self-reported coping behavior in health and disease: assessment with a card sort game. *Behav Med* 1998; 24: 41-44
28. Oort FJ. Using structural equation modeling to detect response shifts and true change. *Qual Life Res* 2005; 14: 587-598
29. Nolte S, Elsworth GR, Sinclair AJ, et al. Tests of measurement invariance failed to support the application of the "then-test". *J Clin Epidemiol* 2009
30. Kallmes DF, Comstock BA, Heagerty PJ, Turner JA, Wilson DJ, Diamond TH, et al. A randomized trial of vertebroplasty for osteoporotic spinal fractures. *N Engl J Med*. 2009 Aug 6; 361(6): 569-79.
31. Buchbinder R, Osborne RH, Ebeling PR, Wark JD, Mitchell P, Wriedt C, et al. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. *N Engl J Med*. 2009 Aug 6; 361(6): 557-68.
32. Alvarez L, Alcaraz M, Perez-Higueras A, et al. Percutaneous vertebroplasty: functional improvement in patients with osteoporotic compression fractures. *Spine (Phila Pa 1976)* 2006; 31(10): 1113-8
33. Diamond TH, Champion B, Clark WA. Management of acute osteoporotic vertebral fractures: a nonrandomized trial comparing percutaneous vertebroplasty with conservative therapy. *Am J Med* 2003; 114(4): 257-65.
34. Voormolen MH, Mali WP, Lohle PN, et al. Percutaneous vertebroplasty compared with optimal pain medication treatment: short-term clinical outcome of patients with subacute or chronic painful osteoporotic vertebral compression fractures. The VERTOS study. *AJNR Am J Neuroradiol* 2007; 28(3): 555-60.

Authors

Steven McPhail is a Research Officer at The Princess Alexandra Hospital and Doctoral Scholar at the School of Health and Rehabilitation Sciences, The University of Queensland. Contact him at steven_mcphail@health.qld.gov.au.

Terry Haines, PhD is the Director of Allied Health Research, Continuing Care, Southern Health, and Director of Clinical Research, Southern Physiotherapy School, Monash University. Contact him at terrence.haines@med.monash.edu.au.