Clinical Measurement Practical Guidelines for Service Providers

Developed By:
Steven Hanna
Dianne Russell
Doreen Bartlett
Marilyn Kertoy
Peter Rosenbaum
Marilyn Swinton

Supported By:
The Jack and Ina Pollock Charitable Foundation

Research Version
Last Revised March 2005

Are you using a clinical measure?

A clinical measure is**:**

- a published measurement tool
- designed for a specific purpose and population
- with instructions for how to administer, score, and interpret the results

For examples of clinical measures used in physical, occupational, and speech-language therapy, see page 21.

Are your assessments as useful as they can be?

Therapists feel a responsibility to ensure that measurements they use are clinically useful and not harmful to children and families. This includes considering the accuracy of assessments. Based on studies done at CanChild and on conversations with therapists, these guidelines were designed to provide therapists with practical ideas for dealing with two measurement problems clinicians commonly face:

- what to do if you feel you need to **modify** the measure to suit your clinical situation
- how to **interpret** a child's **change**.

Research suggests that therapists must deal with both these issues routinely, and that better guidelines about how to handle these problems could improve services to children. We suggest that you keep these guidelines handy when you conduct clinical assessments. They are intended to help you make assessment easier and more effective.

What will this document tell you?

Are you modifying the administration or scoring of a clinical measure?

Sarah: An Example from Occupational Therapy

Sarah is a 26-month old child with cerebral palsy (spastic diplegia, Gross Motor Function Classification System Level III) who has just moved to your area with her family. Her parents are concerned about her fine motor skills, and so you decide to do a baseline assessment using the Grasping and Visual Motor Integration subtests of the Peabody Developmental Motor Scales (PDMS-2).

The PDMS-2 is a standardized, norm-referenced instrument designed to assess the motor abilities of children from birth through 6 years of age using 6 subtests. The *Guide to Item Administration* contains detailed descriptions of every item in each of the subtests, including the child's position, the stimulus, and the procedure for testing, as well as the criteria for scoring. As recommended in the manual, you have read all of the item specifications, and have conducted assessments on three children outside your practice before using it clinically in your setting.

As you progress through your assessment, you note that Sarah does not stay in the position recommended in the manual, that she frequently seems to need more than the recommended 3 trials to be successful (or partly successful) with an individual item, and that she seems to need demonstration in addition to verbal instruction to understand what the item requires.

After 30 minutes of testing, you determine that Sarah's Grasping and her Visual Motor integrations subtest raw scores are 40 and 89, respectively. These raw scores place her at an age-equivalent of 14 and 22 months for grasping and visual motor integration, with corresponding ranking of 16th and 25th percentiles. *Are these inferences correct?*

Modification

Therapists often find that they want to use a clinical measure, but don't want to administer it in the way that the manual suggests. There are lots of reasons that therapists might do this, but therapists say they do it most often to accommodate the special needs of the children they work with.

Are you likely to do any of the following modifications?

ALTERING THE TEST CONTENT OR ITS ADMINISTRATION

- Modifying the sequence of presenting items or materials
- Modifying the presentation of tasks or items by rewording them or by providing demonstrations or cues.
- Modifying starting positions.
- Modifying the test materials or equipment.
- Excluding some items or subscales.
- Modifying the administration or scoring of items to allow a child to do as well as possible.
- Modifying the administration to make the task more appealing to a reluctant child.
- Translating a measure or making other accommodations for language barrier.
- Allowing a child to use an adaptive device during testing.

TIMING

- Allowing a child extra time for a timed test.
- Cutting short an assessment or conducting it over two sessions.

SCORING

- Accepting parent reports of a child's function or behavior when unable to observe it.
- Estimating or assuming the scoring of an item that was not tested.
- Scoring test items by observing the child's play or other activities, if the test normally requires administration using standard prompts.
- Looking up normed, age-equivalent, or scaled scores using an age comparison that doesn't apply to your child.

TEST USE OR APPLICATION

- Using the measure outside the recommended chronological age range.
- Using the measure with a clinical population which is not discussed in the manual.
- Using a measure to assess change but you are not aware of any documentation that discusses the measurement of change using this instrument.

The manual for the measure may support some of these modifications, or you may know about some research in the published literature that supports them.

If you don't know whether these modifications are supported by the manual or other documentation, they may cause problems for your assessment.

It is understandable that you may try to adapt measures to work for your clients in specific situations. After all, meeting the needs of your clients is a part of your job.

The problem is that measures are developed and tested using standardized procedures. What may seem to you to be a small (and harmless!) modification can have unpredictable effects on the meaning of test scores **(**for more about what can happen, see page 8).

The people who develop clinical measures will not always give enough guidance for your testing situation. If you need to decide whether and how to modify a measure, go to page 11 for some easy things you can do to get the best assessment results possible.

What problems are caused by modifying measures?

We understand that modifying the administration or scoring of a clinical measure is sometimes necessary. However, it is important to be clear that there are trade-offs involved with this.

Modification makes it harder to compare this child to other children

If you modify the measure in different ways for different children, it makes it difficult to compare scores between children, or to compare your children to those assessed by other therapists. We cannot know how the comparison is affected by the differences in administration.

Many therapists believe that they can "adjust" their interpretation of scores for the likely effect of modifications they make, but it isn't always clear how they can do this.

Many modifications, such as allowing a child extra time or providing extra prompts seem like they would increase the child's score, but by how much?

For other modifications, such as accepting parent reports, it is not clear whether this would tend to increase or decrease a score, relative to the standard administration.

Modification makes norms or scaled scores difficult to use

Modifications can present serious problems for the interpretations of normed scores, such as age equivalents and percentiles.

The norms were established by administering the test in a standardized way to a large number of comparison children. If you didn't administer the test the same way that it was given to the children in the normative sample, the meaning of the normed scores is unclear.

Some measures use scaled or criterion scores that are calculated using Rasch analysis or some other type of item response theory. Like normed scores, the meaning of scaled scores is established using a comparison to a large number of other children who were administered the test in a particular way, and so modifying a measure can cause similar problems.

What about Sarah?

In the case of Sarah's assessment, the obtained percentile ranks of 16 and 25 are likely higher because of the modifications with respect to position, number of trials, and use of demonstration than they would be without such modifications. A consequence might be that Sarah be denied occupational therapy, if a policy of only working with children below the 10^{th} percentile on a standardized test exists at your centre.

Modification alters the reliability and validity of the test

The degree of reliability and validity which makes a test suitable for clinical use is established in studies that involve standardized administration, rather than adapted use.

It is possible that the reliability and validity of the modified test will no longer be adequate for clinical decision making.

Modification makes it difficult to evaluate change and treatment effectiveness

Some modifications affect the responsiveness of measures to real clinical changes in children. For instance, you may choose an assessment tool that has not explicitly been shown to measure change, or use a test with a child outside the recommended age range. It is often the case that such modifications underestimate the degree to which the child actually changed, and may lead to an inappropriate conclusion about the usefulness of a treatment. Similarly, modifications that affect the reliability of a score may make it harder to see important changes in the child.

What can you do if you think you need to modify a measure?

Some ways to deal with modified administration and scoring of clinical measures are listed below. They are approximately ordered from most to least preferred.

1. **RE-CONSIDER IF YOU NEED TO MODIFY THE MEASURE.**
   Therapists may often modify out of habit, or because they think it won't matter. It is always safer to "play by the book" (the test manual) if you can.

2. **CONSULT THE MANUAL FOR GUIDANCE ABOUT NON-STANDARD SITUATIONS**.
   For measures that are used widely in childhood disability, test developers may surprise you by providing guidance for special populations and circumstances that are similar to yours.

3. **BE HONEST WITH YOURSELF ABOUT THE PURPOSE OF YOUR ASSESSMENT**.
   Therapists often use clinical measures for purposes that don't depend on accurate and meaningful scores. For instance, using a measure may be a way to interact with a client, to observe how a child handles a novel task, or to get a rough idea of what the child can do. If you don't need the scores, then how you administer the test might not matter. In these cases, don't interpret the number and don't record it in the chart.

4. **INVEST THE TIME REQUIRED TO COMPLETE THE MEASURE THE RIGHT WAY.**
   Measurement is part of effective treatment. If time or costs are an obstacle, you may need to be clear with your supervisors and the child's family that accurate assessment is time well spent. It may be important to spend one or more whole sessions on assessment. (For some ideas that can help you make this argument, see page 17).

5. **AVOID USING NORMED, AGE-EQUIVALENT, OR SCALED SCORES WHEN INTERPRETING THE RESULTS OF A MEASURE YOU HAVE MODIFIED.**
   If you must modify any aspect of the administration, **don't use normed, age-equivalent, or scaled scores** when interpreting the results.

   Normative percentiles and other standardized scoring are obtained by using the standard administration, and you can't know whether your child's relative performance is the result of his or her ability or the way you administered the test. The **raw scores** may give you a rough idea of the kinds of tasks the child can do, without implying an accurate comparison to some standard.

6. **TRY TO BE CONSISTENT ON EACH ASSESSMENT.** If you want to assess a child's change over time, and find that you must modify the administration of the measure, carefully record the modifications and try to be consistent on each assessment.

   Doing this may mean that the evaluation of change over time might be valid for this child even if the comparison to other children is not.

7. **ENSURE YOU ARE USING THE RIGHT MEASURE FOR THE RIGHT REASON.**
   Don't count on seeing developmental or treatment-related change using a measure unless you are aware of published documentation that validates the measure for this purpose. (What makes a measure valid for a specific purpose? See page 19)

8. **CAREFULLY RECORD ALL MODIFICATIONS YOU MAKE IN THE CHILD'S CHART.**
   If you must make modifications, carefully record in the child's chart how your administration differs from the standard.

9. **DON'T PROVIDE SCORES OF MODIFIED MEASURES TO CHILDREN AND FAMILIES.**
   If you must make modifications, don't provide the scores to children and families. Be clear that you have done the assessment in a non-standard way, and be honest about the way you are using the measurement. This may be an important ethical issue.

Do you need to measure change?

Often you will assess a child with a particular clinical measure on two or more occasions and you will want to interpret the difference between test scores. This may happen because you want to know if a child is developing at an expected rate, given his clinical presentation, or because you want to evaluate the effectiveness of a program or treatment. There are some tricks to interpreting the difference between two scores that have to do with amount of uncertainty in scores.

The example of Sarah revisited

Recall that Sarah is a 26-month old child with cerebral palsy (spastic diplegia, Gross Motor Function Classification System Level III). After determining Sarah's parents' goals for her, and conducting your baseline assessment of Sarah, you collaboratively decide to improve her fine motor function. You decide to conduct a trial period with motor learning strategies. You use the Grasping and Visual Motor Integration subtests of the Peabody Developmental Motor Scales (PDMS-2) to establish a baseline at your first assessment. You then provide a month of biweekly therapy with home programming and test her again after this period.

Sarah's raw score on the Grasping Subtest was 40 at baseline and 43 at the end of the trial therapy period.

What do you need to consider when interpreting this apparent improvement?

- **Is the PDMS suitable for evaluating change over time (see page 19)?**
  If not, then you may be underestimating the effectiveness of your treatment.

- **If you modified the administration of the PDMS-2, did you do it the same way on both occasions (see page 11)?**
  If not, then some of the observed change in scores may be due to the differences in administration.

- **The PDMS-2 scores on each occasion are estimates of Sarah's hand function.**

  How good an **estimate** they are will depend in part on the **reliability** of the measure. You will need a way to measure the uncertainty in each score.

  The last bullet here is really important when you are interpreting change. Any two scores for the same child are expected to be different to some degree, simply because of chance influences on the test scores that you don't care about. Unless you know how much of this chance variation to expect, you can't know whether the difference you observed was the result of your treatment.

  To understand how to measure the uncertainty in test scores, see page 15.

How is the uncertainty in scores measured?

The most common ways of expressing the uncertainty in scores are explained below. The people who designed the measure may use other methods and the manual should tell you how to interpret them.

Standard Error of Measurement

This is a single number that is a function of the test's reliability and measures how much a child's observed score is expected to vary, even though the child's ability or status has not changed. Approximately 95% of test scores are expected to fall within plus or minus 2 standard errors.

Confidence Intervals

A 95% confidence interval (CI) gives the range within which 95% of multiple test scores are expected to fall, if a child's ability or status has not changed. The 95% CI is calculated as the test score plus or minus 2 standard errors of measurement. Other coverage percentages could be reported, although 90% and 95% are most common.

Minimum Detectable Change

Test designers may report the uncertainty in test scores as the smallest difference between two scores that would be greater than that expected by chance variations. This is normally calculated so that, if the subject was not really changing, the chance that you would observe a difference at least this large would be less than 5% or 10% (the manual should tell you this).

How can you use these quantities to interpret change?

If the difference in scores that you observe is greater than you would expect by chance, as measured by one of the above methods, then there is a high probability that the child is really changing. To get better grasp of this, let's have another look at the example of Sarah.

Did therapy help Sarah?

Sarah's Grasping Subtest score went from 40 to 43 during the 4 weeks of therapy, a difference of 3 points. According to the PDMS-2 manual, the standard error of measurement for the Grasping subtest is 1 point. This means, even though you observed a baseline score of 40, there is approximately a 95% chance that Sarah's typical baseline score could be anywhere in the range 40 ± 2 (the score ± 2 standard errors of measurement).

In other words, the 95% confidence interval at baseline is 38 to 42. Similarly, the 95% interval after therapy would be 41 to 45 (43 ± 2).

The key point is that **these intervals overlap**, so that the upper limit of the baseline interval is greater than the lower limit of the post-therapy interval. The suggestion is that this observed difference could have plausibly occurred even if Sarah's hand function didn't really change, just because of the degree of variation naturally expected given the reliability of the test (i.e. due to measurement error).

This example should make it clear why you can't afford to treat any test score as if it were *the* exactly correct score. Without considering the uncertainty inherent in the two test scores, we may conclude incorrectly about the effectiveness of the treatment.

Why is accurate clinical measurement important?

You may sometimes find it necessary to justify the time you spend on careful assessment. Here are some ideas that you can use.

For therapists…

Assessment is a fundamental part of planning and evaluating services, so does it make any sense to think of it as secondary to treatment? Does it make any sense to conduct an assessment that you feel is probably not accurate?

Because you care about whether the treatments you provide are effective, you will need to care about how you measure whether they are effective. The careful use of clinical measures offers a way of evaluating the services you provide.

Careful clinical assessments offer a reproducible and authoritative basis for explaining and supporting your clinical decisions to children, families, and managers.

For children and families…

Children and families are increasingly regarded as part of the rehabilitation team, rather than passive recipients of services. Accurate assessments of a child's status, prognosis, and treatment outcomes will help children and families participate in goal setting and treatment decisions in an informed and appropriate way. It will offer them a concrete way to understand their child's situation and to understand your approaches to offering services.

Families often ask questions like "how bad is it?", "is he very delayed?", "is she improving?", and "will my child ever walk?" The careful use of clinical measures can help you answer these difficult questions with confidence. Attention to the inherent uncertainty and difficulties in assessing a child can help highlight the times when you can't be confident answering these questions. This uncertainty is a reminder of the ethical responsibility not to be overconfident in measures when they are known to be imprecise.

For administrators and managers…

The routine and conscientious use of clinical measures at your service organization will provide a large body of reproducible, authoritative, and standardized data about the needs and outcomes among the children and families you serve.

This data can inform the difficult decisions you face everyday about how to manage services and allocate resources. Routine assessment results can help you communicate to funding agencies in concrete and credible ways regarding your organization's needs and the effectiveness of the services you provide. Increasingly, funding for services is contingent on having a clear mechanism for evaluating services in terms of client outcomes.

For your profession…

Physical, occupational, and speech-language therapists offer services in a complex health service environment in which professionals from many disciplines compete for the resources to provide the services they think are most important. For better or worse, the status and credibility of your discipline depends in part on the perception that you can provide convincing evidence that what you do is effective.

What are the purposes of clinical measurement?

One framework suggests that there are at least **four different purposes** for clinical measurement.

The purpose of assessment is important, because a measure that is suitable for one purpose may not necessarily be suitable for others.

Test manuals and other supporting documentation should indicate the appropriate purposes and provide explicit evidence to support these uses.

Description

These measures are used to describe the differences among individuals within groups, as when a measure of physical function may be used to identify a child's profile of impairments and identify treatment goals.

A descriptive instrument should measure all the aspects of the condition that are potentially relevant to the clinician.

Discrimination

These measures are designed to distinguish between people with and without a particular characteristic or functional problem. Examples might include speech-language tests designed to identify children who have clinically significant problems of speech or language. Evidence that a measure is suitable for discriminative purposes may include quantities that express the true-positive (sensitivity) and false positive (1 – specificity) rates for the test. In practice, the effectiveness of a test for discriminative purposes depends on the prevalence of the functional problem as well as the sensitivity and specificity. In general, it will be easier to identify clinically significant problems if they are more common.

Prediction

Predictive tests attempt to assess children in terms of their likely future outcomes. For instance, a measure of current function may be used to assess children who are at risk for later functional problems, or may try to predict the limit of later functional status.

To be effective for prediction, measures must have high test-retest reliability and there must be evidence that the measure can predict later outcomes of interest.

Evaluation

Evaluative measures are designed to measure change over time within individuals, often in order to evaluate treatment effects.

To be effective, evaluative measures must be high in responsiveness. More responsive measures show more change in situations in which children are actually changing.

If a measure is to be used to evaluate treatment effects, the test manual or other supporting documents should explicitly report evidence of responsiveness. Sometimes this is established by examining the degree to which scores change in situations where we can safely assume that children really are changing. Frequently used quantitative measures of responsiveness include the standardized response mean and measures of clinically important change.

What clinical measures are most commonly used in pediatric disability practice in Ontario?

PHYSICAL THERAPISTS	
• Goniometer / Range of Motion	• Peabody Developmental Motor Scales (PDMS)
• Gross Motor Function Measure (GMFM)	• Alberta Infant Motor Scales (AIMS)
OCCUPATIONAL THERAPISTS	
• Beery–Buktenica Developmental Test of Visual-Motor Integration (VMI)	• Test of Visual Perceptual Skills (Non-Motor) – Revised (TVPS)
• Peabody Developmental Motor Scales (PDMS)	
SPEECH LANGUAGE PATHOLOGISTS	
• Preschool Language Scale (PLS)	• Clinical Evaluation of Language Fundamentals (CELF)
• Goldman-Fristoe Test of Articulation	• Structured Photographic Articulation Test (SPAT)

Source: Hanna S, Russell D, Bartlett D., Rosenbaum P, Kertoy M, Wynn K (2004). Measurement practices and professional culture in pediatric rehabilitation: A survey of physical, occupational, and speech language therapists in Ontario. Manuscript under review, available by request from Steven Hanna, hannas@mcmaster.ca

How can you find the right measure?

Therapists often choose their measures by relying on what is already available at their treatment centre and by what their colleagues tend to use. This is understandable since you are most comfortable with the measures you and your colleagues already know. However, it might help if you had a resource that could point you to some possible alternatives, particularly if you consistently find that you need to modify the measures you already use. Fortunately, help is at hand.

All About Outcomes is an interactive software package designed by a team from the *CanChild* Centre of Childhood Disability Research. It was designed to help clinicians to identify and select from among the available measures of clinical outcome measures those that are relevant to occupational, physical, and speech-language therapists. Measures are searchable by clinical domains of interest and the target of interventions. Best of all, you already have free access to All About Outcomes, because every Ontario Association of Children's Rehabilitation Services Centre already has a copy!

All About Outcomes is also available commercially from:
Slack Incorporated, 6900 Grove Rd. Thorofare, NJ, USA 08086, phone: (609) 848-1000.

More information about measures is also available from the *CanChild* website:
www.fhs.mcmaster.ca/canchild/

You may see these measurement terms often: what do they mean?

Reliability

The **reliability** of a clinical measure is the degree to which it can reproducibly or consistently detect differences among people. It is common to distinguish among different kinds of reliability.

Test-retest reliability is the degree to which the measured differences among people are consistent over time.

Inter-rater reliability is the degree to which multiple observers agree on the differences among people.

The technical definition of reliability focuses on the proportion of the differences among observed scores that is attributable to real differences among people, and studies of reliability are designed to estimate this.

High reliability is usually considered an important property of a clinical measure because you want to be sure that it measures the same thing each time you use it.

Validity

A clinical measure is **valid** if it measures what you want and expect it to measure. Even when a measure is thought to be reliable, it may not be valid. For instance, a measure of physical function that is strongly influenced by a child's cognitive level may consistently show that children with cognitive deficits do more poorly than children without such deficits (i.e., the measure is reliable). However, since you want to measure physical function, this is not the difference you are interested in, and the measure is not valid for use in clinical samples with large variations in cognitive function.

Responsiveness

A clinical measure is **responsive** if it can detect clinically meaningful changes in the characteristic, or function of interest. The responsiveness of a measure depends on its **sensitivity to change.** For instance, if it is known that children experience developmentally-related increases in some aspect of function, a more responsive measure of that function will show more change than a less responsive measure.

The responsiveness and sensitivity to change are important to consider when evaluating the effectiveness of a treatment or program for an individual person, because a measure of low responsiveness may show no effect of a treatment even if there is one.

Standard Error of Measurement

Any score on a clinical measure is really an *estimate* of the characteristic or ability that you want to assess. Some estimates are better than others, and you will need a way to determine how much uncertainty there is in your estimate, given the measure you have chosen.

The **standard error of measurement** (SEM) is an index of the uncertainty in a score in terms of the degree to which multiple measurements of the same person are expected to vary even if the clinical attribute of interest has not changed.

Conceptually, if a measure could be given many times to a person who is not changing, the SEM is the standard deviation of the scores that would be obtained, although it is not normally calculated this way. The SEM for a measure depends on its reliability and how much scores tend to vary among people generally.

The SEM is related to the probability of observing a score in a given interval. For instance, about 68% of scores will be within plus or minus 1 SEM of the average score, and about 95% will be within plus or minus 2 SEM's. As a consequence, the SEM can be used to construct **confidence intervals** for interpreting the accuracy of a test score.

Confidence Intervals

Confidence intervals (CI) express the uncertainty or accuracy in a measurement in terms of the range of likely scores, given the score the person actually received and the SEM for the measure.

If you could give the same measure to a person many times, it is unlikely that all the scores would be exactly the same, even if the person is not really changing. So, how different would they be?

The CI answers this question. It is usually calculated from the measure's SEM, so that a 95% CI is calculated as the observed score plus or minus 2 standard errors. The usual interpretation of a 95% CI is that upon repeated testing, 95% of a person's scores would fall within the lower and upper bounds of the interval, even assuming that the person had not changed.

The use of the CI emphasizes how the score you observed is actually an *estimate* of the person's characteristic or function, and that repeated testing under effectively the same conditions would give a range of scores, any one of which could be "correct."

More reliable tests will tend to have more narrow intervals, implying that the observed score is more closely representative of the person's function. The 95% CI on a score is sometimes much wider than you would be comfortable with for some clinical purposes, so it is important to consider this.

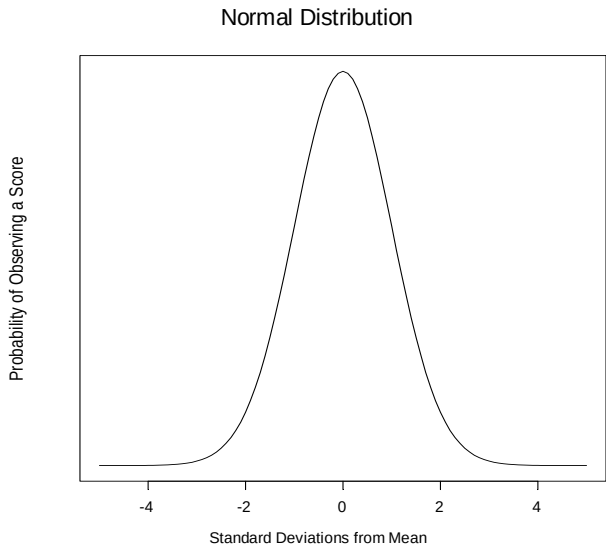
In a research context, confidence intervals can also be calculated for means or other quantities that describe the most likely score in a group of people. However, these confidence intervals are based on standard errors that are calculates in a different way than is the SEM for a single measurement.

Normal Distribution

The **normal distribution** is one of several commonly used mathematical descriptions of the probability of getting any particular score. The normal distribution describes this probability in terms of the average and standard deviation of the scores in the population.

It is sometimes referred to as the "bell curve" because scores near the average are most common, with deviations from the average score becoming increasingly less likely, describing the shape of a bell.

The normal distribution is the usual basis for calculating the SEM and confidence intervals, and is also applied when calculating most types of standardized scores including percentiles, T-scores, Z-scores, and age-equivalent scores.



Normative samples and norm-referenced measures

Measures are said to be **norm-referenced** if individual scores are interpreted by comparing them to the scores obtained for a large number of comparison children. This sample of comparison children constitutes the **normative sample**.

To interpret the measure, scores from the normative sample are used to transform the score for an individual child to a percentile, Z-score, or T-Score to measure how far above or below the child is from the average of the comparison sample.

To adjust interpretation for age- or sex-related differences, it is common to calculate norms within age-bands or genders. In practice, many measures used in pediatric rehabilitation use typically developing children as normative samples, although this is not always a clinically useful comparison.

Z-Scores

By assuming that scores are consistent with the normal distribution and estimating the average and standard deviation for the relevant population, it is possible to transform any obtained score to a **Z-score** or **standard score**. This transformation re-calculates the score as if it came from a population with an average score of 0 and a standard deviation of 1.

Z-scores are therefore a way of expressing a person's score in units of standard deviation and relative to the average. For instance if a girl's score is 1.0, this means that she is 1 standard deviation above average, a score of 0 means she is perfectly average, and a score of –1.0 means that she is 1 standard deviation below average. Z-scores can be used to compare a child's performance on measures with different units that nonetheless can be standardized and compared as Z-scores.

More importantly, Z-scores are directly related to the probability of obtaining scores in a normal distribution, and so Z-scores can be used to calculate confidence intervals and percentiles. For instance, if a child receives a raw score of 25 on a clinical measure, it may not be clear whether this is high or low. However, if the average score in some large relevant comparison population is 20, and the standard deviation is 5, then the child's Z-score is 1. If scores in the comparison population really are normally distributed, then the child has obtained a score which is better than about 84% of similar children. In other words, this is one way of calculating percentiles based on norms.

Notice however, that the meaning of the score has changed in the process of transforming it. The original raw scores may have been based on what the child could actually do (e.g., the number of items correct), whereas the Z-scores and percentiles are only interpretable in terms of how far above or below average the child's performance is.

The usefulness of the Z-score therefore depends strongly on the appropriateness of the comparison sample.

Percentile

A **percentile** is the percentage of people who score below a certain value. Percentiles must therefore range between 0 and near 100. If the scores of a measure approximately follow the normal distribution, a child at the 50th percentile has obtained the average score, and has a score better than 50% of children in the relevant population.

Percentiles are often calculated by administering the measure using standard procedures to a large sample of children who are representative of a reasonable comparison group. These children form the normative sample. Scores for these children can be rank-ordered and tabulated to identify the percentiles.

There are some difficulties with this method, and as an alternative it is common to assume that the scores in the normative sample are normally distributed and estimate the percentiles using Z-scores.

Age-Equivalent Scores

For many measures it is possible to look-up **age-equivalent** or **grade-equivalent scores** from the raw scores.

Age-norms are found by calculating the average score obtained for children in each of several age-bands in the normative sample. The average score for each age-band defines its age equivalency. For example, if the average score for 6 year olds is 40, then any child who scores 40 has an age-equivalent score of 6 years.

There are some problems associated with the use of age-equivalent scores. One problem is that not all children within an age band in the normative sample have the same score. This source of uncertainty in the norms is not usually considered, for instance, by calculating the standard-error of the age-equivalent score.

T-Scores

T-scores are like Z-scores except that they are scaled to have an average and standard deviation other than 0 and 1. The alternative average and standard deviation are usually somewhat arbitrary and often depend on convention. For instance, by long-standing convention, intelligence tests have a mean of 100 and a standard deviation of 15.

Criterion referenced measures

As an alternative to norm-referenced measures, **criterion-referenced** measures do not use comparisons to a normative sample for interpreting scores.

Instead, criterion-referenced scores are interpreted by considering directly whether the child has met age-appropriate functional demands, as reflected in the content of the measure.

Scores are interpreted in terms of whether children can do specific tasks that are important for their age, grade, or clinical context, rather than how "normal", "typical", or "average" their performance is.

In other words, criterion-referenced measures are interpreted in terms of their content, rather than a population. Norms, percentiles, age-equivalent scores and other standardized scores should not be provided for criterion-referenced measures.

Rasch analysis and Item Response Theory

Item response theory (IRT) is a set of related approaches to the reliability and validity of measures. **Rasch analysis** is a particular version of item response theory. In contrast to approaches that define reliability in terms of a measure's ability to detect consistent variations among people, IRT methods use statistical methods to develop separate models of the difficulty of individual items and the functional abilities of children.

Although IRT methods can be more complex to use and interpret than traditional methods, there are some potential advantages. In particular, measures developed using IRT can generally be used for children having a wider variety of functional levels because children can be tested with the subset of items that is most relevant to their abilities. To use an IRT measure, you will be provided tables or a computer program that converts raw scores into IRT scaled scores. You may also be provided with aids such as item maps that describe the estimated difficulty of items by relating scaled scores to the probability of correctly performing the items.

For more information

The *CanChild* website has more information about measures and measurement issues: www.fhs.mcmaster.ca/canchild/. Inquiries can also be directed to:

Steven Hanna
CanChild Centre for Childhood Disability Research
McMaster University, IAHS Room 408
1400 Main St. West, Hamilton, Ontario, Canada L8S 1C7
phone 905.525.9140 ext 27851
fax 905.522.6095 email hannas@mcmaster.ca