J Rehabil Med 2005; 37: 339–345

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

CONTEMPORARY MEASUREMENT TECHNIQUES FOR REHABILITATION OUTCOMES ASSESSMENT*

Alan M. Jette and Stephen M. Haley

From the Health and Disability Research Institute, Boston University, Boston, USA

In this article, we review the limitations of traditional rehabilitation functional outcome instruments currently in use within the rehabilitation field to assess Activity and Participation domains as defined by the International Classification of Function, Disability, and Health. These include a narrow scope of functional outcomes, data incompatibility across instruments, and the precision vs feasibility dilemma. Following this, we illustrate how contemporary measurement techniques, such as item response theory methods combined with computer adaptive testing methodology, can be applied in rehabilitation to design functional outcome instruments that are comprehensive in scope, accurate, allow for compatibility across instruments, and are sensitive to clinically important change without sacrificing their feasibility. Finally, we present some of the pressing challenges that need to be overcome to provide effective dissemination and training assistance to ensure that current and future generations of rehabilitation professionals are familiar with and skilled in the application of contemporary outcomes measurement.

Key words: healthcare evaluation mechanisms, quality of healthcare, outcome assessment, questionnaire design.

J Rehabil Med 2005; 37: 339–345

Correspondence address: Alan M. Jette, Health and Disability Research Institute, Boston University, 53 Bay State Road, Boston, MA 02215, USA. E-mail: ajette@bu.edu

Submitted March 18, 2005; accepted July 5, 2005

INTRODUCTION

As evidence-based practice and initiatives to improve the quality of healthcare have grown around the world, recognition of the need to measure functional outcomes in all healthcare settings has also increased. The rehabilitation field, a leader in functional outcome assessment, has long struggled with a tension between the need for comprehensive and clinically sensitive outcome instruments and the demand from the field for instruments that can be used feasibly in busy clinical settings. Faced with increasing pressure to justify the services they provide in as efficient a way as possible, the rehabilitation field faces an urgent need for more feasible approaches to monitoring relevant clinical outcomes throughout an episode of care and for comparing results across care settings (1, 2).

What exactly do we mean by measures of "functional outcome"? The general term "outcome measures" is used consistently in the rehabilitation literature to refer to assessments of the end results of health service programs and interventions, and does not include process measures of quality of care (i.e. access to services, or measures of satisfaction with a particular healthcare provider) (3). In contrast, there is no clear and commonly accepted definition of functional outcomes, or a clear delineation between instruments that assess functional outcomes, and those that assess other health concepts. As work to define and quantify health concepts has taken place, many different types of instruments assessing overlapping health and functional concepts have been developed. These include instruments of disability, function, activities of daily living, activity performance, advanced activities, physical performance, health, health status, quality of life, health-related quality of life, to name a few. To date, there is no consensus on how these terms should be used (4).

We advocate using the World Health Organization's (WHO) International Classification of Functioning, Disability and Health (ICF) concepts and terminology as a basis for discussion of the application of contemporary measurement technology to rehabilitation functional outcome assessment (5, 6). The ICF portrays human function and decrements in functioning as the product of a dynamic interaction between various health conditions and contextual factors. The ICF identifies three levels of human functioning: functioning at the level of body or body parts, the whole person, and the whole person in their complete environment. These levels are termed: body functions and structures, activities, and participation. The ICF defines an Activity outcome as "the execution of a task or action by an individual." The ICF defines a Participation outcome as "involvement in life situations", the result of a complex relationship between a person, his or her health condition, and the person's environment. In this review, we will use the term "functional outcomes" to include both the individual's ability to carry out specific tasks and activities of daily living

* This article is partly based on a lecture by Alan M. Jette at the international symposium "Measurement and evaluation of outcomes in rehabilitation", in September 2004.

 ISSN 1650–1977
DOI 10.1080/16501970500302793

(Activities) and to participate in life situations and society (Participation) as defined by the ICF (5).

In the first section of this article, we review the limitations existing in current functional outcome instruments in use within the rehabilitation field. Following this, we illustrate how contemporary techniques such as item response theory (IRT) methods and computer adaptive testing (CAT) can meet the challenges that must be addressed to design functional outcome instruments that are comprehensive and sensitive to clinically important change yet do not sacrifice feasibility. Finally, we discuss some of the pressing challenges that need to be overcome to provide appropriate dissemination and training assistance in contemporary functional outcome measurement to the rehabilitation field.

LIMITATIONS IN EXISTING OUTCOME INSTRUMENTS AND MONITORING SYSTEMS

A review of existing functional outcome instruments reveals that over 100 separate instruments have been developed to measure functional outcomes in populations of persons with chronic disease (7). Few functional outcome measures have been considered as a "gold standard," and standardization of functional outcome measures has been uncommon (8).

There are several well-respected, setting-specific functional outcome instruments in widespread use in rehabilitation (9–13). These existing measures have led to important insights into evidence-based rehabilitation practice. Although progress has been made, we believe there are several important deficiencies in current methodology that impede progress in monitoring, managing and improving the outcomes of services provided to patients across the entire episode of care (14, 15). They include: (i) narrowly defined scope of outcome measurement; (ii) the inability of different outcome instruments to talk to each other; and (iii) the classic trade-off between feasibility of existing outcome measures vs their limitations in detecting clinically relevant outcome changes. We will briefly summarize each.

Narrow scope

Among the host of existing outcome instruments, none can be considered as a "gold standard," for monitoring functional outcomes. In the 1960s and 1970s, outcome measures reflected the basic activities of daily living (ADL) needs of patients with chronic disabilities that matched the modest expectations of rehabilitation. With new advances in medical and rehabilitation management, rehabilitation has expanded its goals and consequently is now facing new measurement challenges. With changing professional, consumer and societal expectations, researchers have begun to explore means of documenting broader rehabilitation goals including community integration, patient satisfaction and social participation.

A watershed for standardized functional outcome measures for inpatient rehabilitation care was the introduction of the Functional Independence Measure (FIM^{TM}) (12). The widespread use of the FIM^{TM} in acute inpatient rehabilitation worldwide has made possible the emergence of industry reporting and benchmarking of outcomes across inpatient rehabilitation facilities and has contributed greatly to improving our understanding of the outcomes of inpatient rehabilitation care (16). Substantial limitations of the FIM^{TM}, however, have limited its usefulness outside of the inpatient settings (14, 17).

To meet the needs of the rapidly changing rehabilitation field, broader outcome measures have emerged for use in skilled nursing homes (e.g. Minimum Data Set for Long Term Care (MDS) (18), home care agencies (e.g. OASIS) (11), and in rehabilitation outpatient practices (13). The more recent focus on the importance of community integration is reflected in instruments such as the Community Integration Questionnaire (CIQ) (19), the Craig Handicap Assessment and Reporting Technique (CHART) (20), and the Participation Measure for Post-acute Care (21). Although rehabilitation consumers have made it clear that community integration and participation are key rehabilitation goals, even today, ADL measurement continues to be the dominant outcome focus in rehabilitation functional outcome instruments.

Data incompatibility

While functional outcome measurement advances have undeniably served many useful purposes, a crucial drawback is that the different assessment tools cannot "speak to one another." Data from one setting cannot be compared to another where the assessment of the same outcome trait was achieved using a different set of items. Despite 5 decades of measurement proliferation, each functional outcome measure is its own separate yardstick – each occupies different planes of a space rather than different spots on a common, underlying continuum. Data incompatibility across instruments renders the ability to track relevant outcomes across different care settings almost impossible to accomplish with traditional measurement technology. As a result, providers, clinicians and consumers currently have no reliable way of recording important functional changes that take place across care settings or across an entire episode of rehabilitation (22).

The precision vs feasibility dilemma

Over the past 30 years, we have greatly improved the breadth of measured health dimensions in outcome assessments (3, 7). However, even those functional outcome instruments with excellent breadth still have problems of inadequate depth of measurement (23). Thus, although we now quantify many different dimensions of health, most rehabilitation outcome assessments are imprecise, which restricts their utility to monitor clinical outcomes for quality improvement, benchmarking and research.

The defining signature of most traditional outcome instruments is the use of a standardized set of items for all patients. The advantage of such standardization is that results can be compared. Yet, it is difficult for one instrument to include the

number of items necessary to precisely measure the wide range of ability levels of individuals across care settings. Furthermore, to achieve measurement breadth and precision, patients and/or clinicians are often frustrated by being asked to respond to items that to them are redundant or of low relevance. Regardless of their answers, all individuals are asked the same questions and at least some questions are likely to appear redundant, illogical or unnecessary. The resulting length and complexity of many fixed-form outcome batteries is problematic and raises concerns over respondent burden and administration costs. To be as inclusive as possible, some monitoring systems, such as the OASIS and MDS, have large item sets and have become increasingly burdensome to clinicians and rehabilitation organizations.

The field has responded to this legitimate concern by shifting to shorter fixed-form versions of outcome instruments. The widespread adoption of short forms in outpatient services underscores the importance of practical considerations in determining whether outcome instruments can be used effectively to monitor relevant outcomes. Unfortunately, the very features underlying the popularity of short forms render them less precise. Static short-form questionnaires rely on a fixed set of questions that cannot possibly be the best for all respondents. Many short forms (e.g. FIM^{TM}) achieve their brevity by including questions that define only the lower levels of functioning where individuals with the greatest impairment score. Accordingly, they yield a concentration of scores at higher levels (ceiling problem). Other forms focus on higher levels of functioning, and thus result in a concentration of scores at the bottom of the scale, particularly among those individuals with the most functional limitation. A third short-form strategy (used with the SF-36) is to spread questions over a wider range, resulting in larger gaps and less precision at any one level (14). The "ideal" measure, possessing enough questions to cover the full range with a high degree of precision at all relevant levels, is impractical when using traditional measurement technology.

In summary, the rehabilitation field has long faced the inherent tension between the need for feasible instruments that are comprehensive and sensitive to clinically relevant change in outcome. Collectively, the lack of breadth, unequal precision for all patients, non-comparability across instruments, and the limited feasibility of current systems severely restrict the field's ability to measure and analyze progress across the continuum of rehabilitation care settings (14, 24). Modern test development techniques provide us with a innovative means of solving this measurement dilemma and open the way to monitoring functional outcomes across care settings and through an entire episode of care (2).

CONTEMPORARY METHODS TO IMPROVE OUTCOME MEASUREMENT

We believe 2 contemporary measurement techniques, item response theory (IRT) and computer adaptive testing (CAT), have the ability to overcome the previously discussed limitations in traditional functional outcome instruments and have the potential to transform how functional outcome assessment is done within rehabilitation. Although these advances have been used in educational testing for decades, they have only just recently begun to be applied to functional outcome assessment in rehabilitation and other arenas of healthcare.

IRT techniques

IRT methods examine the associations between individuals' response to a series of items designed to measure a specific outcome domain (e.g. physical functioning) (25). Data collected from samples of rehabilitation patients are fit statistically to an underlying IRT model that best explains the covariance among item responses (26, 27). IRT measurement models are a class of statistical procedures used to develop measurement scales. The measurement scales are comprised of items with a known relationship between item responses and positions on an underlying functional domain, called an item characteristic curve. The form of the relationships is typically non-linear. Using this approach, probabilities of patients scoring a particular response on an item at various functional ability levels can be modeled. Persons with more functional ability have higher probabilities of responding positively to functional items than persons with lower functional abilities. These probability estimates are used to determine the individual's most likely position along the functional dimension. When assumptions of a particular IRT model are met, estimates of a person's functional ability do not strictly depend on a particular fixed set of items. This scaling feature allows one to compare persons along a functional outcome dimension even if they have not completed the identical set of functional items. Since items and functional outcome scores are defined on the same scale, items can be optimally selected to provide good estimates of each outcome at any level of the scale. This feature of IRT creates important flexibility in administering tests in a dynamic and tailored approach for each individual. See reference (28) for a more detailed explanation of IRT methods.

IRT models have been developed for dichotomous and polytomous item response sets, and are manifested in 1, 2 and 3 parameter models. To date, the field of rehabilitation has largely used 1-parameter Rasch modeling because of its relative simplicity, ease of interpretation, and requirement of a smaller sample than more complex models. Rasch models develop item characteristic curves using a 1-parameter logistic function using only the item difficulty parameter. Assumptions are made that items have equal discrimination parameters and that guessing is not a factor in the data, an assumption that holds for most functional assessment applications. See reference (29) for a comprehensive treatment of Rasch analysis.

Increasingly, investigators are applying more complex IRT models in functional outcome development work. As researchers develop larger samples of patients when developing functional instruments, and as computer adaptive testing becomes more widespread, adding a second parameter (discrimination) to the analysis can be an important aspect of item selection within a CAT framework (30, 31).

IRT is currently being applied in rehabilitation outcomes research to develop new measures, to improve existing measures, to investigate group differences in item and scale functioning, to equate different instruments and, as we highlight, to develop efficient test applications, such as computer adaptive tests.

To apply IRT to functional outcome assessment, an appropriate item pool of functional tasks or activities needs to be assembled. An item pool is a collection of outcome items that represent a range of levels of a particular outcome domain. Item pools used in IRT analyses are developed by equating outcome items from different sources so that they can be meaningfully compared on a common underlying scale. IRT methods have been used to calibrate items from existing instruments onto a common scale, thus developing a structure and order of domain-specific items (32, 33). Alternatively, once the structure and ordering of items is determined, items can then be included in short form instruments based on a number of criteria, including comprehensiveness of content, item fit to the construct, item precision, correlation to the total item score, test-retest reliability and practical considerations of length. Within rehabilitation, researchers have linked functional outcome items from an item pool to create a practical yet comprehensive set of short forms that can be applied in different rehabilitation settings

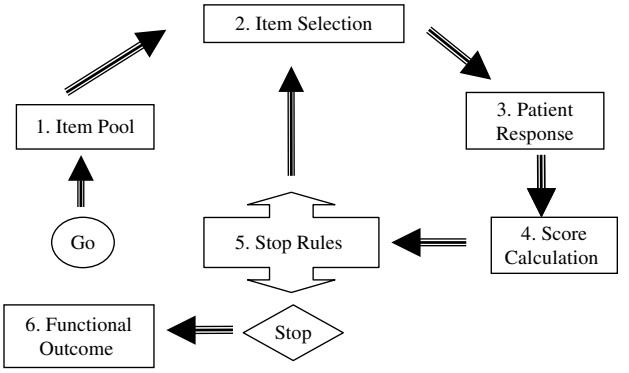

Fig. 1. Illustration of the basic computer adaptive testing (CAT) logic.

(34, 35). IRT methods open the door to understanding the linkages among items used to assess a common functional outcome domain, and in this way serve as the psychometric foundation underlying CAT (36–38).

CAT methodology

CAT programs use a simple form of artificial intelligence that selects questions tailored to the test-taker, and thereby shortens or lengthens the test to achieve the level of precision desired by a user. Functional outcome CAT applications rely on extensive item pools constructed for each outcome area. They contain items that consistently scale along each functional outcome dimension from low to high proficiency, and include rules guiding starting, stopping and scoring procedures. CAT methodology uses a computer interface for the patient/clinician report that is tailored to a patient's unique ability level. The basic notion of a CAT test is to mimic what an experienced clinician does. A clinician learns most when he/she directs questions at the patient's approximate level of proficiency. Administering outcomes items that represent tasks that are either too easy or too hard for the patient provides little information. In contrast to traditional, fixed form functional tests that ask the same questions of everyone regardless of how the respondent answers, CAT instruments, like a skilled clinician, tailor their assessment by asking only the most informative questions based on a person's response to previous questions.

A CAT is programmed to first present an item from the mid-range of an IRT defined item pool, and then direct subsequent functional items to the level based on the patient's (or clinician's) previous responses, without asking unnecessary questions. The selection of an item in the mid-range is arbitrary and the CAT can be set to select an initial item based on other information entered about the patient, such as age, diagnosis, or severity of their condition. By having comprehensive item banks available in each functional outcome domain of interest, the selection of additional items after the initial one is based on responses to the previous items. This allows for fewer items to be administered while gaining precise information regarding an individual's placement along an outcome continuum.

The logic of CATs in outcome assessment is shown in Fig. 1. At step 1, the computer begins with an initial item pool designed to measure a specific outcome domain. At step 2, the CAT is programmed to select and administer an initial item from that selected item pool to provide good discrimination over a wide outcome range. In our CAT models, we have selected an initial item that all patients answer as the first question. In our Physical Movement and Activities item pool, for example, this item might be, "How much difficulty do you currently have bending over to pick up something from the floor without holding on to anything?" with 5 response options, ranging from none to cannot do. On the basis of the response to the first item (step 3), an initial score estimate and confidence interval (CI) are calculated (step 4). CAT stop rules are based either on the size of a pre-programmed confidence interval or the maximum number of items that are to be used to estimate the score. If the CAT program determines that the stop rule (step 5) has not been satisfied, a new item from that same item pool is administered. A patient's response to that first item provides the basis for an initial functional outcome score estimate and the selection of the next item to administer to the patient. After each item receives a response, the score is re-estimated with a new confidence interval, and the stop rule is checked again (steps 2–5 repeated). The underlying CAT algorithm selects each new item to administer based on optimal information functions for each new score estimate. New items continue to be administered iteratively until the stop rule is satisfied (step 6).

We will illustrate how the CAT works using an Activity outcome scale developed in our research group (35). In this Activity outcome scale, we assume that the midpoint of the scale is 50, and this serves as the initial (default) score estimate prior to the CAT administration. For this example, we used data collected in a prospective rehabilitation outcome study (39). We set the CAT precision stopping rule as a 95% CI <3.0. The case is an individual with OA after hip replacement in a community-based outpatient center. The initial item administered is, "How much difficulty do you have coming to sit at side of a bed?" A response of "a little difficulty" yields a score estimate of 38.2 with a large CI (12). A second question is administered, based on the estimate from the first response, "How much difficulty do you have carrying a suitcase?" The person responds "no difficulty." A new score estimate is then calculated ($44.4+9.2$), and the CAT program checks to see if the stop rule has been satisfied. Since the stop rule in this case is a confidence interval of <5, a third item is administered. To the third item, "How much difficulty do you have running to catch a bus?" the person responds, "a little difficulty." A new score estimate is calculated ($44+6$). The stop rule has not yet been satisfied, so a fourth item is administered. The item "How much difficulty do you have doing heavy housework?" is given to the person, and the answer is "a little difficulty," with a new score estimate of $44.2+3$. Since this meets the stop rule, no additional items are administered, and a final score estimate based on 4 items is 44.2 with a confidence interval of $+3$. In this case, the 4 items administered were able to reproduce closely the score of 43.8, which was obtained by the administration of all 101 items in the full item pool. The number of items administered can be increased to achieve the desired level of precision.

EVALUATING CAT APPLICATIONS IN REHABILITATION

Though intuitively appealing on its surface, to be truly innovative and useful in rehabilitation, rehabilitation outcome CATs must meet several standards for acceptance for clinical and research applications. These include: (i) acceptable score accuracy in comparison to the entire item pool; (ii) adequate score precision for group and individual assessments; (iii) sufficient content breadth for application across care settings; (iv) adequate sensitivity for monitoring clinical relevant change; and (v) feasibility with respect to user burden and administration cost for widespread use.

Our research group has been involved in the development of several functional outcome CAT systems, and has undertaken investigations to evaluate their utility as compared to traditional fixed-form instruments. Some of the research we have undertaken is summarized in this section of the paper to illustrate the utility of CATs for rehabilitation applications.

Score comparability

In examining the value of CAT-based scores, an initial question is often: To what extent can a score generated from a few items in a CAT accurately represent a score if all the items were administered? Assessments of accuracy are often conducted initially by using computer simulation methods, followed by

prospective studies, in which patients are given both assessment formats.

In our work, we have found the CAT-generated scores for rehabilitation patients to be remarkably accurate in comparison to scores estimated by either a representative or full set of items from an instrument.

In the simulations studies that we have conducted to date, responses to questions selected by the CAT software were used in a CAT algorithm to imitate the conditions of a CAT assessment. The simulation selects the best item to administer next, re-estimates the score and CI, and decides whether or not to continue testing based on the number of items set for that simulation. These simulated scores are compared through correlation coefficients to scores from either the representative or full-length instrument. In simulation studies, we have found correlations in the range 0.94 to 0.98, indicating that the CAT programs (requiring 10 items or less) can provide comparable scores to those obtained from much larger sets of items for adults and children in rehabilitation programs (35, 40–42). Prospective studies that have been completed to date also confirm a high level of agreement between CAT-based scores and a full instrument (42).

Precision

We define measurement precision as the level of confidence around an individual score estimate. CAT scores based on IRT analyses provide a specific confidence interval for each individual score. In general, measurement precision is optimal when the content of functional items and the patient's abilities are closely matched. In CAT systems, a level of precision can be pre-defined, and items can be administered until a level of precision is obtained. Because of the need to keep items to a minimum number during CAT applications, we have found that there is often a small loss of precision in CAT-based scores from what can be optimally obtained from the full instrument (35, 42). However, the loss in precision is minimal compared to the loss in precision in going from a full item set to a short, fixed-form, in which all persons get the same items (35). The precision of the CAT is superior to fixed forms since the CAT selects specific items to match individual response patterns and a patient's functional outcome.

Content breadth and applicability across settings

In heterogeneous groups, such as are seen in rehabilitation care, an optimal set of items that fits most patients in a particular rehabilitation subgroup may not be relevant for all patients in the larger group. Therefore, any one instrument developed for a specific setting or patient group typically has considerable floor and ceiling effects when used in other post-acute care settings or with other patient groups. To make instruments more practical across groups and rehabilitation settings, large item pools can be developed that limit measurement noise at any level of the scale, or for any patient group. And, because practical limitations in content are minimized, ceiling and floor effects (a high proportion of respondents scoring at the highest or lowest possible score) are nearly eliminated, yet any 1 person only answers a small set of items.

In our current work with measuring Activity outcomes (43), we have an item pool of approximately 130 items that covers physical mobility content from moving in bed to walking long distances in the community. Mobility items incorporate the use of wheelchairs, walking with walking devices, and management of stair and inclines. In a CAT system, items can be filtered so that they are unavailable for any one assessment, based on the setting and patient group. For example, if a person is currently using only a wheelchair for mobility in an inpatient setting, items can be filtered so that for that particular session, only wheelchair items that are applicable for an inpatient setting are made available for administration. As a person no longer needs a wheelchair for mobility and moves back to a community setting, items that are more relevant for the community setting and for ambulation skills are made available for assessment. However, due to prior IRT work, the metric used for measurement of function is constant, so that scores obtained from the first test, even though different items are answered, are on a comparable metric.

Sensitivity to change during rehabilitation interventions

In early computer simulation work, we have found that CAT estimates of pediatric functional outcome measures can detect nearly a similar magnitude of change between admission and discharge administrations as compared to the administration of the full item pool (42). Sensitivity to change in CAT versions closely approximated that of the full-length instrument (59 items), with the 15-item version picking up 99% and 10-item version picking up 98% of the change detected by the full instrument. In an unpublished study in which we tested the effects of a fitness program in 28 children with developmental disabilities, we found that the standardized response mean (SRM) (ratio of mean change to the standard deviation of the change score) was 1.56 for the full instrument (161 items) and 1.00 for the 15-item CAT. Although some sensitivity was lost with the CAT, primarily due to the greater variability in the change scores, the SRM was sufficiently high to detect both clinically meaningful and statistically significant changes in functional mobility. More work of this type is currently underway in our research group within adult rehabilitation programs, using external anchors to help determine minimally clinically important differences of CAT-based scores.

Feasibility for widespread use in research and practice

It is well established that as the number of items or length of time to complete an assessment increases, missing data and data quality diminish. If the burden of assessment is too high for patients and clinicians, we know that assessments will not be administered at all. Therefore, the feasibility and efficiency of instruments must be maintained or even improved if they are to be widely used.

In our experience, CAT programs used in rehabilitation settings generally require 8–15 items per construct to accurately

reproduce full-item pool scores. For example, the mean number of items required for the PEDI-CAT (42) in the computer simulation was 8.2 items (SD 2.7) for patients, which is less than 14% of the number of items in the full item pool. For purposes such as examining change, up to 15 items might be ideal. In prospective administrations, the average time to complete the PEDI-CAT was about 1 minute and 13 seconds. We have found no situations in which a CAT program took more than 3 minutes to estimate a score. All of our CAT programs to date have used no more than 15% of the items for any 1 administration, or have required more than 1/5 of the time needed to administer the full-length test.

FUTURE CHALLENGES

We believe that contemporary outcome measurement techniques such as IRT and CAT methodology present an exciting innovation that has the potential to transform the methods used to administer functional outcome assessment within the field of rehabilitation. Although promising, much evaluation work needs to be done to demonstrate that CAT assessments actually live up to their potential within the healthcare environment.

Contemporary functional outcome instruments based on the WHO's ICF framework are currently being developed and tested by rehabilitation researchers worldwide. The goal is to provide the field with quantitative functional outcome instruments that will replace an earlier generation of ordinal-scaled functional outcome instruments that continue to be the norm.

Once developed and shown to be beneficial for widespread application and use within rehabilitation, the next challenge facing the field will be to develop effective and efficient methods to disseminate these innovations throughout the rehabilitation research and practice communities. It is essential that not only is information about contemporary outcome instruments communicated accurately and efficiently, but that potential users understand what they can offer and have the skill to appropriately assess functional outcomes. Without careful attention to dissemination and training, rehabilitation professionals may not know how to use these innovative tools and, consequently, ordinal-scaled measures are likely to be the norm for years to come.

To accomplish this challenge, many dissemination methods will need to be developed and implemented beyond the traditional methods of dissemination through professional conference presentations and publication in scholarly journals (44). Funding mechanisms will need to be developed that will support these dissemination tasks at every level.

Future users need to be provided with the software needed to apply, analyze, and interpret CAT-based outcome instruments. This may require the development of continuing education seminars or high-quality technical assistance vehicles to assist rehabilitation professionals and organizations in their understanding, application and interpretation of contemporary outcome measurement tools. Accreditation organizations might be able to play a crucial role in this dissemination approach, facilitating the dissemination process. In addition, efforts need to be taken to ensure future generations of rehabilitation professionals are appropriately trained through the development of didactic courses and professional curricula on contemporary outcomes measurement. Specific courses on modern measurement technology can be incorporated into professional curricula as a new basic science in entry-level education across all rehabilitation disciplines. To accomplish this challenge will require efforts to educate rehabilitation faculty in the science of contemporary outcome measurement so that they have the skill to develop and deliver these courses to their future students. All of these dissemination steps are necessary to ensure that future generations of rehabilitation professionals are familiar with and skilled in the application of contemporary outcomes measurement.

For more than a generation, rehabilitation researchers and clinicians have struggled with a difficult choice between applying outcome instruments that were practical but only assessed a limited range of functional outcomes and were unresponsive to meaningful levels of change vs selecting more comprehensive outcome tools that met methodological needs but were excessively long and costly to use. Great frustration has also occurred because no existing functional outcome instrument was appropriate for use across the variety of settings where rehabilitation services were being provided. This made it impossible to monitor rehabilitation outcomes across settings and throughout an entire episode of care. These methodological limitations have impeded the development of evidence-based approaches to rehabilitation care. Research using contemporary measurement methods such as IRT and CAT methods has the potential to help eliminate these limitations. Once developed and fully tested, these contemporary outcome instruments need to be widely disseminated and incorporated into rehabilitation practice and research to improve our understanding of the effectiveness of rehabilitation services.

ACKNOWLEDGEMENTS

Supported by Grant H133B990005 from the Department of Education, National Institute of Disability and Rehabilitation Research to the Boston University Rehabilitation Research Training Center on Measuring Rehabilitation Outcomes.

Also supported in part by an Independent Scientist Award (K02 HD45354-01) to Dr Haley and grant R01 HD043568 (Dr Haley, PI) from the National Institute of Child Health and Human Development(NICHD) and the Agency for Healthcare Research and Quality (AHRQ) and grant RO1 AR 051870 (Dr Jette, PI) from the National Institute of Musculoskeletal Diseases and Conditions, NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NICHD, AHRQ or NIH.

REFERENCES

1. Wilkerson DL, Johnston MV. Clinical program monitoring systems: current capability and future directions. In: Fuhrer M, ed. Assessing medical rehabilitation practices: the promise of outcomes research. Baltimore, MD: Paul H. Brookes Publishing Co.; 1997, pp. 275–306.

2. Latham N, Haley SM. Measuring functional outcomes across post-acute care: current challenges and future directions. Clin Rev Phys Rehabil Med 2003; 15: 83–98.
3. Patrick DL, Chiang YP. Measurement of health outcomes in treatment effectiveness evaluations: conceptual and methodological challenges. Med Care 2000; 38 (9 Suppl): II14–II25.
4. Johnston MV, Steinman M, Velozo CA. Outcomes research in medical rehabilitation: foundations from the past and directions to the future. In: Fuhrer M, ed. Assessing medical rehabilitation practices: the promise of outcomes research. Baltimore, MD: Paul H. Brookes Publishing Co.; 1997, pp. 1–41.
5. World Health Organization. International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.
6. Stucki G, Ewert T, Cieza A. Value and application of the ICF in rehabilitation medicine. Disabil Rehabil 2003; 25: 628–634.
7. McHorney C. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997; 127: 743–750.
8. Fisher AG. Functional measures, part 1: what is functions, what should we measure, and how should we measure it? Am J Occup Ther 1992; 46: 183–185.
9. Harvey R, Jellinek HM. Functional performance assessment: a program approach. Arch Phys Med Rehabil 1981; 62: 456–460.
10. Morris J, Murphy K, Nonemaker S. Long-Term Resident Care Assessment User's Manual. Version 2.0, American Health Care Association. Washington DC, 1995.
11. Shaughnessy P, Crisler KS, Schlenker RE. 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.
12. State University of New York at Buffalo. Guide for the uniform data set for medical rehabilitation (including the FIMTM instrument), version 5.1. State University of New York at Buffalo: Buffalo, NY; 1997.
13. Ware J, Sherbourne CD. The MOS 36-item short form health survey (SF-36) I. Conceptual framework and item selection. Med Care 1992; 30: 473–483.
14. Jette A, Haley SM, Ni P. Comparison of functional status tools used in post-acute care. Health Care Financ Rev 2003; 24: 13–24.
15. Haley SM, Jette AM. RRTC for measuring rehabilitation outcomes: extending the frontier of rehabilitation outcome measurement and research. J Rehab Outcome Meas 2000; 4: 31–41.
16. Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Ostir GV, Granger CV. Trends in length of stay, living setting, functional outcome, and mortality following medical rehabilitation. JAMA 2004; 292: 1687–1695.
17. Coster W, Haley SM, Jette AM. Measuring patient reported outcomes after rehabilitation using the short form activity measure for post-acute care (AM-PAC). Unpublished manuscript.
18. Hamilton BB, Granger CV, Sherwin FS, Zielezny M, Tashman JS. A uniform national data system for medical rehabilitation. In: Fuhrer MJ, ed. Rehabilitation outcomes: analysis and measurement. Baltimore: Paul H. Brookes Publishing Co.; 1987, pp. 137–147.
19. Willer B, Rosenthal M, Kreutzer JS, Gordon WA, Rempel R. Assessment of community integration following rehabilitation for traumatic brain injury. J Head Trauma Rehabil 1993; 8: 75–87.
20. Whiteneck G, Charlifue SW, Gerhart KA, Overholser JD, Richardson GN. Quantifying handicap: a new measure of long term rehabilitation outcomes. Arch Phys Med Rehabil 1992; 73: 519–526.
21. Gandek B, Sinclair JH, Jette AM, Ware J. Development and initial testing of the participation measure for post-acute care (PM-PAC). Unpublished manuscript.
22. Johnston M. Representations of disability. In: Petric K, Weinman JA, eds. Perceptions of health and illness. New York: Hardwood Academic Publishers; 1997, pp. 189–212.
23. Liang M, Lew R, Stucki G, Fortin P, Daltroy L. Measuring clinically important changes with patient-oriented questionnaires. Med Care 2002; 40 (4 Suppl): II45–II51.
24. Petrella R, Overend T, Chesworth B. FIM after hip fracture: is telephone administration valid and sensitive to change? Am J Phys Med Rehabil 2002; 81: 639–644.
25. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Earlbaum Associates; 2000.
26. Hambleton RK. Emergence of item response modeling in instrument development and data analysis. Med Care 2000; 38 (9 Suppl): II60–II65.
27. Thissen D, Steinberg L. Data analysis using item response theory. Psychol Bull 1988; 104: 385–395.
28. Hambleton RK. Principles and selected applications of item response theory. In: Linn RL, ed. Educational measurement (3rd edn) New York: MacMillian; 1989, pp. 147–200.
29. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003; 35: 105–115.
30. McHorney CA, Monahan PO. Postscript: applications of Rasch analysis in health care. Med Care 2004; 42 (1 Suppl): I73–I78.
31. Cook K, Monahan P, McHorney C. Delicate balance between theory and practice: health status assessment and item response theory. Med Care 2003; 41: 571–574.
32. Bjorner JB, Ware JE. Using modern psychometric methods to measure health outcomes. Monitor 1998; April: 12–17.
33. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000; 38 (9 Suppl): II28–II42.
34. Velozo CA, Kielhofner G, Lai JS. The use of Rasch analysis to produce scale-free measurement of functional ability. Am J Occup Ther 1999; 53: 83–90.
35. Haley S, Coster WJ, Andres PL, Kosinski M, Ni P. Score comparability of short forms and computerized adaptive testing: simulation study with the activity measure for post-acute care. Arch Phys Med Rehabil 2004; 85: 661–666.
36. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res 1997; 6: 595–600.
37. Wainer H, Dorans N, Eignor D, Flaugher R, Green BF, Mislevy RJ, et al. Computerized adaptive testing: a primer. New Jersey: Lawrence Erlbaum Associates; 2000.
38. Van der Linden GCAW, Computerized adaptive testing: theory and practice. Dordrecht, Netherlands: Kluwer Academic Publishers; 2000.
39. Jette AM, Keysor J, Coster W, Ni P. Beyond function: predicting participation outcomes in a rehabilitation cohort. Arch Phys Med Rehabil (in press).
40. Andres P, Black-Schaffer RM, Ni P, Haley SM. Computer adaptive testing: a strategy for monitoring stroke rehabilitation across settings. Top Stroke Rehabil 2004; 11: 33–39.
41. Haley SM, Fragala-Pinkham MA, Ni PS, Skrinar AM, Corzo D. An adaptive testing approach for assessing physical functioning in children and adolescents. Dev Med Child Neurol 2005; 47: 113–120.
42. Haley SM, Raczek AE, Coster WJ, Dumas HM, Fragala-Pinkham MA. Assessing mobility in children using a computer adaptive testing version of the Pediatric Evaluation of Disability Inventory. Arch Phys Med Rehabil 2005; 86: 932–939.
43. Haley S, Coster WJ, Andres PL, Ludlow LH, Ni P, Bond TLY, et al. Activity outcome measurement for postacute care. Med Care 2004; 42 (1 Suppl): I49–I61.
44. Farkas M, Jette AM, Tennstedt S, Haley SM, Quinn V. Knowledge dissemination and utilization in gerontology: an organizing framework. Gerontologist 2003; 43 (spec no 1): 47–56.