Comparison of Functional Status Tools Used in Post-Acute Care

Alan M. Jette, Ph.D., Stephen M. Haley, Ph.D., and Pengsheng Ni, M.P.H, M.D.

There is a growing health policy mandate for comprehensive monitoring of functional outcomes across post-acute care (PAC) settings. This article presents an empirical comparison of four functional outcome instruments used in PAC with respect to their content, breadth of coverage, and measurement precision. Results illustrate limitations in the range of content, breadth of coverage, and measurement precision in each outcome instrument. None appears well-equipped to meet the challenge of monitoring quality and functional outcomes across settings where PAC is provided. Limitations in existing assessment methodology has stimulated the development of more comprehensive outcome assessment systems specifically for monitoring the quality of services provided to PAC patients.

INTRODUCTION

A fundamental barrier to fulfilling the emerging health policy mandate in the United States for monitoring the quality and outcomes of PAC is the absence of standardized, patient-centered outcome data that can provide policy officials and managers with outcome data across different diagnostic categories, over time, and across different settings where PAC services are provided (Wilkerson and Johnston, 1997). Recently, the National Committee on Vital and Health Statistics (NCVHS) (2002) made recommendations on the potential for standardizing data collection and reporting for the purposes of quality assurance as well as for setting future research and health policy priorities in the U.S. The NCVHS (2002) was unanimous in stressing two major goals: "…to put functional status solidly on the radar screens of those responsible for health information policy, and to begin laying the groundwork for greater use of functional status information in and beyond clinical care…" The NCVHS project used the term functional status very broadly to cover both the individual's ability to carry out activities of daily living (ADLs) and the individual's participation in various life situations and society.

Within PAC, functional outcome instruments have been developed and are widely used for various applications and for use in specific settings. Examples include the functional independence measure (FIM™) for acute medical rehabilitation (Guide for the Uniform Data Set for Medical Rehabilitation, 1997; Hamilton, Granger, and Sherwin, 1987), the minimum data set (MDS) for skilled nursing and subacute rehabilitation programs (Morris, Murphy, and Nonemaker, 1995), the Outcome and Assessment Information Set for Home Health Care (OASIS) (Shaughnessy, Crisler, and Schlenker, 1997) and the Short Form-36 (SF-36) for ambulatory care programs (Ware and Kosiniski, 2001). If one looks carefully at the content of these

The authors are with Boston University. The research in this article was supported by the National Institute on Aging under Grant Number 5P50AG11669 and the National Institute on Disability and Rehabilitation Research under Grant Number H133B990005. The views expressed in this article are those of the authors and do not necessarily reflect the views of Boston University, the National Institute on Aging, the National Institute on Disability and Rehabilitation, or the Centers for Medicare & Medicaid Services.

instruments, it becomes apparent that substantial variations exist in item definitions, scoring, metrics, and content coverage, resulting in fragmentation in outcome data available for use across different PAC settings (Haley and Langmuir, 2000). Differences in conceptual frameworks used to construct each instrument, the inability to translate scores from one instrument to another, and the lack of outcome coverage and precision to detect meaningful functional changes across settings, severely limit the field's ability to measure and analyze recovery through the period of PAC service provision. If the PAC field is to achieve the goal of comprehensive functional outcome assessment and quality monitoring for different patient diagnostic groups across different PAC settings, efforts are needed to develop functional outcome assessments that are applicable across a continuum of post-acute services and settings.

To our knowledge, no studies exist that have directly compared the content and operating characteristics of functional outcome instruments commonly used in PAC to examine their relative merits for monitoring outcomes across care settings. In this article, therefore, we report the results of a direct empirical comparison of the FIM™, OASIS, MDS, and the physical function scale (PF-10) of the SF-36, focusing on three aspects of each: instrument content, range of coverage, and measurement precision. The objective of this comparative analysis is to evaluate the commonly held assumption that there exist fundamental deficiencies in the current armamentarium of setting-specific outcome instruments that prevents their applicability for more comprehensive patient-centered functional outcome assessment across diagnoses, over time, and across different settings where PAC is provided. In response to identified deficiencies in existing instruments, we also discuss the potential utility of contemporary measurement techniques, such as item response theory (IRT) methods and computerized adaptive technology (CAT), to yield functional outcome instruments better suited for outcome monitoring across PAC settings.

METHODS

Subjects

These analyses use data from a sample of 485 PAC patients drawn from six health provider networks in the greater Boston area. All patients enrolled in the study were recruited by study coordinators within their own health care facility and completed informed consent prior to participating in the study. Patients were recruited from inpatient (199 from acute inpatient rehabilitation and 90 from transitional care units); and community settings (90 from ambulatory services; and 106 from home care). Eligibility criteria included: (1) adults, age 18 or over, (2) recipients of skilled rehabilitation services (physical, occupational, or speech therapy), and (3) English speaking. The sample was stratified by impairment group to include approximately an equal number of subjects within three major patient groups: (1) 33.2 percent neurological (e.g., stroke, multiple sclerosis, Parkinson's disease, brain injury, spinal cord injury, neuropathy); (2) 28.4 percent musculoskeletal (e.g., fractures, joint replacements, orthopedic surgery, joint or muscular pain); and (3) 38.4 percent medically complex (e.g., debility resulting from illness, cardiopulmonary conditions, or post-surgical recovery). To assure good representation of levels of functional severity, the sampling design was stratified to yield a wide distribution of subjects representing three distinct severity levels: slight (35.9 percent), moderate

(44.1 percent), and severe (20.0 percent) based on scores from an adapted modified Rankin scale (vanSwieten et al., 1988).

The sample reflects the racial and ethnic distribution of the greater Boston metropolitan population. The sample contained a wide age range (19-100 years; mean age = 62.7 years.) The majority of the subjects were female (58.8 percent), white (81.6 percent), and non-married (61.1 percent). More than one-half (51.3 percent) had beyond a high school education. The wide range of onset from initial injury/ illness (2.0-3.9 years) characterizes different stages of recovery within both inpatient and community settings. The SF-8 Health Survey (Ware et al., 1999) data indicated that physical functioning of the overall sample (mean=40.3, standard deviation (s.d.)=9.9) was below the U.S. population norms (mean=50, s.d.=10), although mental functioning (mean=50.2, s.d.=10.3) was consistent with U.S. population norms (mean=50, s.d.=10).

Data Collection Procedures

Our overall data collection strategy was to assess items from existing functional outcome tools used in PAC so that they could be combined for analysis into one common scale. Due to practical data collection considerations and potentially high patient response burden, we did not administer items from all four instruments to all 485 subjects. Rather, we linked items together by administering to the entire sample a core set of 58 activity items applicable for patients in both home and community service settings. This method has been referred to as a common-item test equating design, in which a core set of items serve as a scale anchor from which unbiased parameters can be estimated on items with missing data (McHorney and Cohen, 2000).

We collected activity items, when available, from standardized instruments administered and recorded in the medical record. These included the 18-item FIM™ for persons in inpatient rehabilitation settings (*N*=108), 19 MDS items (physical functioning and selected cognitive items) for persons in skilled nursing home settings (*N*=91), and 19 ADL/individual ADL items from the OASIS or persons receiving home care (*N*=103). Since there were no consistent data available from charts in the outpatient setting, we administered the 10 physical functioning items of the SF-36 to individuals receiving outpatient services (*N*=82).

We applied specific rules for handling missing item data within two of the standardized instruments. In accordance with FIM™ scoring rules, items that were not administered are scored as "total dependence." The MDS items have a response option of "activity did not occur." We converted these codes to the lowest score for that item, namely "total dependence." We reasoned that the most likely explanation for the activity not occurring was that the item could not be performed, an assumption that others have made when comparing functional instruments. (Buchanan et al., 2002). Missing data for the PF-10 and OASIS were not systematically recoded.

The 58 core activity items collected on all 485 subjects were used as scale anchors. In order to compare items from instruments across PAC settings, a common link was needed to provide a stable functional base for comparison purposes. The common-item test equating design was used so that every subject had a similar core set of items. Rasch (1980) models (Wright and Masters, 1982) can be conducted with missing data if a core set of items is used to link the setting-specific instruments into a common scale. Core items included: 15 physical functioning, 14 self-care, 12 daily routine, 11 communication, and 6 interpersonal interaction items. Activity questions asked, "How much difficulty do you currently have (without help from another person or device) with the following activities…?" A polytomous response choice included "not at all," "a little," "somewhat," "a lot," and "can not do." We framed the activity questions in a general fashion without specific attribution to health, medical conditions, or disabling factors. For some individuals in the community settings, we also collected 52 additional functional items (*N*=196) and added those to the common linking solution, however the results of these items are not reported here. None of the subjects had difficulty in completing the core set of 58 items.

Data on items from three of the existing instruments (e.g., FIM™, OASIS, and MDS) were recorded from retrospective chart review. Data on the PF-10 items (outpatient programs) and core and community items were collected via subject or proxy interviews. Each interview (approximately 45-60 minutes) was conducted in a quiet atmosphere in an inpatient setting, an outpatient facility, or participant's home. The ability of a subject to take part in the interview self-report was assessed by a treating clinician who determined if the respondent could: (1) understand the interview questions, (2) sustain attention during an interview, and (3) reliably respond to the questions. If the answer to any of these screening questions was no, then the interview was completed by either a clinician or family member proxy participant. A proxy participant completed less than 3 percent of the interviews.

Interviews that included information on core items were timed to coincide with administration of standard outcome instruments. For example, FIM™ is administered in most facilities within 3 days of admission and prior to discharge in the inpatient rehabilitation setting. Thus, the subject interview was arranged to take place within 3 days of admission or discharge such that FIM™ information was collected close to the time of the interview. Likewise, interviews of subjects in transitional care units were scheduled to coincide with the MDS assessment. Subjects receiving home care therapy were interviewed either near to admission or discharge such that the OASIS assessment was performed in close proximity to the subject interview. The interview was fully scripted, with standard instructions and an answer card to help subjects identify the desired response choice. The order of presentation of test sections was randomly assigned to mitigate the possibility of large portions of missing data in any one section due to respondent fatigue.

Interviewers, who were experienced rehabilitation clinicians (mean years of clinical experience=5.7 years), received training and quality assurance from: (1) an initial 3-hour training session, (2) a protocol manual, (3) supervision by the research project staff on all first interviews, and (4) acceptable completion of a procedural checklist and inter-rater reliability on a subsequent interview. A project staff member accompanied interviewers on approximately every 10th interview to check on adherence to study protocol and to assure overall quality control. All subjects, regardless of setting, were administered the SF-8 items to establish a normative description of the sample.

Analyses

Analyses were conducted in three stages. First, we conducted one-parameter Rasch (1980) partial credit analyses for the entire item pool (instruments and core/community items) to develop an overall functional ability scale (Wright and

Masters, 1982). The Rasch partial credit model allows for different number of response categories across items. For example, the FIM™ scale has a seven-point rating scale, while the OASIS incorporates multiple response categories that differ per item. This overall scale was used to estimate parameters for all items in each of the four comparison instruments. We used a method of concurrent calibration, which involves estimating item and ability parameters in the overall subject and item pool simultaneously. This treats items not taken by subjects as missing, as the Rasch program uses information available to estimate person and item parameters. Our goal was to develop a general scale for functional abilities, and specifically to see the location of item content across the four existing instruments. We calculated internal consistency estimates—Cronbach alpha (1951)—for each of the individual instruments and for the overall functional scale to determine the levels of consistency of the items within the overall functional scale. We also calculated goodness of fit tests for all items with the overall functional scale using the standardized infit statistic (+/-2.0) to determine the number of items with poor fit within the overall scale.

Second, we examined the relative amount and range of activity content covered by each of the four existing instruments and made comparisons across instruments. To do this, we used the extreme (highest and lowest) item threshold estimate for each instrument. A threshold represents the point of separation between adjacent item response categories for each item. The Rasch partial credit model estimates thresholds for each item, and the minimum and maximum threshold values per instrument were used to establish the range of content within the overall functional ability scale.

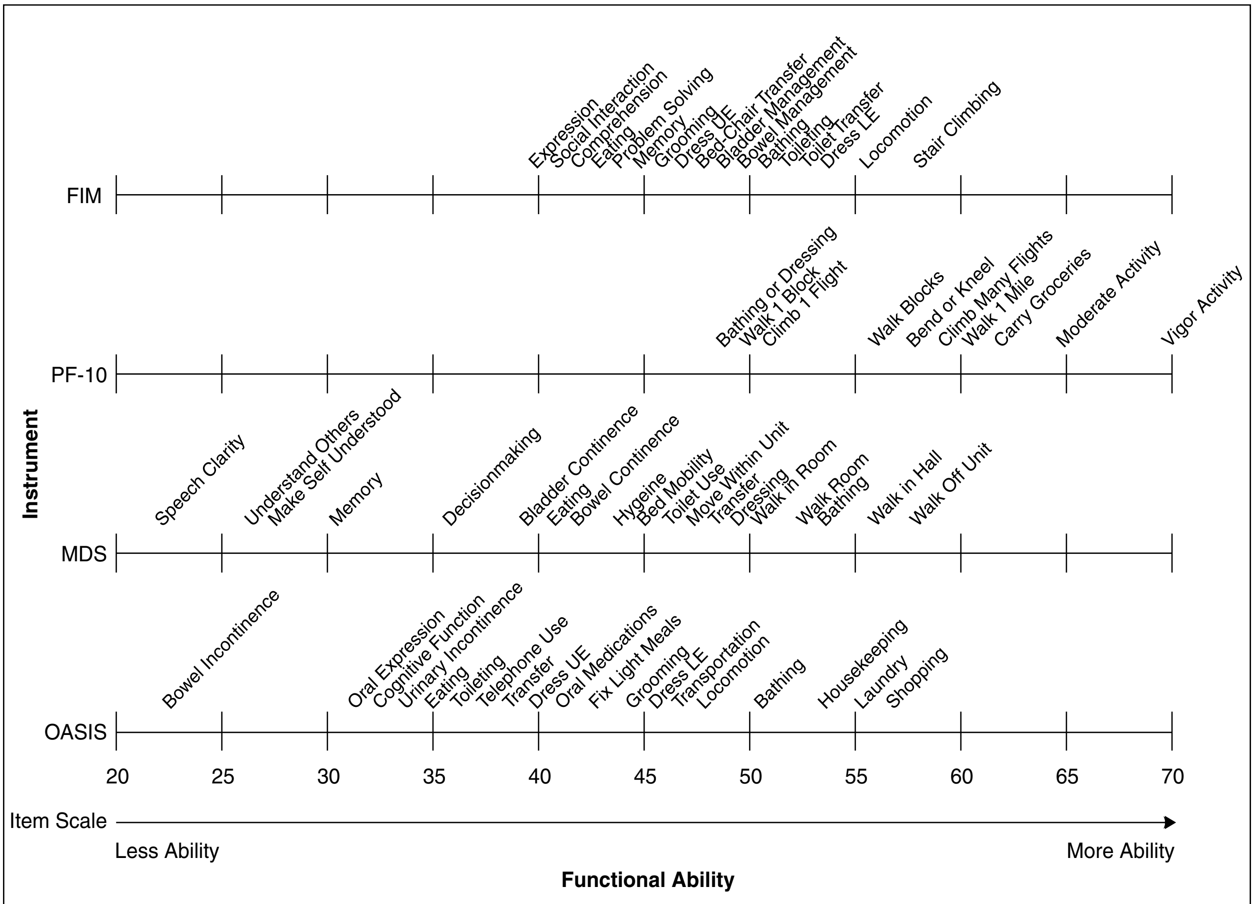
Third, we calculated item and test information functions (Lord, 1980; Dodd and Koch, 1987; Murnki, 1993) to estimate the location of optimal measurement precision of each instrument. The item information function is an index of the degree and location of information that a particular functional item provides for estimating a score along a functional ability scale. Item information functions are related to the location and shape of the item characteristic curve (ICC), which describe the probabilities of responding to particular response options of an item. The steeper the slope of the ICC, the more information about functional ability provided by an item in the scale, thus the greater level of item discrimination and precision associated with estimates of the score at that point along the scale. ICCs that have a broad range of coverage along the scale provide information functions across a wide range of the scale. Therefore, the information function is the relationship of the amount of information of an item at a particular scale level and is described by the ratio of the slope of the ICC and the expected measurement error. Test information functions were calculated by summing the item information functions to obtain an estimate of the measurement precision of the entire test at different levels of the functional ability continuum. We compared the test information function with the ability levels of the sample in each major setting (inpatient, community). We calculated a summary score converted to 0-100 metric for each person based on the overall item pool.

RESULTS

The internal consistency values of the four functional ability instruments were as follows: MDS=0.97, OASIS=0.99, FIM™=0.99, and the PF-10=0.99. The internal

Figure 1

Rasch Model Comparison of the Four Post-Acute Care Instruments



NOTES: Locations are derived as estimates of the average item difficulty parameter generated for each instrument. FIM™ is functional independent measure. PF-10 is physical function index (10 items). MDS is minimum data set. OASIS is Standardized Outcome and Assessment Information Set for Home Health. UE is upper extremity. LE is lower extremity.

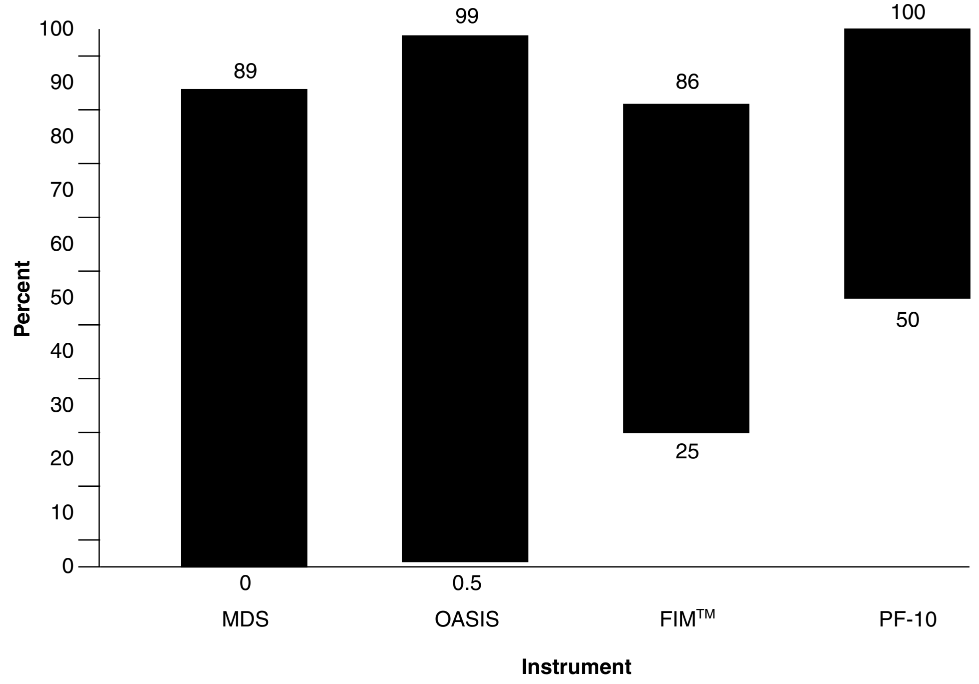
SOURCES: (Hamilton, Granger, and Sherwin, 1987; Morris, Murphy, and Nonemaker, 1995; Ware and Kosinski, 2001; Shaughnessy, Crisler, and Schlenker, 1997.)

consistency of the items within the full functional ability scale was 0.85. Only five of the items within the existing instruments (7.2 percent) exceeded the goodness of fit values. Thus, we felt it was acceptable to combine the items from each of the four functional outcome instruments into an overall functional ability scale for the purposes of directly comparing their range of functional content, breadth of coverage, and measurement precision.

Figure 1 illustrates and compares the location of items in each of the four functional outcome instruments along the broad content dimension of functional ability. These locations are derived as estimates of the average functional ability parameter generated for items in each instrument included in the analysis. Across all four instruments it can be seen that cognitive, communication, and bowel and bladder continence function items achieved the lowest functional ability estimates, indicating that those items were usually less difficult for persons in the sample to perform compared with other items contained in these instruments. In general, the PF-10 scale contained items with the highest item

Figure 2

Comparison of Actual Ranges Covered by the Four Post-Acute Care Instruments



NOTES: Content coverage is calculated by the high and low step estimates for each item's response category. MDS is minimum data set. OASIS is Standardized Outcome and Assessment Information Set for Home Health. FIM™ is functional independent measure. PF-10 is physical function index (10 items).

SOURCES: (Hamilton, Granger, and Sherwin, 1987; Morris, Murphy, and Nonemaker, 1995; Ware and Kosinski, 2001; Shaughnessy, Crisler, and Schlenker, 1997.)

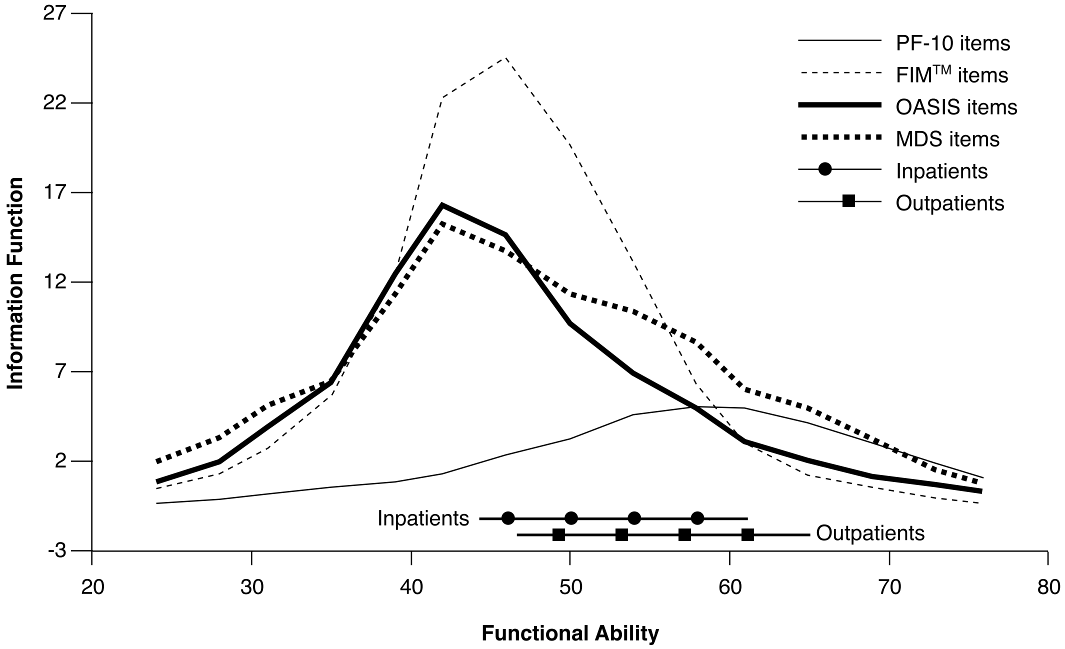
functional ability calibrations compared with the other three. A substantial number of items in the FIM™, OASIS, and MDS instruments achieved average functional ability calibrations in the mid-point of the functional ability continuum. Figure 1 also reveals the substantial overlap of item content across these four instruments.

The actual range of functional ability that is covered by the items contained in each of the four functional outcome instruments is presented in Figure 2. The content coverage is calculated by the high and low step estimates for each item's response categories instead of average estimates, which were the basis of the information presented in Figure 1. Consistent with the general spread of the functional ability parameters illustrated in Figure 1, the range of coverage shown in Figure 2 appears greatest for both the MDS and the OASIS instruments as compared with either the FIM™ or PF-10 scales. Because of the high step estimate for one of the transportation items in the OASIS, the actual range of coverage of the OASIS is nearly the entire range of all four instruments.

Figure 3 displays the measurement precision of each instrument as depicted by the test information function curves. These curves are superimposed with the average score on the functional ability continuum and ±2 standard deviation for the inpatients and community samples. Note the location of the peak amount of precision for each instrument in relationship to

Figure 3

Measurement Precision of Post-Acute Care Instruments, by Test



NOTES: Curves are superimposed with the average score on the functional continuum and ± 2 standard deviation for inpatients and community samples. MDS is minimum data set. OASIS is Standardized Outcome and Assessment Information Set for Home Health. FIM™ is functional independent measure. PF-10 is physical function index (10 items).

SOURCES: (Hamilton, Granger, and Sherwin, 1987; Morris, Murphy, and Nonemaker, 1995; Ware and Kosinski, 2001; Shaughnessy, Crisler, and Schlenker, 1997.)

each sample. Although the OASIS items contain a broad range of content, as was seen in Figure 2, the OASIS items provide a high degree of measurement precision at only the very low end of the functional ability dimension for both the inpatient and community samples. The precision of MDS items is also greatest at the lower functional ability dimension levels, although the MDS items have a greater span of functional ability in which they provide some levels of precision. The FIM™ items peak at the low to moderate end of the inpatient sample. In contrast, the information function of the PF-10 items peaks at the high end of the community outpatient sample with very poor precision for the inpatient sample.

DISCUSSION

The results of these analyses of instrument content, coverage, and measurement precision provide direct evidence of what many have argued are the major limitations of existing functional outcome instruments currently in use within PAC. While each of the four instruments compared in this analysis appears well suited for its primary application, none of them appears well-equipped for the current policy mandate for monitoring the quality and outcome of PAC provided over time and across different PAC settings.

If one looks at the results for the FIM™, the most widely used outcome instrument in PAC, one can see that the FIM™ items cover a very small portion of the functional ability continuum within a narrow range of functional content. The FIM™ is most precise and relevant for PAC inpatients whose function is at the low end of the continuum. All of these characteristics of the FIM™ are acceptable when one considers its primary application is for evaluation of inpatient rehabilitation services. These FIM™ characteristics become severe limitations, however, if the intended application is assessment of outcomes or quality across PAC settings. An instrument such as the PF-10 suffers from a similar type of limitation, as does the FIM™. The PF-10 covers a narrow range of functional content, although, unlike the FIM™, the content covered by the PF-10 is at the higher end of the functional activity continuum. The PF-10 appears most precise for community dwelling outpatients, but much less so for inpatients such as those seen in many rehabilitation facilities. When used with high functioning community patients, the PF-10 covers appropriate content; its content and range is severely limited in application to patients within institutions. The MDS and OASIS instruments, in comparison with the FIM™ and PF-10, cover content from the mid portion of the functional ability continuum with less content covering the low or high ends. Patients functioning at the very high or very low end of the functional continuum would not be as well served by the OASIS and MDS.

Concerns about existing instruments used in PAC, such as the four examined in this analysis, have stimulated the development of more comprehensive functional outcome instruments developed specifically for application across diagnostic groups, and across PAC settings. An example of this type of work is found in the activity measure for PAC (AM-PAC), developed by the Rehabilitation Research & Training Center for Outcomes based at Boston University (Haley and Jette, 2000). In constructing the AM-PAC, its developers used the strategy of combining functional ability items found in existing instruments and from a variety of other sources into quantitative scales that can be employed to assess a wide range of functional content needed to assess quality and outcomes of patients seen across PAC settings.

Although instruments such as the AM-PAC are promising, the continued use of traditional, fixed-form measurement methodology for constructing functional outcome instruments presents the researcher with two common problems that limit their utility in clinical outcomes assessment. One fundamental problem encountered with fixed-form instruments of modest length is floor and ceiling effects where large numbers of individuals who complete these outcome instruments score at either the top or the bottom of the range of possible scores. These ceiling and floor effects severely reduce measurement precision and thus, restrict the utility of the instruments (Andresen et al., 1999; Brunet et al., 1996; DiFabio et al., 1997; Rubenstein et al., 1998). In response to concerns over inadequate measurement precision and inadequate coverage of important outcome domains, some researchers develop more comprehensive and lengthy outcome instruments. Lengthy instruments lead to the frustration and fatigue faced by many subjects and busy clinicians overwhelmed by large and burdensome batteries of instruments (Meyers, 1999).

One promising solution offered to measurement problems faced by traditional fixed form instruments is offered through the combined application of IRT methodology and CAT techniques (Ware et al., 1999; Ware, Bjorner, and Kosinski, 2000; Weiss, 1982; Hambleton, 2000). These techniques of test construction are currently being applied to the development of a new generation of functional assessment instruments designed for use in PAC settings (Haley and Jette, 2000). CAT methodology uses a computer interface for the patient (or a computerized interview/clinician report) that is tailored to the unique functional ability level of the patient. The basic notion of an adapted test is to mimic what an experienced clinician would do. A clinician learns most when he/she directs questions at the patient's approximate level of proficiency. Administering outcome items that are either too easy or too hard provides little information. An adaptive test first asks questions in the middle of the ability range, and then directs questions to the level based on the patient's responses, without asking unnecessary questions. This allows for fewer items to be administered while gaining precise information regarding an individual's placement along a continuum of functional ability. CAT applications require a large set of items in any one functional area (item pools), items that consistently scale along a dimension of low to high functional ability, and rules guiding starting, stopping, and scoring procedures. Large item sets such as the AM-PAC can be readily adapted to a CAT format in future work.

Methods like IRT that make it possible to calibrate questionnaire items on a standard metric (ruler) also yield the algorithms necessary to run the engine that powers CAT assessments. These statistical models estimate how likely a person at each level of function is to choose each response to each survey question. This logic is reversed to estimate the probability of each health score from a particular pattern of item responses. The resulting likelihood function makes it possible to estimate each person's score, along with a person-specific confidence interval. In principle, one can derive an unbiased estimate of an outcome, i.e., an estimate without systematic error, from any subset of items that fits the model. The number of items administered can be increased to achieve the desired level of precision. Most statistical models for estimating such item parameters can be traced to the work of Rasch (1980) or on a second tradition—IRT (Hambleton and Swaminathan, 1985). Both models assume unidimensionality; i.e., that the items included on a particular scale measure only one concept. Whether these techniques, if applied to functional outcome assessment, will solve the problems presented by traditional, fixed form methodology, needs to be carefully evaluated in future research.

There are several limitations in the analyses reprinted in this article. To achieve a direct comparison of these four instruments, we used a liberal interpretation of unidimensionality to combine items from all of the four comparison instruments into a single scale. This simplified the presentation so that an instrument-based rather than a content-based comparison could be made. A more detailed examination of common functional dimensions that underlie the item set was beyond the scope of this article. We also point out that, for practical reasons, we combined data from medical records and from interviews to develop the item calibrations for the instrument comparisons. Although not ideal, we do find only small amounts of error from clinician and self-report interview modes of testing within these functional items (Andres, Haley, and Ni, Forthcoming). More work in this area of combining data across respondents and modes of data collection will be needed as the field advances in assessing functional ability across post-acute core settings.

CONCLUSION

Results of this analysis illustrate some of the inherent limitations in the range of content, breadth of coverage, and measurement precision found in functional outcome instruments currently in use within PAC. Limitations in existing functional outcome methodology has stimulated the ongoing development of more comprehensive outcome assessment systems specifically for monitoring the quality of services provided for patients with different across diagnoses and across PAC settings. The careful use of IRT-based measurement methods coupled with CAT outcome assessment techniques may hold future promise for making outcomes assessment briefer and less burdensome to patients, and thus more acceptable for use in busy clinical settings. What is needed is functional outcome data that is applicable to patients treated across different clinical settings and applications, more efficient and less costly to administer, and, sufficiently precise to detect clinically meaningful changes in functional outcomes. Contemporary measurement methodology may hold considerable promise as a vehicle for advancing PAC outcomes evaluation, thus avoiding the pitfalls of traditional assessment methodology.

REFERENCES

Andres, P.L., Haley, S.M., and Ni, P.S.: Is Patient-Reported Function Reliable for Monitoring Post-Acute Outcomes? *American Journal of Physical Medicine and Rehabilitation*. Forthcoming.

Andresen, E.M., Gravitt, G.W.J., Aydelotte M.E., et al.: Limitations of the SF-36 in a Sample of Nursing Home Residents. *Age and Aging* 28:562-566, 1999.

Brunet, D.G., Hopman, W.M., Singer, M.A., et al.: Measurement of Health-Related Quality of Life in Multiple Sclerosis Patients. *Canadian Journal of Neurological Sciences* 23:99-103, 1996.

Buchanan, J., Andres, P., Haley, S., et al.: *Final Report on the Assessment Instrument for PPS*, MR-1501-CMS. RAND. Santa Monica, CA. 2002.

Cronbach, L.: Coefficient Alpha and the Internal Structure of Tests. Psychometrika 16(3):297-334. 1951.

DiFabio, R.P., Choi, T., Soderberg, J., et al.: Health-Related Quality of Life for Patients with Progressive Multiple Sclerosis: Influence of Rehabilitation. *Physical Therapy* 77:1704-1716, 1997.

Dodd, B., and Koch, W.: Effects of Variations in Item Step Values on Item and Test Information in the Partial Credit Model. *Applied Psychological Measurement* 11:371-384, 1987.

Guide for the Uniform Data Set for Medical Rehabilitation, (Including the FIM™ Instrument, Version 5.1.) State University of New York at Buffalo. Buffalo, NY. 1997.

Haley, S., and Jette, A.: Extending the Frontier of Rehabilitation Outcome Measurement and Research. *Journal of Rehabilitation Outcome Measurement* 4(4): 31-41, 2000.

Haley, S.M., and Langmuir, L.: How Do Current Post-Acute Functional Assessments Compare With the Activity Dimensions of the International Classification of Functioning and Disability (ICIDH-2)? *Journal of Rehabilitation Outcome Measurement* 4(4):51-56, 2000.

Hambleton, R.: Emergence of Item Response Modeling in Instrument Development and Data Analysis. *Medical Care* 38(9 Supplement II); II60-II65, 2000.

Hambleton, R. and Swaminathan, H.: *Item Response Theory: Principles and Applications*. Kluwer Nijhoff. Boston, MA. 1985.

Hamilton, B., Granger, C., and Sherwin, F.: A Uniform National Data System for Medical Rehabilitation. In Fuhrer, M. (Ed.) Rehabilitation Outcomes: Analysis and Measurement. Paul H. Brooks. Baltimore, MD. 1987.

Lord, F.: *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates. Hillsdale, NJ. 1980.

McHorney, C., and Cohen, A.: Equating Health Status Measures with Item Response Theory. *Medical Care* 38(9 Supplement II); II43-II59, 2000.

Meyers, A.: *Enabling Our Instruments: Assuring Access to Health Care Research*. Proceedings of the National Conference on Health Statistics. Washington, DC. 1999.

Morris, J.N., Murphy, K., and Nonemaker, S.: *Long-Term Resident Care Assessment User's Manual. Version 2.0.* American Health Care Association. Washington, DC. 1995.

Murnki, E.: Information Function of the Generalized Partial Credit Model. *Applied Psychological Measurement* 17:351-363, 1993.

National Committee on Vital and Health Statistics. *Classifying and Reporting Functional Status*. Internet address: http://ncvhs.hhs.gov. (Accessed 2002).

Rasch, G.: *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press. Chicago, IL. 1980.

Rubenstein, L.M., Voelker, M.D., Chrischilles, E.A., et al.: The Usefulness of the Functional Status Questionnaire and Medical Outcomes Study Short Form in Parkinson's Disease Research. *Quality of Life Research* 7:279-290, 1998.

Shaughnessy, P., Crisler, K., and Schlenker, R.: *Medicare's Oasis: Standardized Outcome and Assessment Information Set for Home Health Care: Oasis-B*. Center for Health Services and Policy Research. Denver, CO. 1997.

van Swieten, J., Koudstaal, P., Visser, M., et al.: Interobserver Agreement for the Assessment of Handicap in Stroke Patients. *Stroke* 19:604-607, 1988.

Ware, J., Bjorner, J., and Kosinski, M.: Practical Implications of Item Response Theory and Computerized Adaptive Testing. *Medical Care* 38(Supplement II) II73-II82, 2000.

Ware, J.E., and Kosinski, M.: *The SF-36® Physical and Mental Health Summary Scales: A Manual for User's of Version 1, 2nd Edition*. Boston, MA. 2001.

Ware, J.E., Kosinski, M., Dewey, J., et al.: *How to Score and Interpret Single-Item Health Status Measures: A Manual for Users of the SF-8™ Health Survey*. Quality Metric. Lincoln, RI. 1999.

Weiss, D.: Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological* Testing 6:473-492, 1982.

Wilkerson, D., and Johnston, M.: Clinical Program Monitoring Systems: Current Capability and Future directions. In: *Assessing Medical Rehabilitation Practices: The Promise of Outcomes Research*. Paul H. Brookes Publishing Company. Baltimore, MD. 1997.

Wright, B.D., and Masters, G.N.: *Rating Scale Analysis: Rasch Measurement*. MESA Press. Chicago, IL. 1982.

Reprint Requests: Alan M. Jette, Ph.D., Sargent College of Health and Rehabilitation Sciences, Boston University, 635 Commonwealth Avenue, Boston, MA 02215. E-mail: ajette@bu.edu