SPECIAL SECTION: SPECIAL COMMUNICATION

State-of-the-Science on Postacute Rehabilitation: Measurement and Methodologies for Assessing Quality and Establishing Policy for Postacute Care

Pamela W. Duncan, PhD, PT, Craig A. Velozo, PhD, OT

ABSTRACT. Duncan PW, Velozo CA. State-of-the-science on postacute rehabilitation: measurement and methodologies for assessing quality and establishing policy for postacute care. Arch Phys Med Rehabil 2007;88:1482-7.

We present an overview of commonly used postacute outcome measures and review new methodologies for postacute assessment. We question the impact that current measurement has had on improvement of quality of postacute care (PAC) and its utility in informing health policy. We suggest that Donabedian's model of health care quality should be endorsed for measurement. Specifically, measurement of outcomes and process should be used jointly in assessment of PAC.

Key Words: Health policy; Outcome and process assessment (health care); Rehabilitation.

MEASUREMENT OF THE quality and outcomes of postacute rehabilitation care should be the basis on which clinical and policy decisions are made to improve the health of those with disabilities. Despite over 2 decades of measurement in postacute care (PAC) sites (skilled nursing facilities [SNFs], inpatient rehabilitation facilities [IRFs], outpatient rehabilitation, long-term care hospitals, home health agencies [HHAs]), we continue to have inadequate assessment of performance of postacute rehabilitation services. The complicated, inconsistent, and burdensome approaches to PAC measurement and the heterogeneity in timing of assessments are significant obstacles to establishing a policy that ensures appropriate access and quality of PAC.

PAC OUTCOMES MEASURES

In the last 2 decades, we have seen the development and adoption of multiple postacute measures: the FIM instrument for inpatient rehabilitation, the Minimum Data Set (MDS) for SNFs, and the Outcomes and Assessment Information Set (OASIS) for home health services. These 3 instruments dominate PAC measurement. The FIM instrument, MDS, and OASIS were developed for different purposes and have been expanded for off-label use. The FIM instrument was originally developed by a task force from the American Congress of Rehabilitation Medicine and the American Academy of Physical Medicine and Rehabilitation for inpatient rehabilitation settings as a measure of burden of care.[1] It is now the basis for the Center for Medicare & Medicaid Services (CMS)–mandated Inpatient Rehabilitation Facility–Patient Assessment Instrument.[2,3] The MDS was developed as part of a congressionally mandated uniform assessment of nursing home residents to develop resident care plans, but it has emerged over time to be an instrument for assessment of quality and now supports classifications for reimbursement of PAC.[4,5] The OASIS was developed with funding from the Health Care Financing Administration and the Robert Wood Johnson Foundation to improve quality of care in home health services.[6,7] In 1999, Medicare required certified HHAs to collect data to compare outcomes across agencies. In 2000, OASIS data supported the establishment of Medicare reimbursement for home health.

There are a considerable number of psychometric studies of the FIM instrument and MDS with a somewhat lesser number of studies on OASIS. In general, these psychometric studies support the physical functioning domains for monitoring outcomes in health care, but there is less support for their cognitive domains.[8-12] The most serious limitation of these instruments has been the ceiling effect, especially for the cognitive domains.[13,14]

There is little to no consistency in the use of outcomes measures in PAC, beyond the FIM instrument, MDS, and OASIS. Although the Medical Outcomes Study 36-Item Short-Form Health Survey is the most widely used outcomes measure for health care research, this instrument is only sparsely used for monitoring facility-level outcomes. Furthermore, specialized sites, such as long-term acute care facilities, tend to report single-variable outcomes such as survival time, mortality rate, and costs.[15] Recently, CMS recommended (but did not require) 4 measures for monitoring outcomes at outpatient facilities: the National Outcomes Measurement System (NOMS) by the American Speech-Language Hearing Association, the Patient Inquiry by Focus on Therapeutic Outcomes Inc, the Activity Measure for Postacute Care (AM-PAC), and Outpatient Physical Therapy Improvement in Movement Assessment Log (OPTIMAL) by the American Physical Therapy Association.[16] These instruments are all relatively recent developments and there is widespread disparity on their published psychometric support. For example, the AM-PAC[17-19] and patient inquiry[20,21] have several published psychometric studies, but the OPTIMAL[22] has 1 published psychometric study and NOMS has none.

From the Division of Physical Therapy, Department of Community and Family Medicine, Center for Clinical Health Policy Research, Duke University, Durham, NC (Duncan); Department of Occupational Therapy, College of Public Health and Health Professions, University of Florida, Gainesville, FL (Velozo); and Department of Veterans Affairs Rehabilitation Outcomes Research Center, Gainesville, FL (Duncan, Velozo).

Supported by the Department of Veterans Affairs Health Services Research and Development and the Rehabilitation Research and Development Rehabilitation Outcomes Research Center of Excellence, North Florida and South Georgia (grant no. ROC 01-124).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the author(s) or upon any organization with which the author(s) is/are associated.

Reprint requests to Pamela W. Duncan, PhD, PT, Div of Physical Therapy, Dept of Community and Family Medicine, Center for Clinical Health Policy Research, 2200 W Main St, Ste 220, Durham, NC 27705, e-mail: *Pamela.duncan@duke.edu.*

0003-9993/07/8811-00114$32.00/0
doi:10.1016/j.apmr.2007.08.118

CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY

Each of the postacute site-specific measures has different domains of assessment, rating scales, different terminology, and has had different methodology for development. The existing measures have been developed using classical test theory (CTT). Measures that are developed with CTT have a fixed set of questions or items presented to the subject. All items are asked irrespective of relevance or appropriateness of any question. This has led to the burden of respondents taking all items of an instrument or rehabilitation providers administering all items of an instrument, irrespective of their relevance. A person's score is dependent on the items of the particular assessments. With assessments that have challenging items, respondents get lower scores and on assessments that have easy items, they get higher scores. Although the items of the existing instruments have established reliability and validity, they do not reflect the range of function that occurs across all settings of PAC.[23] In addition, the different postacute instruments have different items with different rating scales. There are no clear means to convert scores from 1 instrument to another.

Item response theory (IRT) methodologies and its associated applications (computer adaptive testing [CAT] and equating or linking) offer opportunities to move beyond the limitations of CTT. These methodologies and applications have a long history and acceptance in the field of education. IRT, in contrast to CTT, provides us with a means to avoid test dependency. IRT models are based on the probability of passing or failing items with dichotomous response options, or being rated at specific points on a rating scale on items that use this response option.

IRT models have their benefits; they derive person ability measures that are independent of any particular test, in contrast to CTT models whose scores are connected to a particular assessment. The independence of the measure from the test allows for dramatic possibilities in health care measurement. First, it means that not all of the items of a test are required for measuring a respondent. IRT item difficulty and item discrimination parameters can be used to determine the most relevant items for a subject given his/her ability level. In combination with CAT item administration, IRT provides an efficient method to assess subjects. Although one of the best-known applications of CAT in education is the Graduate Record Exam, CAT is also used by the state licensing boards.[24] CAT is based on an algorithm or set of rules for efficient presentation of items. For example, a simple CAT algorithm consists of first presenting a middle-difficulty item. If the respondent passes that item a more difficult item is presented and if he/she fails the item, an easier item is presented. This procedure continues until a stopping criterion is reached, such as reducing measurement error or the confidence interval to a prescribed level.[25] CAT item administrations are highly efficient. Using CAT simulations, Haley et al[19] showed that precise person measures can be achieved with 4 to 10 items (ie, CAT measures derived from only 4–10 items correlate .90–.98 with measures derived by taking all 101 items of their assessment). This feature has the potential to dramatically reduce respondent burden and health care data collection time and costs.

In addition to efficiency of measurement, IRT and CAT applications offer other benefits to measurement. Because IRT provides item-level information, subsets of items representing a particular construct (eg, physical function) can be selected for particular purposes.[26] For example, short forms can be designed to cover the full range of a trait (ie, low ability to high ability), or can be designed to focus on a particular subset of respondents (eg, low ability people). Furthermore, CAT applications have the advantage of immediate data entry and instantaneous output of results.[26] In contrast to paper-and-pencil forms, CAT participant responses are immediately entered into a database. This database can be localized to the particular computer that presents the CAT or can be sent through a local area network or through the web to a larger server. In addition, the output of a participant's responses can be quickly produced in a table or graphical output for the participant or administrator.

Another area of measurement that is well established in education, but has had very little attention in health care, is equating or linking. With the surge of national testing in public schools in the 1950s, there emerged a need for equivalence between different versions of test batteries.[27] With the advantage of large national datasets in education and the need to create test versions that have similar item structures (eg, dichotomous rating scales), CTT methodologies have been commonly applied to equating in education. IRT methodologies facilitate and extend equating by allowing the linking of instruments that represent an identical or similar trait, but that have different item structures (eg, 1 instrument a 5-point scale, another instrument a 7-point scale).

In health care, there have been and continue to be efforts to create new measures of the same construct, which have produced over 85 instruments to measure basic and instrumental activities daily living (IADLs) and over 75 instruments to measure quality of life.[28] Despite extensive statistical designs and methods to equate instruments in education, equating in health care has been challenged due to the lack of item-to-item match between instruments; early attempts at linking have focused on rescaling items to get an item-to-item match.[29-31] Only within the last 12 years have IRT approaches been applied to linking measures in health care.[32-38] Although crosswalks between physical function measures of health care instruments have been shown, there are only 2 validation studies (ie, investigating the similarities of converted scores and actual scores) for item-to-item linked instruments and no validation studies of instruments linked through IRT methods.[30,31]

IRT, CAT, and linking in health care clearly represent exciting and promising methodologies to advance outcomes measurement in health care; yet we should be cautious and thoughtful about the results that these methodologies can offer. Of these methodologies, studies of IRT and CAT appear to be exponentially increasing, especially over the last 5 years. Fayers,[39] in a recent review, cautions that the "excessive enthusiasm" over these methodologies may lead to a "corresponding lack of circumspection." He notes a number of areas that need considerable development in using IRT in health care, such as: reliability and easy-to-use IRT software, guidelines for selecting the appropriate IRT model, and guidelines for sample size determination. Further, he notes other questions need to be answered, such as whether IRT offers advantages for some instruments and not others, when to use multidimensional versus unidimensional IRT models, and the effects of differential item functioning (ie, obtaining different item difficulty calibrations across different groups; eg, males vs females). Finally, he cautions that IRT does not supplant the need for rigorous instrument development using qualitative methods.

In addition to the statistical and methodologic questions that surround the use of IRT in health care, CAT applications must overcome multiple barriers before it is accepted in clinical settings. CAT has considerable software and hardware requirements if implemented across postacute health care venues. Although the CAT methodology is fairly straightforward, it requires on-site technology investment. Software data collection systems would have to be modified or adapted for CAT