Research in Physical Medicine and Rehabilitation

XII. Measurement Tools with Application to Brain Injury

MARK V. JOHNSTON, PhD, THOMAS W. FINDLEY, MD, PhD,
JOHN DELUCA, PhD, AND RICHARD T. KATZ, MD

There are basic principles and techniques of measurement that are relevant across biomedical disciplines. The purpose of this article is to explain some of the most important of these for medical rehabilitation, to illustrate how to use them to choose assessment instruments, and to describe the nature of measurement in medical rehabilitation by examples in brain injury rehabilitation. Reliability is basic to any scientific measure. Validity, the ultimate criterion, is closely associated with the purpose of the measure. Content validity, criterion validity and construct validity are explained. Sensitivity to rehabilitative interventions and significance in patients’ real lives (ecological validity) are emphasized. Measures of functional outcomes (disability) may show improvement after rehabilitation even when impairment measures do not. An extensive but selected list of measures of coma, global status, disabilities, communicative and cognitive impairments, and handicaps is presented, and their main uses are illustrated. Examples illustrate how to choose measures to study comprehensive program-level outcomes, to study learning-based interventions, and to develop a general-purpose database. Although there are many measures of activities of daily living and mobility, little published evidence of reliability and validity could be found even for some well-known scales. Ecologically valid and sensitive outcome measures are especially needed. Studies of the clinical utility of measures were also scarce. Clinical researchers with limited resources can span many of these gaps. Physical medicine and rehabilitation will benefit from formal studies of the reliabilities and validities of both its old and its new measurement instruments and by increased sophistication in choice of measures.

The field of physical medicine and rehabilitation is distinguished in medicine by concern not just with organ system function but with “function” in the sense of abilities that are evident in the clinical setting and generalize to the person’s real life after discharge. As a result, it encounters particular problems in development, validation and choice of measures of its processes and outcomes. Reviews of measurement in rehabilitation have noted that the field greatly needs more formal study of the reliability and validity of its measures and assessment procedures. Without evidence of reliability and sensitivity to change, studies reporting no treatment effect are moot. Without valid, sensitive measures, the field will be hampered in attempts to identify what it is in rehabilitation that is most effective, and also for whom it is most effective.

OBJECTIVES

This article will:

- Explain principles of measurement that need to be understood to choose and use instrument rehabilitation research (and clinical practice).
- Review outstanding measures of use to medical rehabilitation research in brain injury. The review is intended to be illustrative of major domains, but not exhaustive within each category.
- Illustrate how to use measurement principles in conjunction with knowledge of rehabilitation processes, to choose optimal measures for study of different types of rehabilitation programs and different types of research.

This article aims to take you a step or two beyond common sense towards a more scientific level of understanding of measurement tools. You can then exclude many instruments without even trying them out, and you are likely to discover useful instruments that you might otherwise have dismissed.
TERMINOLOGY

We will refer to “brain injury” rather than traumatic brain injury because the measures treated here refer to a wider diagnostic group than merely traumatic cases. “Assessment” will refer to a formal clinical evaluation resulting in numeric or categorical description of the patient. “Measure” or “measurement” is used generically.

We will use the World Health Organization’s (WHO) *International Classification of Impairment, Disability, and Handicap*. An *impairment* is “any loss or abnormality of psychological, physiological, or anatomical structure or function.” Impairments are measured by mechanical means (e.g., blood pressure cuff) or by very specific observations of performance (e.g., range of motion, abnormality of syntactic function). Some measures of impairment are structural (e.g., CT scan, amputation), whereas others are more functional (e.g., V0, Spectamine test). A *disability* is “any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being.” Disabilities are measured by behavior or performance in defined situations or environments of limited variability. Handicap is the highest, broadest and least specific level. *Handicap* is “a disadvantage for a given individual, resulting from an impairment or disability, that limits or prevents the fulfillment of a role that is normal . . . for that individual.” For medical research, a fourth level can also be essential: *pathology*, an abnormality at the cellular or biochemical level.

Functional assessment in medical rehabilitation cannot be understood without understanding these different levels of function. Disability reduction is central to medical rehabilitation, but impairment measures are essential to understand results and because interventions may aim proximally to reduce impairment. Handicap is crucial to research on the worth of medical rehabilitation, for handicap measures assess the real world value of outcomes to individuals served. The aim of rehabilitation, we believe, is to decrease patients’ handicap by improving their function at the level of disability.

BASIC MEASUREMENT CONCEPTS
AND PRINCIPLES

Reliability and Validity

In choosing a measure, we start with standards developed by the American Psychological Association: “Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from the test scores. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores.” Strictly speaking, it is inferences about a measure of function that are validated, not the measure itself. The validity of these inferences is not all or none, but is limited to a wide or narrow range of diagnoses, organ systems, human activities, range of severity and uses.

Reliability

Reliability is essential to scientific measurement and to assessment in medicine. Reliability may be simply explained as the degree to which an instrument is consistent or reproducible when administered by appropriately trained individuals. It may be thought of as the ratio of “signal” to “noise” in a measure. A validity coefficient cannot be greater than the square root of the reliability coefficient.

*Interrater reliability* is the degree to which two or more trained raters agree and is the most basic form of reliability for rating scales and observational measures. Test-retest reliability or *stability* over time is also a basic feature needed to evaluate a measure. *Accuracy* is the relevant concept for measures of a physical quantity. It is typically determined by comparing a clinically practical measure with a highly exact and expensive criterion. Reliability of a scale tends to increase in proportion to the square root of the number of items comprising it. Longer scales can more easily be reliable than short ones.

Reliability differs from *bias*. Strictly speaking, unreliability is the degree of random error or noise in a measure. Unreliability then increases the margin of error or uncertainty of a measure but has no effect on the estimated average score. Bias is whether a measure gives systematically too high or too low a number. Self-report questionnaires, judgmental rating scales, and attitudinal measures often have substantial biases, but if well-developed
(e.g., proper phrasing), correctly administered, or if special procedures (e.g., single or double blinding) are used, they can have negligible biases.

An implication that whether an instrument is an “observational rating” or even a “self-report” rather than a measure of a physical quantity is not the most essential scientific question. The question is whether two observers, following the assessment protocol, come up with the same number, and whether useful inferences about other measures can be made from this number. Measures of physical aspects of function (e.g., blood pressure, time or speed of performance) can of course be extremely reliable, but one cannot assume that impairment measures are more reliable or valid than measures of disability or handicap. Ratings of physical disability in the activities of daily living (ADLs) and mobility by experienced physical therapists, for instance, have been shown to be more reliable than ratings of impairments, particularly of pain and range of motion.9

Internal consistency or homogeneity is essential to the reliability of a multi-item assessment when the items are added up or labeled as assessing the same underlying factor. If you add up unrelated items, you should expect them to fail to predict anything very well, even if a few of them are in fact highly correlated with the outcome of interest. Adding unrelated items and summing them muddies the waters. On the other hand, a rater may be self-consistent even in mood ratings that have no appreciable interrater reliability, so internal consistency is no substitute for interrater agreement or test-retest stability. Homogeneity is an important concern in functional assessment. Improved independence in one mobility activity, for instance, may be poorly related to improved independence in others (e.g., wheelchair mobility vs ambulation). Does it make sense then to speak of “independence in mobility?” Studies of the internal homogeneity of scales answer such questions.

Statistics

The proper statistic to characterize reliability depends upon the level of measurement and the consequences of error.7 Percent agreement works only for dichotomous measures. Kappa is an excellent measure for categorical measures, for it corrects for the lessened a priori probability of correctness in a multipoint scale. Weighted Kappas and intraclass correlation coefficients are relevant to ordinal measures. Fleis provides a terminology to describe reliability figures in common language.10 A Kappa greater than 0.75 is labeled as high; 0.40 to 0.75 is fair to good; and less than 0.40 is labeled as low or poor. We will use his terminology in this article. To assess homogeneity, other specialized statistics are used (e.g., Cronbach alpha).7,8

The degree of reliability you need depends upon your sample size, the importance of the factor, and the consequences of error. You can substantially compensate for a smaller sample size by using more reliable (precise) measures.

Content Validity

Content validity is whether a scale includes items that cover the domain of interest. For instance, to assess community reintegration skills, a scale that includes basic self-care and mobility skills but excludes out-of-house mobility, driving, use of public transport, stairs and rough ground, speed, and endurance would lack content validity. Content validity cannot always be judged on simple logical grounds. For instance, to make a functional assessment shorter, domain sampling is required: one must select those items that are most frequent and critical for the patient population of interest, excluding redundant and infrequently relevant items.

Face validity is the degree to which a measure seems applicable or useful upon first examination. You need to look at an actual scale and perhaps try it out to know if it is practical and relevant to your purpose. If a scale is in widespread use, others have found it to be practical and to have at least face validity or sensibility.11 Whenever possible, therefore, you should use scales that others have found to work, rather than devising new ones. Not only does this save you work, but it makes your results more meaningful in the context of other studies.

Feinstein points out that considering face validity only begins an invaluable process of logical analysis and experience with an instrument.11 A measurement scale in medicine is properly judged not only according to formal psychometric criteria but also according to logical criteria rooted in experience and knowledge of physiological processes. Physicians reason from knowledge of underlying physiological mechanisms to a far greater degree than does the psychometrician, and this knowledge may alter or overturn the interpretation of behavioral or disability measures. The problem with face validity or sensibility is not that they are logically unimportant but that too often
validity is not assessed in other ways. Many measures in rehabilitation have gained widespread use even though
formal reliability and validity studies are nonexistent, scant, or unpublished, and normative data are unavailable.¹

Applicability of measures to the impairment group or diagnosis of interest is a fundamental issue in medical
rehabilitation. It is known that measures of self-care and mobility like the Barthel Index predict burden of care and
institutional discharge across most impairment groups commonly seen in rehabilitation.²,³,¹²,¹³ So some measures of
functional disability are applicable across a range of impairments. Alternatively there are assessment tools designed
for specific diagnoses or impairments (e.g., Minimal Record of Disability for multiple sclerosis,² tests of memory in
Table 4).

Internal Structure

Some measures may purposely be constructed to have several subscales. The issue is the internal structure
of the scale. Is the scale really measuring one thing or several? Any set of items that you treat as one scale should
cluster or correlate together. Advanced statistical techniques (e.g., factor analysis) are used to determine how to
group items together. They help you to empirically decide whether it is sensible to add a group of self-care and
mobility items together and call them “physical independence.”

Scales formed by adding up individual items scores should have the related property of additivity.⁷,⁸,²⁴ Additivity is the mathematical property that enables one to add up or average multiple (related) items and treat them
as if they were a single meaningful thing. This is no problem with equal interval scales (e.g., degrees Fahrenheit,
time), but it is a serious problem when you are trying to measure something that is naturally ordinal or imperfectly
interval. Measures of such things can be treated as equal-interval scales and meaningfully added together only
provided they can be shown to be additive and unidimensional. The point can be illustrated by an evaluation of food
supplies for a picnic. You can add up 2 apples and 10 oranges to get 12 fruits. But it is ludicrous to combine
strawberries and watermelons without converting to a common scale such as weight or number of “servings.” There
are several advanced techniques for determining the additivity of items, all based on some common scale (e.g.,
difficulty in Rasch analysis).⁷ In a Rasch or logistic scale,²⁴ an improvement of one point is equal in difficulty at all
points in the scale. Quantifying the relative difficulty of tasks is essential to quantitative functional assessment.

Range of Difficulty

Rehabilitation deals with people with a great range of abilities, from quadriplegic and even locked-in
patients to individuals attempting to return to work and those with sports injuries. Existing functional measures are,
however, sensitive to improvement only in a limited part of this range. Measures of physical independence (e.g.,
functional independence measure (FIM), Barthel) are optimized for the typical lower midrange of disabilities seen in
inpatient medical rehabilitation facilities. They are insensitive to improvements in speed and endurance among
“high” level clients who need to return to productive community activities (ceiling effects). They may also display
“floor” effects; they are, for instance, insensitive to the subtle improvements in neuropsychological function that
herald emergence from coma or minimally responsive states.¹⁴-¹⁷ It is critical to distinguish between lack of
improvement due to limitations of the measure and that due to lack of change in patient performance. Continued
funding of patient care may depend on this.

Criterion-Referenced Validity

Useful measures predict important events in the future (predictive validity). A measure of independence or
burden of care like the FIM should, for instance, predict hours of care required.²² Even a brief cognitive screening
assessment might be expected to discriminate between patients who have cognitive problems verified by a full
neuropsychological examination and those who do not and between patients who have better v poorer general
outcomes. A detailed neuropsychological examination ordered for a patient facing problems in return to work might
be expected to predict likelihood of success in intellectual v manual jobs. Clinical professionals demand high levels
of predictive validity to justify use of a test in rehabilitation planning for individuals. Most single measures cannot
provide this.

Concurrent validity is the degree to which a measure correlates with events occurring at the same time. Study of concurrent validity is practical for the clinical researcher. If two instruments purport to measure the same
thing, you can apply both of them to the same sample, analyze the correlation between them, and judge whether one
substitutes for the other. You may test whether an instrument that claims to assess a distinctive, new functional
domain in fact correlates so highly with other domains that the two are for practical purposes the same. Or you may
test whether a short version of an instrument correlates very highly with the full version. You may find that an
instrument correlates with things it should not, which would raise serious questions about interpretation (validity).
 Unless concurrent validity figures are very high (e.g., greater than r = 0.90), however, concurrent validity is no
substitute for predictive validity. For instance, Rancho Los Amigos Level of Cognitive Function scores correlate with
Disability Rating Scale (DRS) scores, but that does not mean that they both predict independent living outcomes
with the same degree of accuracy.

The availability of norms greatly enhances the interpretability of a measure. Norms on the basis of the
general population should be used if you think a functional limitation may be common in the age- and sex-adjusted
general population. Comparison to such a norm tells us how disabled or handicapped the person is, which is useful
for insurance and legal purposes but does not help with prognosis. With measures of patient performance at several
points in time, one can see progress toward the norm. But the progress itself can be charted without reference to the
norm, which is meaningful only if the patient has a reasonable chance of getting close to it. Often in rehabilitation,
there is no serious question that the patient’s function is far below that of the general population. Norms based on
other rehabilitation hospitals or on the basis of all patients hospitalized with the diagnosis are needed in medical
rehabilitation.

Generalizability (or external validity) is essential to useful functional assessment instruments. Ecological
validity refers to the degree to which a measure is meaningful or useful in the person’s real life outside of the clinical
setting. Ecologically valid measures are needed to assess the degree to which function assessed in therapy generalizes
to real life. For example, neuropsychological tests, designed to test for brain dysfunction, commonly lack ecological
validity. They are difficult to translate into terms of real world performances and goals. Functional limitations of
brain injured patients may decrease greatly in rehabilitation, whereas intelligence or memory scores change little. For
rehabilitative purposes, we might replace the term “ecological validity” with “outcome validity,” since the aim of
rehabilitation is to improve the person’s function and life not only outside the clinic but also after discharge and for a
sustained period.

Construct Validity

Measures of many important constructs (e.g., memory, motor function, independence) cannot be validated by
correlation with a single ideal criterion. Construct validity is the degree to which an instrument measures the
theoretical construct it was designed to measure. One tests construct validity by seeing whether a measure displays
a pattern of converging relationships (convergent validity) and an absence of confounding relationships (discriminant
validity). The expected pattern is defined by well-grounded, thought-out theory. Construct validation is a central,
ongoing process in the development of a science.

The notions of reliability and validity may be simply summarized. Imagine that you are testing a bow and
arrow in a range with a line of targets. You mechanically clamp down the bow, insert an arrow, and pull back the
same distance every time you shoot. Reliability is whether the arrows fall in a nice tight pattern. Validity is whether
you have aimed at the right target.

USES AND PURPOSES OF MEASUREMENT INSTRUMENTS

Level of Resolution

Measures legitimately differ in level of resolution or sensitivity to detail. Level of resolution figures
prominently in choice of measures for a given purpose. To understand the following explanation, you should be
familiar with the concept of “power” or precision of a research design.

Large Samples

A low-resolution instrument may be designed for mass epidemiological surveys of a large population (e.g.,
n > 1000). Such an instrument must have a very limited number of items, broadly applicable to all subjects. Such
instruments answer questions such as what percentage of hospitalized patients say they are unable to engage in their
primary activity a year later? The WHO coding scheme is an example of a scale designed for mass epidemiological applications.\textsuperscript{2}

**Moderate Samples**

To assess function in a small or moderately sized sample (e.g., minimum n = 20-100), moderately detailed instruments are needed. These instruments are suitable for studies of program outcomes and program evaluation. Examples include the Level of Rehabilitation Scale (LORS)\textsuperscript{21} and the FIM.\textsuperscript{22} The latter is more detailed and commonly used as part of a clinical assessment.

Both of these types of instruments rely on the notion of sampling of a domain of related items. This is the idea that you do not have to measure every item in a domain, but only the most outstanding, characteristic ones. The most frequently used functional measures in medical rehabilitation sample only the most common of routine human activities (e.g., measures of physical independence in self-care and mobility such as the FIM,\textsuperscript{22} the Program Evaluation Conference System\textsuperscript{23, 24} and the Barthel Index.\textsuperscript{3})

At about the same level of resolution, an instrument may be used for purposes of screening, that is, to identify problems that need special attention. For instance, a physician may do a quick cognitive examination and decide whether in-depth specialized neuropsychological examination is worth prescribing. A useful screening device needs to include the most frequent problems found in a patient population.

**Small Samples**

To study a small number of subjects (including an n of one), one needs the highest level of resolution or detail. This is found in instruments designed for use as a comprehensive clinical assessment. Traditionally, these instruments have been developed as forms and procedures in rehabilitation hospitals and include a mix of numeric measures and language. To show that such an instrument is valid, evidence is needed that choice of treatment strategy depends in practice upon results of the assessment and that outcomes to the patient are thereby improved.

Specific assessment devices have been developed for delimited aspects of function (e.g., standard neuropsychological tests, tests of visual acuity, gait analysis systems). Somewhat paradoxically, formal reliability and validity studies are most frequently found at this level. Though they may be reliable, these measures may have very limited generalizability to real world outcomes because of their narrowness.

**Sensitivity to Change**

Sensitivity to the change that is observed in rehabilitation settings is a minimum criterion for a measure of patient progress or outcomes in rehabilitation.\textsuperscript{4} A measure that has not been demonstrated to change in association with the rehabilitative treatment of interest, even in an uncontrolled study of improvement, is a poor candidate as an outcome measure. For instance, uncontrolled studies of improvement among aphasic patients have found that practical communication improves following participation in speech therapy, but aphasia impairment scores may not.\textsuperscript{25} The functional, situational measure is then essential as an outcome measure. This is instructive, for many rehabilitative interventions alter the social or physical environment and train persons to use their strengths to compensate for their weaknesses. Rehabilitation may be successful even if impairment measures do not change. Knowledge of sensitive, relevant measures is prerequisite to an enlightening controlled study.

On the other hand, if your program has a truly wrong effect on patient outcomes, you do not need an exquisitely sensitive measure. For instance, an expensive comprehensive rehabilitation program may claim to restore patients to independent living and productive vocational activity. These outcomes can be measured by simple, available measures (e.g., Tables 1, 2 and 5). Extreme sensitivity may not be needed because isolated improvements in function would not justify the great expense of these programs anyway.

A distinction is often made between patient tracking or monitoring instruments and one-time measures of patient status or comprehensive assessments. To be a practical tracking instrument, a measure needs to be: (1) clearly sensitive to change; (2) so inexpensive that it can be given repeatedly; and (3) not invalidated by test-retest learning effects. To be clinically useful, it also has to assess an aspect of patient function that is expected to be responsive to treatment. Measures also need specificity to the process or construct of interest. Global outcome measures, for instance, are affected by many factors other than rehabilitation, and hence are rather nonspecific in studies of rehabilitation effectiveness.
Practicality and Cost

Feasibility and cost are in practice dominant considerations. Though more information is generally desirable, costs increase in proportion to the amount of data collected. Your research will fail if you have so many items to measure that you cannot obtain complete data on your calculated sample size. Your research will also fail if in the interests of practicality you choose a simplistic, unreliable measure. So you must budget your resources. The main tool for doing this is precise definition of the research question.

MEASURES FOR BRAIN INJURY REHABILITATION

Tables 1 to 6 list illustrative measures in brain injury rehabilitation and comment on the contents and reliability of selected measures. Illustrative measures have been chosen in each domain relevant to brain injury rehabilitation. This makes the list more relevant to other areas of rehabilitation and conveys a sense of the complexity of the topic. For comprehensive review of measures for brain injury rehabilitation, see articles on assessment in brain injury\(^\text{25-29}\) and books on neurobehavioral measures\(^\text{6,30-33}\) and measures in rehabilitation\(^\text{3,13}\) and health care.\(^\text{12,34,35}\)

Most existing measures were developed before the WHO conceptual framework and consequently many have items that are at mixed levels or that attempt to reduce one level or another. For instance, global scales assume that outcomes are wholly determined by the impairment, ignoring the person’s compensating strengths and socio-environmental supports. Predictive validity is limited by the simplistic assumptions of these scales.

Coma and Minimally Responsive Patients

(Table 1)

The Glasgow Coma Scale (GCS) is the standard instrument. In the first 24 hours to one week post-injury, it is the best simple predictor of general outcomes.\(^\text{15,16,17,31}\) Predictive utility declines somewhat after the acute stage. Addition of brain stem reflexes\(^\text{38}\) and auditory evoked potentials may\(^\text{14}\) improve outcome prediction. The GCS is fairly insensitive to subtle changes in responsivity in long-term comatose patients.\(^\text{17,39}\)

More refined instruments involving more neural response systems have recently been developed to track recovery of minimally responsive patients in coma stimulation programs.\(^\text{39}\) The new Coma Recovery Scale is more sensitive than GCS to subtle changes in responsivity of long-term coma and vegetative patients.\(^\text{17}\) It has high concurrent validity with GCS and DRS and higher predictive validity than either during acute rehabilitation.

TABLE 1

<table>
<thead>
<tr>
<th>Measures of coma and global functioning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coma and Near Coma</strong></td>
</tr>
<tr>
<td>- <strong>Glasgow Coma Scale (GCS)</strong>(^\text{14,17,36,37,64}) Simple scoring of three visual, motor and verbal responses (impairments). High reliability.</td>
</tr>
<tr>
<td>- <strong>Coma Recovery Scale (CRS)</strong>(^\text{17}) New. Fairly high reliability. Assesses a wide range of behaviors— arousal, visual, auditory, tactile, communication, motor responses.</td>
</tr>
<tr>
<td>- <strong>Others</strong>. Rappaport Coma-Near Coma. Western Neuro-Sensory Stimulation Profile (WNNS).(^\text{39})</td>
</tr>
<tr>
<td><strong>Global Measures</strong></td>
</tr>
<tr>
<td>- <strong>Glasgow Outcome Scale (GOS)</strong>(^\text{16,18,38}) A single 3-point or 6-point (EGOS) rating. Good reliability in general postneurosurgical sample; less reliable for restricted range of severity (e.g., rehabilitation patients, greater than 3 months after injury).(^\text{41})</td>
</tr>
<tr>
<td>- <strong>Rancho Los Amigos Level of Cognitive Function</strong>(^\text{16,40}) Good reliability at levels 1-4; fair at higher levels due to confusions (e.g., scoring of aphasic but otherwise cognitively intact patients).</td>
</tr>
<tr>
<td>- <strong>Disability Rating Scale (DRS)</strong>. Frequently used. Eight general functional items rated on 0-4 scale. Reliability: good, better than Rancho.(^\text{18}) More sensitive than GOS for rehabilitation.(^\text{41})</td>
</tr>
<tr>
<td><strong>Others</strong>. See also screening measures (Table 4).</td>
</tr>
</tbody>
</table>


Global Measures (Table 1)

Early research on outcomes after brain trauma attempted to assess patient function in terms of a single, global assessment of level. The Glasgow Outcome Scale (GOS) is the earliest of these. It was used to assess global outcome in pioneering long-term neurosurgical outcome studies in the 1970s. Its virtue is its simplicity. Its sensitivity to rehabilitative interventions has not been established, but it is still valuable for comparison with neurosurgical studies of long-term outcomes after brain trauma.

When patients are clearly out of coma, the relevant dimension or construct is global environmental responsibility. The Rancho Los Amigos Level of Cognitive Function Scale is based on a description of typical stages of recovery that patients experience after severe traumatic brain injury. It is easy to use, clinically useful as an initial description of the global cognitive level and function relevant to brain injury, and sensitive to change in rehabilitation. Moderate levels of predictive validity have been shown. Though reliability (and consequently, validity) is limited, its widespread use makes it a standard. It is valuable to quickly convey a sense of the patient level and to determine if one’s study sample is roughly comparable to that of another study.

Rappaport’s disability rating scale is more reliable than Rancho due to its multiple, simple items. It has been shown to be more sensitive to change than the GOS in both acute and post-acute rehabilitation programs. Among all global measures here, its reliability and sensitivity make it the best candidate for a singular global outcome measure. The DRS cannot be expected to be sensitive to the effects of treatments designed to decrease specific disabilities (e.g., behavioral programs) or to restore clients to specific roles (e.g., return to work). It is clearly valuable to characterize sample severity and to compare results to others in the literature.

TABLE 2
Disability Measures

<table>
<thead>
<tr>
<th>Basic Disability Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Functional Independence Measure (FIM)¹²</td>
</tr>
<tr>
<td>• Program Evaluation Conference System (PECs)¹³,¹⁴</td>
</tr>
<tr>
<td>• Others.¹,³,⁴³</td>
</tr>
<tr>
<td>Disability Management Skills</td>
</tr>
<tr>
<td>• Items in instruments dealing with understanding of disability, adjustment to disability, self-medication, safety management, etc.²⁷,²⁸,⁶¹ Reliability unclear to moderate.</td>
</tr>
<tr>
<td>Instrumental ADLs and Community Skills</td>
</tr>
<tr>
<td>• Community Living Assessment Scale (CLAS),⁶¹</td>
</tr>
<tr>
<td>• Others.²⁷,²⁸</td>
</tr>
<tr>
<td>Social Skills, Behavioral Problems and Psychosocial Disability</td>
</tr>
<tr>
<td>• Scales used in TBI include: Katz Adjustment Scale (Relatives’ Form) and Bond’s Social Scale⁴⁴; ratings of prosocial behavior and counts of maladaptive social behavior;²⁷,²⁸ Child Behavior Checklist and Vineland Adaptive Behavior Scale.⁷⁷ Several others.²⁹,⁶²</td>
</tr>
<tr>
<td>• Neurobehavioral Rating Scale.³²</td>
</tr>
<tr>
<td>• Frequency of behaviors. Count of behavioral symptoms of emotional distress and maladaptive social behaviors.²⁷,²⁸</td>
</tr>
<tr>
<td>• Others.²⁹</td>
</tr>
<tr>
<td>Vocational Skills</td>
</tr>
<tr>
<td>• Agitated Behavior Scale.⁷⁸ Fourteen items with moderate internal consistency (alpha=0.8).</td>
</tr>
</tbody>
</table>
• **Functional Assessment Inventory (FAI)**.\textsuperscript{47} Factor analysis revealed eight scales: cognitive function, motor function, personality and behavior, vocational qualifications, medical condition, vision, hearing, and economic disincentives.

• **Others.** Employability ratings.\textsuperscript{79} Many others.\textsuperscript{46}

---

**Disability Measures (Table 2)**

**Activities of Daily Living**

Most brain-injured patients enter acute rehabilitation hospitals with disabilities in basic self-care and mobility skills, and therapeutic interventions are directed at these ADLs. At this level of care, progress in these basic disability measures is the most frequently used means of marking patient progress.

The FIM is currently, the most-used measure of disability reduction, being used in several hundred medical rehabilitation hospitals. The FIM clearly predicts burden of care in a multiple sclerosis population (criterion validity),\textsuperscript{22} and work on its validity in other populations is ongoing. Excellent norms are available from the Uniform Data System for participating rehabilitation hospitals, but reliability data remains unpublished. The PECS is useful both for evaluation and in team conferences. The Glasgow Assessment Scale deserves mention there as well as in Table 5. There is a surplus of competing scales in this area.\textsuperscript{1,43} Additional reliability and validity work is needed to apply these to brain injury, but use of a nonstandard measure of physical independence and mobility is no longer justifiable in rehabilitation research.

*Instrumental ADLs* and *community skills* are critical to restoration of the individual to a productive life. These measures are highly suitable as outcome measures for post-acute programs, higher level patients and, perhaps, long-term follow-up studies.

**Social Skills, Behavioral Problems and Psychosocial Disability**

Psychosocial and behavioral problems are well-established sequelae of head injury. Different measures of psychosocial disability and behavior problems may not intercorrelate well, so this domain may readily be several functional domains.\textsuperscript{44} Behavioral competency (disability) needs to be distinguished from social reintegration in the actual social environment of the person (handicap).

Levin’s Neurobehavioral Rating Scale\textsuperscript{45} is used to track psychiatric patients and sometimes to describe the nature of neurobehavioral problems with moderate predictive validity. Since it is comprised of judgmental ratings, procedures to lessen bias (e.g., blinding, meticulous training of raters) would be needed to use it as an outcome measure. Actual counts of the frequency of behavioral problems could potentially be more reliable and less subject to bias. The problem is how to aggregate counts of heterogeneous behaviors into more generic dimensions.

The literature of *vocational* rehabilitation has several instruments with promising application to brain injury.\textsuperscript{6,46} The Functional Assessment Inventory is one of the simpler, more embracing of measures for use in rehabilitation settings. Clear content and concurrent validity have been shown.\textsuperscript{47}

**Communicative Function (Table 3)**

Standardized measures of communicative impairment and disability are presented in Table 3. These were developed before the WHO framework, and they sometimes blur distinctions between impairment and disability. Instruments for assessment of aphasia and linguistic function are listed first. These assess aspects of linguistic function and largely more psychosocial, ecological, and to some degree even cognitive aspects of communication. They greatly underestimate the communicative success (handicap) of adult aphasics in their natural settings but overestimate the communicative success head-injured clients.\textsuperscript{25} Moreover, impairment measures (including these batteries) may be insensitive to improvements made in real-life communicative situations associated with treatment.
(e.g., via compensatory strategies) and therefore lead to false conclusions that rehabilitative interventions are ineffective.25, 48, 49

The above batteries are difficult to use as measures of treatment goals or success.25 For example, a goal such as “error rate in this specific test will decrease to 5%” can be used as a short-term clinical goal but has a questionable relation to real-life needs. In contrast, a goal such as “client will communicate with close caregiver regarding self-care needs 95% of the time” has face validity as a real-world outcome measure. In sum, these instruments have limited or unclear utility as outcome measures but are very useful indices of processes during therapy.

Sarno’s Functional Communication Profile (1969) is primarily a measure of communicative disability of different types.50, 51 Many of its items appear suitable as measures of long-term therapeutic goals (e.g., understand verbal directions). Haffey and Johnston provide an example of an ecologically relevant measure in terms of frequency of problems in various real-world communicative situations; reliability, however, is unknown. Reliable measures of motor speech disorders (dysarthrias) are available, including not only assessments of type of dysarthria but also the communicative disability produced by decreased intelligibility. In general, more work needs to be done on the reliability and validity of measures of communicative outcome at the level of disability and handicap.

### TABLE 3

*Measures of communicative function*

- **Boston Diagnostic Aphasia Examination.** Comprehensive, lengthy assessment of “components of language.” Experienced examiner needed. Reliability unclear.
- **Other measures of linguistic dysfunction:** Multilingual Aphasia Examination (MAE). Token Test. Porch Index of Communicative Ability. Western Aphasia Battery, Word Fluency.
- **Functional Communication Profile (FCP)** Forty-five items on verbal and written comprehension and expression for adult aphasics. A few functional impairment items, many on whole communicative tasks (disabilities). Reliability fair but predicts functional communication.
- **Communicative Abilities in Daily Living (CADL).** Broader range of communicative behaviors than FCP. Uses structured conversation and role playing (e.g., doctor’s visit). Simple 3-point scoring. Good reliability.
- **Others** Assessment of Intelligibility of Dysarthric Speech. Good reliability and normative data. Summarizes communicative function by “communication efficiency ratio.”

### Measures of Cognitive Function (Table 4)

Measures of central nervous system dysfunction are central to brain injury rehabilitation. Neuropsychological measures are the primary measures here. Brain functions are highly interconnected, and traumatic brain injury is more commonly diffuse than focal. With insightful choice of measures, deficits and compensating strengths can often be identified. Such information is critical to planning a remediation program.

Difficulties with memory are typical with traumatic brain injury. The Wechsler Memory Scale-Revised (WMS-R) is the most frequently used instrument to measure impairment. Good-to-excellent data is available on its content, predictive, and construct validity. Current work on memory in brain-injured patients distinguishes several types of memory subsystems (e.g., verbal and visual immediate and delayed). Knowledge of affected relatively preserved memory systems is vital to effective remediation.

### TABLE 4

*Cognitive function measures*

Memory Impairments
- **Wechsler Memory Scale-Revised (WMS-R).** Subscales for Verbal, Visual, Delayed and General memory. Well standardized and normed. Moderate to high reliability depending on subscale.
• **Busche Selective Reminding Test.**\(^{32}\) Verbal list learning task. Attempts to distinguish between encoding, memory, and retrieval. Moderate test-retest reliability.


• **Galveston Orientation and Amnesia Test (GOAT).**\(^{53}\) Widely used measure of post-traumatic amnesia.


**Memory Disabilities**

• **Rivermead Behavioral Memory Test (RBMT).**\(^{54}\) Attempt at memory test related to everyday memory problems. Fairly good reliability. Questionnaires/ratings of memory problems in everyday life (e.g., Sunderland questionnaire).

**Attention and Concentration**

• **Wechsler Digit-Span.**\(^{30,33}\) Popular. Easily administered. Well normed.


• **Paced Auditory Serial Addition Test (PASAT).**\(^{30,33}\) Very sensitive to information processing deficits. Reliability and validity not well studied.

• **Others.**\(^{33}\) Vigilance tests, cancellation tests, symbol digit modalities.

**Perceptual Impairments**

• **Judgement of Line Orientation.**\(^{30,33}\) Sensitive test of visual-spatial processing. Highly sensitive to posterior right brain damage (right parietal); no motor requirements. High reliability and validity.

• **Hooper Visual Organization Test.**\(^{30,33}\) Test of visual analysis and synthesis of elements into a whole. Reliability: satisfactory. Moderate to good construct and concurrent validity.

• **Other visual.**\(^{30,33}\) Raven’s Progressive Matrices, Visual Form Discrimination, Facial Recognition Test, prosopagnosia (inability to recognize familiar faces), color agnosia, Line Bisection, visual inattention.

• **Auditory perceptual impairment.**\(^{30,33}\) Wepman Auditory Discrimination Test; Seashore Rhythm Test; Speech-Sounds Perception Test.

**Concept Formation and Executive Functions**

• **Wisconsin Card Sorting Test (WCST).**\(^{30,33}\) Sensitive to impaired concept formation. Moderate to good sensitivity to frontal lobe dysfunction. Reliability not adequately researched.

• **Category Test.**\(^{30,33}\) Good test-retest stability. Valid and sensitive measure of generalized cerebral dysfunction but data limited. Very time consuming to administer.

**General Intelligence**

• **Wechsler Adult Intelligence Scale-Revised (WAIS-R).**\(^{30,33}\) Most frequently used measure of intelligence. Well standardized and normed. High reliability. Good content, criterion and construct validity. Trained staff (e.g., psychologist) should administer.

• **Others.**\(^{30,33}\) Raven’s Progressive Matrices. Very good sensitivity. Susceptible to reasoning and visual perceptual deficits. Good for nonverbal population.

**Screens**

• **Barry Rehabilitation Inpatient Screening of Cognition (BRISC).**\(^{55}\) Mini-Mental State, Strub and Black.\(^{33}\)

---

The Galveston Orientation and Amnesia Test (GOAT) is a more reliable and valid measure of post-traumatic amnesia than a simple judgment of duration of PTA.\(^{32,53}\)

The Rivermead Behavioral Memory Test has shown good correlations with functional memory, problems rated by therapists in clinical setting.\(^{54}\) More independent evaluation of the instrument is needed to fully assess its utility as an ecologically valid measure of memory disability.

Dysfunctions in **attention** and **concentration** are also common after head injury. Problems in **perception** are frequent both in traumatic brain injury and stroke. Visual perception is most widely studied. Existing instruments are, by and large, impairment measures. The few attempts at measures of the impact of attentional and perceptual impairments on meaningful daily task performances lack reliability studies. This will hamper study of rehabilitative interventions to remediate these problems.

Diffuse axonal injury and frontal-temporal lobe dysfunction are hallmarks of traumatic (automobile accident) brain injury. These impairments require measures of **concept formation** and **executive integration.**\(^{33}\)

General intelligence measures are also relevant, although intellectual skills may well be preserved after frontal lobe
injury. Training interventions in rehabilitation may be expected to have only a small impact on these impairments but a larger impact on disability.

Cognitive screening instruments in Table 4 range in length from 10 min to 30 min (the BRISC\textsuperscript{55}) or more.\textsuperscript{33} These instruments do not have the reliability coefficients needed to distinguish between types of brain dysfunction for individuals, but they should distinguish general cognitive level or may be used as a guide to likely problem areas.

Many neuropsychological tests were originally validated as tests of presence and type of brain damage. They were not developed to be functional outcome measures. Their clearest use is in understanding processes of recovery and function. More work on how more effective treatment strategies can be chosen, depending on neuropsychological test results, would be valuable.

**Measures of Handicap and Real World Outcome**

(Table 5)

Table 5 presents alternatives for measurement of ultimate outcomes-handicap. The six WHO dimensions of handicap have not been well tested for reliability and validity, though they are an invaluable conceptual resource.\textsuperscript{2} The CHART,\textsuperscript{56} an adaptation of the WHO taxonomy for rehabilitation, is new, and reliability and validity information should be published in the next year. It was originally validated for individuals with spinal cord injury. Most commonly, occupational handicap is measured by ratings of full or partial return to work. Handicap in orientation to surroundings overlaps with global measures of cognitive function. Handicaps in “physical independence” (including self-care) and in mobility seem conceptually to be very general disability measures and, empirically, they correlate highly with self-care and mobility disability measures, respectively, whereas other handicap dimensions may not correlate well with disability measures.\textsuperscript{37}

Social integration is a dimension of handicap, for social life is highly dependent on the social environment. Methods of measurement used in outcome studies include psychosocial rating scales\textsuperscript{42} and ratings of the frequency of social activities in actual life,\textsuperscript{58,60} and the Community Living Assessment.\textsuperscript{61} Except for the last, reliability data is not easily available for these.

Alternative methods of assessing rehabilitation outcomes have been suggested.\textsuperscript{62} Haffey and Johnston have suggested that perceived stress is the relevant dimension for this valuation, as studies have documented high stress levels in brain injury survivors and especially in their families.\textsuperscript{27, 28} Based on findings with such ratings in the general population, ratings of stress are likely to have only fair reliability and stability over time.\textsuperscript{6}

The Glasgow Assessment Schedule (GAS) is very promising. It is much more sensitive to improvement than is the GOS and much more descriptive of dysfunctions following brain injury.\textsuperscript{63}

**TABLE 5**

**Measures of handicap and general outcomes**

<table>
<thead>
<tr>
<th>Broad Multidimensional Handicap Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WHO Handicap Codes.</strong>\textsuperscript{2} Ratings of general disadvantage (handicap) in six survival roles: orientation, physical independence (self-care ADLs), mobility, occupation, social integration, economic sufficiency. Reliability and validity unclear.</td>
</tr>
<tr>
<td><strong>Craig Handicap and Reporting Technique (CHART).</strong> Adaptation of WHO scheme for rehabilitation.</td>
</tr>
<tr>
<td><strong>Others.</strong> Role attainment ratings, Willer’s Community Integration Questionnaire.\textsuperscript{60}</td>
</tr>
</tbody>
</table>

Other Approaches Congruous with WHO Approach

- **Disability scales at follow-up.** FIM or others in Table 2 applied at follow-up may measure mobility and physical independence handicap (Table 2). Reliability: good. Good connection to abilities nurtured in the rehabilitation setting. The most common basis for program evaluation in rehabilitation.
- **Social integration.** Community Integration Questionnaire,\textsuperscript{60} Katz Adjustment Scale, or others in Table 2.\textsuperscript{62} GCS, Rancho, GOAT, perhaps other global cognitive measures (Table 1).
- **Vocation and economic self-sufficiency.** Percentage of time employed, wages, others.\textsuperscript{83} Blishen Quantitative Social Economic Index.\textsuperscript{84} |

Other Approaches to Outcomes
• **Glasgow Assessment Schedule (GAS).** Forty items, 3-point rating scale. Six subscales: personality change, subjective complaints, occupational functioning, cognitive functioning, physical examination, and activities of daily living. Very good interrater reliability. Correlates with GOS. Mix of narrow disability/impairments with broader measures.

• **Activity Pattern Indicators (APIs) at follow-up.** What persons do with their time: type of activities, frequency, duration, and diversity. Reliability: good. Norms and validation needed. Laborious. Part of comprehensive system (Rehabilitation Indicators).

• **Global status measures.** See DRS and other global ratings in Table 1.

• **Stress, psychological well-being, quality of life.** Bradburn. Questionnaire on Resources and Stress. Ratings of problem severity/stress related to different functional domains.

• **Health status indicators.** Sickness Impact Profile.

• **Economic approaches.** Aggregate monetary and near-market costs and benefits into dollar terms.

---

**Other Domains (Table 6)**

Structural measures of brain damage (e.g., CT scans) reveal gross brain abnormalities but generally are not sensitive to function. CT scan data do not predict outcomes as well as GCS or age. Measures of brain metabolism (e.g., positron emission tomography) may prove to be better predictors. Assessment of pathologies in isolation describe little about cognitive function, but are important in many studies (e.g., on pharmacological interventions).

Rehabilitation patients commonly, perhaps usually, have comorbidities. Patient progress in hospital settings may be due to resolution of infection, taking a cast off, or control of blood sugar level, change of medication, or reduction of hydrocephalus. Heterogeneity inhibits study of these factors using conventional methodologies. Attempts have been made to use measures of trauma severity in Table 6 (ISS, AIS) to assess severity of comorbidities, but though these scales predict probability of death, survivors with the greatest trauma severity scores do not tend to become the most disabled. The GCS is a better predictor of head injury outcomes.

*Sensory problems* (vision, hearing, touch, balance, etc.) are found in a significant fraction of a brain injured population, usually associated with higher brain dysfunction. Instruments for assessment of sensory impairment are well developed (e.g., touch, vision and hearing), but assessment of associated disabilities lag.

Patient history, length of illness, and demographic variables are usually treated as descriptors of the sample population, but they are also known to affect outcomes. As shown in Table 6, a few attempts have been made to develop measures of the social and physical environment, but far more work is needed here. These variables need to be added to measures of degree of dysfunction to estimate case severity.

**TABLE 6**

**Other measures**

---

**Structural and Functional Brain Impairment and Pathology**

- Structural brain damage. Diagnoses from CAT scan, angiography, etc.
- Functional brain impairment. Evoked potentials, PET scan, etc. Reliability varies. Predictive validity not proven, but potentially good.
- Pathologies. Neurotransmitter levels, cellular abnormalities, etc.

**Comorbidities and Organismic Health**

- Comorbidities. Second and later-listed diagnoses, innumerable more specific medical and nursing impairment measures.

**Musculo-skeletal Function**

**Sensory Impairments and Disabilities**

- Tests of impairments in vision, hearing, touch, proprioception, smell, balance.

**Other Measures of Severity Before Rehabilitation**

- Pre-injury history. Reliability in records varies greatly.
• Information from acute injury.\textsuperscript{64} Trauma severity measures: the Abbreviated Injury Scale (AIS) and Injury Severity Scale (ISS).\textsuperscript{56, 65, 90}

• Chronicity. Onset-admission days. Very high reliability.

• Other measures of prerehabilitation severity correlating with poor outcome. Duration of coma and, in some samples, of post-traumatic amnesia (PTA) predict recovery.\textsuperscript{58} Reliability: poor to fair in hospital records; far better with sequential GOATs.

Environmental Measures

• Work disincentives, presence of family, marital status, Family Help Available,\textsuperscript{91} Family Adaptability and Cohesion Evaluation Scales (for family structure), Social Network List (for behavioral support received), Social Support Questionnaire, Perceived Social Support from Friends and Family, Questionnaire on Resources and Stress, the Canadian Health and Activity Limitation Survey,\textsuperscript{29} McMaster Family Assessment Device.\textsuperscript{92}

_CHOOSING MEASURES_  

**What to Do if You Don’t Have a Theory**

Traditional methodology makes few assumptions regarding treatment processes. However, without such assumptions, one would have no basis for choosing content areas from among measures in the preceding list, or indeed from among the universe of measures in all of medicine, public health, and psychology. The most recent work in research methodology emphasizes the need for at least simple theories of treatment in studies of effectiveness.\textsuperscript{68} At minimum, one needs to specify:

1. Approximate kinds of outcomes that are likely to be affected by the treatment of interest. This can be done on the basis of knowledge of rehabilitative processes and past studies.
2. Approximate magnitude of effect. If one does not have a rough idea of the magnitude of effect expected, one has no basis for choosing a sample size or for deciding how sensitive a measure is needed.
3. Expected timing of any effects. Should effects be evident immediately after administration of the treatment, after a few weeks, or only years later?

If you can’t answer the above three questions, you need to think about the research design more. Read the first article in this series to sharpen your sense of how to ask research questions.\textsuperscript{68} Study previous research. Whatever you are planning to do, something similar has probably been done before in another field. If you cannot answer one of these questions, that question becomes your first study. Such a study might assess a wide array of measures, at different time periods, and use statistical techniques to distinguish between dimensions (groups of variables) that are highly correlated with the treatment of interest, and those which are not. Such dimensions are the outstanding input and output variables for the next step, a controlled study.

A common mistake made by clinicians and administrators is to expect very large treatment effects.\textsuperscript{70} This expectation is implicit in studies that use gross outcome measures and a small sample (less than 30). Controlled studies of inpatient hospital medical rehabilitation programs, for instance, have shown only moderate or even modest effects\textsuperscript{58, 71} making it unlikely that you will find a massive group treatment effect. We must design most of our research with the expectation that fairly sensitive, specific measures will be needed in order to have sufficient power to quantify effects. Fairly sensitive measures (in an otherwise powerful research design) are needed to give us the ability to discover new treatments that, when fully developed, are effective and cost-effective.

Without a strong theory, clinically relevant studies should concentrate on study of factors that are clearly observable by ordinary persons. Use your skills in examining the measure, not looking for subtle changes in the patient.

**WHAT TO DO IF YOUR THEORY IS BIGGER THAN YOUR SAMPLE SIZE**

Comprehensive medical rehabilitation programs involve medical, nursing, psychological, direct learning and educational interventions with the patient and family, and modification of the physical environment. The number
and type of interventions are so great and complex that it is difficult to specify a generic treatment theory to cover all or even most patients. Discussion of how to study these complex programs in their totality is beyond the scope of this article. See Haffey and Johnston\textsuperscript{27,28} for an example of a database and methodology designed to untangle complex relations in a head injury rehabilitation program. Most readers would do well to study more delimited problems.

Some rules of thumb can, however, make it possible for you to study aspects of comprehensive rehabilitation programs. First, strong effects are needed to justify the expense of operational programs to the public, so simple outcome measures should be useful for policy, program evaluation, or estimating effect size. A sustained effect on generalized function is an expected outcome, so measures of disability and handicap in the person’s real life after discharge are needed. The rehabilitation plans sent to third party payers define in practical terms the level and type of effects expected of these programs. Pick measures from National TBI Model Systems Data Set, to be released by the National Institute of Disability and Rehabilitation Research in 1991, whenever possible, so that comparative data will exist.

Another rule of thumb is to base your measure on the level of difficulty or stage of recovery of patients in your sample. Consider the specific life problems (handicaps) that will be faced by program participants shortly after discharge. Decreased burden to society, and in some cases, medical stabilization and reinstitutionalization are issues. Basic disability measures (Table 2) are well established as indicators of the improved function that occurs in hospital-based rehabilitation programs. These measures are related directly to the aims of many interventions by therapists and indirectly to good medical management and nursing. For post-acute programs, only a minority of clients may have dependencies in activities of daily living (ADLs).\textsuperscript{59} Key measures here include household and community skills, behavioral problems, social integration measures, and certainly paid and unpaid productive activities. For residential post-acute programs, total daily support hours is a critical outcome dimension.\textsuperscript{72}

The most outstanding methodological difficulty here is distinguishing improvement specific to the intensive rehabilitative interventions from that due to the natural healing and learning that would occur with unspecialized care. A control group is the primary solution to this problem. One might vary frequency or duration of treatment sessions among groups and analyze relationships to outcomes. Measurement of outcomes most associated with specialized treatments strengthens even a controlled study.

IF YOU HAVE A LITTLE THEORY

We already know enough about many areas of rehabilitation to specify the needed simple theory of treatment.

Learning-Based Interventions

It is commonly said that learning is the \textit{sine qua non} of rehabilitation. If the process you want to study involves repeated practice by the patient, you are in the domain of learning theory and conditioning. You can therefore use learning theory to select measures. For instance, training an individual in the specific skills needed for subsequent success in role performances, and then training for generalization and flexibility, is more likely to be successful than nonspecific training that merely hopes for generalization to generic life demands. There is a forgetting curve, and learned skills deteriorate without practice. Perhaps the biggest difficulty in training brain injury survivors has been the lack of generalizability of what has been learned. We know that certain teaching strategies increase the likelihood of continued use.\textsuperscript{73,74} If we are studying practice-based interventions, we should choose a performance measure related to the skills being specifically trained. A sensible research strategy would be to first study fairly short-term outcomes (e.g., success in improving performances in the clinic) and then to study longer-term outcomes (generalization and long-term maintenance in the community after discharge). If these treatments are ineffective in the short-run, they will not be effective in the long-run.

If one is studying practice-based interventions, important effects are usually at the level of disability or performance rather than at the level of impairment. Impairment measures are best used to understand why the intervention worked, or did not, rather than as an outcome measure. The role of impairments is of course much more important in study of more “medical” interventions. For instance, in a study of improved methods of management of hydrocephalus, decreased ventricular size would be a necessary measure, but long-term functional outcomes would still need to be assessed.

Brain Recovery and Healing Processes
Ill-understood healing and compensatory processes are involved in long-term recovery after brain injury. Bach-y-Rita has argued persuasively that neuroplastic changes contribute to the improvement noted in brain injured patients. Undamaged and less damaged regions of the brain may assume functions previously performed by damaged areas. Although novice researchers may be attracted to this topic, a very high level of rigor and sophistication is required here. To legitimately examine these processes, the most precise of brain impairment measures (e.g., positron emission tomography) and the most discriminating of neuropsychological tests would be required.

**Cognitive Remediation**

Cognitive remediation is not cognitive restoration but an attempt to augment compensatory processes of the recovering and relearning person. Testing whether a new technology such as this has any effect even in laboratory or single-setting tasks is scientifically defensible. However, it is also valid to ask whether the technology produces a practical reduction in disability. The effectiveness of interventions involving environmental aids is highly studyable. Interventions designed to provide internal cognitive orthoses by compensatory learning (e.g., teaching the person to habitually repeat what was just said to avoid forgetting it, using procedural memory to compensate for deficiencies in declarative memory) should also affect specific functional domains. To do this research, first try to choose established disability measures or even items (cf. tables 2-6) logically related to the intervention used (e.g., items from the Rivermead Behavioral Memory Test might be used to study the effectiveness of training involving reminder systems). The dearth of reliable, ecologically valid measures of cognitive disability may impede your research, and you may find that your attempt to test the effectiveness of a treatment “merely” answers one of the three questions above (e.g., it may exclude the possibility of a large effect and suggest maximal effectiveness in certain ways). In this research, you will need impairment measures (e.g., of cognitive function) to understand why the remediative intervention did or did not work.

How you define “success” will influence your results. If “success” is defined only as return to work, success will be mild to moderate. If success is stratified (i.e., for one subgroup, goals are return to work, whereas for another, the goal is return to minimally supervised living), success will improve. It is often incorrect to apply a single outcome criterion to all individuals in a cognitive remediation program, or for that matter, in other rehabilitation programs. Goals are individualized. If as a clinical researcher you do not have the resources to do outcome research with long-term follow-up, you can still do invaluable work on rules for selecting goals for outcome research.

**DATABASES FOR FUTURE STUDIES**

It is not uncommon for enthusiastic rehabilitation professionals to attempt to create a generalized database to study questions relating to outcomes and/or processes. Registries are common in medical care. If you want to create a large general purpose database for future (unknown) studies, the best advice we can give you is, stop and think first! Collecting all sorts of data on all patients without a hypothesis is expensive and wasteful. Generate a research hypothesis and make sure that you include the measures needed to test the hypothesis.

One type of general purpose database is a registry designed to assist future retrospective studies and grant proposals. Contents would typically include: full patient name, identification numbers (medical record number, perhaps social security number), all diagnoses (accurately coded), enough information to contact the patient for future follow-up (addresses, at least two phone numbers plus the name and number of “someone who will always know where you are”), age, sex, occupational status, marital status, discharge destination and addresses, principle physician(s), admission and discharge dates, and discharge status.

The main use of most databases is descriptive. Longitudinal databases can be excellent vehicles for study of predictive relationships. Here the key is to decide on the critical outcome issues and the most likely predictors. If necessary, sacrifice variables related to treatment effectiveness, for observations of improvement, however numerous, do not tell what is responsible for that improvement. Causation is determined from research designs involving departures from base-line expectations or control groups, not from measures alone.

Another more complex database is one designed for program evaluation. Start with the outcomes that would be found in marketing claims (as is required by CARF regulation) and items required by regulations. Most program evaluation designs collect disability measures that are the frequent issues in the progress of patients seen. Though these databases exclude many details of rehabilitative processes and outcomes, they are the backbone of rehabilitation
outcome studies. The next step up in complexity is to include process measures necessary to make the database useful as a quality assurance system.\textsuperscript{27,28}

Whatever the level of complexity of your planned database, try to choose measures for which comparative data are available. Most of all, we advise you to specify the precise use for your database.

**DISCUSSION AND CONCLUSIONS**

Reliable and valid measures are essential to the progress of any scientific field. Biometric and psychometric principles are highly applicable to medical rehabilitation. Choice of measures for research depends both on formal reliability and validity data and on assumptions regarding treatment processes, a matter in which clinical experience can help immensely.

Our review of instruments has revealed some important features of measures for brain injury rehabilitation:

- There are a few areas with a surfeit of more or less similar, widely used scales (e.g., basic disabilities in activities of daily living, areas of neuropsychology, global outcome measures).
- Excluding the best-developed neuropsychological instruments, formal data on reliability, validity, and norms is absent or very difficult to obtain even for commonly used scales in medical rehabilitation.
- Other functional domains have a paucity of scales with anything but face validity (e.g., ecologically valid outcome measures, measures of cognitive disability, measures for use as long-term outcome goals by therapists, and measures of handicap, of the environment, or of situational difficulty). There is also a need to develop measures that are clinically useful, i.e., economical and proven to result in more logical rehabilitation plans and better outcomes.

Full development and validation of a new measure is extremely time-consuming and not for the novice, but it is practical and useful for clinical researchers to provide critical data on the reliability and validity of instruments already in use. Comparison between alternative instruments and techniques is also practical and useful. Practitioners will need to collaborate with a professional in psychometrics or biostatistics to do these studies.

More standardization in measurement is needed. We do not need more home-baked scales lacking basic reliability and validity coefficients. We do need more well-developed measures that, because of their proven reliability and validity, can become standards in the field. No small set of simple instruments will do, given the number and complexity of impairments, disabilities, stages of recovery, and environments. We need a full tool kit of standardized measures, optimized for different functional domains.

Medical rehabilitation is ripe for measurement standards. The American Psychology Association has had measurement standards since 1966\textsuperscript{4}. The American Physical Therapy Association and Joint Task Force of the American Academy of Physical Medicine and Rehabilitation and the American Congress of Rehabilitation Medicine are now attempting to write such standards, using psychometric standards as a model and starting point. Increased effort at formal measure development and validation will put the measurement of “function” on a scientific basis and propel physical medicine and rehabilitation into a leadership position in medicine.

**ADDENDA**


One database unintentionally omitted from Part X in this series is RECAL, a database of approximately 20,000 bibliographic references in prosthetics and orthotics from the past 10 years. It is available from Recal Information Services, University of Strathclyde, National Center for Training and Education in Prosthetics and Orthotics, Curran Building, 131 St. James Road, Glasgow G4 OLS, Scotland.

**REFERENCES**

86. Bradburn N: Structure of Psychological Well Being. Chicago, Aldine, 1969. 10 item true-false questionnaire that is the most widely used instrument for assessment of psychological well being. Norms for every adult age sex race group. Three subscores: positive affect, negative affect, balance score.