DOI: 10.1111/j.1475-6773.2005.00447.x

Proxies and Other External Raters: Methodological Considerations

A. Lynn Snow, Karon F. Cook, Pay-Shin Lin, Robert O. Morgan, and Jay Magaziner

Objective. The purpose of this paper is to introduce researchers to the measurement and subsequent analysis considerations involved when using externally rated data. We will define and describe two categories of externally rated data, recommend methodological approaches for analyzing and interpreting data in these two categories, and explore factors affecting agreement between self-rated and externally rated reports. We conclude with a discussion of needs for future research.

Data Sources/Study Setting. Data sources for this paper are previous published studies and reviews comparing self-rated with externally rated data.

Study Design/Data Collection/Extraction Methods. This is a psychometric conceptual paper.

Principal Findings. We define two types of externally rated data: proxy data and other-rated data. Proxy data refer to those collected from someone who speaks for a patient who cannot, will not, or is unavailable to speak for him or herself, whereas we use the term other-rater data to refer to situations in which the researcher collects ratings from a person other than the patient to *gain multiple perspectives* on the assessed construct. These two types of data differ in the way the measurement model is defined, the definition of the gold standard against which the measurements are validated, the analysis strategies appropriately used, and how the analyses are interpreted. There are many factors affecting the discrepancies between self- and external ratings, including characteristics of the patient, the proxy, and of the rated construct. Several psychological theories can be helpful in predicting such discrepancies.

Conclusions. Externally rated data have an important place in health services research, but use of such data requires careful consideration of the nature of the data and how it will be analyzed and interpreted.

Key Words. Validity, proxy, self-report

Health services researchers often need to collect information about patients from external raters (as opposed to *internally* referenced self-reports) because a patient is sick/impaired/unavailable/deceased or because another perspective

on the assessed construct is needed. Reports from external raters are needed when

(1) (a) the construct being measured does not have a well-established objective indicator (e.g., quality of life, pain) or (b) objective measurement exists but cannot be obtained (e.g., assessing functional disability in a national telephone survey)

and

(2) (a) the patient is unwilling to report, unable to report, produces a suspect report, or cannot be reached to provide a report or (b) other perspectives would be valuable for a full understanding of the construct.

An example of a situation in which an external report is needed because a patient is unable to report would be trying to diagnose depression in a stroke victim with severe receptive and expressive language impairments. An example of a situation in which an external report is needed because the patient's report is suspect would be trying to determine the extent of disability in a car accident victim who is considered at risk for malingering because the disability decision directly affects chances of being awarded disability compensation.

The need for external reports is quite common in health services research; a search of the word "proxy" in the most recent 8 months of Pubmed abstracts resulted in 50 publications in which external reports methodology was used or in which this issue was considered, and all of these publications were related to health services research. This frequency is probably because of the large number of health services research constructs studied for which there is no objective indicator. Some of the more commonly assessed of these are functional disability (Yasuda et al. 2004), affective states (Snow et al. 2004), pain (Boldingh et al. 2004), social functioning (Lam et al. 2004; Snow et al.

Address correspondence to A. Lynn Snow, Ph.D., 2002 Holcombe Blvd., (152), Houston, TX 77030. A. Lynn Snow and Robert O. Morgan, Ph.D., are with the Houston Center for Quality of Care and Utilization Studies, Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX. Dr. Snow is also with the Veterans Affairs South Central Mental Illness Research, Education, and Clinical Center (MIRECC), Houston, TX. Dr. Snow is also with the Medicine and Psychiatry and Behavioral Sciences Departments, Baylor College of Medicine, Houston, TX. Dr. Morgan is also with the Medicine Department, Baylor College of Medicine, Houston, TX. Karon F. Cook, Ph.D., is with the Department of Physical Medicine, University of Washington, Seattle, WA. Pay-shin Lin, Dr. PH., M.S., L.P.T., is with the Department of Physical Therapy, Graduate Institute of Rehabilitation Science, Chang Gung University, Taiwan. Jay Magaziner, Ph.D., is with the Department of Epidemiology, University of Maryland at Baltimore, Baltimore, MD.

2004), quality of life (Snow et al. 2004; von Essen 2004), health information (Skinner et al. 2004), health preferences and values (Buckey and Abell 2004), and utilization reports (Langa et al. 2004). Further, it is very common to study populations that might not be able to provide accurate self-reports, including infants and small children (Sudan et al. 2004), persons with dementia (Porzsolt et al. 2004), brain injuries (Kuipers et al. 2004), chronic mental illness (Becchi et al. 2004), or severe physical illness (Gnanadesigan et al. 2004). The need for external reports is particularly high in the Veterans Affairs health care system user population because this population consists of large numbers of individuals who are unable to report or cannot report because of both the aging of the WWII, Korean, and Vietnam veteran cohorts (leading to large numbers of persons with dementia and other types of cognitive impairment as well as very frail older adults), and to the incentives for false disability reports to receive compensation. An inherent difficulty of externally rated reports is their intrinsic subjective nature. That is, regardless of who serves as the rater, whenever a report is subjective the judgment is colored by the way the rater perceives and processes information. Memory, emotion, information processing, and motivation to respond accurately all affect self-reports (Stone et al. 1999). Thus, it is not surprising that a large body of literature has established that there are no one-to-one correlations between self-rated and externally rated reports (Neumann, Araki, and Gutterman, 2000). Only by achieving a clear understanding of the factors affecting the discrepancy between self- and external reports can researchers accurately interpret and use externally rated data.

The purpose of this paper is to introduce researchers to the measurement and subsequent analysis considerations involved when using externally rated data. We will define and describe two categories of externally rated data, recommend methodological approaches for analyzing and interpreting data in these two categories, and explore factors affecting agreement between self-rated and externally rated reports. We conclude with a discussion of needs for future research.

Conceptual Framework

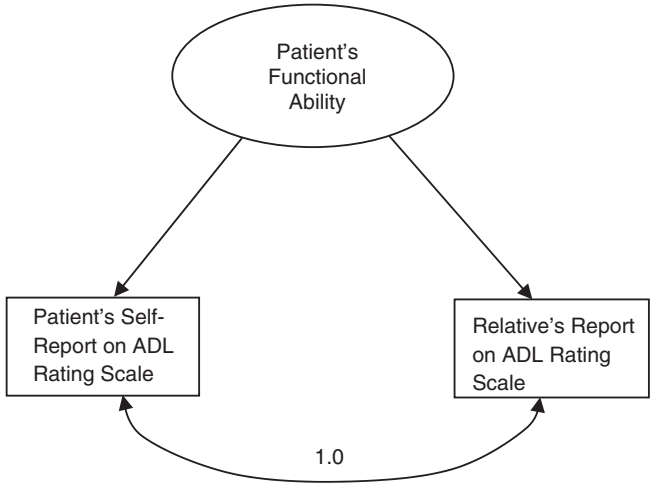
We define two types of externally rated data: proxy data and other-rated data. Proxy data refer to those collected from someone who speaks for a patient who cannot, will not, or is unavailable to speak for him or herself. In the case of a true proxy, the researcher should be indifferent to whether the proxy or

subject is used. There is an assumption that barring measurement error, both will give the same report. The researcher expects that the proxy's report can truly *substitute* for the patient's report. For example, in a longitudinal study of health, a researcher may initially gather health self-ratings from subjects, but as the study proceeds and subjects become too ill or cognitively impaired to complete self-reports, the researcher may substitute proxy ratings of the subject's health, and thus avoid missing data cells. We use the term other-rater data to refer to situations in which the researcher collects ratings from a person other than the patient to *gain multiple perspectives* on the assessed construct. For example, a researcher may choose to supplement a child's self-reports about classroom behavior with teacher reports to develop a richer understanding of the child's classroom performance. The term proxy has often been used for this kind of data, but this is technically incorrect because in this case the external rater is not standing in, or substituting, for another, but adding a second voice regarding the question at hand. We suggest the term other rater to denote this particular situation.

Thus, the purpose of the data collection is to determine whether the data will be in the proxy or other-rater data category. The researcher should answer the question, "Are these data a substitute for, or a supplement to the patient's report?" If the data are a substitute, then it is in the proxy data category, if they are a supplement, then it is in the other-rater data category. To answer this question, the researcher must clarify the reasons for gathering data from raters other than the patient. This clarification is a foundational step because these two categories of external-rater data each have different purposes with distinct implications for statistical analysis and interpretation of results.

The first difference between the two types of externally rated data (proxy versus other rater) is how the measurement model is defined (i.e., how the measures are expected to map onto the theoretical constructs). If the external-rater data are collected for proxy data purposes, then they will be used to model the patient's self-report. In this case, external-rater reports and self-reports are considered as two manifest indicators of another manifest variable—the patient's self-report. For example, in a study of 206 stroke survivors, a report of functional disability was elicited from a relative for the 10 participants who were too cognitively or communicatively impaired to complete the questionnaire for themselves (Perry and McLaren 2004). The researchers consider relative and patient reports to be interchangeable and analyzed the data together. This example is illustrated in Figure 1, in which both patients' and relatives' reports are modeled as manifest indicators of the patient's functional status. In this example, the researchers were implicitly assuming that in

Figure 1: Example of the Proxy Data Measurement Model



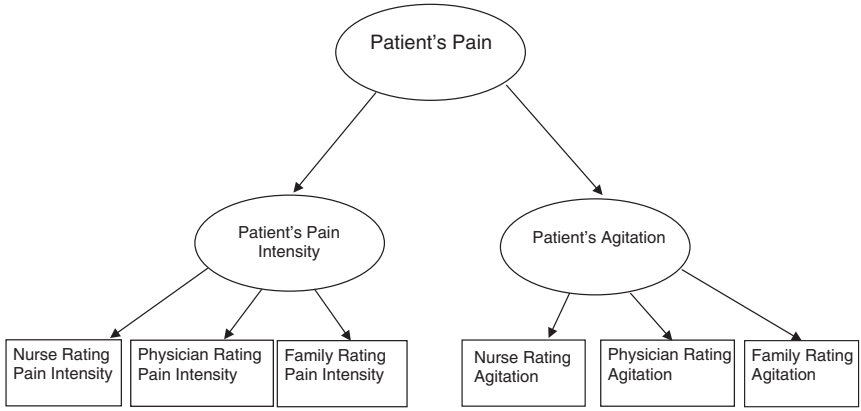
the absence of measurement error, the two reports should correlate perfectly with one another, and thus patient- and proxy-rated data could be combined.

In contrast, when the external-rater data are collected for other-rater data purposes, the measurement model changes. For example, consider a study to determine the efficacy of a pain management intervention for severely demented individuals. Expected outcomes include a decrease in pain intensity and a decrease in agitation. Pain intensity and agitation ratings are collected from nurses, physicians, and family members. Multiple raters are included because it is expected that each has a valid and unique perspective because they all interact with the patient in different ways. This measurement model is illustrated in Figure 2; the rater's reports of pain intensity and agitation are modeled as manifest indicators of first-order pain intensity and agitation latent variables, subsumed under a second-order latent variable representing the overall pain construct.

Another concept that needs to be considered in this discussion is the definition of the gold standard against which ratings are compared. If the patient's report is considered the desired gold standard, then the proxy measurement model is indicated. If the gold standard is considered to be an overarching construct of which patient and external-rater perspectives are components, then the other-rater measurement model is indicated.

The type of externally rated data (proxy versus other rater) also affects how discrepancies between external-rater reports and a patient's self-report

Figure 2: Example of the Other-Rated Data Measurement Model



are conceptualized. With proxy data, discrepancies are defined as bias. Design and statistical methods should be used to minimize bias. Example design approaches would include: (a) stringent proxy inclusion criteria to assure that the chosen proxy is the person most likely to be able to replicate the patient's perspective for proxies, and that proxies are comparable in terms of amount of current and historical contact with patient and type of relationship with patient; (b) use of matching to assure that patient proxies are comparable across intervention and control groups. Alternatively, proxy comparability could be achieved by controlling for proxy characteristics in the analyses. With other-rater data, conceptualization of the discrepancy is more complex. While some of the discrepancy may be because of bias, some of the discrepancy may be because of perspective differences, which is acceptable, expected, and often desirable given the goal of using other-rater data.

Another relevant issue is whether or not it is desirable for the patient and external rater to use different information sources to complete ratings of the assessed construct. The use of different sources would be undesirable for proxy data because the reports are meant to capture the patient's response (i.e., the patient is the gold standard, and discrepancies between the external-rater report and the patient's report are undesirable). In contrast, discrepancies between external-rater and patient reports because of differences in perspective are desirable for other-rater data. For example, the physician calling the nurse to check on how comfortably his bed-bound patient is breathing wants the nurse to tell him how comfortable the patient is (proxy report), not how

comfortable the nurse thinks the patient is (other-rater report). In contrast, the organizational psychologist who is hired to make the CEO more effective wants to know not just how effective the CEO thinks she is, but how effective her employees think she is (other rater).

Several researchers have reported that discrepancies between self- and external raters decrease as the objectiveness of the assessed construct increases. For example, discrepancies are low when observable phenomena such as physical symptoms and levels of functioning are rated. They are high when less observable phenomena such as depressive symptoms and quality of life are measured (Magaziner 1997; Long, Sudha, and Mutran 1998). This certainly is related to information source—both the patient and proxy can observe the patient's physical symptoms. The proxy relies on outward signs of depression (e.g., irritability, low activity) and must interpret and then report them. In contrast, the patient has access to unobservable, internal experiences of depression to interpret and report. Thus, more objective constructs appear to be more amenable to the proxy data approach.

ANALYSIS

Categorizing externally rated data as proxy or as other-rater data clarifies the analysis approach. The two primary foci of analysis of externally rated data are response precision and response bias (Magaziner 1992). Response precision refers to the degree to which the external rater's response agrees with the patient's response. Response bias is a systematic over- or underreporting by the external rater in comparison with the patient's report. Both response precision and bias have been found to be functions of the nature of the construct being assessed, characteristics of the patient, characteristics of the proxy, characteristics of the assessment instrument/item/method, and the criterion against which proxy reports are being compared (Magaziner 1997; Nekolaichuk et al. 1999).

There are three standard statistical approaches to evaluate response precision between patients and external raters: calculation of percent/proportion agreement, κ or weighted κ, and correlations (Magaziner et al. 1996). These approaches apply to both proxy and other-rated data. Each has disadvantages that are compensated by one of the others: percent/proportion agreement is affected by chance; κ and weighted κ are affected by low prevalence of condition of interest; and correlations are affected by low variability, distribution shape, and mean shifts. All three of these statistics can be

conceptualized as indicators of interrater reliability; this conceptualization is useful for proxy data. Alternatively, these statistics can be conceptualized as indications of concurrent validity; this conceptualization is useful for other-rater data. There are three approaches to evaluate the amount and direction of response discrepancy between the patient and external rater (the *amount* of response discrepancy is an indication of response precision, the *direction* of response discrepancy is an indication of response bias [Magaziner et al. 1996]). The researcher can calculate (a) a discrepancy score (external-rater score − patient score) or standardized discrepancy scores (external-rater score − patient score/$SD_{patient}$) (Sherer et al. 1998); (b) percentage bias ([external-rater − patient]/patient) × 100 (Magaziner, Hebel, and Warren 1987); and (c) a general linear model (e.g., linear regression, MANOVA) in which a discrepancy variable is the dependent variable predicted by independent variables which are factors hypothesized to affect the discrepancy.

The effect of information source can be evaluated by directly comparing results from different external raters or by comparing one or more external ratings to self-ratings. There are three approaches to evaluate the effect of information source on the ratings: (a) stratifying analyses by information source; (b) using regression models in which information source is included as an independent variable and interaction terms between information source and other independent variables of interest are also included; and (c) performing generalizability studies using latent variable models in which each information source is included as a separate manifest variable contributing to the latent variable of the construct of interest (Magaziner 1992; Byrne 1994).

INTERPRETATION

Although available analytic approaches, as described above, may be similar across proxy and other-rated data types, the processes of choosing an analytic strategy and interpreting results are informed by the measurement model, and thus differ between proxy and other-rated data types. How the criterion, or gold standard is conceptualized, and how discrepancies between patient and external rater are conceptualized affects the researcher's choice of analytic approaches and how results are interpreted.

For example, Teresi and Holmes (1997) point out that measures of association such as weighted κ, sensitivity, and specificity are more appropriate to use when the patient's report is seen as the gold standard (proxy data type),

because such techniques assume that one measurement (in this case, the external rater's) is being compared with a standard (in this case, the patient's). In such cases, the Landis and Koch (1977) κ interpretations could be applied to guide interpretation of patient/proxy agreement as poor (<0.20), fair (0.21–0.4), moderate (0.41–0.60), or substantial (0.61–0.80). Intraclass correlations are more appropriate to use when the patient and external rater's responses are considered to be two manifest indicators of a single latent variable (other-rater data type) (Landis and Koch 1977; Teresi and Holmes 1997). The interpretation guidelines above for κs can also be applied to correlation coefficients (Magaziner 1992). In such cases, use of κ, sensitivity, or specificity do not make much sense because both the patient rating and external rating are weighted equally, and neither is assumed to be the more correct representation of "truth."

Discrepancy scores have typically been reported with the implicit assumption that the external rater is the gold standard (Snow et al. 2004). Percentage bias has typically been reported with the implicit assumption that the patient is the gold standard (Bassett, Magaziner, and Hebel 1990). However, neither of these assumptions of patient versus external rater as gold standard need to be made in the interpretation of these analyses. If deviations in scores are interpreted purely as discrepancies (rather than indicators of patient inaccuracy or rater bias, for example), then the scores can become useful quantifiers of the gap between the two raters.

Another concept relevant to this discussion is how self- and external-rater reports are integrated together. The approach to this integration will vary with the purpose of the assessment. The decision of how to weight the self-report relative to reports of the external rater (or several external raters) will be different if the purpose of the assessment is to develop a treatment plan than to fully understand a concept. As an example of the first case, if the purpose is to decide whether a person with dementia needs antidepressant therapy, the physician will weight his or her own observations and opinions more heavily than the reports of the patient, because of the cognitive impairment of the patient. Ideally, the physician will weight his or her observation more heavily as the severity of the patient's cognitive impairment increases. Or, if the purpose is to decide if a cognitively intact person in her or his third day of recovering from surgery needs a higher analgesic dose, best practices dictate almost total reliance on the report of the patient. However, if the purpose is to fully explore factors affecting health care utilization, the researcher may choose to weight reports from the patient, relative, and physician equally.

Factors Affecting Discrepancies

The finding that self-rated and externally rated reports are never perfectly correlated is consistent across multiple constructs, patient populations, and types of raters (Rubenstein et al. 1984; Sprangers and Aaronson 1992; Burke et al. 1998; Pierre et al. 1998). Several factors affect the magnitude of these discrepancies, including characteristics of the construct, the patient, the external rater, the relationship between the patient and external rater, and the assessment method (Neumann, Araki, and Gutterman 2000).

As previously noted, the more subjective a construct is the larger the discrepancy between self- and other ratings. Good agreement between self- and external-rater reports has been found for levels of functioning, overall health, less private chronic physical conditions and symptoms (e.g., limb functioning, skin condition), and preferences for type of health care setting (Magaziner 1992; Neumann, Araki, and Gutterman 2000). Moderate agreement is reported for ratings of cognitive status; family (as opposed to clinical) external raters tend to provide higher estimates of cognitive functioning compared with the patient (Magaziner 1992; Neumann, Araki, and Gutterman 2000). Finally, moderate to low agreement is reported for depressive symptoms, psychological well-being, quality of life, and pain (Bassett, Magaziner, and Hebel 1990; Gilley et al. 1995; Logsdon et al. 1999; Neumann, Araki, and Gutterman 2000; Bruera et al. 2001). Such findings may be because of the variance in access to information different raters have. The rater may be able to ascertain much more information about a condition that has less private, easily observable symptoms compared with a condition with primarily internal symptoms. Constructs that can be rated using more objective procedures like frequency counts (e.g., how often was the patient able to make a bed last week) may be easier for the rater to assess than constructs that require judgments based on the rater's own store of experience (e.g., "rate the intensity of the patient's pain, from no pain at all to worst possible pain"). Indeed, Magaziner (1997) found that questions that asked for a presence/absence judgment about a symptom produced higher agreement than questions that asked for a symptom intensity rating.

Both patient and external ratings are affected by personal characteristics. Demographic characteristics are an important factor; for example, older men appear less likely to endorse symptomatology than younger men and women (Bassett, Magaziner, and Hebel 1990; Neumann, Araki, and Gutterman 2000). Regarding patient characteristics, cognitive impairment often introduces a negative response bias (i.e., denial of symptoms when they exist) (Gilley et al.

1995). Similarly, patients with decreased awareness of symptoms because of dementia are less likely to endorse other problems, such as depression or pain (Ott and Fogel 1992). Depression also affects ratings, usually by introducing a positive response bias (i.e., depressed people tend to endorse symptomatology or disability at higher levels than nondepressed people [Miller 1980; Brustrom and Ober 1998]). Finally, there is evidence that ratings are affected by some personality variables; for example, one study reported that those patients in psychiatric treatment who continued to self-report high levels of symptomatology after their psychiatrists had rated them significantly improved were more likely to rate high on a scale of external locus of control and on a scale of hostility (Castrogiovanni, Maremmani, and Deltito 1989).

Some rater characteristics can specifically affect the ratings of external raters. Ratings are affected by both overall level of education and by specific knowledge about the assessed construct. It has been demonstrated that experts have much more refined and developed cognitive schemas for constructs in their area of expertise than lay people (Hardin 2003). It is reasonable to expect that varying schemas could yield different interpretations of the same data, thus affecting ratings. The context of the situation in which the rater is asked to produce a report affects that report. For example, Fordyce, Roberts, and Sternbach (1985) have produced extensive evidence regarding the positive effects of secondary gain on self- and external ratings of pain. In one study, female chronic pain inpatients were randomly assigned to a condition in which every day the interventionist talked to the patient for 5 minutes systematically reinforcing verbal complaints about pain with sympathy, verbal attention, or praise, or an identical condition except the interventionist reinforced reports of feeling better, coping more effectively, or following through with treatment recommendations such as exercise. Pain ratings taken immediately after these 5 minute encounters were significantly (20–25 percent) higher for the sympathy condition versus the feeling better condition (White and Sanders 1986).

The nature of the relationship between the patient and external rater affects the discrepancy between their reports. Discrepancies decrease as the amount of time an external rater spends with the patient increases, probably because the external rater's access to information about the patient increases (Magaziner 1997). Further, clinician and family external raters differ in their response precision (the amount of response discrepancy), although bias is consistently in the positive direction. These effects are surely due in part to differing access to information and cognitive schemas for processing that information. For example, relative to family raters, clinicians tend to rate

patients with dementia as more depressed, and both clinicians and family raters rate the patient as more depressed than the patient's self-ratings (Teri and Wagner 1991).

Finally, the more the assessment methods of patients and external raters vary, the more discrepancy between their reports might be expected. By definition the patient and external rater use different assessment methods because the patient relies in large part on inward exploration whereas the external rater must rely on observations of external signs. However, the potential for discrepancy increases when different assessment instruments are used. For example, discrepancy would be expected if a patient's level of depression is quantified based on responses to a symptom rating scale such as the Beck Depression Inventory, and the clinician's judgment of the level of the patient's depression is based on a structured clinical interview and guidance from the Major Depression diagnostic criteria of the ICD-9 (Boyd et al. 1982). As another example, different results are achieved with performance-based measures of functional ability as compared with self-report measures (Gotay 1996; Zanetti et al. 1999).

Psychological Theories for Predicting Discrepancies

As we explored above, there now exists a body of research exploring variances between self- and external ratings across a variety of settings, constructs, and moderating and mediating variables. There is also a growing body of knowledge about factors affecting self-report, as informed by the theories of social and cognitive psychology (Carroll, Fielding, and Blashki 1973). However, there has been little or no effort to integrate these two areas. As a first step toward integration, we now explore several theories that may be useful in predicting discrepancies between self- and external ratings.

The response-shift theory posits that as people adapt to their conditions, they perceive them differently (Daltroy et al. 1999; Gibbons 1999). One implication of this theory is that the chronicity of a condition will decrease ratings of that condition. Thus, both a patient who has suffered for 5 years from progressive dementia and his spouse may provide higher ratings of cognitive and functional ability compared with a newly assigned clinician or a daughter visiting from another state.

There is evidence that people tend to rate themselves more favorably than is truly the case. Research on the attitudes and beliefs of depressed

individuals and controls has indicated that psychologically healthy individuals are actually *less* accurate in their estimations of their performance and tend to be more optimistic and positive than is warranted by the facts (Layne 1983). Thus, a discrepancy between self- and external ratings is always to be expected because people are inherently biased in their self-ratings.

Cognitive dissonance theory asserts that when an individual holds two conflicting beliefs, this creates an uncomfortable state of psychological tension. This tension is relieved as the person relinquishes the belief that is held less firmly (Festinger 1957). Consider a patient who believes he is healthy and strong, but also knows that he is having arthritic pain in his knees that makes it hard to walk up stairs and do other activities. One way to relieve the uncomfortable dissonance between these competing beliefs is for the man to downplay the severity of pain and functional limitation caused by his arthritis. This theory helps us understand why firm evidence about a person's condition may sometimes not be well integrated into the rater's reports about the condition.

Similarly, self-awareness theory asserts that when a person comes across a self-focusing cue in the environment (e.g., another person's comment regarding one's behavior), this cue creates a state of self-awareness, which leads to a comparison of current thoughts and behavior to one's internal standards or expectations for oneself. If they do not match, this will cause discomfort and the person may flee from self-awareness to relieve the discomfort (Carver and Sheier 1981).

Finally, ratings are affected by one's self-identity, also called self-schemas. People have complex implicit internal theories about which aspects of their lives are stable, and which are subject to change. People reorganize their memories of events and experiences to accommodate their beliefs about the stability or instability of a particular component of their self-schema (Markus and Zajonc 1985). Thus, if being healthy and high functioning is considered to be a stable part of one's self-schema, this theory would predict that when experiencing disability such a person would tend to downplay those experiences, and be less likely to endorse those negative symptoms compared with an external rater. The implications for the patient's self-schema are greater than the implications for the external-rater's self-schema. External raters also may be disinclined to attribute negative symptoms to the patient when there are pertinent implications for the external rater. For example, caregivers of children or impaired adults may have a self-identity that includes the concept of being a good caretaker, and this concept may be threatened by the patient's symptomatology. Conversely, a caregiver who believes that his or her failings

in other areas of life are because of the sacrifices he or she makes as a caregiver might produce elevated patient symptom ratings to support this self-concept.

Conclusion and Future Directions

In this paper, we have defined two types of approaches to external-rated data, the proxy and other-rater approaches, and have explored how these approaches vary in their conceptualization of the measurement model and development of data analysis and interpretation strategies. We have also reviewed research on factors affecting discrepancies between self- and external-rated data and theories that might help to predict such discrepancies.

In conclusion, we strongly caution against any interpretation of the existing theory and research that would lead to a conclusion that either self- or external-rated data is undesirable for all situations in a certain category. Rather, the intent is to develop a more sophisticated understanding of factors that positively and negatively affect the accuracy of both self- and external-rated data. Such an understanding should guide data analytic approaches and allow for (a) better control of undesirable characteristics of the external rater when using the proxy approach and (b) better integration of self- and external-rater data to provide the best estimate of "truth" for the assessed concept when using the other-rated approach.

Acknowledgments

This material is based upon work supported by the Veterans Affairs Health Services Research and Development Service (RCD 01-008-1 awarded to Dr. Snow) and the VA HSR&D Measurement Excellence and Training Resource Information Center (METRIC; RES 02-235); Drs. Snow, Cook, and Morgan are METRIC core investigators. Partial support for Dr. Magaziner's effort was provided by the Claude D. Pepper Older Americans Independence Center, NIH P60 AG12583.

No authors have any conflicts of interest to disclose. This material reflects the opinions of the authors, and does not necessarily reflect the opinion of the Veterans Affairs Research and Development Service.

REFERENCES

Bassett, S. S., J. Magaziner, and J. R. Hebel. 1990. "Reliability of Proxy Response on Mental Health Indices for Aged, Community-Dwelling Women." *Psychology and Aging* 5: 127–32.

Becchi, A., P. Rucci, A. Placentino, G. Neri, and G. de Girolamo. 2004. "Quality of Life in Patients with Schizophrenia—Comparison of Self-Report and Proxy Assessments." *Social Psychiatry and Psychiatric Epidemiology* 39: 397–401.

Boldingh, E. J., M. A. Jacobs-van der Bruggen, G. J. Lankhorst, and L. M. Bouter. 2004. "Assessing Pain in Patients with Severe Cerebral Palsy: Development, Reliability, and Validity of a Pain Assessment Instrument for Cerebral Palsy." *Archives of Physical and Medical Rehabilitation* 85: 758–66.

Boyd, J. H., M. M. Weissman, W. D. Thompson, and J. K. Myers. 1982. "Screening for Depression in a Community Sample. Understanding the Discrepancies between Depression Symptom and Diagnostic Sales." *Archives of General Psychiatry* 39: 1195–200.

Bruera, E., C. Sweeney, K. Calder, L. Palmer, and S. Benisch-Tolley. 2001. "Patient Preferences versus Physician Perceptions of Treatment Decisions in Cancer Care." *Journal of Clinical Oncology* 19: 2883–5.

Brustrom, J. E., and B. A. Ober. 1998. "Predictors of Perceived Memory Impairment: Do They Differ in Alzheimer's Disease versus Normal Aging?" *Journal of Clinical and Experimental Neuropsychology* 20: 402–12.

Buckey, J. W., and N. Abell. 2004. "Validation of the Health Care Surrogate Preferences Scale." *Social Work* 49: 432–40.

Burke, W. J., W. H. Roccaforte, S. P. Wengel, D. McArthur-Miller, D. G. Folks, and J. F. Potter. 1998. "Disagreement in the Reporting of Depressive Symptoms between Patients with Dementia of the Alzheimer Type and Their Collateral Sources." *American Journal of Geriatric Psychiatry* 6: 308–19.

Byrne, B. 1994. "Testing the Factorial Validity of a Theoretical Construct." Chapter 3. Application 1. In *Structural Equation Modeling with EQS and EQS/Windows*, pp. 41–72. Thousand Oaks, CA: Sage Publications.

Carroll, B. J., J. M. Fielding, and T. G. Blashki. 1973. "Depression Rating Scales. A Critical Review." *Archives of General Psychiatry* 28: 361–6.

Carver, C., and M. Sheier. 1981. *Attention and Self-Regulation: A Control-Theory Approach to Human Behavior.* New York: Springer-Verlag.

Castrogiovanni, P., I. Maremmani, and J. A. Deltito. 1989. "Discordance of Self Ratings versus Observer Ratings in the Improvement of Depression: Role of Locus of Control and Aggressive Behavior." *Comprehensive Psychiatry* 30: 231–5.

Daltroy, L. H., M. G. Larson, H. M. Eaton, C. B. Phillips, and M. H. Liang. 1999. "Discrepancies between Self-Reported and Observed Physical Function in the Elderly: The Influence of Response Shift and Other Factors." *Social Sciences and Medicine* 48: 1549–61.

Festinger, L. 1957. *A Theory of Cognitive Dissonance.* Stanford, CA: Stanford University Press.

Fordyce, W. E., A. H. Roberts, and R. A. Sternbach. 1985. "The Behavioral Management of Chronic Pain: A Response to Critics." *Pain* 22: 113–25.

Gibbons, F. X. 1999. "Social Comparison as a Mediator of Response Shift." *Social Sciences and Medicine* 48: 1517–30.

Gilley, D. W., R. S. Wilson, D. A. Fleischman, D. W. Harrison, C. G. Goetz, and C. M. Tanner. 1995. "Impact of Alzheimer's-Type Dementia and Information Source on the Assessment of Depression." *Psychological Assessment* 7: 42–8.

Gnanadesigan, N., D. Saliba, C. P. Roth, D. H. Solomon, J. T. Chang, J. Schnelle, R. Smith, P. G. Shekelle, and N. S. Wenger. 2004. "The Quality of Care Provided to Vulnerable Older Community-Based Patients with Urinary Incontinence." *Journal of American Medical Directors Association* 5: 141–6.

Gotay, C. C. 1996. "Patient-Reported Assessments Versus Performance-Based Tests." In *Quality of Life and Pharmacoeconomics*, 2d ed., edited by B. Spilker. Philadelphia: Lippincott-Raven Publishers.

Hardin, L. E. 2003. "Problem-Solving Concepts and Theories." *Journal of Veterinary Medical Education* 30: 226–9.

Kuipers, P., M. Kendall, J. Fleming, and R. Tate. 2004. "Comparison of the Sydney Psychosocial Reintegration Scale (SPRS) with the Community Integration Questionnaire (CIQ): Psychometric Properties." *Brain Injury* 18: 161–77.

Lam, T. H., S. Y. Ho, A. J. Hedley, K. H. Mak, and G. M. Leung. 2004. "Leisure Time Physical Activity and Mortality in Hong Kong: Case–Control Study of all Adult Deaths in 1998." *Annals of Epidemiology* 14: 391–8.

Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33: 159–74.

Langa, K. M., E. B. Larson, R. B. Wallace, A. M. Fendrick, N. L. Foster, and M. U. Kabeto. (2004). "Out-of-Pocket Health Care Expenditures among Older Americans with Dementia." *Alzheimer Disease and Associated Disorders* 18: 90–8.

Layne, C. 1983. "Painful Truths about Depressives' Cognitions." *Journal Clinical Psychology* 39: 848–53.

Logsdon, R., L. E. Gibbons, S. M. McCurry, and L. Teri. 1999. "Quality of Life in Alzheimer's Disease: Patient and Caregiver Reports." *Journal of Mental Health and Aging* 5: 21–32.

Long, K., S. Sudha, and E. J. Mutran. 1998. "Elder-Proxy Agreement Concerning the Functional Status and Medical History of the Older Person: The Impact of Caregiver Burden and Depressive Symptomatology." *Journal of American Geriatric Society* 46: 1103–11.

Magaziner, J. 1992. "The Use of Proxy Respondents in Health Studies of the Aged." In *The Epidemiologic Study of the Elderly*, edited by R. B. Wallace and R. F. Woolson, pp. 120–9. New York: Oxford University Press.

———. 1997. "Use of Proxies to Measure Health and Functional Outcomes in Effectiveness Research in Persons with Alzheimer Disease and Related Disorders." *Alzheimer Disease and Associated Disorders* 6 (11, suppl): 168–74.

Magaziner, J., S. S. Bassett, J. R. Hebel, and A. Gruber-Baldini. 1996. "Use of Proxies to Measure Health and Functional Status in Epidemiologic Studies of Community-Dwelling Women Aged 65 Years and Older." *American Journal of Epidemiology* 143: 283–92.

Magaziner, J., J. R. Hebel, and J. W. Warren. 1987. "The Use of Proxy Responses for Aged Patients in Long-Term Care Settings." *Comprehensive Gerontology [B]* 1 (3): 118–21.

Markus, H., and R. Zajonc. 1985. "The Cognitive Perspective in Social Psychology." In *Handbook of Social Psychology*, vol. 1, edited by G. Lindzey and E. Aronson. Hillsdale, NJ: Erlbaum.

Miller, N. E. 1980. "The Measurement of Mood in Senile Brain Disease: Examiner Ratings and Self-Reports." *Proceedings of the Annual Meeting of the American Psychopathological Association* 69: 97–122.

Nekolaichuk, C. L., E. Bruera, K. Spachynski, T. MacEachern, J. Hanson, and T. O. Maguire. 1999. "A Comparison of Patient and Proxy Symptom Assessments in Advanced Cancer Patients." *Palliative Medicine* 13: 311–23.

Neumann, P. J., S. S. Araki, and E. M. Gutterman. 2000. "The Use of Proxy Respondents in Studies of Older Adults: Lessons, Challenges, and Opportunities." *Journal of the American Geriatrics Society* 48: 1646–54.

Ott, B. R., and B. S. Fogel. 1992. "Measurement of Depression in Dementia: Self versus Clinical Rating." *International Journal of Geriatric Psychiatry* 7: 899–904.

Perry, L., and S. McLaren. 2004. "An Exploration of Nutrition and Eating Disabilities in Relation to Quality of Life at 6 Months Post-Stroke." *Health and Social Care in the Community* 12: 288–97.

Pierre, U., S. Wood-Dauphinee, N. Korner-Bitensky, D. Gayton, and J. Hanley. 1998. "Proxy Use of the Canadian SF-36 in Rating Health Status of the Disabled Elderly." *Journal of Clinical Epidemiology* 51: 983–90.

Porzsolt, F., M. Kojer, M. Schmidl, E. R. Greimel, J. Sigle, J. Richter, and M. Eisemann. 2004. "A New Instrument to Describe Indicators of Well-Being in Old-Old Patients with Severe Dementia—The Vienna List." *Health and Quality of Life Outcomes* 2: 10.

Rubenstein, L. Z., C. Schairer, G. D. Wieland, and R. Kane. 1984. "Systematic Biases in Functional Status Assessment of Elderly Adults: Effects of Different Data Sources." *Journal of Gerontology* 39: 686–91.

Sherer, M., C. Boake, E. Levin, B. V. Silver, G. Ringholz, and W. M. High Jr. 1998. "Characteristics of Impaired Awareness after Traumatic Brain Injury." *Journal of International Neuropsychology Society* 4: 380–7.

Skinner, K. M., D. R. Miller, A. Spiro III, and L. E. Kazis. 2004. "Measurement Strategies Designed and Tested in the Veterans Health Study." *Journal of Ambulatory Care Management* 27: 180–9.

Snow, A. L., M. P. Norris, R. Doody, V. A. Molinari, C. A. Orengo, and M. E. Kunik. 2004. "Dementia Deficits Scale." *Alzheimer Disease and Associated Disorders* 18: 22–31.

Sprangers, M. A., and N. K. Aaronson. 1992. "The Role of Health Care Providers and Significant Others in Evaluating the Quality of Life of Patients with Chronic Disease: A Review." *Journal of Clinical Epidemiology* 45: 743–60.

Stone Arthur, A., S. Turkkan Jaylan, A. Bachrach Christine, B. Jobe Jared, S. Kurtzman Howard, and S. Cain Virginia. 1999. *The Science of Self-Report: Implications for Research and Practice.* Hillsdale, NJ: Lawrence Erlbaum Association.

Sudan, D., S. Horslen, J. Botha, W. Grant, C. Torres, B. Shaw Jr., and A. Langnas. 2004. "Quality of Life after Pediatric Intestinal Transplantation: The Perception of Pediatric Recipients and Their Parents." *American Journal of Transplantation* 4 (3): 407–13.

Teresi, J. A., and D. Holmes. 1997. "Reporting Source Bias in Estimating Prevalence of Cognitive Impairment." *Journal of Clinical Epidemiology* 50: 175–84.

Teri, L., and A. W. Wagner. 1991. "Assessment of Depression in Patients with Alzheimer's Disease: Concordance among Informants." *Psychology and Aging* 6: 280–5.

von Essen, L. 2004. "Proxy Ratings of Patient Quality of Life—Factors Related to Patient-Proxy Agreement." *Acta Oncologica* 43: 229–34.

White, B., and S. H. Sanders. 1986. "The Influence on Patients' Pain Intensity Ratings of Antecedent Reinforcement of Pain Talk or Well Talk." *Journal of Behavior Therapy and Experimental Psychiatry* 17: 155–9.

Yasuda, N., S. Zimmerman, W. G. Hawkes, A. L. Gruber-Baldini, J. R. Hebel, and J. Magaziner. 2004. "Concordance of Proxy-Perceived Change and Measured Change in Multiple Domains of Function in Older Persons." *Journal of the American Geriatrics Society* 52: 1157–62.

Zanetti, O., C. Geroldi, G. B. Frisoni, A. Bianchetti, and M. Trabucchi. 1999. "Contrasting Results between Caregiver's Report and Direct Assessment of Activities of Daily Living in Patients Affected by Mild and Very Mild Dementia: The Contribution of the Caregiver's Personal Characteristics." *Journal of the American Geriatrics Society* 47: 196–202.