Innovation and technology have had a profound influence on our everyday lives, as well as on our health-related quality of life (QOL) through advances in medical and surgical treatments. Measurement of spine surgery outcomes has undergone an evolution. Millennial and postmillennial surgeons came into the field inheriting an established practice of collecting patient-reported outcome measures (PROMs), often referred to as “legacy” measures. In the pre-PROM era, outcomes from surgery were evaluated without the patient’s perspective and, hence, they lacked responsiveness and direct relevance.

The purpose of this article is to review the current state of outcome measurement in spine surgery, with an emphasis on patient-reported outcome measures (PROMs). The commonly used generic and disease-specific outcome measures used in spinal surgery and research will be discussed. The authors will introduce the concepts of response shift and appraisal processes, which may affect the face validity of PROMs, as well as their interpretation over time. It is not uncommon for there to be a discrepancy between the observed and expected outcome, which is not wholly explainable by objective measures. Current work on understanding how appraisal affects outcome measurement will be discussed, and future directions will be suggested to facilitate the continued evolution of PROMs.

There has been an evolution in the way clinicians measure outcomes following spinal surgery. In moving from purely physical, objective measures to a growing emphasis on the patient’s perspective, spine surgery outcomes are better able to integrate the impact at multiple levels of relevant change. Appraisal concepts and methods are gaining traction as ways to understand the cognitive processes underlying PROMs over time. Measurement of appraisal is a valuable adjunct to the current spine outcome tools.

Yesterday to Today: a Swinging Pendulum

Lee et al. published a notable article in 1978 about the surgical treatment of spinal stenosis. This article was significant for two reasons: 1) it outlined the goals and indications for surgical treatment of spinal stenosis, and 2) it introduced a new concept for evaluating functional disability in patients with chronic low-back pain. The authors...
TABLE 1. Evaluation system for patients with chronic low-back pain

<table>
<thead>
<tr>
<th>Category (total point value)/Type</th>
<th>Point Value for Each Type</th>
<th>Basis of Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain (60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe, incapacitating</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Moderate, continuous pain, worse at times on exertions, not relieved by pain medications (non-narcotics)</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Mild, pain relieved by medications and has slight restriction of activities</td>
<td>30</td>
<td>Evaluation based</td>
</tr>
<tr>
<td>Very mild pain but not requiring continuous medication and no restriction of activities</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>No pain, no restriction of activities</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Sitting endurance (10)</td>
<td></td>
<td>Performance based</td>
</tr>
<tr>
<td>Less than half hour at a time</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Half hour to 1 hour at a time</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1 hour to 2 hours</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>No restriction</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Walking distance (10)</td>
<td></td>
<td>Perception based</td>
</tr>
<tr>
<td>Less than half block at a time</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Half block to 3 blocks</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>More than 3 blocks but with pain</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>No restrictions</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous (20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Night pain (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleeping frequently disturbed by pain and muscle spasms</td>
<td>0</td>
<td>Evaluation based</td>
</tr>
<tr>
<td>Occasionally disturbed</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Ambulation (5)</td>
<td></td>
<td>Performance based</td>
</tr>
<tr>
<td>Limping, requiring a cane or crutches</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Limping but not requiring external support</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>No restriction</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Sphincter function (5)</td>
<td></td>
<td>Perception based</td>
</tr>
<tr>
<td>Sphincter disturbances or impotence</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Ability to lift (3)</td>
<td></td>
<td>Performance based</td>
</tr>
<tr>
<td>None</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>More than 25 pounds</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Muscle weakness (2)</td>
<td></td>
<td>Examination based</td>
</tr>
<tr>
<td>Significant motor weakness on examination</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>


remarked that “an accurate and objective evaluation of patients with chronic low-back pain is difficult, since most of their complaints are subjective in nature.” They further observed that “the objective findings on examination such as motor or sensory change, reflex change, or mobility of the lumbar spine do not reflect significantly in the total function of the patient.” In contrast to the prior emphasis on physical factors, this system included objective measures determined by physical examination as only a minor component of outcome. The new evaluation system introduced in this article placed the greatest weight for determining patient outcome on performance-, perception- and evaluation-based measurements (Table 1).

This distinction among these constructs is illustrated in Fig. 1. Performance-based measures reflect a quantifiable score that is independent of judgment, whereas perception-based measures include items that involve judgment, where a knowledgeable “other’s” rating, such as that given by a spouse or close friend, would be expected to converge with the patient’s own rating. Evaluation-based measures rate experiences as positive or negative compared with an internal standard. Unlike for the perception-based measure, a knowledgeable “other” would not be expected to give the same answer as the patient because both parties would have unique evaluative criteria and standards of comparison. Some evaluative QOL measures may be confused with performance- (e.g., functional status) or perception-based (e.g., perceived health) measures. Evaluative QOL measures tap one’s subjective experience of health and well-being.

The prominence of the patient’s perspective grew in the 1980s. Symptom relief, daily functioning, and work status were increasingly recognized for their relevance to patients and society. Classic test theory psychometrics were used to construct numerous health status scales. The pendulum was shifting in the evaluation of treatments, from a sole focus of physician-evaluated physical factors to an increasing emphasis of subjective factors from the patient’s perspective.

### Measuring Functional Outcomes in Spinal Care

Validated measures in spinal care are necessary for clinicians and policymakers to compare the effectiveness of different treatments for the same condition. Measures are also critical for patients and clinicians to assess patients’ responses to treatment by changes in their health-related QOL.

Having a comprehensive measurement framework or conceptual model that captures the relevant domains of health-related QOL is increasingly recognized as important. The International Classification of Functioning, Disability, and Health (ICF) provides a framework for documenting the manifestations of health conditions that result from the complex interactions of the person with the physical, social, and psychological environment (Fig. 2). According to the ICF, disability is a state resulting from activity limitations and participation influenced by personal factors and one’s environment.

No single measure comprehensively covers all aspects...
of a patient’s functioning, disability, and health as represented in the ICF framework. A combination of measures is often applied in clinical studies of spinal disorders, which include disease-specific measures that assess those aspects of health that are affected by the disease and generic health measures that assess multidimensional measures of health-related QOL.

Legacy Measures

Disease-Specific Functional Outcome Measures

Several questionnaires are available for assessing functional outcomes related to back pain. The most commonly used measure in both operative and nonoperative patient cohorts is the 10-item Oswestry Disability Index (ODI). The ODI assesses the level of pain and interference with physical activities, sleeping, personal care, social life, sex life, and traveling. Each item is scored from 0 to 5 (0, indicating little disability, to 5, indicating severe disability). The ODI yields a spine-specific disability (scale) score between 0 and 100, with the higher score reflecting more perceived disability.

The psychometric properties of the ODI were initially determined by using classic test theory. This approach examines relatively simple statistics to characterize psychometric performance. For example, it examines item (i.e., question) and score distributions to characterize the relevance of the items (i.e., ceiling and floor effects). It investigates correlations among items and related measures to characterize reliability and construct validity, respectively. Using these methods, the ODI demonstrated good internal consistency, reliability, and construct validity. It also exhibited good test-retest reliability, meaning the scores did not change when the patient’s clinical state was stable. The measure’s responsiveness to clinically important change is, however, lower than other low-back outcome measures, such as the Roland-Morris Disability Questionnaire. A large range in ODI change scores has been considered clinically significant, suggesting a concern with this measure (see Minimal Important Difference). Recent work using more sophisticated psychometric methods, such as item response theory (IRT), has suggested further weaknesses of the ODI. Although the ODI was developed to be a unidimensional scale (i.e., it measures a singular construct of back disability), recent research has been variable in ascribing unidimensionality to the ODI. If the measure is not unidimensional but is treated as unidimensional in analyses, the outcome
assessment may be “noisy” and thus less able to detect relevant differences cross-sectionally and over time; hence, its interpretability will be affected. It also evidenced significant floor and ceiling effects, which prohibit differentiation of patients at the low- or high-disability ranges of the scale, respectively. The ODI is a perception-based measure so psychological factors may bias its scores.

Visual Analog Scale and Numerical Rating Scale

Two other commonly used single-item outcome measures are the visual analog scale (VAS) for leg pain and for back pain, and the Numerical Rating Scale. The original description of the VAS is a line 10 cm in length ranging from no pain to unbearable pain. The respondent is asked to mark the line to indicate his/her pain severity. This method is readily understood and has been demonstrated to be reliable, sensitive, and concurrently valid with other methods. A commonly used variation of the VAS is as a 100-point Likert scale, where patients are asked to grade their back pain and leg pain. The scores are compared at the follow-up time point. A difference of 18–19 points is considered clinically important. The Numerical Rating Scale asks patients to rate their average pain intensity in the past week on an 11-point scale, with 0 indicating no pain and 10 indicating worst possible pain. A 2-point change is considered as clinically important.

Generic Measures

Generic measures assess health, disability, and QOL in ways that are applicable across a wide spectrum of conditions. They are less responsive to changes in a specific condition than disease-specific instruments, but they are valuable for comparing disease burden in different patient populations. The Medical Outcomes Study Short Form–36 (a.k.a. the Rand-36 or the SF-36) is the most commonly used PROM in general and in the evaluation of low-back pain. The SF-36 has 36 items comprising 8 health-domain scores (subcales) assessing evaluative functional health. The tool yields two summary scores: Physical Component Summary (PCS) and Mental Component Summary (MCS) scores. A 5-point change is considered clinically important.

The SF-36 asks the patient 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. Unlike the case for the perception-based ODI, knowledgeable “others” would not be expected to give the same answer as the patient because both parties would have unique evaluative criteria and standards of comparison. Changes in the way patients appraise their QOL will affect their own evaluative process and can be a source of measurement error in PROMs. This will be discussed below.

The Promise of IRT

In 2004 the National Institutes of Health initiated a multicenter cooperative group referred to as the Patient-Reported Outcomes Measurement Information System (PROMIS). The goal was to create a new system of PROMs that would enable a common measurement ap-

FIG. 2. The ICF model. This framework reflects the manifestations of health conditions that result from the complex interactions of the person with the physical, social, and psychological environment. According to the ICF, disability is a state resulting from activity limitations and participation influenced by personal factors and one's environment. Geertzen HH: Moving beyond disability. Prosthetics and Orthotics International 32(3):276–281. ©2008 ISPO. Reprinted by Permission of SAGE Publications, Ltd.
proach across National Institutes of Health–funded studies. This approach built on notable advances in measurement science afforded by IRT.\textsuperscript{14,26} PROMIS minimized patient burden using calibrated item banks that enabled the development of “static” short forms that selected the most informative items within a domain or individually tailored short forms using computerized adaptive test technology.\textsuperscript{1,22}

Similar to the PROMIS system, the Neuro-QoL system was developed with similar funding sources, methods, and goals to provide a measurement system for neurological disorders. The PROMIS and Neuro-QoL systems have led to a large number of measurement domains relevant to spine surgery (see www.healthmeasures.net). Depending on the domains selected, clinicians can create diseasespecific or generic assessment tools to fit their research goals.

In patients with musculoskeletal and spinal disorders, the PROMIS physical function–related domains have been the most thoroughly studied.\textsuperscript{5,28} Patel et al. utilized the PROMIS computerized adaptive tests for physical function, pain interference, and pain behavior to assess surgical outcomes in patients with lumbar spinal stenosis. They demonstrated convergent validity, discriminant (known-groups) validity, and comparable responsiveness to the legacy measures with minimal floor and ceiling effects. Completion time tended to be shorter, and thus PROMIS measures could be more efficient than legacy measures.\textsuperscript{46} It should be noted that if a comprehensive assessment of QOL is desired, numerous domains would need to be captured (e.g., physical, social, and emotional functioning, pain interference and pain behavior, positive affect, and well-being, etc.), and thus the completion time would likely be similar to existing multidimensional outcome measures or combinations of such.

Patient-Surgeon Perspectives Do Not Always Agree

Despite the best of the surgeon’s intentions, evidence-based decision-making, and technical prowess, it is not uncommon to have a mismatch between the surgeon’s perspective (“successful surgery”) and the patient’s perspective (“not-so-successful surgery”). Physician-patient mismatch in perceived outcomes was seen in 24% of patients in a study by Schwartz et al.\textsuperscript{54} In this study, the surgeon was blinded to the postoperative PROMs but had the ability to examine the patient postoperatively and to predict the expected outcome based on clinical examination and chart notes prior to and following surgery. Of 172 patients, 76% of patients agreed with their surgeon’s subjective assessment. In 24% of the cases, the patient perceived a worse outcome than the surgeon. Predictors of subjective disagreement included demographic variables (lower education level and compensation), worsened mental health scores on the SF-36, and worsened VAS leg pain.\textsuperscript{53} In another study that looked at expectations and preoperative PROM variables, Yee et al. showed that in a subset of patients in which expectations were not met from surgery, these patients reported lower mean preoperative SF-36 general health, vitality, and mental component scores.\textsuperscript{57}

More recently, Khor et al.\textsuperscript{29} utilized patient characteristics, diagnosis, and baseline PROM scores to develop a personalized PROM prediction model for improvement in function, back pain, and leg pain following spinal surgery. They found that patients who were least likely to improve with lumbar fusion included those with Medicaid or workers’ compensation as their primary insurance source, current and previous smokers, and those with lower (i.e., better) preoperative pain disability scores. The ability to predict who may do well from surgery and what modifiable patient factors affect outcome is useful in decision-making and preoperative patient education.\textsuperscript{29} These 3 studies illustrate the importance of patient characteristics and, when modifiable, how they might be used to improve outcomes.

Changing Values, Changing Outcomes: Response Shift and Appraisal of QOL

Patient-centered research embodies an imperative to understand individuals’ ways of appraising and evaluating their health-related QOL. Aristotle noted in the *Nicomachean Ethics* that “what constitutes happiness is a matter of dispute…. Very often the same man says different things at different times: when he falls sick, he thinks health is happiness; when he is poor, wealth.” These perspective differences can reflect experience-induced changes in the ways one thinks about symptoms and functional health, in one’s values, and the way one conceptualizes QOL. This psychological phenomenon, or “response shift,”\textsuperscript{25,49,56,59,64} has been shown to influence estimates for treatment benefit in numerous medical conditions and to undermine the underlying assumptions of psychometric science.

Response shift is a psychological process whereby over time and in response to a catalyst (e.g., health-state change, significant life event, growing older, etc.), a patient will alter his/her self-evaluation of the target construct (e.g., health-related QOL, disability, pain, etc.). Undergoing spine surgery for back and/or leg pain would be the catalyst for a patient with a spinal disorder.\textsuperscript{49} Response shift is defined as the changes in the meaning of a target construct reflecting: 1) changes in internal standards (i.e., recalibration); 2) changes in values (i.e., reprioritization); or 3) redefinition of the concept (reconceptualization). Response shift will be mediated by cognitive processes (i.e., how one appraises QOL). The Rapkin and Schwartz model for response shift conceptualizes how individual characteristics (antecedents), behaviors (mechanisms), and cognitive processes (appraisal) will reflect response shift in QOL\textsuperscript{50} (Fig. 3). There are a number of response shift detection methods, which vary in statistical complexity, reliance on large samples, and interpretability.\textsuperscript{52} While a detailed discussion is beyond the scope of this article, we will briefly describe a series of response-shift studies in spine surgery patients.

Response Shift Research in Spine Surgery

The earliest response-shift detection method used was the retrospective-pretest design (“then-test”), borrowed from management science and educational research.\textsuperscript{21,27} This method was initially attractive due to its simplicity and feasibility with small to moderate sample sizes. However, research has demonstrated that the method is con-
founded by recall bias and implicit theories of change. Consequently, the then-test is no longer commonly used for detecting response shift.

In brief, the then-test method uses the standard PROMs at baseline and at follow-up. At follow-up, the respondent is asked to complete an additional questionnaire to retrospectively reevaluate his/her level on the items at baseline, but from his/her current perspective (then-test score). Recalibration response shift was operationalized as the then-minus-pre difference score. The treatment effect was operationalized as the then-minus-post difference score since it is assumed that both of these scores, which are collected at the same time, share the same internal standards.

In an effort to tease apart the competing possible meanings of the then-test discrepancy score (i.e., response shift vs implicit theories of change), Finkelstein et al. explored discrepancy scores as a function of judgment or evaluative criteria in a sample of spinal surgery patients. Figure 1 delineates the types of discrepancies considered in this study. A discrepancy in scores measured by the ODI would reflect a discrepancy in perception and reflect response bias of implicit theories of change as the measurement process involves judgment. Implicit theories of change refer to the idea that patients do not remember their initial state and instead extrapolate backward from their present state. A discrepancy in evaluation-based items (e.g., as measured by the SF-36) would reflect response shift. This study found that implicit theories of change were prevalent and prominent across domains and follow-up time points. Recalibration response shift in physical functioning occurred in the sample only at 6 weeks postsurgery but not at 3 months postsurgery. These findings suggest that implicit theories of change bias the estimation of postsurgical outcomes, but that recalibration response shift biased only the estimation of physical functioning at 6 weeks postsurgery and not at 3 months.

There are a number of statistical approaches for response-shift detection that utilize secondary data analysis. These include structural equation modeling, latent trajectory analysis, relative importance analysis, item response theory, classification, and regression tree modeling. Two studies have investigated response shift in spinal disorders using some of the above methods.

FIG. 3. The Rapkin and Schwartz appraisal model, which posits that when individuals undergo a health-state change (e.g., spinal surgery), their response to surgery is affected by stable characteristics of the individual ("antecedents") and by cognitive, affective, or behavioral "mechanisms" employed by the individual. These factors impact the individual's cognitive processes ("appraisals"), which comprise one's frame of reference, experience sampling, standards of comparison, and patterns of emphasis, which can then impact the individual's perception of QOL over time. In this model, response shift is inferred if changes in appraisal—either directly or in interaction with the catalyst—explain the discrepancy between expected and observed QOL change over time. In other words, after adjusting for sociodemographic and medical covariates in predicting QOL change scores, response-shift effects are inferred when a change in appraisal explains significant variance in the model. Modified from 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 2:14, 2004. With permission from Carolyn Schwartz.
Relative importance analysis investigated differences in patients who underwent lumbar decompression without fusion between those who reported no effect or improvement after surgery and those who reported improvement. This method used a discriminant analysis or logistic regression modeling to identify differences in how (SF-36) domain scores are weighted or ranked to differentiate groups. Change in VAS back and leg pain scores following surgery was used to create the improved and no-effect groups. Although there were no differences between groups presurgically in the relative importance of QOL domains, the two groups diverged in their priorities at follow-up (i.e., re-prioritization response shift). Postoperatively patients in the improved group placed a greater emphasis on bodily pain and physical function and a lesser emphasis on role physical than the patients in the no-effect group. A further difference was that for the improved patients, bodily pain was primarily reflective of physical item response post-surgery, while in the no-effect patients, bodily pain was reflective of physical, social, and emotional domains. This study demonstrated that there were differences in how patients valued or prioritized the domains of the SF-36 over time as a function of surgical outcome. These changes in values may reflect “moving goalposts” in outcome assessment and make it challenging to interpret change over time on standard spine outcome measures.

Oort’s variation of structural equation modeling was used to investigate response shift effects in 1,125 people with spinal cord injury followed for up to 5 years post-injury. This method evaluates measurement invariance over time in PROMs; in other words, it examines how item-scale relationships vary over time. This study detected recalibration and reconceptualization response shift effects at 1- to 5-year follow-up. Despite stable motor and cognitive function, people with spinal cord injury are adapting to their condition. This adaptation reflects a progressive disconnection between symptoms and physical or mental health and a real improvement in their (new conceptualization of) their physical functioning.

**One Size Does Not Fit All: The Importance of Appraisal**

PROMs are created with the assumption of measurement invariance over time. In other words, the same item means the same thing to a respondent each time s/he reads and answers it, and the relationship among items does not change over time. This assumption may be undermined if different people or groups interpret items consistently in a different way (i.e., differential item functioning) or if individuals change in their interpretation of an item (i.e., recalibration, re-prioritization, and/or reconceptualization response shifts). Recent work on the development of practical appraisal measures enables the direct study of these response-shift effects over time.

Appraisal was discussed above as the cognitive process that mediates a response shift. The most comprehensive method for assessing appraisal is the QOL Appraisal Profile (QOLAP). This tool comprises 4 domains, frame of reference (i.e., the meanings individual attaches to questions, experience sampling (i.e., subjective retrieval strategy of relevant experiences), standards of comparison (i.e., relevant subjective standard each sampled experienced is judged against), and combinatorial algorithm (i.e., patterns of emphasis used to summarize their evaluation). It includes open-ended and multiple-choice questions to assess these 4 appraisal parameters. Other measures that have been developed are the QOLAPV2, which contains 85 closed-ended items to assess these same parameters, representing a less resource-intensive approach to measuring appraisal. Finally, the Brief Appraisal Inventory contains 23 closed-ended items to assess higher-order constructs reflecting these 4 appraisal processes. These tools have been tested and used to assess response shift in empirical studies of approximately 7,000 people with chronic illness (HIV, multiple sclerosis, spinal surgery, rare diseases, and bladder cancer). These appraisal studies illustrate how appraisal assessment facilitates the interpretation of QOL change over time, and understanding the backstory of treatment outcomes.

**Minimal Important Difference**

In providing treatment to patients with spinal conditions, it is important to consider what is a minimal important difference (MID). MID is defined as the smallest change that patients perceive as clinically important. MID determination is based on an external anchor, such as a patient-reported global assessment of change. MID can also be distribution dependent—i.e., based on the statistical characteristics of the obtained sample (e.g., SEM, SD, and effect size). In designing clinical trials, an MID value is an important consideration in study design in order to appropriately power the study (i.e., the smaller the MID effect size, the larger the required sample). However, what is an MID may not be constant nor uniform. In considering the ODI, the MID has been demonstrated to be anywhere between 5.9 and 20. Variability in the MID has been shown to be dependent on the degree of baseline disability; the higher the baseline severity of pain and disability, the higher the improvement has to be for the patient to perceive it as clinically important. The trajectory of the outcome also affects MID. Hägg et al. used an external criterion for the patient’s global assessment of treatment effect, and they noted that the MID for those who deteriorated was smaller than that in those who improved in their global assessment.

Variability in the MID was also noted by Schwartz et al. in a study of 167 patients undergoing spinal surgery. The authors investigated changes in patients’ MID over recovery and how cognitive appraisal processes were implicated in the change trajectories. Standard spine outcome measures were used, along with the QOLAP Standards of Comparison items. The MID was calculated at two time points: comparison to last time point and comparison to baseline. Trajectories of change were determined by grouping patients by recovery trajectory: those who consistently got better or worse and those whose global assessment of change varied substantially over the recovery period (“bouncers”). This study showed two important findings. First, moderate to large changes were recognized as clinically important in the early stages of recovery (i.e., 6 weeks postsurgery). Over time, smaller changes become important. Second, global assessment of change groups differed in the relationships between QOL change.
(measured using standard spine outcome measures) and appraisal change over time. Depending on the trajectory of the outcome, patients focused on different standards of comparison. This study illustrated that underlying differences in appraisal may influence how patients experience the same change over time.62

Future Directions

There has been an evolution in the way clinicians measure outcomes following spinal surgery. In moving from purely physical, objective measures to a growing emphasis on the patient’s perspective, spine surgery outcomes are better able to integrate the impact at multiple levels of relevant change. While all measures are susceptible to measurement error, the measurement error of perception- and evaluation-based measures of outcome implicates differences in how patients use the same PROM at two different points in time. Consequently, appraisal concepts and methods are important for understanding the cognitive processes underlying PROMs. Characterizing appraisal can help to clarify the impact of spine surgery18 and help understand when there is an observed vs expected mismatch in characterizing the outcome between the surgeon and the patient.54 Appraisal can also be used as a clinical tool supporting realistic expectations40 and better coping.19 Measurement of appraisal is a valuable adjunct to the current spine outcome tools.37,55

Acknowledgments

We are grateful to the research staff and patients from the Sunnybrook Health Sciences Centre. This work was funded by the Feldberg Chair in Spinal Research, Sunnybrook Health Sciences Centre.

References

28. Hung M, Hon SD, Franklin JD, Kendall RW, Lawrence BD,


44. Oort FJ: Using structural equation modeling to detect response shifts and true change. *Qual Life Res* 14:587–598, 2005


66. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of...
Disclosures
The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions
Conception and design: both authors. Analysis and interpretation of data: both authors. Drafting the article: both authors. Critically revising the article: both authors. Reviewed submitted version of manuscript: both authors. Approved the final version of the manuscript on behalf of both authors: Finkelstein. Administrative/technical/material support: both authors.

Correspondence
Joel A. Finkelstein: Sunnybrook Health Sciences Centre, Toronto, ON, Canada. joel.finkelstein@sunnybrook.ca.