2008 Eleanor Clarke Slagle Lecture

Embracing Ambiguity: Facing the Challenge of Measurement

KEY WORDS
- assessment
- evidence
- measurement
- methodology
- research

Assigning numbers to variations in the behaviors, experiences, or beliefs of individuals and groups is a firmly established tradition in Western culture. In many areas of occupational therapy practice, quantitative measures are required to document need for services and, increasingly, quantitative measures are required to document the value of these services in terms of the outcomes achieved. In order to meet these expectations, occupational therapy frequently has adopted both methods and instruments from other disciplines including psychology, medicine, and education. However, the assumptions and modes of thinking about people that accompany these methods are not always compatible with the values and practice of occupational therapy. This paper explores tensions inherent in the assessment process in a profession that is holistic and humanistic in its orientation. I propose that in order for assessment to serve our goal of supporting health and participation through engagement in occupation we must accept the uncertainty and be vigilant about the biases in thinking that are inherent in our measures.


I will introduce the topic of this paper with a story:

The summer that I turned 10 was very full. After living in Naples, Italy, for 3 years, my family and I were moving north to Milan. I was leaving the American school I had attended and would be start attending a new Italian school in the fall. In between, we returned to the United States for the summer. We traveled by ship each way, each trip lasting over a week with stops along the way in exotic places such as Gibraltar, Morocco, Majorca, and Cannes.

In the United States, we traveled around in a big, old, used Cadillac my father had bought to visit family and friends across New York and around Washington, DC. We celebrated several birthdays and the christening of my baby sister, went for a long hike with my uncle and dad and got lost, and watched hours of Saturday morning cartoons—a novel experience for us because Italy didn’t really have television at that time.

Upon arrival in Milan my brothers and I had a few weeks of tutoring in Italian and then started at our new school. The first day my teacher gave the class what must be the universal first-week-of-school assignment: to write about “what I did over the summer.” I picked up the unfamiliar dip pen, dipped it in the inkwell, and scratched out (not very neatly): “Io sono andato a America.” “I went to America. I went on a boat. I came to Milan.” At that point I had exhausted my knowledge of Italian grammar and vocabulary and stopped. I knew that what I had written looked like a first-grader’s essay, including the blotches that came from my unruly, unfamiliar pen. For a previously competent student it was embarrassing and demoralizing not to be able to do better. However, getting a bad grade was not my worst fear. What was more important to me was that my teacher might think that those few simple sentences I had managed to produce told the whole story of my adventurous summer—that she would think that essay was ME.

Facing the Challenge of Measurement

My teacher was actually very kind and understanding on that day long ago, but the experience became like a grain of sand in my consciousness, an irritant that never
quite went away. I had realized the discrepancy between what we can tell or show and what the experience really is, between the measure of a person’s ideas or abilities and reality. Eventually that insight led to the question that has fascinated and challenged me for many years: How can we reconcile the need to design and use measures in our research and practice with the knowledge that the information they give us is inadequate, often ambiguous, and sometimes misleading?

In our society today there are very strong forces pressing us to treat the data from standardized measures as the person’s “full story.” From reimbursement decisions based on scores on the Functional Independence Measure (FIM; Uniform Data System for Medical Rehabilitation 1997) to high-stakes achievement tests in schools, there is pressure to simplify very complex decisions through the application of numbers. Occupational therapy’s concern for the whole person is being challenged daily by this pressure in our practice and in our research.

An effective response to this challenge will need to go beyond selecting instruments with the best reliability or predictive accuracy, or the application of modern methods such as Rasch analysis. We also must examine and challenge some of the assumptions underlying the current use of measures and the conclusions being drawn from this use. The phenomena at the core of occupational therapy’s concern are complex and, as our attention moves from the domain of body structures and functions to activity and participation, they also are increasingly abstract. Our concerns encompass both directly observable events—the doing—and experiences that can be conveyed only through some intermediary mechanism—the phenomena we call meaning, feeling, being, and quality of life.

To try to capture a picture of these phenomena for use in our practice or research, occupational therapy often has turned to the methods and instruments developed in other disciplines whose concerns, priorities, and knowledge may not be the same as ours. One source has been medicine, a discipline whose primary concern is observable phenomena such as the integrity of body functions and performance of physical tasks. Another has been psychology, which focuses more on abstract unobservable processes such as the cognitive, social, and emotional dimensions of experience. Although each of these bodies of knowledge has contributed in valuable ways to the tools we use, each has also brought influences and assumptions that have often gone unexamined for their compatibility with occupational therapy. This paper examines several of these important issues:

- How we define what we are measuring;
- How we derive and interpret quantitative data from our instruments; and
- How the social nature of the assessment process influences the results we obtain.

I will end with some thoughts about how a better understanding of these issues can help us to achieve the ideals of occupational therapy practice and to advance our research.

**Power of Words**

Words connote reality. When we have extracted a pattern from the array of stimuli we experience, we mark the pattern with a name—a word. Almost immediately the word takes on the power to influence our thoughts and feelings. We know from cognitive science, for example, that speaking or seeing the word first makes it more likely that we will perceive a particular stimulus (Bueno & French-Mestre, 2002). Words reduce ambiguity to enable us to live socially in a world of objects. I can show you the object I give the name to, and we can agree to use that name whenever we speak about that object. Cultures vary in the extent to which they differentiate within particular categories, but they all have ways of marking or pointing out with words the features that differentiate the categories that are meaningful within their culture.

The sciences vary in the degree of precision in their naming processes. For example, in physics the phenomenon given a particular name may have a very precise mathematical reference such as an equation (Lightman, 2005). In biology some terms have very precise referents (e.g., fern, poodle) that can be readily identified from their observable features. However, when we begin to study human experience scientifically, terms begin to appear that have varying degrees of uncertainty or ambiguity in their referents. We cannot demonstrate pain with a microscope or point out health as clearly as we can identify a cell.

Metaphor often comes into play to express aspects of phenomena that are less easily pointed out or defined (Brown, 2003). These metaphors are shaped by a culture’s values and orientation to human affairs (Lakoff & Johnson, 1980). For example, the metaphors related to disease in U.S. culture evoke images of disease as an enemy: Viruses attack our cells, we fight off a cold, and we beat or succumb to an illness. Our metaphor of disease describes it as a foreign invader, something with distinct boundaries that is separate from ourselves.

However, other cultures may hold different metaphors about disease. For example, disease may be understood as a storm within us, caused by disturbance of natural harmony between body and soul, or as a disturbance in the balance between opposing forces (Karaz, 2005; Storck, Csordas, & Strauss, 2000). In these metaphors, disease is on a continuum with health and is not exclusively caused by outside influences.
The important point is that different approaches to diagnosing and treating illness will appear correct or “right” to cultures that hold these different metaphors. If the cause of disease is believed to be external to ourselves, then it makes sense to search for a primary cause using measures that focus on external observable factors. However, if illness and disorder reflect a disturbance of balance among internal and external forces, we are more likely to consider multiple causal forces as relevant and to use a combination of objective and subjective measures to examine these possibilities.

Implicit meaning associated with words is not limited to metaphors. Consider the word recovery, a term used frequently in medical rehabilitation outcome studies. In common usage recovered means “restored fully to health,” such as recovered from the flu. However, the same term is currently applied in research with a very different meaning. Duncan and her colleagues illustrated this problem in an analysis of results from stroke outcome studies (Duncan, Jorgensen, & Wade, 2000). Twenty-seven of these studies used the Barthel Index as their outcome measure, which examines the person’s need for assistance to perform basic activities of daily living, including eating, dressing, grooming, and walking. All studies used a cut-off score to identify patient groups whom they labeled as either recovered or not recovered. However, Duncan et al. (2000) found that, across the 27 studies, 7 different cut-off scores had been used to define the recovered and not recovered groups, and the choice of cut-off point was not explained in several of the studies. This variability in the definition of “recovered” affected the conclusions drawn about the proportion of patients likely to improve over a particular period of time or as a result of intervention. Conclusions about the effectiveness of intervention would have been different in several studies if a different cut-off score had been used.

Whose Definition Is This?

The impact of variability in the operational definition of recovery is not the only concern raised by Duncan et al.’s (2000) findings. Regardless of which cut-off score they used, by choosing “recovery” to describe the focus of their research, these investigators all accepted the implicit meaning of the term as defined by the measure, which is “not needing physical assistance with most basic ADLs.” Whose definition is this? Most likely it is that of the payers, who are concerned with how many days the person must be treated in an expensive facility and when the person can go home without need for specialized or extra support. Is that a meaningful focus of concern? Of course it is. Health care is an expensive resource, and it is reasonable for the compa-

nies and agencies involved in financing it to be concerned about efficient allocation of this resource.

However, using “recovery” to describe the results of this research also pulls in the usual meaning of the term to most people reading it. The implicit message being communicated to the reader is that the participants are now “well,” back to the way they were before, and their major health issues are resolved. Perhaps, according to a very narrow physical health standard, they are. But occupational therapy practitioners, family members, and the clients themselves know that this isn’t the case. As Radomska expressed so well in the title of her article, “There is more to life than putting on your pants” (Radomska, 1995).

The use of a cut-off score on a measure to define recovery rests on the assumption that recovery of function can be marked distinctly in the same way that we can define whether someone does or does not have a fever. This approach has roots in medicine’s focus on success as measured by cure rate. However, in this context it denies the ambiguity inherent in processes, such as functional recovery, that are slow and continuous and that often vary depending on contextual factors present at a given moment. The medical researchers reporting these studies are applying the same framework to a very different type of experience and treating it as if it is the same. This research evidence may then be used to determine whether treatment is authorized, or a service is deemed medically necessary.

In the field of mental health, introduction of the term recovery in the 1990s represented a radical reframing of assumptions about services for people with serious mental illness. In this case the leaders in the field argued forcefully for a dynamic conceptualization of recovery as a process focused on meaningful participation in life even though the illness may not be cured (Anthony, 1993). Recovery as defined here clearly requires consideration of more than whether the person is able to complete basic ADLs, and it cannot be evaluated using cut-off scores on a single measure.

Meaning Depends on the Measure

Duncan et al.’s (2000) review of stroke research is a rare example of a scientist looking carefully at the impact of decisions about measures on the results of research and the conclusions drawn from these results. The paucity of such critical reviews stands in striking contrast to the level of scrutiny and amount of discussion one can find in the literature on sample selection methods, controls for bias in the administration of measures, or selection of appropriate data analysis methods. Although considerable attention is paid to evaluating the psychometric properties of the instruments used, little is paid to the appropriateness of the
measures selected for the question being investigated. In many research reports it seems sufficient to report that the measures selected are “reliable and valid” before moving on to other weightier issues. Yet the validity of even the best-designed randomized clinical trial ultimately depends on whether the outcome measure used is appropriate for the question and responsive to the expected amount of change (e.g., Matson, 2007).

This word recovery, as well as other terms we frequently use, such as function, disability, activity, and participation, share the common feature that their definition (their meanings) depend heavily on the measure used in the particular context. Our habits of thought and communication lead us to expect that a word like function always refers to approximately the same construct or has the same meaning. However, depending on the measures chosen, the reality at present is that there may be very little overlap in content and often quite variable degrees of association between clinical instruments that purport to measure the same thing (e.g., Coster et al., 2004). As a consequence, when a study reports that a treatment is or is not effective, it is impossible to draw any conclusions about the implications of these results until we know how the outcomes were measured and the criteria used to define effectiveness.

Borrowed Ideas

Medicine meets the behavioral sciences in the arena of disability and rehabilitation. Here the medical orientation to objective phenomena and preference for clear-cut distinctions meets the complexity, unpredictability, and ambiguity of people’s daily behavior. We can see the influences of both of these disciplines played out in the design of various measures used in the field (Streiner & Norman, 1995). Medicine has a pragmatic focus on “what works” in terms of differentiating groups with different diagnoses or predicting outcomes of professional interest. It has not concerned itself particularly with identifying or defining underlying constructs. Not surprisingly, the discipline tends to emphasize physical performance in measures of function and to emphasize signs, symptoms, and diagnosis during the assessment process.

In contrast, psychology has consistently concerned itself with abstract constructs presumed to reflect processes underlying observed behavior. As a science, it also views itself as seeking facts or truths about persons, but the primary objects of its theory and research—constructs such as memory, self-efficacy, and attention—are not fully present in nature. Instead, their existence and influence is inferred based on what can be observed. Definitions of abstract concepts such as these are particularly likely to reflect cultural orientation and values. They represent what a given group considers distinctive, worth knowing about, or real.

For example, Westerners believe that we can identify and measure a set of personality traits that influence the behavior of a person across situations (Ozer & Benet-Martínez, 2006). This view is consistent with a cultural value and philosophy that view the person as an active, autonomous agent and historically has had difficulty acknowledging how much the environment (both social and physical) influences our behavior (Danziger, 1997). Considerable research in social psychology has been done to identify and name these traits and to investigate the relations between measures of these traits and other phenomena of interest.

However, in many Eastern systems of thought individuals are considered to be an inextricable part of a larger whole and their characteristics can only be understood in relation to the social context of which they are a part (e.g., Iwama, 2003; Nisbett, Peng, Choi, & Norenzayan, 2001). This alternative view challenges the Western approach, asserting that it does not reflect all reality and cannot be assumed to capture universal truths. Although Western measures might be translated and administered to Asian people, this does not guarantee that their data can validly be interpreted using the same Western framework.

From Numbers to Measures

In its early days, in order for psychology to be considered a science rather than remain a branch of philosophy, it had to find a way to assign numbers to the abstract qualities of interest or find things to count that were accepted as representative of these qualities. Numbers were the language of science and the means of measurement.

Another story will serve well to introduce some important features of measurement. There is a bridge in Boston crossing the Charles River named the Harvard Bridge, although it crosses from Boston over to MIT. If you walk across the bridge you will notice that the pavement is marked at regular intervals with lines indicating that this distance is equal to so many “smoots.” The bridge is 364.4 smoots—“plus or minus an ear”—long. This fact was discovered when the Lambda Chi Alpha fraternity used the body of their shortest freshman pledge, Oliver Smoot, to measure the bridge back in 1958. The fraternity, with the support of the city, has maintained this unique measuring system ever since (reported in Tavenor, 2007).

This story is a humorous reminder of the original relation between measures and the form of the human body, which is still preserved in our “foot” ruler, and the need for measures to solve practical problems such as determining
the length of a field or the correct height for a doorway. Systematic application of measures was necessary for order and harmony in early societies (e.g., to ensure that a square house was built with equal corners and parallel walls of the same height). From the beginning measures also had important social value as they were needed (e.g., to mark boundaries of land ownership or to determine appropriate charges for the weight of goods being sold; Tavenor, 2007).

Since ancient times secret qualities and powers have also been associated with numbers and mathematics (Livio, 2002). For the Greeks, numbers and the proportions described in geometry reflected the structure and harmony of the universe. They could be used to represent the systematic relations of musical notes and the patterns seen in nature. Thus the study of mathematics revealed important truths about the universe.

As subsequently discovered by Newton, Galileo, and other scientists, mathematics also could be used to express precisely the laws governing many physical phenomena. Descartes subsequently summarized the view of the scientific age, arguing that because qualities are the product of our unreliable senses, quantity is a more reliable measure of reality than quality (Tavenor, 2007). From that point on in Western history, it was not a large leap to begin to view quantity as a measure of ultimate value. This view was also consistent with the industrial era’s valuing of productivity and standardization of units of industrial products (Danziger, 1997).

We don’t have to look far to see the expression of this thinking in psychology. In many of the instruments designed to measure human abilities, the capacity to do more (e.g., complete more puzzles, solve more arithmetic problems) is the means used to rank persons. The numbers obtained from these measures are believed to provide a more objective way to determine individuals’ standing on a culturally important dimension such as intelligence. A subtle but important conclusion that has followed from this reasoning is that if these numbers were obtained through rigorous and standardized procedures, then they must reflect reality (Gould, 1981).

Some constructs of interest to the behavioral and social sciences are not as readily measured by observation of performance. An alternative approach to generating the numbers needed for scientific analysis is needed. Here another leap of reasoning is made. If qualities of experience, like self-efficacy, or confidence, goal-orientation, or disability, are assumed to exist on a continuum, then we can use ordered response scales to locate each person on that continuum. So we ask the person whether each of several statements about his or her confidence is true or not on a scale from 1 to 7, from “not at all true” to “definitely true,” or we ask the person who is recovering from a stroke how much difficulty, on a scale from 1 to 4, from “cannot do” to “no difficulty,” she has performing a set of daily activities. If we add or average the resulting item scores, we now have a number we can use to rank people on their confidence or their function or to correlate with scores from other measures.

There are several of problems with this approach to developing measures. For one, it rests on the assumption that the items and ratings on the instrument divide the dimension of confidence, disability, or self-efficacy into equivalent units like the inch markings on a ruler (Wright & Linacre, 1989). Therefore, a score of 20 is interpreted as indicating that the person has twice the confidence, function, or self-efficacy as the person whose score was 10, and half the confidence, function, or self-efficacy as the person who scored 40. Or, as another example, we assume that achieving a positive change in one’s confidence from a score of 1 to a score of 2 is equally significant as achieving an improvement from a score of 3 to a score of 4.

We know from clinical practice that not all tasks or levels of performance are equally challenging or meaningful, and yet this is what we assume when we add these kinds of scores together. In reality when we subject our measures to modern analytic approaches, such as item response theory or Rasch analysis (Bond & Fox, 2001), the picture often look quite different: Sometimes it is a short step to improve from one rating level to the next, and sometimes it is a very big step. In other words, our untested measures of these complex constructs may well give a distorted picture of reality. They may give equal weight to easy and difficult achievements or underestimate the degree of progress a person has made toward important goals.

The Social Context of Measurement

The construction of a measure is a human process. Therefore, by definition, the process is embedded in a social system of values and ideas about people (Danziger, 1990). This social influence often is quite hard to see when we are part of the same system, but it becomes apparent if we ask, Who is not well-described by this measure? For example, practitioners know that many standardized tests do not provide a differentiated profile of their clients with significant disabilities. As a psychology intern I saw this clearly when I tested several community-living adults with developmental disabilities. It was apparent from observing and interacting with these clients that their profiles of cognitive strengths and limitations were very different, but they all achieved identical (low) scores on the IQ test I administered.
This odd situation results from the fact that the primary objective for the developers of these instruments was to maximize the differentiation among the majority of people in the population: those who score between ±2 standard deviations from the mean in the distribution of the ability or trait (Anastasi & Urbina, 1997). Therefore, most of the items provide useful information only about individuals within that range. Unfortunately, that is not the part of the population we typically serve. The choices the developers made when selecting the items and creating the scoring system may have made it impossible for the person with a motor or communication impairment to obtain a score above 0. The implication of such a score is that this person’s performance reveals nothing of significant interest, a value judgment that is embedded in the design of the instrument.

The practitioners who were involved in pilot testing of the School Function Assessment (SFA; Coster, Deeney, Haley, & Haltiwanger, 1998) recognized—and resisted—the negative social implication of 0 scores. When an earlier version of the SFA used performance ratings that were on a scale of 0 to 3, they almost never gave a student a score of 0, even though it was obvious from other data in the form that a score of 0 would have been the appropriate rating. When we asked why, they explained that they hated the connotations of giving a child a 0 score because it seemed so pessimistic and because others often interpreted these scores to mean that the student couldn’t do anything.

**Impact of Our Choice of Measurement Lens**

Our instruments provide a way to extract a pattern from the performance of an individual for some purpose. However, the complexity of a person’s behavior can be viewed through many lenses, each of which may detect a different pattern. In turn, the choice of lens has a profound influence on the picture that the user forms of the person who is being assessed. It can emphasize deficit, as measured by standard deviations below the mean, or it can call attention to achievements, as measured by a score that reflects the current repertoire of daily life skills.

More importantly, the type of picture constructed by an instrument often leads to very different kinds of dialogues about the person’s needs, potentially useful interventions, and likely outcomes. To illustrate, a study by Linehan, Brady, and Hwang (1991) presented teachers with two different assessment reports on the same 12-year-old student with severe disabilities. One report summarized the student’s performance on standardized test items, such as standing on one foot for 5 seconds, cutting out a circle, or drawing a cross. The other report provided descriptions of how the student accomplished various tasks during the school day, indicating that he could dress himself independently except for tying shoes and could travel independently from his classroom to the lunchroom. When asked to project the student’s likely level of achievement of goals on his individualized educational plan for the year, those who read the second description expected significantly higher achievement than those who read the first one.

Since Rosenthal and Jacobson (1968) published their classic book *Pygmalion in the Classroom*, there have been ample other demonstrations of the power of evaluative information to influence expectations and, in turn, to influence outcomes in situations that include classrooms, research labs, physicians’ offices, and social encounters (Rosenthal, 1976). Our measures are a major source of information in our practice and our research.

In the past decade the United States has increasingly moved toward decision making on the basis of instrument numbers. In national surveys to guide policy decisions, a person is counted as having a disability based on whether he or she has difficulty performing two or more specific daily activities (e.g., Walsh & Khatutsky, 2007). In some settings one must qualify for services by scoring sufficiently low on a particular test (e.g., at least 1.5 standard deviations below the mean). The argument often made in support of these approaches is that an objective method is being applied because numbers from standardized measures are being used and that objective methods are more trustworthy, reasonable, and fair. But is this true? Behind these numbers is a human decision to select a particular instrument that emphasizes certain tasks or abilities and minimizes the importance of others. It is a human decision to set or accept a particular criterion. That decision is based on pragmatic or economic reasons and not on science or on an understanding of the strengths and limitations of measures. Nevertheless, these decisions affect our practice and the services our clients can obtain.

The situations I have been describing fall in the domain of what Messick (1980, 1989) has termed the *consequential aspects of validity: Are the social consequences that follow from administration of the test appropriate, given the nature of the test? Messick and others (e.g., Cronbach, 1988) have argued that both meaning and values are always involved in the validation of measures and therefore validators must examine whether an application of a measure has appropriate consequences for individuals and institutions. It is hard to resist the apparent legitimacy of numbers, but we need to examine existing practices by asking questions, such as*

- Is this an appropriate measure to use for identifying clients whose functional limitations are of a type and degree that require intervention and support services?
• Is requiring that a student must score minus 1.5 SD deviations on a specific standardized test in order to obtain occupational therapy services consistent with the legal definition of students to whom these services should be provided?
• Does this measure sample the appropriate content using appropriate methods to identify whether occupational therapy services have helped the client progress toward important goals?

If the answer to any of these questions is no, then I believe we have a professional and ethical obligation to challenge these misapplications of measurement and to advocate strongly for more appropriate alternatives.

Sources of Bias in Measurement

Assessment is always a social process. For one, assessment is most often conducted in some kind of face-to-face exchange between practitioner and client. In addition, the majority of clinical assessment tools require that a person (the practitioner or the client) determine the appropriate quantitative measure (a score or rating) to assign for a given item. These social features of the assessment context may exert more influence over the data obtained from measurement than we realize. Most efforts to ensure the quality of assessment data focus on reducing random sources of inconsistency across occasions or raters. These random influences are the sources of potential error that traditional reliability studies examine. But there is a substantial literature demonstrating that consistent biases also may influence the outcomes of measurement (Gilovich, Griffin, & Kahneman, 2002).

One source of bias is our susceptibility to influence by elements of a situation of which we are not even aware, or which we believe (wrongly, as it turns out) we are able to resist. For example, two studies done with the FIM (UDS, 1997) showed that raters were systematically influenced in their own ratings by seeing the ratings of other items that had already been completed by other team members (Doctor, Wolfson, McKnight & Burns, 2003; Wolfson, Doctor, & Burns, 2000). Studies with other instruments have shown that respondents are systematically influenced by the anchor or range of the scale. When asked to rate an experience such as the frequency of feeling irritated or sad, respondents gave different answers depending on whether the scale extended over a short or longer period (e.g., “in the past week” vs. “in the past month”; Chapman & Johnson, 2002). Nevertheless, raters consistently claim that they were not influenced by variations such as these when making their judgments.

Respondents also appear to use the structure of items and scales (and features of the assessment context) to make inferences about what the examiner is really most interested in (Redelmeier, Schull, Hux, Tu, & Ferris, 2001; Redelmeier, Tu, Schull, Ferris, & Hux, 2001; Schwartz, 1999). Then they respond according to this inferred purpose, perhaps by emphasizing certain types of experiences and minimizing others. Thus, cues suggesting that the examiner is most interested in physical function may lead respondents to under-report experiences reflecting their emotional well-being, or vice versa. These cues may be as subtle as the pictures on the wall of the room where the assessment is conducted.

Instruments developed without input from people with disabilities frequently present quandaries of interpretation. For example, the well-known SF–36 (Ware & Sherbourne, 1992) and other health-related, quality-of-life measures introduce questions about a person’s positive or negative daily experiences with the phrase, “Does your health limit you...?” The person with a recent stroke may include stroke-related impairments as part of his or her definition of current health state when answering these items. However, the teenager with cerebral palsy or an athlete with a decade-old spinal cord injury may not consider his or her disability to be a health problem and may wonder how to answer a question such as, “During the past 4 weeks were you limited in the kind of work or other regular daily activities you could do as a result of your physical health?”

If a study compares the quality-of-life of these groups using the SF–36, can we be sure that the responses from each group are capturing the same experiences? (Hays, Hahn, & Marshall, 2002) The SF–36 is considered the gold standard among measures of health-related-quality-of-life and is applied widely in clinical research. But is an item that asks whether or not the person is “limited in walking more than 1 mile” a valid indicator of quality of life (Meyers & Andresen, 2000)?

Clinical assessment also is a judgment process that requires a complex integration of information from multiple sources. Studies in cognitive psychology have documented that in situations with complex processing demands people often reason by applying heuristics, which are thinking short-cuts that help reduce the complexity of information processing by applying a general guideline to arrive at a judgment (Tversky & Kahneman, 1974). These shortcuts actually work quite well in many problem-solving situations of daily life. However, they also make us susceptible to systematic errors, particularly when we try to synthesize results from multiple sources of assessment information (Croskerry, 2003; Garb, 1998). For example, we are susceptible to confirmatory bias, which is the tendency to notice only information that supports our working hypothesis about the source or nature of a person’s problem, so we fail to seek out or account for con-
Thinking differently also leads to different decisions being made in the context of intervention or research. For example, when the PEDI was first used in early studies of dorsal rhizotomy surgery for children with cerebral palsy, it revealed that meaningful functional changes could occur even without significant changes in impairment-level measures (Dudgeon et al., 1994; Nordmark, Jamlo, & Hagglund, 2000). This was a significant challenge to existing approaches to evaluating interventions, which had assumed that changes in impairment must precede changes in function. Now, use of functional outcome measures is routine in clinical trials of surgical and pharmacological interventions for children with cerebral palsy.

Here is an incredible power for positive change if one can design an instrument that simultaneously fits well enough within the existing system to be adopted but incorporates enough differences to change thinking in a positive way—in other words, making test development a subversive activity. To be an effective subversive one must first thoroughly know the existing system—its rules, policies, priorities, resources, and ways of thinking—so that the new instrument is designed to meet the system’s essential criteria well enough to be given serious consideration. The instrument developer also must have a vision of a new way to bring the client’s story forward and must be able to persuade the powers-that-be that this new alternative is just what they need. Who knew that instrument development also might require political strategy and skills? But it does. And it must so that we can persuade powerful entities, like the Centers for Medicare and Medicaid Services and other policy-making bodies, to adopt or accept the new measures we develop.

An effective, positive subversive must be willing to try a different way to capture the client’s story. Occupational therapy’s holistic and client-centered philosophy, values, and practice provide excellent preparation for this creative role. We have seen many wonderful examples of this creativity in the development of instruments such as the Activity Card Sort (Baum & Edwards, 2008), the Children’s Assessment of Participation and Enjoyment (King et al., 2004), and the Occupational Self-Assessment (Baron, Kielhofner, Jenger, Goldhammer, & Wolenski, 2002), which have influenced practice and research both within and beyond occupational therapy by enabling a richer portrait of the client’s life and concerns to emerge.

Routinely using instruments such as these in occupational therapy practice is one way that we can change the dialogue with clients, family members, and other professionals. But we also need to acquire and use sophisticated knowledge about measures to challenge current assessment practices that are overly narrow in focus or require use of

Where Should We Go From Here?

The discovery that the tools and processes we have thought of as objective, scientific, or sound are, in fact, fraught with uncertainty can be disconcerting. Confronted with this evidence of pervasive ambiguity we may want to throw up our hands and say “Oh, well, it is what it is” and continue with our usual habits. Alternatively we may consider abandoning the whole enterprise as hopeless. However I would like to urge another alternative, which is to use the power that a deeper understanding of measures gives us to work toward more positive outcomes.

Becoming a Positive Subversive

The story at the beginning of this paper introduced the idea that measures have the power to shape the story that others hear about a person. I have presented several examples where that power has resulted in a narrow, truncated, impoverished, or misleading view of the person. However, that power also can be used to positive ends to challenge these limitations and bring the larger story to life. We can all play a part in this important enterprise.

For years I had a title ready for a talk I hoped to give someday: “Test Development as a Subversive Activity.” In this talk I would tell about how I discovered, from working on the Pediatric Evaluation of Disability Inventory (PEDI; Haley, Coster, Ludlow, Haltiwanger, & Andrellos, 1992) and School Function Assessment (Coster et al., 1998), that the design of an instrument could actually cause the people using it to think differently about the children they were assessing. After using one of these assessments it was common to hear comments from parents, teachers, and therapists like the one from a mother who said, “This was the first team meeting where we talked about Jeremy’s strengths, not just his deficits.”

tradictory information. Another well-known cognitive bias is search satisfying, the tendency to call off a search (e.g., for an explanation of assessment results) once a plausible answer has been found. This tendency may lead us to conclude our consideration of the data prematurely, before we have completed a full examination of all the findings.

These cognitive biases are an outcome of how our human brain functions, and they operate outside of our awareness. They have been studied extensively in psychology as well as in medicine, and their impact on diagnostic reasoning is described in the recently popular book *How Doctors Think* (Groopman, 2007). The reality of their influence is yet another reason we should draw conclusions from our measures with caution.
instruments in inappropriate ways, such as when qualifying students for services. These practices not only limit the client’s ability to tell his or her full story, but they also restrict our profession.

We need to challenge interpretations of research evidence that draw inappropriate conclusions from the measures that were used, particularly when those interpretations are used to restrict occupational therapy practice or to establish overly narrow service guidelines. If a study or a systematic review concludes that a therapy program “does not improve function,” then we must examine whether the outcome measures examined more than basic physical function and challenge the conclusions if they do not. If a study concludes that little further recovery is seen after the first six months following a traumatic brain injury, then we must examine how recovery was defined and whether the measure used to do so is sensitive to smaller amounts of functional change. And if a study purports to examine participation, then we should make sure that the content of the outcome measure examines more than basic ADLs or whether the person can walk a mile and asks about social relationships and engagement in family and community life, work, play, and leisure. A life of quality is about so much more than buttoning a shirt or tying shoes. We must make sure that our measures capture its richness and complexity. ▲

Acknowledgments

There are many, many people whose support, encouragement, and belief in me helped to bring this work to fruition. I would especially like to thank my students at Boston University for the energy and intelligence that makes teaching them such a joy; Ellen Cohn and Sue Berger for nominating me and the rest of the creative and generous faculty at Boston University for being the best colleagues one could ever hope to have; and my family, whose love and good humor keep me going.

References


