Health and Quality of Life Outcomes

Review

Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal

Carolyn E Schwartz*[1,2,3,4] and Bruce D Rapkin[5]

Address: [1]QualityMetric Incorporated, Waltham, MA, USA, [2]Health Assessment Lab, Waltham, MA, USA, [3]Division of Preventive and Behavioral Medicine, Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA, [4]DeltaQuest Foundation, Inc., Concord, MA, USA and [5]Department of Psychiatry and the Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Email: Carolyn E Schwartz* - carolyn.schwartz@deltaquest.org; Bruce D Rapkin - rapkinb@mskcc.org

* Corresponding author

Published: 23 March 2004

Health and Quality of Life Outcomes 2004, **2**:16

Received: 21 July 2003
Accepted: 23 March 2004

This article is available from: http://www.hqlo.com/content/2/1/16

Abstract

The increasing evidence for response shift phenomena in quality of life (QOL) assessment points to the necessity to reconsider both the measurement model and the application of psychometric analyses. The proposed psychometric model posits that the QOL true score is always contingent upon parameters of the appraisal process. This new model calls into question existing methods for establishing the reliability and validity of QOL assessment tools and suggests several new approaches for describing the psychometric properties of these scales. Recommendations for integrating the assessment of appraisal into QOL research and clinical practice are discussed.

Studies examining response shift phenomena suggest that underlying processes of appraisal differ across people and over time and can greatly affect how people answer questions on QOL measures. The current generation of QOL measures were, however, not designed to account for response shift phenomena [1], but are based on the assumptions that people use measurement scales consistently and that QOL scale scores are directly comparable across people and over time. As Bjorner, Ware and Kosinski [2] point out, both classical and modern psychometric theories view individual differences in scale usage as sources of error. In these nomothetic approaches, the psychometric properties of QOL instruments are framed in terms of the estimation of underlying QOL "true" scores or "latent" scores. This concept is essentially identical to the approach taken in the measurement of constructs like personality and abilities.

We argue that the idea of the true score and related psychometric concepts have been misapplied in QOL measurement because QOL phenomena cannot be appropriately understood in the classical nomothetic measurement paradigm. Rather, we contend that critical properties of QOL measurement are overlooked or relegated to error variance because they do not fit within prevailing psychometric models. Following our companion piece [3], which introduces a new model for conceptualizing and measuring QOL appraisal, we discuss the implications of response shift and appraisal for the psychometric assessment of QOL measures. In sum, our position is that individual differences in cognitive appraisal processes should not be viewed as sources of error in QOL research. Instead, these processes are intrinsic to all QOL measurement. We propose that psychometric models of QOL must be expanded to take differences in appraisal into account by positing the notion of the "contingent true

score". We discuss the implications of this concept for assessing the reliability, validity and responsiveness of QOL scales existing and for the development of new QOL measures.

Some operational definitions

QOL refers to the broad-based construct described by Smith [4] that reflects physical health, role performance, functional ability, and adaptability (coping efficacy), as well as existential aspects of QOL that relate to psychological well-being [5]. 'Response shift' refers to a change attributable to changes in the meaning of that construct, as understood or experienced by a respondent. Response shifts can reflect change in the respondent's internal standards of measurement (scale recalibration), change in the respondent's values regarding the relative importance of component domains of QOL (reprioritization), or a redefinition of meaning of QOL itself (reconceptualization) [1,6]. 'Appraisal' refers to the psychological processes involved in rating a QOL item.

Although often ignored, all QOL assessment involves some process of appraisal [7]. Response shift studies of intra-individual change in the meaning of QOL led to an operational definition of the appraisal process including four parameters: 1) induction of a frame of reference; 2) recall and sampling of salient experiences; 3) use of standards of comparison to appraise experiences; and 4) use of subjective algorithm to prioritize and combine appraisals into a QOL rating [3]. These components of appraisal may be related to culture, personality, and situation, and may vary across persons and over time. It follows that any QOL score is ambiguous without attention to this process. By explicitly addressing differences in QOL appraisal, it is possible to more accurately interpret and compare QOL ratings and to gain a more clinically relevant understanding of the impact of illness and treatment (see [8,9] for example).

Psychometric constructs and QOL measurement

The implications of the appraisal argument are useful in considering psychometric properties of QOL measures. If the meaning of a QOL rating depends upon underlying appraisal processes, the relationship between the observed item and the underlying latent true score is far more complicated than that assumed in current psychometric models. To develop this line of thought, we draw distinctions among three types of constructs: performance-based, which yield measures reflecting the quantity and quality of effort; perception-based, which yield measures of individual judgment concerning the occurrence of an observable phenomenon; and evaluation-based, which yield measures rating experience as positive or negative compared with an internal standard (Figure 1). Some QOL measures may be confused with performance (e.g., functional status) or perception (e.g., perceived health) measures. Measuring the time it takes individuals to walk up a flight of steps provides a performance measure; asking them to recall how many steps they can walk up unassisted is a perceptual measure; asking them to rate how difficult it is for them to walk up a flight of steps is an evaluative measure. QOL research is most often interested in individual evaluation: e.g., that an individual rated going up steps as difficult, even if she performed well against some external standard or expectation. Although some evaluative QOL constructs can be linked to observable performance or self-monitored activity, the purpose of QOL measurement is to tap the subjective experience of health and well-being.

Observed scores on psychological scales are understood to be estimates of a "true score". This construct is derived from classical test theory and is fundamental in all psychometric models, including the "universe score" in generalizability theory [10], the "factor score" in factor analysis [11], the "latent variable" in structural modeling [12], and "theta" in item response theory [13]. Assessment is founded on the assumption that observed scores can consistently and unambiguously convey information about the latent variable of interest. Whether a scale purports to measure performance, perception, or evaluation, the assumption is that an individual's actual status is reflected by responses to some array of items. Errors of measurement can occur, and "noise" can overwhelm "signal", but the fundamental relationship of the item to the construct is presumed static and unchanged.

This assumption holds up well for constructs of performance. Items on a math test are chosen to convey information about "math ability". Of course, this measure is subject to sources of error, such as test anxiety, ambient noise, or cheating. The test may even be biased, with items inadequately representing the underlying construct of math ability (thus underestimating ability in certain subgroups). We can debate which component skills constitute math ability, but however the construct is specified, the meaning of these items and their relationship to math ability are presumed not to change from person to person or occasion to occasion.

Like performance measures, perception measures are presumed to include items consistently related to the phenomenon of interest. Unlike measures of overt performance, scores on perception measures are highly dependent upon the individual making the rating. Observers may attend to behaviors of interest differently or they may intentionally or unintentionally distort responses due to social desirability. However, attentiveness or desirability are not intrinsic to the constructs being

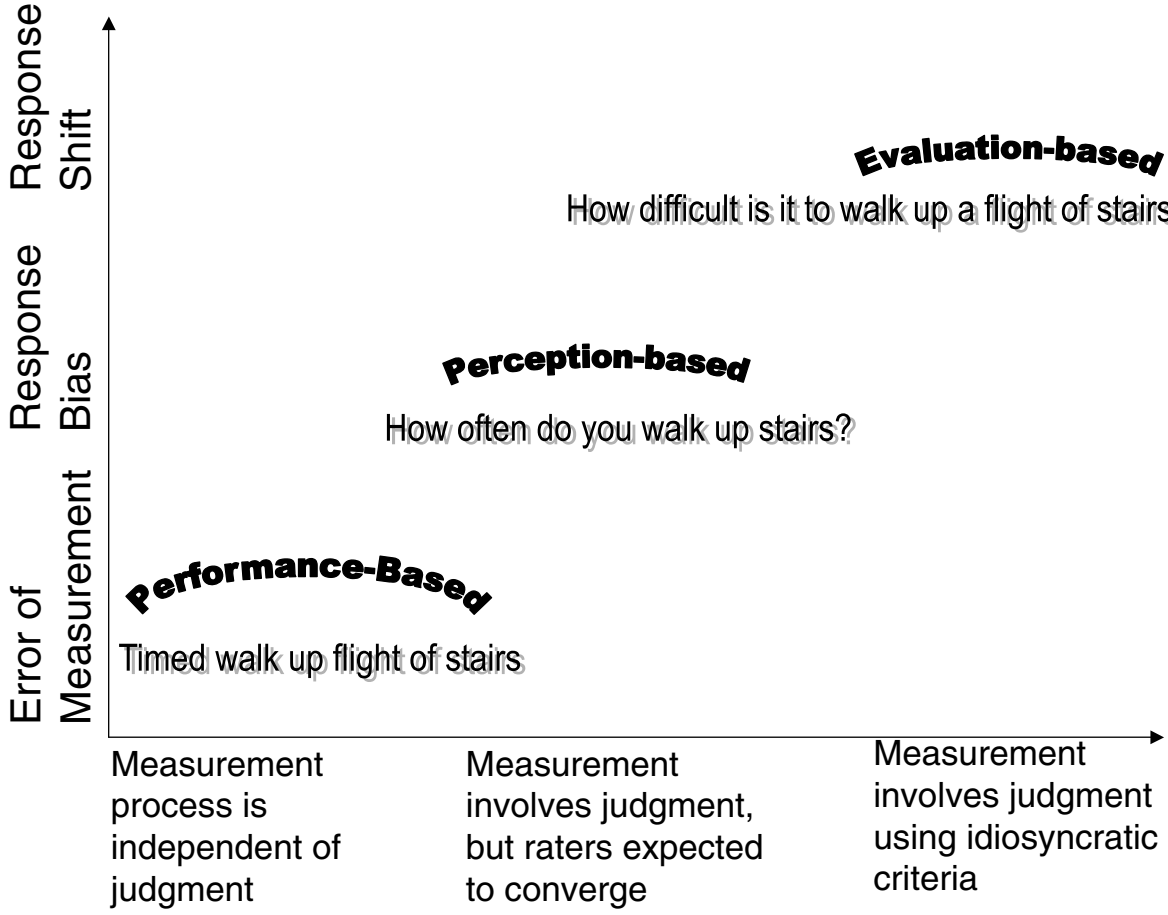

Figure 1
Clarifying the discrepancy in performance-, perception- and evaluation-based methods.

measured. For instance, the meanings of personality traits like extroversion or aggression are defined theoretically. Personality measurement items are linked to these constructs by theory. Observer characteristics may add variance to these variables. Observers may even bring a particular bias or response style to their ratings. However these characteristics of observers are not intrinsic to the definition of the perceptual phenomenon being measured. We expect measures of these phenomena to converge across different observers and modes of measurement. We also expect these measures to have relationships with external criteria that are theoretically determined (e.g. extroverts should have large social networks; aggressive people should have more conflicts with spouse or coworkers, etc.). Such theoretically predictable relationships are the basis for establishing the convergent and discriminative validity of a measure of a construct. In short, with perceptual measures there is a "right" answer. Psychological factors may obscure or bias perception or distort reports, and these are appropriately understood as sources of error.

For evaluative constructs, unlike performance or perception, the subjective perspective of the observer is absolutely intrinsic to the phenomenon of interest. Cancer patients may interpret the item, "How do you rate your health?" to mean, "How do you rate your health relative to other cancer patients?" or "compared with your ideal health?" or "compared with when you were really sick last week?" They may consider health narrowly – "side effects of cancer treatment" – or inclusively. Their answers are no more or less "true" if they adopt one of these standards over the others. There really is no "wrong" answer. We cannot say a QOL measure is inaccurate if someone uses one of these standards or another. Nonetheless, such differences in standards make convergence across observers

or modes of measurement unlikely, and we certainly do not require high inter-rater reliability as a psychometric property of QOL measures. Similarly, construct validation of measures against external criteria cannot be achieved if raters can reasonably interpret items in idiosyncratic ways.

Thus, standard ways of thinking about psychological measurement are not appropriate in the case of QOL assessment because it is misleading to conceive of one true score. For any QOL item, an individual's rating is completely contingent upon how s/he appraises QOL at that time. There is no single right way to think about QOL and no one standard of comparison that is intrinsically more valid or truer. Unlike performance or perceptual measures, individual differences in appraising QOL and attendant problems in discerning the meaning of QOL scores cannot simply be dismissed as measurement error, bias in perception, or misuse of scales. The individual's particular vantage point is not arbitrary; it is intrinsic to rating of QOL. Unfortunately, in the usual case, QOL measurement does not preserve any information about the psychological process used to arrive at a particular score or the "calibration" of the rater when he or she makes a rating. There are many ways to arrive at a rating, but QOL ratings in and of themselves convey no "backstory" about the process of appraisal. Without this information, it is impossible to understand what a score means or how to validate it.

Operational definition of the contingent true score

New psychometric theory and methods are needed to describe the properties of measures of evaluative constructs, such as QOL. In the remainder of this paper, we develop the implications of psychometric theory of QOL based on the notion of a "contingent true score". In this formulation, any rating of a QOL item reflects a latent QOL true score that is completely contingent upon processes of QOL appraisal.

In our companion piece [3], we developed a model that explicitly represents QOL scores as contingent upon four parameters of appraisal: individual's frame of reference $\{FR_t\}$; their strategies for recalling and sampling specific experiences related to these concerns S_{kt}; their reference groups and standards of comparison used to evaluate these experiences, R_t; and the salience weights they associate with different experiences when arriving at an overall rating of QOL $[W_t]$. Each of these terms is marked with a subscript (t) indicating that they are subject to change over time.

$$Q_t = q_t|\{FR_t\},S_{kt},R_t,[W_t]+e$$

In a sense, these appraisal parameters represent the "calibration" of the respondent as the instrument of QOL measurement. Change these parameters and the value of q_t may change; not simply Q_t (the observed QOL score at time t), but the true score itself. We have included an error term in the equation to show that the estimate of the contingent true score is also subject to errors of measurement akin to perceptual measures. Even with all appraisal parameters held constant, separate estimates of observed Q_t (across items, parallel test forms, occasions, or observers) may not be identical, due to the usual sources of error that can enter into measures. However, this model is markedly different from other psychometric theories, which focus primarily on parsing sources of error variance that distort or mask the true score. Measures of stair-climbing performance may be subject to error if the stopwatch is running fast or slow, but there is an actual true time. Alternatively, we contend that there is no QOL without an individual's appraisal. Establishing psychometric properties of QOL measures requires distinguishing differences in estimates due to appraisal parameters versus actual sources of measurement error.

Sources of variance in QOL assessment

In any given QOL assessment, parameters of appraisal may be greatly influenced by characteristics of tests. Instructions on a QOL measure may direct respondents to consider a particular time frame, standard of comparison, or type of experience. Similarly, item content partially dictates the individual's frame of reference, constraining the types of experience that enter into any given QOL rating. Cognitive interviewing may be helpful to ensure that items are used more consistently by different respondents [14]. However, the issue of the contingent true score cannot be reduced to a problem of writing better items or clearer instructions. Instructions may be more or less successful in focusing the respondent's attention (i.e., constraining parameters of the contingent true score) in the ways that researchers intend. Common phrases like "bodily pain" or "some help" are highly subject to interpretation. Efforts to introduce more precise terms may reduce variance in appraisal parameters, but narrower concepts like "headache" or "unaided every time" can still connote different meanings. Even ratings of very specific functions ("difficulty lifting your arm over your head") may be affected by individual differences in standards of comparison ("compared with how I used to be?" or "compared to my mom after she had the same surgery?") and salience ("how often do I really need to lift my arm like that?"). We suspect that there is no practical way to write a nomothetic measure (a single set of items) that is invariant on all parameters of appraisal for all persons over all times.

The current generation of QOL measures used widely in health research and clinical trials do not have items that

were refined through cognitive interviewing to ensure the narrowest possible variation in individual interpretation. Rather, items are often phrased in terms that are as general as possible, to apply to a wide audience. Such items are widely subject to individual differences in appraisal. These differences clearly contribute to the variance of QOL measures in complex, non-linear and dynamic ways [2,7].

Cognitive techniques to refine survey methods may succeed in bounding individual differences in the appraisal process, but cannot eliminate them. Indeed, it is likely that we would not want to develop instruments that so constrain appraisal processes that individuals would be forced to think about QOL in invariant, investigator-determined ways. Most people may be capable of following instructions to focus and generate their responses in a certain way, but such an approach to measurement would run counter to the ways that individuals experience QOL. The general and broadly applicable language common in many standard QOL items makes them subject to a wide variety of appraisal processes, but this also permits individuals to formulate responses in a manner that is most natural for them. QOL items are "projective tests" in the sense that individuals respond by reading their own reality into the item. Rather than writing items that attempt to steer people to think about QOL in certain ways, it is more desirable to try to understand native differences in appraisal as meaningful sources of variation. We would not hand an individual an inkblot and ask, "What does this butterfly look like to you?"

In sum, individual and temporal variance in QOL appraisal may be unavoidable but not undesirable. The contingent true score theory does not imply that we need to scrap existing instruments or re-design them from scratch. Rather, understanding how these sources of variance affect existing QOL measures will help us to select measures, compare groups, and interpret study findings.

Experience with new psychometrics based on QOL appraisal assessment may ultimately provide information necessary to design new instruments with known appraisal properties from the outset. For example, we may find that some assessment approaches are far better than others in capturing changes in QOL during periods of crisis, while others are better at measuring chronic problems and deteriorating health. We could systematically determine which measures elicit equivalent appraisal processes across cultural and ethnic groups or different age cohorts. In the contingent true score model, psychometric equivalence does not mean that groups have similar distributions of QOL scores or factor structures; rather, equivalent measures must elicit similar processes of appraisal from group to group or time to time. To achieve these advances in QOL assessment and determine parameters associated with contingent true score estimates, we must establish psychometric properties that incorporate direct measures of appraisal. Our companion piece [2] refers to studies of response shift that have used various measures, and we also introduce a Quality of Life Appraisal Profile developed to assess the parameters outlined above. Similarly, Jobe [7] discusses think-aloud techniques and other methods for directly measuring individual differences in the interpretation and use of QOL items. In the following sections, we discuss how direct measures of appraisal such as these can be incorporated into studies to establish the psychometric properties of QOL measures consistent with a contingent true score model.

Psychometric reliability

Evidence for reliability in measures is commonly operationalized in terms of inter-item homogeneity, agreement between raters, and/or stability over time. It will be useful to consider each aspect of reliability in turn, as it applies to the contingent true score paradigm.

Internal consistency

The most common evidence of psychometric soundness in QOL measures is internal consistency reliability, such as Cronbach's alpha coefficient [15]. At any point, one would expect individuals to apply self-consistent frames of reference, sampling strategies, standards, and priorities, yielding high internal consistency within a given set of QOL items or scales. High internal consistency or cross-measure correlations would generally be expected under the contingent true-score model.

The question of internal consistency calls attention to a far more challenging concern for QOL assessment: the issue of "bandwidth" versus "fidelity". Five to ten items that are variations of similar questions would almost always yield a high internal consistency coefficient (high fidelity, narrow bandwidth). Alternatively, some aspects of QOL involve multiple types of behavior that need not co-vary: individuals may have several different symptoms on a checklist, but all symptoms need not be elevated simultaneously (broad bandwidth, low fidelity). High correlations among very different symptoms might suggest an actual syndrome, a "response set" (e.g., the tendency to respond identically to all items) or lack of item specificity. Similarly, multi-attribute measures of QOL at times demonstrate high correlations among distinct subscales. Such correlations may reflect actual interdependence among different aspects of QOL or response set, perhaps related to method variance [2]: these are indistinguishable without further assessment of QOL appraisal. In short, high internal consistency and cross-scale correlations demonstrate that people answer a set of items in a similar way. They provide little psychometric information about the

perspective people bring to the assessment situation that draws these ratings together.

Direct assessment of appraisal can help address the bandwidth versus fidelity issue that implicitly underlies general assessments of QOL. In the context of our discussion of appraisal, bandwidth-fidelity refers to the underlying homogeneity or heterogeneity of appraisal processes that are elicited by a QOL measure, including a single item. Consideration of QOL appraisal can help to address the problem of how broad or narrow the assessment needs to be, particularly in domains that are more subject to idiosyncratic interpretation (general health, social support, spiritual fulfillment, or sense of purpose). It may be desirable to include items designed intentionally to induce different frames of reference, suggest several ways of identifying experiences, or pose different standards of comparison – an approach that ensures a robust QOL measure designed to span a broader "bandwidth" than most individuals might otherwise consider. Alternatively, QOL measures could be tested to determine the concepts that they elicit. Instructions and procedures could be adjusted to achieve the desired degree of bandwidth and fidelity (e.g., "only consider days that you were able to get out of bed" or "support includes help that you asked for as well as ways that others helped you out spontaneously").

Inter-observer agreement

For inter-observer agreement to be high, observers must share a frame of reference, sample the same experiences, apply the same standards, and give experiences equal priority. Patient and spouse may identify the same examples of "physical limitation" as relevant to a performance rating and use the same standards of comparison but disagree about the ratings they give. This kind of disagreement, and only this kind, would represent "unreliability" in the classical sense: e.g., we determine agreement only after we ensure that different raters are using the same "scale" to rating the same observed phenomena. This is clearly a limited case, since observers may differ in many aspects of QOL appraisal. When raters use different ways of thinking about events or apply different standards of comparison, more profound disagreements about QOL must be considered in establishing inter-rater convergence. For example, patients may rate performance on difficulties they had in trying to function independently, while caregivers may prioritize episodes when they had to intervene. Such differences in appraisal are not due to "measurement error" or "unreliability" but to the application of different appraisal criteria. Both patient and proxy measures may be valid (that is, accurate, reproducible, correct, true, representative) representations of two people's perspectives; these perspectives can only be understood by explicitly measuring observer differences in appraisal.

Conversely, mere numerical agreement in QOL scores does not guarantee that observers arrived at their responses in the same way. Doctor and patient may agree that a patient is doing poorly, but base their conclusion on different, albeit complementary, observations. Understanding these differences in appraisal fosters better communication between patient and provider, and increases our ability to predict or explain QOL scores.

Test-retest stability

As with inter-rater agreement, the QOL appraisal model suggests that high test-retest reliability must be subject to several strict assumptions: individuals must use a consistent frame of reference, consider the same types of experiences, maintain their standards for evaluating these experiences, and prioritize experiences in the same way. Over brief periods (e.g. without incidents that impact QOL), change in these parameters may be easily disregarded. However, demonstrating high reliability over such brief intervals is relatively trivial. The real challenge in outcome research is in having measures that are sensitive to changes of interest but are otherwise stable over practical time intervals.

Recent research shows that low test-retest reliability among chronically ill patients on the self-reporting of a dynamic construct such as coping behavior is not simply measurement artifact but reflects a meaningful behavioral pattern [16]. Thus, low reliability on a measure subject to dynamic appraisal processes may have meaning with respect to the appraisal process and does not simply reflect random error. Changes in appraisal may also contribute to the stability of QOL measures. Response shift studies suggest that people attempt to maintain a constant QOL level by selectively changing appraisal parameters and standards of comparison through a process of feedback [6]. If we know the extent of such changes, we can adjust QOL scores to equate parameters at different times. For example, apparent remission in symptoms may reflect habituation; statistical adjustment for this recalibration effect could distinguish constant symptoms from improvement. This suggests a more refined analogue to the popular retrospective pretest-posttest methodology: rather than simply asking people to re-rate their baseline status using "today's criteria", we can assess their appraisal process to make those criteria explicit at each time in order to characterize qualitative change.

Psychometric validity

Construct validity

Validation of psychological measures involves determining whether a measure behaves in a way that is consistent

with theoretical expectations [17,18]. For example, a measure's ability to discriminate among groups known to be in different states of health may be taken as evidence of validity. However, if that same measure does not differ between groups, it is not necessarily invalid. Rather, lack of discrimination may be desirable because we recognize that QOL is something more than physical health. Years of experience with QOL measures demonstrate both scenarios: QOL can be directly related to or independent of health status (see review in our companion piece [3]).

The problem of whether, when, and how QOL measures should distinguish known groups is just one theoretical "loose end" that makes it difficult to determine how to validate QOL measures. Consider: How much should different instruments and measures of different QOL domains "hang together?" Should a "global well-being" scale correlate with a "pain" scale? What about measures of "physical" and "psychological" symptom distress? Do correlations among such measures indicate convergent validity? Does lack of correlation indicate discriminant validity? In each case, there is no *a priori* theory to dictate how scales ought to relate to one another and how groups ought to differ. All such findings are interpretable, but how can they be reconciled?

Introducing appraisal parameters in QOL validity studies makes it possible to frame *a priori* questions about how measures ought to behave and to develop stronger and more consistent evidence about construct validity [3]. In light of the QOL contingent true score formulation, it makes sense to distinguish two steps to the construct validation of QOL instruments. 'Internal construct validity' examines whether a QOL measure elicits the desired process of appraisal from respondents. For example, we might hypothesize that a given measure causes respondents to focus on recent changes in health. Direct assessment of experiences people consider in answering QOL items can determine whether these instructions were indeed successful in constraining the appraisal process as desired.

Evidence of marked individual differences in appraising a set of QOL items does not necessarily suggest low internal construct validity. As our discussion of bandwidth-fidelity suggested, in some instances a measure that permits this breadth of perspective may be desirable. Alternatively, investigators may want to impose a narrower or more consistent set of criteria to appraise QOL. In either case, the internal construct validity question concerns whether the measure performs as intended. Internal construct validity can be expressed in terms of the observed range of appraisal parameters elicited by a specific measure, relative to the theoretically-specified or expected range. For example, on a measure for advanced cancer patients in treatment, we may craft an instrument so that 100% of patients consider treatment in responding to physical health items. If subsequent cognitive assessment indicates that only 30% explicitly recalled recent treatment experiences in making their ratings, this is evidence of low intrinsic validity.

'External construct validity' of QOL measures involves the relationship of QOL ratings to objective criteria or other QOL measures, in light of established appraisal parameters. For example, among individuals who actually do consider "recent treatment events" in appraising their QOL, their ratings would be expected to correlate with a measure of the toxicity of their current treatment regimen. This correlation should be evident in those patients who consider this experience, whether that includes 30% or 100% of the sample. Alternatively, one would expect toxicity of treatment regimen to be less highly correlated with QOL ratings among individuals who based their ratings on "their overall history with this chronic disease" or "their satisfaction with their treatment team" or "their fear that the treatment may stop working". Again, the expectation that everyone responding to a particular QOL measure will consider "recent treatment experiences" is an internal construct validity issue. The expectation that QOL ratings will be highly correlated with treatment toxicity among those patients who do emphasize "recent treatment experiences" is an external construct validity issue. In short, once we have established an individual's criteria for appraising QOL, external construct validity means that QOL ratings correlate with other measures or external phenomena in a manner consistent with and dictated by that appraisal process.

As this example demonstrates, internal and external construct validity may be considered distinct features of QOL measures. In studies of internal construct validity, appraisal parameters are dependent variables. In examining external construct validity, appraisal parameters ought to moderate relationships between QOL scales and other measures. Although existing QOL measures may have been written with little attention to the specific appraisal processes they elicit, these measures may still demonstrate high external construct validity once appraisal parameters are specified. Indeed, disaggregating a diverse population according to ways of appraising QOL should yield higher validity coefficients then correlations in the full, undifferentiated sample[19]. Subsequent generations of QOL instruments may be written to constrain appraisal parameters or allow them to vary, depending on the goals of assessment [3].

Responsiveness

The psychometric properties of QOL measures also break down with respect to measurement of change, although here the problems differ somewhat from those involving

convergence across perspectives and with external criteria. As noted above in our discussion of temporal stability, QOL scores can remain stable in the face of marked changes in health status and well-being. Adding or reducing stress does not lead to predictable linear decrements or increments in people's QOL ratings. Rather, in some studies, QOL responses seem subject to adaptation and may return spontaneously to some provisionally stable set-point, despite a constant level of stress [20-22], due to habituation [23] and/or active coping [24]. Change in QOL may also follow a pattern described by engineers and economists as hysteresis [25]. That is, stress may be added without inducing apparent change in a system (or a person), up to a certain level of tolerance, beyond which the system may undergo permanent and profound change that makes it impossible to returning to earlier tolerances. The human experience of QOL may demonstrate both homeostatic and hysteretic properties: coping and habituation may help to maintain stability in QOL unless and until an individual becomes overwhelmed, requiring the establishment of a new adaptive state with its own frame of reference [26].

QOL is not merely a stable and fixed disposition or capacity, nor does it change in lock-step reaction to events. QOL is dynamic [27,28]. Individuals always construct the overt scores that we observe based upon the recall and appraisal of relevant experience. High stability in a QOL measure may mean not that QOL is static but that patients are "running as fast as they can" just to stay in place. Patients may have to accommodate by making major changes in values, priorities, or conceptions of health to achieve a sense of well-being, or they may raise the standards they use to assess their QOL and view what used to be acceptable QOL as no longer adequate. Some evidence from the literature suggests that response shifts modify both the direction and magnitude of change [29,30]. Movement up or down a QOL scale tells us little about processes underlying that change. Adequate QOL assessment must distinguish patients who are feeling better from those who have changed their mind about what it means to feel terrible.

The implications of appraisal for the concept of clinical significance are substantial in regard to how existing QOL measures would be used. We are not suggesting that one reconsider the criteria for clinical significance (e.g., 1/2 standard deviation). Rather, it may be necessary to recalibrate scales so that determination of effect size is made after statistically adjusting appraisal parameters. This will likely lead to increased sensitivity to clinical change over time, and may even lead to smaller changes (i.e., less than 1/2 standard deviation) being clinically meaningful.

Missing data

The problem of missing data in QOL assessment troubles many clinical researchers and has led to a battery of approaches for estimating values for missing items so that patient data can be included for analysis. To our knowledge, the role of appraisal processes in generating missing data has not been addressed in published research. It is possible that people skip items because they do not recognize themselves or their experience in the item, are unsure which response option is most appropriate, or do not understand the item. Cognitive interviewing approaches would likely be effective in elucidating the underlying causes of missing data. Exploring appraisal processes in the context of cognitive interviewing would be a useful foundation for increasing our understanding of the role of such processes for missing data.

Limits and possible extensions of existing psychometric theories

For the sake of unfolding our argument in the most straightforward fashion, we have used the psychometric terminology of classical test theory (e.g., that the observed score is a combination of true score plus error). Although this basic psychometric model is most familiar, several other psychometric theories have been applied to QOL assessment and several have gained popularity in recent years.

Cronbach et al.'s [10] elegant extension of classical test theory, generalizability theory, assumes that observations are randomly sampled from a universe of possible observations, along different facets of measurement. Generalizability theory primarily deals with issues of reliability. Items, observers, and occasions are all treated as random effects in order to identify sources of variation in measurement. However, the notion of the "universe score" in generalizability theory is similar to the true score concept in classical test theory. Observations may vary, but they are all estimates of a single true score. There is no analogue to appraisal in generalizability theory, although differences in appraisal might be expressed in terms of a person-by-occasion interaction with each facet of assessment. For example, estimates of inter-item variation (internal consistency) might differ from person to person because relationships among items might depend upon differences in appraisal processes. Generalizability theory methods would be able to detect certain ramifications of differences in appraisal, but additional appraisal constructs would be required to explain these differences.

An integration of generalizability theory and the appraisal paradigm might be accomplished by techniques for random effects models with nested data, such as hierarchical linear modeling (HLM) [31]. As in generalizability theory, items would be treated as random effects, nested within

individual respondents. Individual differences in appraisal parameters (level 2 independent variables) could be used to account for differences in variance among QOL items as well as correlations of QOL item ratings with item characteristics discussed by Bjorner, Ware and Kosinski [2] such as positive or negative valence, specificity, or type of rating scale (level 1 independent variables). The two-level HLM model could be further generalized to a three-level model to account for items nested within occasions within persons. The three-level model could incorporate changes in appraisal over occasions to determine whether response shifts affect relationships among items.

Structural equation modeling (SEM) also provides several useful ways to incorporate appraisal parameters in studies of the psychometric properties of QOL measures. SEM estimates of true scores on "latent variables" are based on the convergence of observed variables. One interesting approach to examining the internal consistency or factor structure of items on a QOL measure would be to disaggregate a sample according to appraisal parameters of interest and compare relationships among QOL items using confirmatory factor analysis. For example, appraisal assessment could identify one-year post-treatment cancer survivors who use different standards of comparison to judge QOL. Diverse QOL items could load on a single factor ("I now have less pain, more energy, am less worried, have gone back to work, and my mood is better") or yield a more complex factor structure ("I get tired more easily and I have had to slow down at work, but I don't have pain anymore and I don't worry about the small stuff"). This sort of difference in factor structure as a function of appraisal has implications for the use of QOL measures. As factorial complexity increases, the sensitivity of a full-scale score that combines all the items decreases. This is especially problematic in intervention studies [29] where the goal of treatment is to foster rehabilitation and reentry into normative roles and relationships. Selection of QOL outcome measures for such studies might be based on analyses demonstrating factorial invariance against differences on key appraisal parameters. Scale construction could be optimized to take into account the impact of anticipated group differences or individual changes in appraisal.

An analogous SEM approach could also be used to examine differences in the structural relationship between QOL measures and various antecedents and catalysts (see discussion of "Appraisal and Response Shift in the Regression Paradigm" in our companion piece [3]). We might predict that the correlations among indicators of functional impairment and overall well-being are significantly greater among those most concerned about maintaining highly active roles at work or in the community. SEM could be used to compare these relationships among groups identified as having relevant differences in their frames of reference, as a test of external construct validity of the QOL measures.

Over the past decade, item response theory (IRT) has been applied to the psychometric evaluation of QOL measures. IRT identifies characteristics of items in terms of changes in the probability of responses along a latent dimension, "theta", which is analogous to the underlying "true score" in classical test theory. Items vary in their ability to discriminate people with higher and lower values of theta. One advantage of IRT is the ability to select sets of items that discriminate along a continuum of levels of difficulty. For example, it is easier to say yes to "I am uncomfortable" than to "I am in agony".

IRT depends on the ability to identify coherent monotonic relationships between item responses and underlying latent dimensions. IRT cannot work if responses to items are ordered using different and unidentified underlying criteria. As such, IRT leads researchers to exclude items that do not clearly discriminate people at different levels of an underlying theta.

If we view variability in appraisal as an intrinsically meaningful aspect of QOL as opposed to a source of measurement error, the exclusion of items based on lack of fit with the IRT paradigm may be problematic. Limiting QOL items to those that can be precisely and consistently ordered may unduly constrain variability in QOL appraisal. Indeed, Bjorner, Ware and Kosinski [2] suggest that cognitive assessment could be used as an adjunct to IRT, to further assess individuals whose responses do not fit the parameters established for item difficulty in an IRT model. This could lead to expansion of item content or revised assumptions about the dimensionality of the QOL construct under investigation. We would further suggest that the specific inclusion of QOL items shown to be sensitive to differences in appraisal parameters may be quite desirable.

Assessment of appraisal parameters may complement IRT in development of computerized adaptive testing systems to estimate an individual's level of QOL. By incorporating measures of appraisal parameters, adaptive testing could be guided by individual information on the meaning of QOL and criteria for self-appraisal, as well as by estimated thresholds on different QOL items based solely on item characteristics in the aggregate. Appraisal assessment may lead CAT systems to focus on different areas of QOL for some people and to alter the ways that items are ordered and combined. Sufficient reason exists on theoretical, clinical and empirical grounds to argue that individuals can differ widely in their interpretation and use of QOL

Table 1: Reconsidering the psychometrics of QOL assessment in light of response shift and appraisal

Psychometric property	Standard conceptualization	Consequences of neglecting appraisal	Appraisal-based conceptualization	Appraisal-based operationalization
RELIABILITY				
Internal consistency	High homogeneity. Items on a scale are chosen to demonstrate high inter-correlation.	Item content may be too narrow; fail to capture important aspects of QOL. QOL items that are necessarily general and unspecified remain difficult to interpret.	Determine what frames of reference and sampling strategies are systematically induced by specific items and measurement approaches.	Tune items and instructions to achieve desired appraisal parameters. Adjust analyses for differences in appraisal.
Inter-observer agreement	Convergence in QOL ratings made by two or more observers (i.e., self, family, provider).	Differences in perspective so pervasive that this is often ignored or not considered to be a psychometric issue.	Direct measurement of all appraisal parameters to determine whether they explain differences in perspectives.	Ask raters to assume criteria for appraisal used by other observers to calibrate agreement in ratings.
Test-retest stability	High stability over short periods of time. Low stability indicative of measurement error.	QOL is contingent upon appraisal, so low stability may represent change in the appraisal process rather than error of measurement.	True test-retest reliability requires individuals to use a consistent frame of reference, to consider the same types of experiences, to maintain their standards for evaluating these experiences, and to prioritize experiences in the same way.	Impose and test strict assumptions about similarity of appraisal parameters. Establish test-retest stability over a timeframe in which changes in appraisal would not be expected.
VALIDITY				
Construct validity	High correlation with some other QOL measures.	Equivalent to internal consistency reliability. Content may be too narrow. There is no a priori theory to dictate how such measures ought to relate to one another.	Makes it possible to frame a priori questions about how measures ought to behave, and to develop stronger and more consistent evidence about construct validity.	QOL Contingent True Score: 'Internal construct validity' examines whether a QOL measure elicits the desired process of appraisal from respondents. 'External construct validity' involves the relationship of QOL with objective criteria or other QOL measures, in light of established appraisal parameters.
Responsiveness	QOL changes in conjunction with health state changes.	Movement up or down on a QOL scale tells us little about processes underlying that observed change.	Observed overt scores are always constructed by individuals based upon the recall and appraisal of relevant experience.	Must be able to distinguish patients who are feeling better from those who have changed their mind about what it means to feel bad.

items. Adaptive testing to estimate all parameters of the contingent true score would provide information on patients sensitive to the full range of variation and diversity in QOL.

Recommendations and future directions

The problem of appraisal in QOL psychometrics has a strong theoretical base that builds on substantial experience in implementing QOL studies. The practical implications of the problem are likely to change as research on these processes matures. At this juncture, empirical data must be collected to show how these processes matter in measurable and important ways for clinical outcomes research. This demonstration will involve developing further measurement models and tools that are easy to use and interpret. It is our hope that these two companion pieces will provide a foundation for realizing these will be goals. We recognize that substantial development work will be involved in arriving at standard definitions and measures of QOL appraisal parameters, and that each aspect of appraisal raises psychometric issues in its own right. Nonetheless, it is more intellectually acceptable to take on this complexity then to settle for QOL measures that are ambiguous in meaning and that perform inconsistently from study to study.

Our solutions to the problem of appraisal are threefold: design QOL measures with known appraisal parameters, use appraisal measures as stratification or screening variables for certain studies or certain analyses, and include explicit assessment of appraisal constructs in studies to function as mediators or moderators of effects of interest. Some increase in sample size may be warranted to incorporate new variables, but more must be known about appraisal before it is possible to say how many variables must be added and how independent they are. Appraisal assessment may also serve to reduce error variance in QOL scales, improve the specificity of tests, and increase power so that in the long run studies require fewer participants to demonstrate effects related to QOL outcomes.

Acknowledgements

We gratefully acknowledge Mirjam Sprangers, Ph.D. and Kathleen Wyrwich, Ph.D. for their helpful comments and discussions as the ideas in this manuscript evolved. We also want to acknowledge the participants in the University of Hull symposium on "Assessing Health-Related Quality of Life – What Can the Cognitive Sciences Contribute?" (December, 2000) and the subsequent special issue of *Quality of Life Research* (Volume 12, Number 3, May 2003), organized and edited by Ivan Barofsky, Ph.D., Keith Meadows, Ph.D. and Elaine McColl, Ph.D. These opportunities for scholarly exchange encouraged us to formulate the ideas in this paper in much greater depth.

References

1. Schwartz CE, Sprangers MAG: **Methodological approaches for assessing response shift in longitudinal quality of life research.** *Social Science and Medicine* 1999, **48:**1531-1548.
2. Bjorner JB, Ware JE, Kosinski M: **The potential synergy between cognitive models and modern psychometric models.** *Quality of Life Research* 2003, **12(3):**261-274.
3. Rapkin BD, Schwartz CE: **Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift.** *Health and Quality of Life Outcomes* 2004, **2:**14.
4. Smith JA: **The idea of health: a philosophical inquiry.** *ANS Adv Nurs Sci* 1981, **3:**43-50.
5. Ryff CD: **Happiness is everything, or is it? Explorations on the meaning of psychological well-being.** *Journal of Personality and Social Psychology* 1989, **57:**1069-1081.
6. Sprangers MAG, Schwartz CE: **Integrating response shift into health-related quality-of-life research: A theoretical model.** *Social Science and Medicine* 1999, **48:**1507-1515.
7. Jobe JB: **Cognitive psychology and self-reports: Models and methods.** *Quality of Life Research* 2003, **12(3):**219-227.
8. Schwartz CE, Coulthard-Morris L, Cole B, Vollmer T: **The quality-of-life effects of interferon beta-1b in multiple sclerosis. An extended Q-TWiST analysis.** *Arch Neurol* 1997, **54(12):**1475-1480.
9. Rapkin B: **Personal goals and response shifts: Understanding the impact of illness and events on the quality of life of people living with AIDS.** In: *Adaptation to changing health: Response shift in quality of life research*. Edited by: Schwartz CE, Sprangers MAG. American Psychological Association, Washington D.C; 2000:53-71.
10. Cronbach LJ, Linn RL, Brennan RL, *et al.*: **Generalizability analysis for performance assessments of student achievement or school effectiveness.** *Educational & Psychological Measurement* 1997, **57:**373-399.
11. Fayers PM, Machin D: **Factor analysis.** In: *Quality of life assessment in clinical trials: Methods and practice*. Edited by: Staquet MJ, Hays RD, Fayers PM. Oxford University Press, Oxford; 1998:191-223.
12. Meredith W: **Latent variable models for studying differences and change.** In: *Best methods for the analysis of change*. Edited by: Collins LM, Horn JL. American Psychological Association, Washington DC; 1991:149-163.
13. Rausch G: **An item analysis which takes individual differences into account.** *British Journal of Mathematical and Statistical Psychology* 1966, **19:**49-57.
14. Collins D: **Pretesting survey instruments: An overview of cognitive methods.** *Quality of Life Research* 2003, **12(3):**229-238.
15. Cronbach LJ: **Coefficient alpha and the internal structure of tests.** *Psychometrika* 1951, **16:**297-334.
16. Schwartz CE, Daltroy LH: **Learning from unreliability: The importance of inconsistency in coping dynamics.** *Social Science and Medicine* 1999, **48:**619-631.
17. Nunnally JC, Bernstein IH: *Psychometric Theory*. 3rd edition. McGraw-Hill, Inc., New York; 1994.
18. Cronbach LJ, Meehl PE: **Construct Validity in Psychological Tests.** *Psychological Bulletin* 1955, **52:**281-302.
19. Rapkin BD, Fischer K: **Framing the construct of life satisfaction in terms of older adults' personal goals.** *Psychology and Aging* 1992, **7(1):**138-149.
20. Schwartz CE, Sprangers MAG, Carey A, *et al.*: **Exploring response shift in longitudinal data.** *Psychology and Health* 2004, **19(1):**51-69.
21. Carver CS, Scheier MF: **Scaling back goals and recalibration of the affect system are processes in normal adaptive self-regulation: understanding "response shift" phenomena.** *Social Science and Medicine* 2000, **50:**1715-1722.
22. Helson H: *Adaptation Level Theory*. New York: Harper & Row; 1964.
23. Folkman S: **Positive psychological states and coping with severe stress.** *Social Science and Medicine* 1997, **45:**1207-1221.
24. Brandtstädter J, Renner G: **Tenacious goal pursuit and flexible goal adjustment: Explication and age-related analysis of assimilation and accommodation strategies of coping.** *Psychology and Aging* 1990, **5:**58-67.
25. Mayergoyz ID: *Mathematical models of hysteresis*. New York: Springer-Verlag; 1991.
26. Scherer KR: **Emotions as episodes of subsystems synchronization driven by nonlinear appraisal processes.** In: *Emotion, development, and self-organization: Dynamic systems approaches to emotional development. Cambridge studies in social and emotional development*. Edited by: Lewis MD, Granic I. Cambridge University Press, New York, NY, US; 2000:70-99.
27. Leventhal H, Colman S: **Quality of life: A process view.** *Psychology and Health* 1997, **12:**753-767.
28. Allison PJ, Locker D, Feine JS: **Quality of life: A dynamic construct.** *Social Science and Medicine* 1997, **45:**221-230.
29. Schwartz CE, Feinberg RG, Jilinskaia E, Applegate JC: **An evaluation of a psychosocial intervention for survivors of childhood cancer: paradoxical effects of response shift over time.** *Psychooncology* 1999, **8(4):**344-354.
30. Rees JE, Waldron D, O'Boyle CA, *et al.*: **Response shift in individualized quality of life in patients with advanced prostate cancer [abstract].** *Clinical Therapeutics* 2002, **24(Supplement B):**33-34.
31. Bryk AS, Raudenbush SW: *Hierarchical Linear Models: Applications and Data Analysis methods*. Sage, Newbury Park; 1992.