I have asserted that occupational therapists have a unique perspective on function, and that perspective provides a context for the interpretation of our evaluation results (Fisher, 1992). This perspective recognizes function as a process as well as an outcome, and it gives consideration to the volitional nature of our clients. Our challenge is to develop and use functional measures that reflect our unique perspective. While the developer of new assessments has the potential to use innovative and creative methods to develop tools that meet our needs, the clinician and the researcher are faced with the dilemma of choosing from among an almost endless array of options for the most appropriate measures for evaluating each client.

This is not an easy task. The 1980s heralded an increased concern with state-of-the-art functional assessments (Forrer, 1982; Keith, 1984; Spitzer, 1987). Functional assessment, for the purposes of the present paper, refers to the evaluation of basic self-care, that is, activities of daily living (ADL) and instrumental activities of daily living (IADL) (e.g., homemaking, traveling outside the home, managing finances). In large part, the concern with functional assessment has arisen from the need to be able to evaluate the appropriateness, necessity, and effectiveness of intervention programs. Certainly, occupational therapists share this need.

Moreover, the occupational therapist faced with selecting appropriate evaluation tools must consider several critical issues related to the reliability, validity, and usefulness of existing functional (ADL and IADL) assessments that have been reported recently in the literature. As will become readily apparent, these issues are interrelated; their separation into distinct issues serves only to clarify the scope of the problem. Some of these issues are described below.

Test Purpose

“Occupational therapists use measurement to describe the client’s problem, to formulate a prognosis, and to evaluate the effects of occupational therapy intervention” (Law & Letts, 1989, p. 522). Functional assessments have many uses and can be grouped into general categories, based on their intended purpose. It is critical that the occupational therapist choose from among available assessments one that is designed for the purpose that he or she intends.

Kishner and Guyatt (1985) categorized tests as (a) discriminative (e.g., baseline data collection, diagnosis), (b) predictive (e.g., decision making regarding provision of care, including treatment, support services, and placement), and (c) evaluative (e.g., efficacy research, quality assurance). Occupational therapists require measures that are directive or prescriptive as well. A well-designed, criterion-referenced hierarchical scale in which items (e.g., ADL or IADL tasks) have been calibrated along a continuum of increasing ability such that each task is more challenging than the previous task can provide a guide to the progression or sequence of intervention.

Many standardized functional assessments exist, most of which were framed from the perspective of epidemiological research, health care policy, and determination of need. Therein lies the problem: Although these assessments are sometimes helpful in identifying who is in need of therapeutic services, they tell us little about specific areas to target for intervention. That is, they tell us what a person can and cannot do (e.g., dress, eat), but not why or why not (e.g., limited dexterity, memory loss, apraxia) or whether the client wants or needs to do the task.

Another caveat is that practice often is driven by measurement, and measures have the potential to guide practice in inappropriate ways. Because most existing standardized functional assessments fail to inform us about the process or volitional aspects of task performance, occupational therapists emphasize the use of measures that are focused on more discrete aspects of function. Often, the result is to focus treatment at that level.

Standardization

Although occupational therapists routinely evaluate ADL and IADL by direct observation, most therapists use home-grown evaluation tools that lack known validity or reliability. That is, there is general recognition that many programs have developed their own ADL and IADL assessments without adequate attempts to establish the validity and reliability of the instruments; no instrument has become a gold standard.
Rater Reliability and Severity

The influence of rater judgment is another frequently cited area of concern, especially for IADL assessments (George & Fillenbaum, 1985; Lawton, 1987; Rubenstein, Schairer, Wieland, & Kane, 1984). Rubenstein et al. found that clients rate themselves higher on ADL and IADL performance than do nurses or community proxies. Other factors that may affect the reliability of ratings include the competence of the client doing the self-rating, the client's physical capacity, the rater's familiarity with the subject, the rater's professional discipline, the function being rated, and the purpose of the assessment (Lawton, 1987).

Another major reason for lowered reliability of IADL assessments is that the complexity of IADL tasks requires that greater degrees of the rater's judgment be used in scoring; what constitutes adequate performance is highly variable and reflects the rater's personal bias (Lawton, 1987). This bias occurs at two levels. The first level pertains to the rater's judgment regarding what specific IADL task constitutes an acceptable representation of the global IADL category. For one rater, the preparation of a full meal may mean roast and potatoes; whereas to another rater, the warming of a frozen TV dinner may be acceptable. The second level pertains to the rater's bias in judging the quality of the performance. Two raters simultaneously observing the same person prepare a meal may not agree on whether the performance or the outcome of meal preparation was acceptable. In both instances, the raters differ in severity; they do not agree because each has a different definition of the level of challenge of the task or of the difficulty of the test items that describe performance.

Evidence suggests, however, that while severity of scoring varies among raters, the degree of severity remains stable within individual raters (Lunz & Stahl, 1990). Moreover, rater training and standardized test administration procedures serve "only to direct the attention of the [raters], not to control the severity of their assessments" (Lunz & Stahl, 1990, p. 426). Therefore, because of our emphasis on performance evaluations, we must develop measures that enable us to account for differences in rater severity in the determination of the client's score.

Global Versus Discrete Skills
Approach to Functional Assessment

Standardized evaluation of ADL and IADL abilities, implemented to evaluate at the level of disability, is generally global and commonly emphasizes assessment of whether or not the person can perform independently a number of ADL and IADL tasks (e.g., eating, dressing, bathing, housekeeping, preparing meals) and, if not, what level of assistance is required. Such standardized global assessments provide an indication of what a person can or cannot do but no information regarding why the person might be having difficulty or whether limitations in performance are affecting the person's ability to assume desired roles.

Discrete skills evaluation, typically implemented to evaluate at the level of impairment, emphasizes thorough assessment of the distinct underlying constituents of ADL and IADL performance (e.g., strength, range of motion, perception, vision, mental status) (Fisher, 1992). The basic assumption made is that if the underlying cause of the ADL and IADL limitations can be identified and treated, the effects will generalize to improved functional performance across a wide range of ADL and IADL tasks. Although this approach has logical appeal, research has not demonstrated a sufficiently strong relationship between underlying constituents and ADL and IADL performance so as to enable us to make valid predictions about the abilities of a person in daily life task performance based on his or her discrete test scores (Bernspang, Asplund, Eriksson, & Fugl-Meyer, 1987; Jongbloed, Brighton, & Stacey, 1988; Pincus et al., 1989; Recd, Jagust, & Seab, 1989; Skurla, Rogers, & Swender, 1988; Teri, Borson, Kiyak, & Yamagishi, 1989).

It may be inappropriate to extrapolate performance on a specific task to an individual's abilities in daily life. Furthermore, performance tests may not give specific information on whether the identified limitations have any relevance to the actual activities or needs of the individual, or how well an individual with a limitation in a specific test item might have adapted to his or her individual environment. (Guralnik et al., 1989, p. M145)

Assessment Method: Self-Report Versus Performance Evaluation

Most standardized global assessments of ADL and IADL are of a self- or proxy-report interview format (Fillenbaum, 1988; Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963; Lawton, 1971). ADL and IADL are typically measured globally by 8 to 30 items administered in less than 30 min. Although the interview remains the foundation for diagnostic assessment, there is increasing concern about the frequently cited assumption that self- or proxy-report is necessary because direct observation is too time-consuming and of limited usefulness in the assessment of IADL abilities (cf. Harris, Jette, Campion, & Cleary, 1986; Spitzer, 1987). Moreover, there is increasing recognition that observational assessment of IADL may be preferred in many instances (Consensus Development Panel, 1988; Guralnik et al., 1989).

Occupational therapists are recognized for their expertise in performance evaluation (Guralnik et al., 1989) and for their ability to effect comprehensive task analyses that lead to the appropriate adaptive equipment or compensatory strategies that will enhance the client's functional ability (Faletti, 1984). Thus, we have the basic knowledge necessary to develop much-needed standardized performance assessments.

Sensitivity and Range of Scales

Recognition is growing of the need for more sensitive scales that can be ap-
plied across a wider range of ability levels. Scales defined by small increments of change are more sensitive; longer scales make it possible to measure change over a broader ability range (Spector, Katz, Murphy, & Fulton, 1987). Existing global scales, however, tend to be narrow in range, and the increments in challenge between global task categories are often large. As a result, the summary ratings obtained from global IADL assessments lack the sensitivity needed for the detection of small to modest improvements in IADL functioning (George & Fillenbaum, 1985). Furthermore, most existing self- or proxy-report scales are better able to “identify those at the disabled end of the spectrum but, within the large group of nondisabled, have less ability to distinguish high levels of functioning” (Guralnik et al., 1989, p. M145).

Ordinal Scores Versus Interval Measures: The Problem of Summing Rank Scores

Virtually all functional assessments yield ordinal raw scores. Even timed tests typically yield ordinal data. The person who can get dressed in half the time it takes another person to get dressed is not necessarily twice as able; dressing in half the time means only that the person is twice as fast (cf. Merbitz, Morris, & Grip, 1989; Wright & Linacre, 1989). By comparing dressing times or adding ordinal counts, we gain an understanding of the qualitative differences between people but learn little about their quantitative differences.

In contrast, measurement requires that we conceptualize ability as a magnitude or unidimensional continuum of greater or lesser ability and maintain equal-interval calibrations along the length of the scale. When this occurs, we have units of measure that are additive (Michell, 1999; Silverstein, Kilgore, & Fisher, 1989; Wright & Linacre, 1989).

To help clarify this problem of the addition of ordinal counts, one can consider a three-item test that examines feeding, dressing, and meal preparation. We might rate each item on a 4-point scale, as follows: independent (4), requires verbal assistance (3), requires physical assistance (2), and unable to perform (1). Not only might the distance between 2 and 3 be different for each of these items, but also, requiring assistance with feeding means something very different than requiring assistance with dressing or meal preparation. These tasks vary in difficulty, and the ease with which assistance is provided is distinctly different (e.g., helping someone cut meat is easier than preparing someone’s dinner). To add these ordinal counts and interpret them as being representative of equal interval, quantitative units is not meaningful.

This problem is exacerbated when we begin to add item scores from diverse dimensions of function. Imagine, for example, two clients, one who has an upper limb amputation, lives alone, and has not worked for the past 2 years and another who has sustained a spinal cord injury, lives with his wife, and is a university professor. The first client might obtain high scores on self-care and mobility but low scores on social functioning. The second client might obtain high scores on social functioning, moderate scores on self-care, and low scores on mobility. These two people could easily obtain equal total scores. How are we to determine which of these people is more able? Although it is not meaningful to sum scores that reflect different dimensions of function (e.g., social, physical), we do just that when we sum scores from multidimensional scales (cf. Merbitz et al., 1989; Silverstein et al., 1989; Wright & Linacre, 1989).

Knowledge of Task Hierarchy

Another important limitation of existing functional assessments is that the relative challenge of most ADL and IADL tasks is unknown. Knowledge of the relative difficulty of tasks can provide the therapist with a sequential guide for planning and grading intervention programs.

Many investigators have demonstrated a Gurman hierarchy among global ADL and IADL task categories (Fillenbaum, 1985; Katz et al., 1963; Lawton, 1971; Spector et al., 1987). For example, meal preparation is generally regarded as easier than heavy housework. Each global task category, however, comprises many specific tasks of varying challenge that overlap in difficulty with tasks that define other global categories. That is, although we can safely speculate that preparing a meal consisting of roast and potatoes is more challenging than warming a frozen TV dinner in the oven or than preparing a breakfast of cold cereal and juice, we currently can only guess at the relative challenge of these three meal preparation tasks. Moreover, we can only guess at whether they are easier or harder than the heavy housework task of vacuuming.

Criterion-referenced hierarchical scales can potentially serve as a guide to the progression of treatment and to target specific areas for intervention. They also have the potential to help us to determine if a client has the ADL and IADL performance ability to live independently or needs to live in a more supportive environment.

Conclusion

Existing standardized assessments pose many limitations, and no gold standard exists. Occupational therapists developing new assessments must heed these limitations and seek to overcome them. We must also remember that one test will not serve all purposes. When selecting existing tests for clinical or research purposes, we must be clear about what we will know given the results of a test, why we want to know it, and whether it is what we really want to know. We must then choose the best test for the job and clarify the relationship between the test results and the occupational behaviors and roles that the client needs and wants to perform.

Acknowledgments

This paper is based on a keynote address presented at the American Occupational Therapy Foundation minicourse entitled "Assessing Adults: Functional Measures and Successful Outcomes" at the 71st Annual Conference of the American Occupational Therapy Association, Cincinnati, June 1991. Preparation of this manuscript was supported, in part, by funding for the Occupational Therapy Center of Research and Measurement at the University of Illinois at Chicago, College of Associated Health Professions, from the American Occupational Therapy Foundation and the American Occupational Therapy Association.

References


