Interpreting Change Scores of Tests and Measures Used in Physical Therapy

Over the past decade, the methods and science used to describe changes in outcomes of physical therapy services have become more refined. Recently, emphasis has been placed not only on changes beyond expected measurement error, but also on the identification of changes that make a real difference in the lives of patients and families. This article will highlight a case example of how to determine and interpret "clinically significant change" from both of these perspectives. The authors also examine how to use item maps within an item response theory model to enhance the interpretation of change at a content level. Recommendations are provided for physical therapists who are interpreting changes in the context of clinical practice, case reports, and intervention research. These recommendations include a greater application of indexes that help interpret the meaning of clinically significant change to multiple clinical, research, consumer, and payer communities. [Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. *Phys Ther.* 2006;86:735–743.]

Key Words: *Health status, Measurement: applied, Outcome assessment (health care), Pediatrics, Physical disability, Professional issues*

Stephen M Haley, Maria A Fragala-Pinkham

This article presents a selected perspective on how physical therapists can interpret clinical changes both at the individual and group levels.

Consider the following case:

Mario is a 7-year-old boy who was admitted to an inpatient rehabilitation facility after a sledding accident in which he acquired a nondisplaced fracture at C1 and a closed traumatic brain injury. He has been in the inpatient rehabilitation unit for 3 weeks, and the facility routinely administers standardized functional assessments at admission, during periods of rapid change, and at discharge. Mario has progressed from being very dependent at admission to rapidly attaining some basic motor skills. He is now medically stable, cooperative, and appears ready to make changes in motor function that will allow him to return home with some transfer and self-mobility skills. He recently started sitting by himself and is standing with minimal support. The physical therapist has administered a functional test of mobility at admission and recently to determine progress. The child has a score of 6.1 at admission and then most recently a score of 35.9 at 3 weeks after admission.

The physical therapist providing intervention (and others) may have a number of questions regarding how to interpret the functional test results described in the case. For example: What do the summary scores from the outcome measures mean? How do we interpret the change score? Has the child achieved "clinically significant change" up to this point in the hospitalization and physical therapy episode of care? Is the change meaningful? Is the change score beyond measurement error that would typically occur in the routine administration of this measure? How can these scores be used to help examine the patterns of mobility changes that have taken place? Because the meanings of scores on a standardized instrument are not intuitively apparent,[1] there is a need to provide meaning to scores that result from tests and measures used in physical therapist practice.

Physical therapy and other health care fields are beginning to explore, in increasing depth, the proper interpretation of tests and measures and the clinical changes that score improvements represent. Measures to detect important effects related to physical therapy intervention must be valid (ie, measure what is intended), responsive (ie, able to detect an important change, even if that change is small), and interpretable (ie, the intended audience must understand the magnitude of effect).[1,2] At the center of this issue of "interpretability" is the attempt to have a better understanding of a "clinically significant difference" (CSD).[3,4] Understanding CSD can be a bewildering endeavor, particularly with the myriad of terms and anachronisms that are used across different fields and traditions. A number of terms to describe the phenomenon of CSD have been proposed, but different terms often have a similar meaning, such as "reliable change index" (RCI) and "minimal detectable change" (MDC), or "minimal clinically important difference" (MCID) and "minimal important difference" (MID).

Various audiences may have very different perspectives on CSD. For example, from a patient's point of view, a clinically significant change could result from greater

SM Haley, PT, PhD, is Associate Director, Health and Disability Research Institute, Boston University, Boston, Mass, and Director of Research, Research Center for Children With Special Health Care Needs, Franciscan Hospital for Children, Boston, Mass. Address all correspondence to Dr Haley at Health and Disability Research Institute, Boston University, 53 Bay State Rd, Boston, MA 02215 (USA) (smhaley@bu.edu).

MA Fragala-Pinkham, PT, MS, is Research Associate, Health and Disability Research Institute, Boston University, and Clinical Researcher, Research Center for Children With Special Health Care Needs, Franciscan Hospital for Children.

Dr Haley provided concept/idea/research design and data analysis. Both authors provided writing, Ms Fragala-Pinkham provided subjects. The authors extend their appreciation to Elaine Rosen, who provided the case study, and to Erika Wright for assistance with the Figure.

This work was supported by an Independent Scientist Award (K02 HD45354-01) to Dr Haley and grant R01 HD043568 (Haley, principal investigator) from the National Institute of Child Health and Human Development (NICHD) and the Agency for Healthcare Research and Quality (AHRQ). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or AHRQ. This manuscript also was partially supported by the Noonan Foundation (Fragala-Pinkham, principal investigator). Dr Haley has stock interest in CRE Care, LLC, which distributes the Pediatric Evaluation of Disability (PEDI) products.

This article is based on a lecture by Dr Haley at the III STEP Symposium on Translating Evidence Into Practice: Linking Movement Science and Intervention; July 19, 2005; Salt Lake City, Utah.

Physical Therapy acknowledges that the authors retain the right to provide a copy of the final article to NIH upon acceptance for publication or thereafter, for public archiving in PubMed Central as soon as possible after publication by the Journal.

This article was received July 9, 2005, and was accepted December 21, 2005.

freedom to resume previous activities; for a physical therapist, however, CSD may provide an indication to change the course of intervention. For other audiences such as payers, CSD may have a broader definition relating to a reduction in costs and utilization of future health care dollars. Crosby et al[5] and Wells et al[6] provided comprehensive reviews of CSD and its associated terminology.

In this article, we will present a selected perspective on how physical therapists can interpret clinical changes both at the individual and group levels. Our presentation will adopt a deductive approach toward identifying the meaning of clinical change by using information from group studies and applying these findings to individual patients. Cella et al[3] provided a detailed discussion of the merits of both deductive and inductive (starting with the individual and applying findings to group analyses) approaches toward defining meaningful changes. We also will highlight some remaining challenges that will need to be solved, particularly with the accelerating use of instruments designed with item response theory (IRT) methods, so that the meaning of CSD can be more readily understood by physical therapists, patients, and other interested parties.

We approach the topic of CSD by identifying 2 complementary but distinct methods. *Distribution-based* methods rely on expressing change scores in terms of an underlying sampling distribution, whether in between-person standard deviation units, within-person standard deviation units, or some variation of the standard error of measurement (SEM). These methods are based on statistical significance, sample variability, and measurement precision. In contrast, *anchor-based* approaches require an external, independent standard to "anchor" the meaning of clinical importance, one that is itself interpretable and at least moderately correlated with the test or measure. We will highlight an example of both distribution-based and anchor-based methods for interpreting the functional outcome data in the physical therapy case presented above. Eton et al,[7] Wyrwich,[8] and Schmitt and Di Fabio[9] provided a more comprehensive review of both distribution- and anchor-based methods.

Minimal Detectable Change

One of the more common distribution-based change indexes is the minimal detectable change (MDC), also called the reliable change index.[10,11] The MDC is based on the SEM and is calculated using the following formula:

$$\text{MDC} = z\text{-score}_{\text{level of confidence}} \times \text{SD}_{\text{baseline}} \times \sqrt{(2[1 - r_{\text{test-retest}}])}$$

where the *z*-score represents the confidence interval (CI) from a normal distribution, SD is the standard deviation of the baseline or pre-intervention scores (in our case example, the admission scores), and *r* is the coefficient of the test-retest reliability. In recent literature, the traditional Pearson product moment correlation (*r*) used to calculate test-retest reliability is more commonly estimated using a form of the intraclass correlation coefficient (ICC). The multiplier of $\sqrt{2}$ is to account for the additional uncertainty introduced by using difference scores from measurements at 2 points in time.

The MDC is considered the minimal amount of change that is not likely to be due to chance variation in measurement. For the case example, we will use a CI of 90%, because that level seems to be the most common standard used in the literature; however, an MDC at a 95% CI or other values could be selected, depending on the precision needed for the score estimate.

A vital choice in calculating the SEM is whether one uses internal consistency or test-retest reliability to calculate MDC. Although Wyrwich and colleagues[8,12] argued for using internal consistency (Cronbach alpha), we favor the more conservative approach of using test-retest reliability. The size of the reliability coefficient that is used is a very critical element in the equation; therefore, instruments that cannot demonstrate good stability across repeated tests will have sizable MDCs.

It is interesting to note that the use of a form of the SEM for understanding the extent of estimated measurement error is not new in physical therapy. Hinderer et al[13] proposed using the SEM (with a test-retest correlation estimate) to determine the extent to which the Peabody Developmental Motor Scales were stable in the context of determining clinical change in pediatric patients. We should not be fooled that recent updates of terms by authors, or minor changes in error calculations, are something new to the field of physical therapy tests and measures. Perhaps we may not have fully appreciated the importance of estimating distributional errors in tests and measures used in physical therapist practice; however, the approach toward estimating distributional measurement precision has been recommended for more than a decade.

In the case example, we use the Pediatric Evaluation of Disability Inventory (PEDI) as a broad functional measure in an inpatient rehabilitation setting. The PEDI[14] is designed to measure functional status in children and youths between the ages of 6 months and 7.5 years in 3 content domains: self-care, mobility, and social function. The PEDI is routinely used in the physical therapy, occupational therapy, and speech-language-hearing departments of many hospitals to generate numerical scores that reflect children's functional change from inpatient admission to discharge. We will just use the Functional Skills Mobility Scale in our example. For use in the case study, we have determined that the MDC_{90} for the PEDI in an inpatient setting, populated largely by children with severe brain injuries, is 5.1 points.[15] This is a value obtained by using a 90% CI, a standard deviation value of 15.4 for children seen at hospital admission on the PEDI Functional Skills Mobility Scale (0–100 scale), and a test-retest reliability (ICC) value of .96, based on a previously published report[16] and our own internal testing. Based on these results, the change score across the 2 time points (6.1 at admission, 35.9 at 3 weeks) exceeds the MDC_{90} value, and the change is not likely due to chance variation or random measurement error.

We have found the MDC useful for interpreting changes in a case report recently published in *Physical Therapy* describing the changes observed after a 26-week fitness intervention for children with disabilities.[17] By using the MDC, we were able to identify reliable changes in function, strength, and walking efficiency in 6 of the 9 children following a twice weekly group strength and endurance training program.

MDC Proportion

In a follow-up to this case report, we conducted a group fitness intervention study for children with disabilities in community settings.[18] We used the concept of an MDC to determine the proportion of the study group that achieved at least the minimal amount of reliable change (ie, not likely due to measurement error). For example, the mean change in knee extensor force production after the 16-week intervention period in this single group, pretest-posttest study was 2.14 kg ($SD=2.98$, $t=3.90$, $P<.01$). From inspecting the mean change value, it is not accurate to assume that all study participants achieved the mean change value, because change scores always form some type of distribution.

The variability in individual responses highlights the fundamental problem of summarizing treatment effects as a difference in means. In this example, 50% of the children achieved a positive change in knee extensor force production that exceeded the MDC_{90} value of 1.8 kg. A further subgroup analysis also can be conducted using the MDC, which indicated that, of the children who exceeded the MDC_{90} value, 59% were from the developmental disability group (children with intellectual disabilities, pervasive developmental disorders, or genetic disorders with intellectual or behavioral components) and only 28.6% were from the neuromuscular group (children with cerebral palsy, Duchenne muscular dystrophy, or traumatic brain injury). Reporting the proportion of patients achieving a degree of improvement that is beyond measurement error is a more informative method for describing the effects of the intervention than overall mean change.

In the case example at the beginning of this article, change between the 2 administrations of the PEDI exceeded the MDC_{90} value for the PEDI used in the inpatient setting. Some would argue, however, that, although the change noted is likely not due to measurement error, the MDC by itself does not provide us with an answer as to whether the change is clinically significant. (We will explore anchor-based indexes of change to address this concern later in the article.)

For most applications of the MDC, we assume that the amount of measurement error is constant along the entire functional scale. If one does not want to accept this assumption, Stratford et al[19] provided a solution by demonstrating the usefulness of the conditional SEM with a common measure of physical disability. The MDC is based on a summary score metric; little to no attention is given to the pattern of changes at the item level with the MDC. The inability to take into account changes in responses to individual items is a limitation of the classical test theory approach, which is the basis of the MDC calculation. In summary, because the MDC is one of a family of distribution-based methods, it is easy to generate (because it requires no additional data collection) and can serve as an important adjunct for estimating reliable change in a wide variety of tests and measures used routinely by physical therapists in clinical practice. It is somewhat limited in its interpretation, however, because it assumes that detectable changes are uniform at any point along the scale. In contrast, as is discussed below, we will see that measurement error will vary at different points along the scale. A strong advantage of IRT is that standard errors can be calculated at each point along the scale, and as will be highlighted in the case discussion, these standard errors are usually larger at the score extremes and smaller in the middle of the scale.

Item Response Theory Maps

Outcome instruments that incorporate IRT as a basis for modeling the probability of item and test scores may afford important advantages in the interpretation of clinical change, particularly at the individual patient level. Item response theory methods examine the associations between individuals' response to a series of items designed to measure a specific outcome domain (eg, physical functioning). Data collected from samples of physical therapy patients are fit statistically to an underlying IRT model that best explains the covariance among item responses and are used to build measurement scales.[20,21] The measurement scales are composed of items with a known relationship between item responses and positions on an underlying domain. Using this approach, probabilities of patients scoring a particular response on an item at various ability levels can be modeled. People with more functional ability have higher probabilities of responding positively to items than people with lower functional abilities.

These probability estimates are used to determine an individual's most likely position along the scale. When assumptions of a particular IRT model are met, estimates of a person's ability do not strictly depend on a particular fixed set of items. This scaling feature allows one to compare people along a functional dimension even if they have not completed identical sets of items. Because items and scores are defined on the same scale, items can be optimally selected to provide good estimates of the domain 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. Hambleton[22] provided a more detailed explanation of IRT methods. Item response theory is currently being applied in physical therapy research to develop new measures, improve existing measures, investigate group differences in item and scale functioning, equate different instruments, and, as we highlight, develop better approaches to understanding the meaning of differences in scores. Jette and Haley[23] and Ware and colleagues[24] reviewed recent applications of IRT to rehabilitation tests and measures.

In order to better interpret change in an individual patient, most physical therapists have an interest in the types of items that make up a total score on the measure. Using a one-parameter IRT model in its simplest form, in which item difficulty is used to locate dichotomous items along a scale, the clinician can examine the test from the perspective of a hierarchic set of items that serves as a representation of an underlying variable.[25,26] Item response theory procedures take full advantage of modeling of individual items; therefore, one can examine changes in item responses from serial assessments at an item level.

The PEDI scoring profile and summary scores are based on Rasch IRT[27] measurement technology. This approach provides an important hierarchical framework in which the construct validity and clinical utility of summary scores can be determined. A hierarchic scale defines a set of sequential tasks that represent increasingly more difficult functional items along a single dimension. The scales of the PEDI were specifically constructed to meet the objective of forming independent hierarchic dimensions. Each scale can be used to identify which functional items are relatively easy or more difficult for a child to achieve.

In the Figure, we have constructed an "item map" for the PEDI Functional Skills Mobility Scale corresponding to the case example, which allows us to define the specific items for which the child has shown capability and the items that he has yet to master. Because a child is expected to move along the continuum of hierarchically defined items, a summary score provides a clear indication of the child's performance level in that content domain, thus leading to an unambiguous interpretation of a summary score. Knowledge of specific content and location of items along the Functional Skills Mobility Scale can contribute to a richer understanding of the nature of mobility skill development and the interpretation of individual scores. For illustrative purposes in the case, we have arranged the entire set of items into 2 subsets: Transfers and Locomotion.

On admission to inpatient rehabilitation, Mario used a recliner wheelchair because he could not sit up in a regular wheelchair. He was able to sit in a tub seat, which provided back and leg support during bathing (item 20). Note that because his score at the low extreme of the scale, not many items populate the scale in this low-score region; therefore, the SEM is relatively large (6.1) (Figure). He required assistance for all transfers and was nonambulatory. Three weeks after being admitted for rehabilitation, Mario's functional abilities were reevaluated. His improvements in upper-extremity, lower-extremity, and trunk strength and control were reflected in his ability to roll and creep on hands and knees without assistance (item 25). Improved trunk strength and balance were reflected in his ability to independently sit in a wheelchair (item 6) and on a bench (item 7) and get on and off the bench by himself without assistance (item 8). He can independently move around in a room using a wheelchair but with decreased speed (item 28). He walks with one hand held for household distances and is beginning to learn how to climb stairs but needs moderate assistance (item 26). Note that this progress score, in contrast to the initial score estimate, has a relatively smaller SEM (2.5), because a number of items are in this area of the scale. Using the item map concept, a therapist can more readily interpret the meaning of the nearly 30-point change within the context of the new skills that have been achieved.

The potential utility of using item maps to track progress is to examine the specific item changes that are occurring during a physical therapy intervention program. This can be used to understand summary score changes, provide information to physical therapists about the pattern of skill changes, and perhaps suggest new items that might be the focus of revised patient goals. As IRT models of tests and measures used by physical therapists become more complex, such as tests using more than 1 parameter for estimation and response scales with more than 2 response choices (polychotomous), the item maps will become more complicated, but should still be informative. In addition, with the emerging use of computer-adapted testing applying the complex IRT models,[28–30] it will be imperative that considerable thought goes into the development of computer-generated item maps in order to help clinicians interpret the summary scores from an item response level for an individual patient. Some work of this kind has been completed in educational applications,[31] but has yet to be fully adapted to the test and measures used by physical therapists.

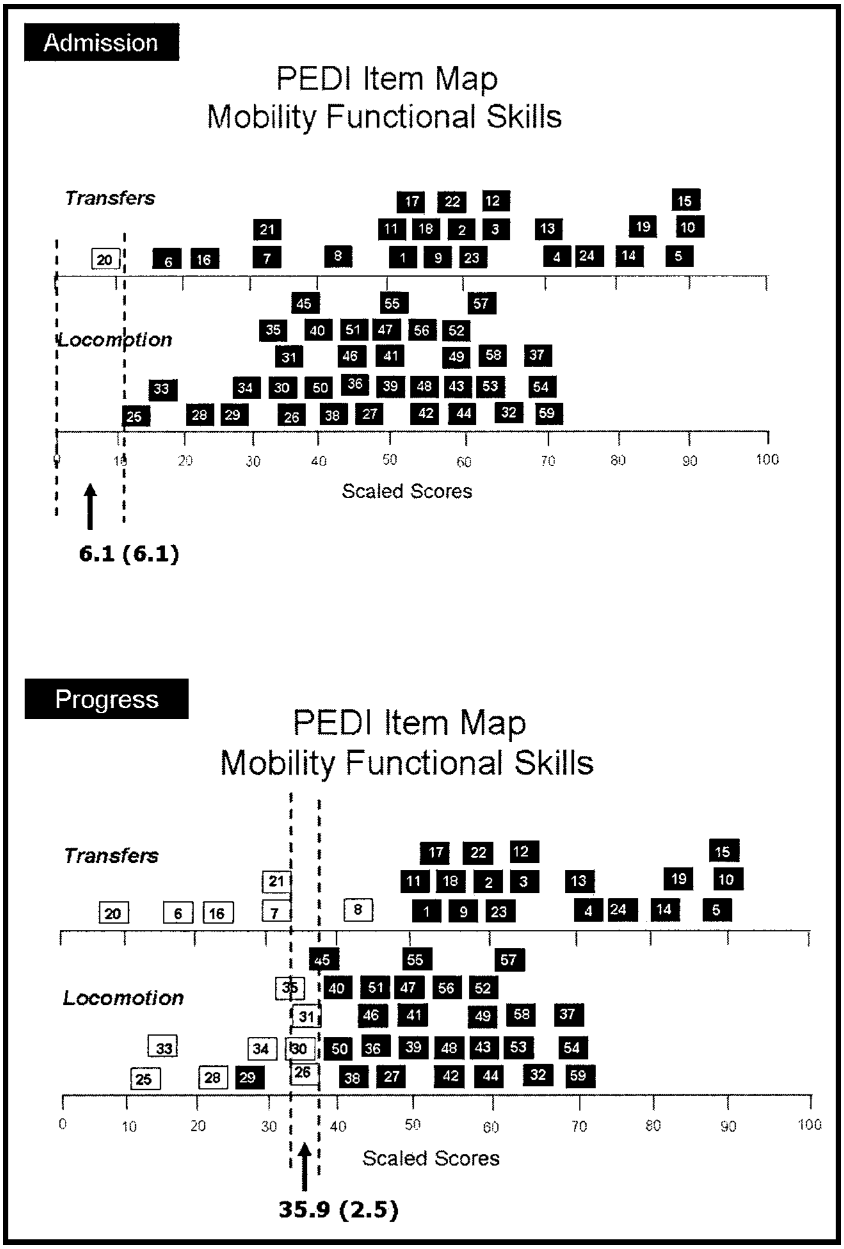

Figure.
An example of an item map constructed for the Pediatric Evaluation of Disability Inventory (PEDI) Functional Skills Mobility Scale. Each box represents a mobility item from the PEDI. The item map defines the specific mobility items for which Mario, the child in the case example, has shown capability (unshaded boxes) and the items that he has yet to master (shaded boxes) at admission and 3 weeks after admission for rehabilitation. Mario's score on the PEDI Functional Skills Mobility Scale was 6.1 at admission and 35.9 three weeks after admission (standard error of measurement in parentheses). Vertical dashed lines indicate the standard errors of measurement. Key to items that Mario is capable of doing 3 weeks after admission: 6=sits in chair if supported; 7=sits unsupported on chair or bench; 8=gets on and off low chair or furniture; 16=raises to sitting position in bed or crib; 20=supported sitting (tub); 21=sits unsupported and moves in tub; 25=rolls, scoots, crawls, or creeps on the floor; 26=walks, with support using wall and furniture; 28=moves within a room but with difficulty; 30/31=moves between rooms; 33=changes physical location purposefully; 34=moves objects along floor; 35=carries objects small enough to be held in one hand. Item 20 is the only item Mario could do at admission.

Minimal Important Difference

One of the essential problems of most tests and measures used in physical therapy practice is the lack of a clear external criterion (or anchor) to help with the interpretation of scores. For instance, what can a person do differently if he or she is able to lift an additional 5 kg in knee extension? What does it mean to a person's involvement in sports if his or her energy expenditure index improves 0.5 beat per meter walked? And for our case example, what does it mean for the child to improve 5 to 10 points on the PEDI Functional Skills Mobility Scale? Can these changes be grounded with some external criterion that can help us make sense of the tests and measures we use in practice?

What are some possible anchors that would help us understand scores or score change on a test and measure? Anchors might include self-reported opinions of individuals, including patients, family members, clinicians, or uninvolved judges. They often are collected by asking respondents to rate the amount of change in a particular area of health or function that has occurred during an episode of care. For example, if Mario's parents were asked to rate how much change had occurred during the current hospital episode, we would expect that the parents would identify that Mario has made a noteworthy change in function. We discuss some advantages and limitations of this approach below. Anchors also might include more objective indicators, such as laboratory values or disease markers. Other anchors may include return to expected recovery events, such as walking, wheelchair mobility, sports activity, work, school, independence in home, safety, or other important life activities or roles.

One of most apparently obvious, but controversial, approaches to understanding change scores is to get information from the patient regarding his or her perception of change. For certain content areas, such as functional gain, pain, fatigue, quality of life, and others, the patient appears to be a good selection to provide a global anchor for measures, even though this reasoning is fraught with a certain element of circularity. One of the limitations of anchor-based methods that rely on global ratings from a patient (eg, how much have you improved during your physical therapy treatment episode?) is that these retrospective ratings, particularly those focusing on an extended period of time, are susceptible to recall bias. For patients who are followed over long periods of time, longitudinal anchor-based methods are preferable to cross-sectional methods because the former are more temporally linked with change.[32] In addition, global change questions often have unknown reliability and validity.[3]

Clinicians also may be appropriate candidates to provide an external assessment of patient change, although without proper training and rigor in making judgments about change, large variations may occur. Iyer et al[33] recently reported an anchor-based study to determine the MID (also called "minimal clinically important difference" [MCID]) in the PEDI scales using physical therapists and others in an inpatient pediatric rehabilitation hospital as external anchors. An important difference is described as a "clinically important" change in patient function that is perceived as beneficial and that would change the patient's management.[1] The "minimal important difference" is the smallest change in what is measured that is considered to be worthwhile or important to a patient.[34]

In the clinician-based anchor study by Iyer et al,[33] the authors provided significant training and evaluated clinicians' performance on case examples before recording their global judgments of patient change. They asked therapists to "indicate how much this child changed from admission to discharge in capability to perform mobility skills (that were important to home/community functioning)." The therapists used a 15-point Likert scale and a visual analog scale to indicate how much better (or worse) the child was at discharge than upon admission. The authors collapsed the original clinician rankings into 4 categories (worse/no change, minimal change, moderate change, and large change). The minimal change category included original Likert scale points of "somewhat better" and "a little better." The average change in PEDI mobility scores for the group of children in the minimal change category was 8.7 points. Thus, using clinicians as an anchor for describing changes during inpatient rehabilitation programs, Iyer and colleagues defined a change of 8.7 points as representing a clinically meaningful level of change. In contrast, children who were identified as having moderate change on average had an admission-to-discharge change of 28.4 points, and those who were classified as making large changes had an average change of 58.7 points on the PEDI Functional Skills Mobility Scale. In the case of children who are admitted to an inpatient rehabilitation program, most children are at a very low functional level when admitted, and thus do not change in the negative direction often. For many other acquired or progressive conditions, however, an analysis of change both in the positive and negative direction is warranted.

In our case example, using this anchor-based criterion, the child has changed from an admission score of 6.1 to 35.9 points (a difference of almost 30 points), and he has certainly exceeded an MID so far in his episode of care. He has even reached a point where one might consider his change to be more at the moderate level. Although this information is helpful in interpretation of the summary scores for this case, there are some important caveats to consider when using MIDs.

Minimal important differences have been shown to vary across patients and patient groups and to have limited generalizability.[35] Different MID values may be obtained by using alternate anchors and methods; therefore, corroborating results across methods and multiple anchors will be important in future research. Furthermore, any estimate of the MID will be associated with a degree of uncertainty and variation. In the study by Iyer et al,[33] although they report mean MIDs, there is considerable variability within each of the global change categories, so that reporting a range of MIDs might be preferable. Another limitation in the determination of an MID is the effect of initial placement of the patient on the scale. Patients who have very low initial scores at baseline (or admission) may have a greater ability to achieve an MID than those who start at a higher level on the scale. An additional concern is estimating an MID in certain groups with expected loss in function. Many studies simply use the absolute value of change scores, rather than separately evaluating improvement and deterioration.[36]

MID Proportion

We can use the MID value applied to an individual patient and extend its use to group analyses to identify how many patients exceeded the minimum values of clinically important change. This is analogous to using the MDC to determine a proportion of people who exceeded likely measurement error. In this case, we will define the *MID proportion* as the percentage of patients who exceed a minimal standard of change that is considered clinically important. Dumas et al,[37] in a recent study that examined intervention intensity and functional outcome in 80 children with traumatic brain injury, determined the MID proportion to include 74% of the cases. This is in the context of also reporting a mean change in PEDI Mobility Functional Skills Scale score of 37.1 (SD=27.0). For the purposes of understanding which children with traumatic brain injury gained from their inpatient rehabilitation stay and which children made clinically significant changes in mobility skills, the MID proportion was much more interpretable than only knowing the average change (albeit with relatively large standard deviation). The MID proportion may have important interpretive advantages when physical therapists are examining group-level data, conducting program evaluations, and participating in quality assurance activities.

Combining Distribution- and Anchor-Based Methods

There is recognition of both the value and limitations of distribution-based (eg, MDC) and anchor-based (eg, MID) methods in defining CSD. Triangulation of measurement error and external anchor studies may be taken collectively to support the identification of change scores that are clinically meaningful. The process of arriving at a standard for CSD is cumulative and, therefore, enhanced by aggregating evidence from multiple perspectives. Until we know more, it appears that we cannot easily extrapolate and assume that the same criterion value for a particular test or measure applies to all types of patients, or even across the full range of scores of a single test or measure. Approaches directed toward combining the distribution- and anchor-based methods have been the recent focus of a number of clinical reports.[7,38]

In our case example, we add to the interpretation of the child's change scores by using the following empirical information collected in previous studies of children with traumatic brain injuries admitted to the inpatient rehabilitation program. An amount of change that is not likely to be measurement error at a 90% CI (MDC_{90}) is 5.1 or at a 95% CI (MDC_{95}) is 6.0. The MID, based on a clinician anchor, is 8.7 points.[33] These values are in the range of what might be expected from approaches that have defined one half of a standard deviation[39,40] (7.7 in the case of the PEDI Functional Skills Mobility Scale) to identify CSD. From this collective work, we might consider the changes seen in the case example to have met requirements for a CSD to be in the range of 5 to 9 points on the PEDI Functional Skills Mobility Scale.

An additional set of information regarding the kinds of items that are changing is provided by the item map. Combining total score information within the content of MDC and MID values, and inspecting patterns of item changes using IRT methods, may yield the most informative data for physical therapists who are attempting to use tests and measures for the examination of individual patients.

Recommendations

As physical therapists increase the routine use of reliable and valid tests and measures in clinical practice, we hope that journals such as *Physical Therapy* encourage the reporting of MDC and MID values with the same regularity as statistical significance or effect sizes. The interpretation of clinical significance will become transparent and more commonly accepted if we are informed about MDC and MID proportions in group studies and MDC and MID for application to case reports and individual patients. Physical therapist investigators should be persuaded to increase research output concerning the relationship between measurement error and anchor-based estimates of change for commonly used tests and measures. We encourage test developers to use IRT as a basis for test development and scoring. We also encourage greater use of item map concepts to help users interpret change scores from a content perspective. If this is accomplished, clinically significant change will be much more fully accepted in audiences that are demanding an interpretable characterization of improvement associated with physical therapy intervention.

References

1 Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc.* 2002;77:371–383.

2 McHorney CA. Health status assessment methods for adults: past accomplishments and future challenges. *Annu Rev Public Health.* 1999;20:309–335.

3 Cella D, Bullinger M, Scott C, Barofsky I; Clinical Significance Consensus Meeting Group. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clin Proc.* 2002;77:384–392.

4 Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research: how meaningful is it? *Pharmacoeconomics.* 2000;18:419–423.

5 Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol.* 2003;56:395–407.

6 Wells G, Beaton D, Shea B, et al. Minimal clinically important differences: review of methods. *J Rheumatol.* 2001;28:406–412.

7 Eton DT, Cella D, Yost KJ, et al. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. *J Clin Epidemiol.* 2004;57:898–910.

8 Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? *J Biopharm Stat.* 2004;14:97–110.

9 Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol.* 2004;57: 1008–1018.

10 Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Phys Ther.* 1998;78:1186–1196.

11 Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. *J Clin Epidemiol.* 2001;54:1204–1217.

12 Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care.* 1999;37: 469–478.

13 Hinderer KA, Richardson PK, Atwater SW. Clinical implications of the Peabody Developmental Motor Scales: a constructive review. *Phys Occup Ther Pediatr.* 1989;9:91–106.

14 Haley SM, Coster WJ, Ludlow LH, et al. *Pediatric Evaluation of Disability Inventory: Development, Standardization and Administration Manual.* Boston, Mass: Trustees of Boston University; 1992.

15 Dumas HM, Haley SM, Ludlow LH, Rabin JP. Functional recovery in pediatric brain injury during inpatient rehabilitation. *Am J Phys Med Rehabil.* 2002;81:661–669.

16 Nichols DS, Case-Smith J. Reliability and validity of the Pediatric Evaluation of Disability Inventory. *Pediatric Physical Therapy.* 1996;8(1): 15–24.

17 Fragala-Pinkham MA, Haley SM, Rabin J, Kharasch VS. A fitness program for children with disabilities. *Phys Ther.* 2005;85:1182–1200.

18 Fragala-Pinkham MA, Haley SM, Goodgold S. Evaluation of a community-based group fitness program for children with disabilities. *Pediatric Physical Therapy.* In press.

19 Stratford PW, Binkley J, Solomon P, et al. Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Phys Ther.* 1996;76:359–365.

20 Hambleton R, Robin F, Xing D. Item response models for the analysis of educational and psychological test data. In: Tinsley HA, Brown SD, eds. *Handbook of Applied Multivariate Statistics and Mathematical Modeling.* San Diego, Calif: Academic Press; 2000:553–581.

21 Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika.* 1986;51:567–577.

22 Hambleton RK. Principles and selected applications of item response theory. In: Linn RL, ed. *Educational Measurement.* 3rd ed. New York, NY: American Council on Education, Macmillan Publishing Co; 1989:147–200.

23 Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. *J Rehabil Med.* 2005;37:339–345.

24 Ware J Jr, Gandek B, Sinclair S, Bjorner B. Item response theory in computer adaptive testing: implications for outcomes measurement in rehabilitation. *Rehabil Psychol.* 2005;50:71–78.

25 Haley SM, Ludlow LH, Coster WJ. Pediatric Evaluation of Disability Inventory: clinical interpretation of summary scores using Rasch rating scale methodology. *Phys Med Rehabil Clin North Am.* 1993;4:529–540.

26 Coster W, Ludlow L, Mancini M. Using IRT variable maps to enrich understanding of rehabilitation data. *J Outcome Meas.* 1999;3:123–133.

27 Wright BD, Masters GN. *Rating Scale Analysis.* Chicago, Ill: MESA Press; 1982.

28 Haley SM, Raczek AE, Coster WJ, et al. 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.

29 Haley SM, Ni P, Fragala-Pinkham MA, et al. A computer adaptive testing approach for assessing physical functioning in children and adolescents. *Dev Med Child Neurol.* 2005;47:113–120.

30 Haley SM, Coster WJ, Andres PL, et al. 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.

31 Bradlow ET, Weiss RE. Outlier measures and norming methods for computerized adaptive tests. *Journal of Educational Behavioral Statistics.* 2001;26:85–104.

32 Liang MH, Lew RA, Stucki G, et al. Measuring clinically important changes with patient-oriented questionnaires. *Med Care.* 2002;40: II45–II51.

33 Iyer LV, Haley SM, Watkins MP, Dumas HM. Establishing minimal clinically important differences for scores on the Pediatric Evaluation of Disability Inventory for inpatient rehabilitation. *Phys Ther.* 2003;83: 888–898.

34 Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10:407–415.

35 Santanello NC, Zhang J, Seidenberg B, et al. What are minimal important changes for asthma measures in a clinical trial? *Eur Respir J.* 1999;14:23–27.

36 Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual Life Res.* 2002;11:207–221.

37 Dumas HM, Haley S, Carey TM, Ni PS. The relationship between functional mobility and the intensity of physical therapy intervention in children with traumatic brain injury. *Pediatric Physical Therapy.* 2004;16:157–164.

38 Cella D, Eton DT, Lai JS, et al. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *J Pain Symptom Manage.* 2002;24:547–561.

39 Norman GR, Sloan JA, Wyrwich KW. Is it simple or simplistic? *Med Care.* 2003;41:599–600.

40 Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Med Care.* 2003;41: 593–596.