Effectiveness and Efficiency of Criterion Testing Versus Interviewing for Collecting Functional Assessment Information

Connie Klein-Parris, Thea Clermont-Michel, John O’Neill

Key Words: mental retardation • functional assessment • costs and cost analysis

This exploratory study investigated two factors believed to influence the efficiency (cost/time) and effectiveness (accuracy/validity) of functional assessment, client’s level of functioning, and complexity of the skill items. The results show that cost depends on the mode of assessment; criterion testing was found to be more than twice as expensive as interviewing the client’s caretakers. The effectiveness or accuracy of the assessment was found to be enhanced under two sets of circumstances: (a) when clients were functioning at a higher level and (b) when assessment items were less complex. The findings have implications for developing client assessments that are both efficient and effective.

Administrators, supervisors, and clinicians need to make informed decisions regarding the trade-offs between efficiency (cost/time) and effectiveness (accuracy/validity) of different functional assessment modes. To make these decisions, it would be useful to know the conditions or factors that affect efficiency and effectiveness. This study investigated two such factors: the client’s level of functioning and the complexity of test items.

Prior to 1959, intelligence tests were used to assess individuals with mental retardation. Among the tests used were the Bayley Scales of Infant Development, the Cattell Infant Intelligence Test, the Stanford Binet, and the Wechsler Adult Intelligence Test (1–3). Infant intelligence tests were used because testers assumed that the functioning level of severely and profoundly handicapped persons was similar to that of children (1). Tests standardized on children and used with an older population rendered the normative data useless, and only a rough developmental level could be derived (3). However, it soon became apparent that intelligence tests are not optimal assessments for the developmentally disabled. They do not delineate the client’s needs or motivation. In addition, they fail to look at the total person assessing that person’s physical abilities and limitations, verbal and nonverbal communication skills, and adaptive behavior and interpersonal skills (1, 2, 4). These tests were not given in the client’s natural environment and seldom provided information as to the process used by the therapist to induce the client to perform various tasks (5).

During the 1950s, the realization that environmental factors often contribute to mental retardation changed the way mental retardation was conceptualized (2). As a result of this change, the American Association on Mental Deficiency, in 1959, changed its definition of mental deficiency (2, 4). Intelligence quotient (IQ) alone was no longer the major determinant; a person had to have both a low IQ and an impairment in adaptive behavior (2). This change in definition led to the development of adaptive behavior measurements. Adaptive behavior refers to the individual’s ability to successfully accomplish tasks demanded by the environment; the individual must show appropriate social adjustment and accomplish societal roles and responsibilities (2, 4).

Two types of adaptive behavior scales were developed: an inventory scale that looks at skills and responsibilities (i.e., the Adaptive Behavior Scale) and another scale that measured adaptive behavior according to age norms (i.e., the Vineland Social Maturity Scale) (2). These tests functionally assess activities of daily living (ADL) and are useful in diagnosing and classifying. However, they have limited

Connie Klein-Parris, MA, OTR, is Coordinator, Occupational Therapy Services, Kennedy Child Study Center, New York, New York 10021.

Thea Clermont Michel, MA, OTR, is in private practice on Long Island, New York.

John O’Neill, PhD, is Coordinator, Rehabilitation Counseling Program, Hunter College, New York, New York.
use for developing programs and monitoring treatment because the individual's effectiveness in coping with the demands of community living is not assessed in sufficient detail (2, 6).

The trend toward deinstitutionalization increased the need for an effective and efficient assessment. This trend can be attributed to several factors. Both the nature-nurture debates of the 1930s and 1940s and sociologists' findings about the importance of the environment in the 1950s intensified the scrutiny of the large institutions (2). Articles appeared in the lay and professional press that exposed the appalling conditions in institutions (2). This led to the 1972 television broadcast on Willowbrook State School, which presented an emotional appeal for reform to the public (7). The 1962 report by the Kennedy administration's Panel on Mental Retardation, which advocated community care, further aided the move to deinstitutionalization (2, 8). The courts added a legal mandate that stressed the right of the mentally retarded to appropriate and effective treatment (2, 9, 10).

Normalization, the use of culturally valued means to enable people to live culturally valued lives, became the philosophy of care for the mentally retarded during the 1970s (2, 11-15). This led to changes in the treatment philosophy, as the focus shifted to integrating individuals in community environments. Society realized that the mentally retarded had been overly protected and should be allowed to assume the risks associated with everyday life (2, 13, 15).

The personal needs of individuals should be considered to ensure the effectiveness of the normalization and of the deinstitutionalization process (9). This is done by assessing the factors necessary for successful community integration (12, 15). A comprehensive assessment of the client can aid in coordinating the services provided for the client; it is also an important method of documenting the effectiveness of community programs and thereby meeting accountability demands (12, 9, 16-18).

Thus detailed functional assessments were developed to meet the needs of deinstitutionalization programs. One example is the adaptation of the rehabilitation indicators developed by New York University Medical Center Rusk Institute for Rehabilitation Medicine for the United Cerebral Palsy Associations of New York State (19-22). This adaptation is a criterion-referenced testing approach used to determine the actual competencies of the individual. The criterion testing approach is a nonnormative test that breaks down skills into component parts. It uses a hierarchy of prompts (graded from no assistance to verbal or physical assistance) to assess what an individual can do or learn. The approach identifies problem areas, evaluates progress, and offers an overall assessment of the strengths and weakness of the client (1, 23). It yields precise and specific information for the formation of educational objectives that can be used by paraprofessionals who provide many of the services to clients in community-based programs (i.e., ADL training). This use of paraprofessionals is critical if community programs are to function economically (24, 25). However, certain aspects of this testing approach have been criticized: it is both time-consuming and expensive. Lack of cooperation from clients also occasionally creates difficulties in getting accurate measurements (26).

Interviewing the client's caretaker is an alternative method for determining the client's performance on functional assessments. Interviewing is less time-consuming than criterion testing, making it less expensive. Verbal reports have generally been considered less reliable or valid than criterion testing; however, there are no definite data for concluding that parents or caretakers are unreliable informers (27). In fact, others suggest that parents or caretakers may be reliable reporters, particularly in regard to the self-care area (28).

Deinstitutionalization places increased demands on the community service system (29). Community programs typically are decentralized and require a more efficient use of professional time than centralized congregate living programs (i.e., institutions). Professionals responsible for conducting client evaluations in the community need to use time more efficiently because (a) their need to travel limits the number of clients they can see and (b) the clients, who are participating in other programs, are less available.

Performing functional assessments via criterion testing is highly desirable; yet it may prove inefficient in community programs. Interviewing staff about a client's performance on functional skills may be a more efficient method than criterion testing, but it may be less accurate (i.e., valid). The purpose of this study, therefore, was to test the relative efficiency and accuracy of performing functional assessments.

Method

Participants

Twenty-two mentally retarded adults (male and female) were randomly selected from a group of 63 mentally retarded individuals living in community residences in the New York City boroughs of Manhattan and Queens. These various residences are either apartments or group homes, and they are administered by the United Cerebral Palsy Associations of...
Table 1
Self-Care Skill Areas and Randomly Selected Skill Items

<table>
<thead>
<tr>
<th>Serving Meals and Washing Dishes</th>
<th>Setting table</th>
<th>Cleaning table</th>
<th>Cleaning cooking and eating areas</th>
<th>Washing dishes</th>
<th>Washing glasses</th>
<th>Preparing Meals</th>
<th>Opening cans</th>
<th>Turning on water faucet in kitchen</th>
<th>Turning off water faucet in kitchen</th>
<th>Opening jars</th>
<th>Obtaining ingredients from storage</th>
<th>Eating and Drinking</th>
<th>Spreading food with a knife</th>
<th>Cutting food with a knife</th>
<th>Chewing food well before swallowing</th>
<th>Drinking from a glass</th>
</tr>
</thead>
</table>

New York State. The average age of the 22 participants was 34 years (SD = 9.66), and their mean IQ was 35.27 (SD = 15.34).

Procedure

For each participant, information was gathered regarding skill performance using the skill indicators of the Rehabilitation Indicators Project (22). For the purposes of this study, three skill areas were selected from the self-care domain, and one third of the skill areas (see Table 1) were chosen because (a) they included skills that are necessary parts of residents' daily routine and (b) the staff would be familiar with the participants' level of performance.

Two methods were used to assess each participant's skills: criterion testing and interviewing. Criterion testing involved having the participant perform the skill. All materials necessary for performing a skill were present before a participant was verbally requested to perform the skill. If the participant did not perform the skill within 15 seconds, a verbal and gestural prompt was provided, and increasing levels of prompting (see Table 2) were administered until the participant completed the skill.

The questionnaire used in the interview was designed to obtain the same type of information as was obtained from criterion testing (i.e., level of prompting required to complete a skill). The program counselor, who is responsible for the residents' programs, is asked to describe how a participant performed each skill. From this description the evaluators determined the level of prompting typically required to complete a skill. If the program counselor provided insufficient information on the first request, further probing was used by the evaluator to obtain a step-by-step description of a skill performance.

Two master's level occupational therapists served as evaluators. At no time did an evaluator perform both types of assessment on the same participant. Each evaluator was ignorant of the other's assessment results until both methods had been completed.

Analysis

Two dependent variables were generated for each participant. A discrepancy score was obtained by subtracting the prompting levels obtained on interviewing and criterion testing and taking the absolute value (i.e., ignoring the number's sign). The discrepancy score measures the accuracy of interviewing relative to criterion testing, which is assumed to be the most accurate functional assessment method.

The second dependent variable was the cost of interviewing and criterion testing. The overall cost of administering these two methods reflects their efficiency. Cost was determined by first calculating how long it took to administer each assessment method. Length of time for administration was then multiplied by the appropriate hourly wage rates for all staff members involved, and these costs were added to arrive at the overall cost of an assessment method.

Two factors related to the accuracy of interviewing were evaluated: level of IQ and complexity of

Table 2
Prompting Levels

<table>
<thead>
<tr>
<th>Prompting Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Fully independent—Initiates and completes activity when material is available and/or setting is appropriate.</td>
</tr>
<tr>
<td>2. Requires one gestural and/or verbal prompt—one verbal and/or gestural request, presented without undue emphasis, is necessary for client to complete the task.</td>
</tr>
<tr>
<td>3. Requires demonstration—a single demonstration of appropriate behavior is necessary for client to complete the task.</td>
</tr>
<tr>
<td>4. Requires repeated gestural or verbal prompts—More than one verbal and/or gestural repetition of a request is necessary for client to complete task.</td>
</tr>
<tr>
<td>5. Requires repeated demonstration—More than one demonstration of appropriate behaviors is necessary for client to complete task.</td>
</tr>
<tr>
<td>6. Requires partial physical prompt—Some level of physical guidance is necessary for client to initiate or complete the task.</td>
</tr>
<tr>
<td>7. Requires full physical prompt—Continuous full physical guidance is necessary for client to complete the task.</td>
</tr>
<tr>
<td>8. Fully dependent—Client is fully dependent on others to complete the task and cannot be guided through steps.</td>
</tr>
</tbody>
</table>

Table 3
Split-Plot Analysis of Variance for Cost With Two Within (type of assessment) and Two Between (IQ level) Factors

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subjects</td>
<td>7.77</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IQ</td>
<td>4.6</td>
<td>1</td>
<td>4.6</td>
<td>1.24</td>
</tr>
<tr>
<td>Subject within groups</td>
<td>7.31</td>
<td>20</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>Within subjects</td>
<td>25.92</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of assessment</td>
<td>20.22</td>
<td>1</td>
<td>20.22</td>
<td>77.77*</td>
</tr>
<tr>
<td>IQ x type of assessment</td>
<td>0.54</td>
<td>1</td>
<td>0.54</td>
<td>2.10</td>
</tr>
<tr>
<td>Type of assessment x subject within groups</td>
<td>5.16</td>
<td>20</td>
<td>0.26</td>
<td></td>
</tr>
</tbody>
</table>

*p < .01
The Cost of Performing Functional Skill Assessments as a Function of Client IQ Level (high, low) and Method of Assessment (criterion testing, interviewing)

A median split was performed creating a low IQ group (30 participants) and a high IQ group (31 participants). In addition, the authors dichotomously classified the 18 skill items as either difficult or easy with 88% agreement. The effects of these two factors on the accuracy of interviewing were evaluated via a split-plot factorial 2.2 analysis of variance (30).

Two factors were also evaluated in terms of their effects on efficiency: level of IQ (low, high) and type of assessment method (criterion testing, interviewing). As with accuracy, a split-plot factorial 2.2 analysis of variance was used to evaluate the effects of these two factors on efficiency.

Results

In terms of efficiency, the results (see Table 3) indicated a significant main effect for mode of functional assessment, but not for client IQ level. In addition, there were no significant interaction effects between IQ level and mode of assessment. These results show that cost depends on the mode of functional assessment, but not the client's level of functioning (see Figure 1). Criterion testing the clients on each functional skill item was more than twice as expensive as interviewing the direct care staff regarding the client's functional skills. This finding is even more poignant considering that interviewing involved the time of two staff members (a professional and direct care staff member) whereas criterion testing required only one staff member (a professional).

However, it appears that interviewing provided considerably less accurate information than criterion testing. There was an average discrepancy of 2.3 rating levels between interviewing and criterion testing on 30% of the items tested. The range of inaccurately assessed items was from 5% to 72%, and the rating level discrepancies ranged from 1 to 7. This variability in accuracy indicated that there might be client and/ or instrument factors affecting the level of discrepancy between interviewing and criterion testing. The results (see Table 4) showed significant main effects for client IQ level and complexity of test item. However, there was no significant interaction effect between IQ level and item complexity. Thus, effectiveness or accuracy of assessment is enhanced under two sets of circumstances: when the clients are functioning at a higher level and when the assessment items are less complex (see Figure 2).

Discussion

Two factors, level of IQ and type of assessment method (criterion testing, interviewing), were evaluated. Table 4

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subjects</td>
<td>13.25</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IQ</td>
<td>3.33</td>
<td>1</td>
<td>3.33</td>
<td>6.79*</td>
</tr>
<tr>
<td>Subject within groups</td>
<td>9.92</td>
<td>20</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>Within subjects</td>
<td>16.86</td>
<td>22</td>
<td></td>
<td>1.77</td>
</tr>
<tr>
<td>Complexity</td>
<td>4.96</td>
<td>1</td>
<td>4.96</td>
<td>8.41*</td>
</tr>
<tr>
<td>IQ x complexity</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>.20</td>
</tr>
<tr>
<td>Complexity x subject within groups</td>
<td>11.78</td>
<td>20</td>
<td>.59</td>
<td></td>
</tr>
</tbody>
</table>

Table 4
Split-Plot Analysis of Variance for Discrepancy With Two Within (item complexity) and Two Between (IQ level) Factors

Figure 2: The Discrepancy Between Two Methods (criterion testing, interviewing) of Functional Skill Assessment as a Function of Item Complexity (simple, complex) and Client IQ Level (high, low)
Hypothesized Model for Choosing Assessment Mode Based on Client's IQ Level

<table>
<thead>
<tr>
<th>Item Complexity</th>
<th>Mode of Assessment</th>
<th>Item Complexity</th>
<th>Mode of Assessment</th>
<th>Mode of Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>Interview</td>
<td>Simple</td>
<td>Interview</td>
<td>Criterion test</td>
</tr>
<tr>
<td>Moderate</td>
<td>Interview</td>
<td>Moderate</td>
<td>Interview</td>
<td>Criterion test</td>
</tr>
<tr>
<td>Complex</td>
<td>Criterion test</td>
<td>Complex</td>
<td>Criterion test</td>
<td>Interview</td>
</tr>
</tbody>
</table>

at in terms of their effects on efficiency. Irrespective of the level of IQ, criterion testing was found to be less cost-effective than interviewing. The realization that criterion testing is more accurate than interviewing puts practitioners in the difficult situation of having to choose between an accurate assessment or an efficient assessment. A practitioner would then need to know when to use the most efficient assessment method (interviewing) without jeopardizing accuracy. The dilemma could be partially resolved by understanding the factors that influence accuracy.

This study showed that two factors influence the accuracy of interviewing: IQ level and complexity of the question. The accuracy of interviewing is greater for clients with a higher IQ than for clients with a lower IQ. Additionally, interviewing is more accurate for simple items than complex items. The results suggest that clients with a high IQ can be interviewed on simple items without sacrificing accuracy. For clients with a low IQ, it is preferable to criterion test on difficult items and to interview on simple items.

The results of this study begin to outline some alternatives for obtaining accurate assessment information that is efficient and cost-effective. The functional assessment of clients could be divided into two forms. One form would contain simple items that should be used to interview caretakers of clients with a high IQ. The other form would contain difficult items to criterion test clients with a low IQ. In this way both accuracy and cost-effectiveness could be achieved. The results of this study do not offer any definitive guidelines on the method of testing to be used for clients with a low IQ on simple items and for clients with a high IQ on complex items. However, we hypothesize that a more adequate model for choosing an assessment method would have evolved (see Figure 3) had the study included a wider range of clients, in terms of their IQs, and a wider range of questions. By using such a model, a practitioner would have a set of procedures for determining whether to interview or criterion testing, depending on the level of functioning of the client and the complexity of the question.

There are probably other factors, in addition to IQ level and item complexity, that may affect the relative accuracy of interviewing and criterion testing. These include the client's or interviewee's mood, his or her response to the interviewer or tester, and disruptions in the environment or routine.

Several weaknesses are apparent in this research. First, there is always tension between the results of research and their clinical applications. In this study, the assessment process had to be standardized for criterion testing, and this may interfere with the generalization of results to the typical assessment situation. Other weaknesses were that (a) the clients selected for each group did not represent a wide enough range of IQ levels and (b) the skill items selected did not adequately represent a broad enough range from simple to complex. In future studies, a systematic a priori selection of clients should be made to elicit the full range of functioning (low to high IQs). Criteria should also be determined to select an adequate range of skills from simple to complex. This could include such factors as the number of steps necessary to complete the skill, the developmental level of prehension used to perform the skill, and/or the use of bilateral hand movements. A wider range of questions from the other ADL areas, and from gross and fine motor areas could be selected.

Future research could also look at the validity and reliability of the self-interview form. The self-interview form may greatly increase the cost-effectiveness of functional assessment. Various situational factors, which probably enhance the validity of the self-interview, could be studied: paraprofessionals' motivations, the number of clients for whom paraprofessionals are responsible, and their level of education.

References