WHAT IS MEASUREMENT?

Carl V. Granger, MD and Paulette Niewczyk, MPH, PhD   November 2010
John Michael Linacre, PhD provided editorial assistance

Thoughts based on review of -- Measuring Behaviors and Perceptions: Rasch Analysis as a Tool for Rehabilitation Research; Luigi Tesio: J Rehabil Med 2003; 35: 105-115

Historically, physical measurement developed from counting objects. For instance, oranges are countable entities. However, in the marketplace, since all oranges are not created equal, in order to price oranges fairly, an intangible but measurable entity, weight, had to be invented. Curiously, weight is "objective" even though it is a mental construct. Weight remains constant and allows for fair commerce with respect to trading oranges for money. In other words, measuring weight, not counting, makes oranges fungible.

In the same way, we want to develop measurement for behavioral phenomena. Again, this would start with counting observable activities. And again, intangible but measurable entities, now called "latent traits", will need to be invented. By "measurable" we mean what physical scientists usually mean, linear measurement of a continuous variable so that each additional equal amount adds exactly the same extra quantity of the trait to the total.

Before the 19th century, it was not considered possible to apply measurement to internal human experiences. Then "psychophysics" was developed to quantify touch pressure, sound pitch, etc. It became accepted that mental constructs could be accessible to measurement. So, measurement was extended to intelligence, depression, suffering, attitudes, and knowledge about various topics. However, it was thought that only the physical expression could be measured, such as crying as evidence for depression. This evolution of thinking about measurement continued to the point that allowed that a behavior need not be observed or stimulated directly; rather, a question might be asked to reveal "a latent trait or variable."

However, psychophysics sought to count the positive responses to successive questions as if each response had the same weight. This is the deterministic approach to measurement. On the other hand, the latent trait approach is probabilistic, in that it involves inferences. Since the possibilities are infinite, only a sample of behaviors manifesting the same construct need be used. Test items must represent varying quantities of a shared underlying construct and be clinically pertinent to the pathophysiologies of persons being tested.

Cumulative raw scores are only counts of discrete observations, no matter what numbers or labels are assigned to them. Thus, counting is not the same as measuring. Ordinal assignment, for example, 1 through 7, tells us that there is more of a quantity as

the number get larger, but not how much more. Measurement requires that the intervals between levels be uniform, that is, adding one more always adds the same amount, making the interval between 2 and 3 the same as that between 5 and 6. This illustrates the property of additivity.

The latent trait paradigm also requires that the items of a measure present a range of challenges from less to more for the individual for whom the test is being applied. It is expected that easy items are passed by almost anyone but difficult items are passed only by the more able or high-scoring subjects. This demonstrates the property of unidimensionality. Both additivity and unidimensionality are necessary components of the Rasch model.

When an attempt is made to "widen" a measure by adding an item that does not fit the construct, the result is that the summed scores misrepresent the true circumstance with respect to the main construct. Another distortion of measurement occurs when a subject of lesser ability is passing a difficult item. This cannot be explained except by the presence of an error in administering the test or some misrepresentation of ability. These are examples of violation of the principles of unidimensionality and additivity.

Raw ordinal scores are bounded by upper and lower limits, such as 1 to 5, but do not reveal the distance between levels. Therefore, one cannot assume that intervals between levels are equal.

Sometimes, ratio scores (proportions) or z-scores are used to form continuous scoring entities. However, these are not necessarily linear nor objective, thus, they are not satisfactory solutions for measurement. Logarithmic or "logit" transformation provides the solution to linearization of proportions.

Objective measurement is the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured. See http://www.rasch.org/define.htm

Objectivity is satisfied when the calibrations of the measure are independent of the sample from which they are derived. Thus, the difficulty of the items and the ability of subjects are independent of each other. Jack Stenner expands on Georg Rasch's use of "specific objectively" as basically two types, depending upon the level of theory underlying the construction of the particular measurement instruments. Local objectivity is the case in which relative measures are empirically discovered to be independent of which instrument is actually used to take the measurement, by estimating Rasch item measures from observations.

Local objectivity varies in precision and must be confirmed by further sampling. On the other hand, general objectivity is theory-based with fixed precision. This term means

that amounts being measured are independent of which instrument (within the appropriate class) is employed. An example of an instrument with absolute calibration is a thermometer. See http://www.rasch.org/rmt/rmt83e.htm

The Rasch measurement model (developed by Georg Rasch, a Danish mathematician) is the solution -- probabilities are dictated by a model imposed on the data, rather than having the model fit the data. The Rasch model is based on "item response" theory (meaning that the model tests whether items cooperate to form consistent calibrations forming a measure) and conjoint additivity (meaning that calibration of the items and measurement of subject ability are independent of each other although they are measured on the same metric). Therefore, it is the only system of true measurement of a latent trait. It is applicable to physical and psychosocial sciences.

Thoughts based on review of -- A Rasch Primer: The Measurement Theory of George Rasch by Ronald Mead, March 2008 of the Data Recognition Corporation

----

Rasch modeling is the core method for building and testing measures. The distinction between Rasch measurement and the more complex IRT (item response theory) models, is that Rasch intended to extract all information from the data that was *relevant* to the construct of interest, not how to reproduce the data most precisely. IRT models tend to focus on fitting to the data, while the older term *latent trait models* fits better the Rasch perspective, with the focus on the attribute to be measured. Rasch modeling takes more work and careful planning up front. Two commonly cited motivations for using Rasch models are (a) the mathematics are easy to apply compared with IRT models and (b) useful results can be gotten with relatively small samples and the estimation algorithms converge readily unless the data are pathologically bizarre. However, the fact that mathematics of Rasch modeling may be *easy to apply* does not necessarily mean that Rasch modeling is *simple to apply.* Instead, meeting Rasch requirements can be hard, particularly if prior planning has been neglected or misdirected, or data are poor.

The purposes of RA are to *maximize the homogeneity* of the trait and to allow greater *reduction of redundancy* at no sacrifice of information by decreasing items and/or scoring levels *to yield a more valid and simple measure*. At times this requires extracting from messy data measures that conform to a homogeneous latent variable and/or identifying for removal features of the data (e.g., bad items, mis-categorization) which contradict homogeneity of the measure.

Measurement theory refers to a body of principles, ideas, rules, and techniques for quantifying some interesting aspect of an object. The goal is to facilitate understanding with a valid instrument that is the best and most useful way to measure. Measurement precedes analysis. Typically, the intent is to make inferences based on the measures; however, it must be appreciated that *analysis is a distinctly separate process from measurement*. Measurement does not care if we simply collect and file the measures or

use them to achieve world peace! Rasch models provide a structure for using and not misusing the information contained in the responses. It is concerned with the very narrow problems of specifying the group of objects, defining the interesting aspect of the objects, determining relevant evidence, and transforming the evidence into a measure. It is about building rulers. The numbers themselves are meaningless until they are associated with objects that we are interested in and mileposts that mean something.

For Rasch, reproducing the observed item responses is not measurement. The construct is paramount, defined operationally by the items. Any empirical weights will change the definition from the one provided by the designers of the instrument. If different selections of data define the construct, through empirical weights, we no longer have a firm grasp on what we are trying to measure. If different subdivisions of the sample give different orderings of the items, then we do not know what *more or less* means. A measurement model should be sounding alarms in this situation, not adapting to it.

Measurement is *the process of quantifying an aspect of an object in a way that is general, reproducible, and amenable to analysis*. *General* means there is a broad class of objects and agents over which comparisons are valid and interesting. *Reproducible* means competent observers using appropriate instruments, perhaps of their own devising, will obtain statistically equivalent results. *Amenable* to analysis means standard statistics (e.g., means, variances, differences) will suffice.

The ultimate goal is to develop measurement scales that are well-defined and useful in the classroom and clinic. Thurstone and Guttman, and others, defined what must be. Rasch provides the mechanism to achieve it and the structure to know when we have it. It depends upon rigorous development of the agents (items) from a substantive theory and careful verification of performance based on data. When accomplished, we will have measurement instruments that we can place along side thermometers, rulers, and scleroscopes (a device to measure hardness of an object), no apologies needed.
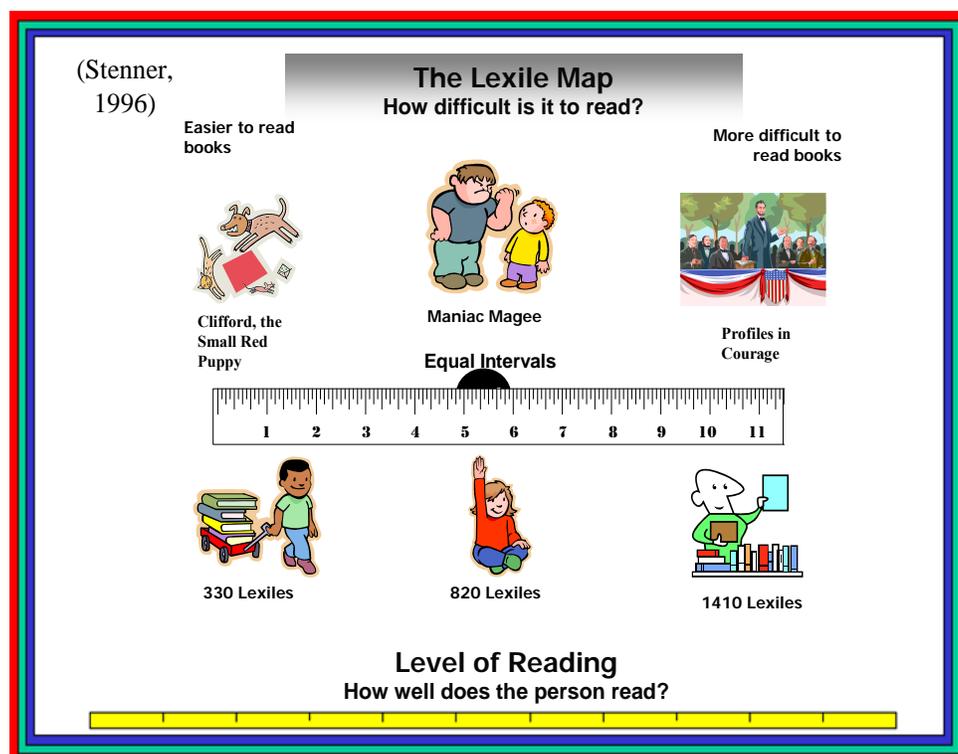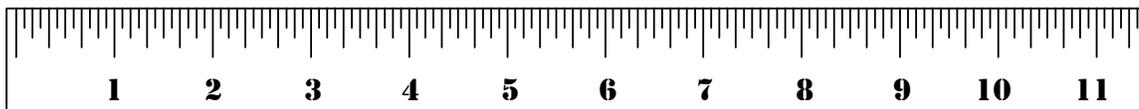
Proper use of Rasch modeling requires building instruments with items that are valid expressions or components of the idea being measured. They need not be perfect but they must be sufficiently appropriate to be useful. Rasch seeks to apply better methods for manipulating good data rather than better methods for manipulating poor data. Development of the construct must be "theory-driven and data verified." If the data do not verify the theory, then you need to rethink everything including, perhaps especially, the theory. Validity trumps reliability.

Initial applications of Rasch analysis have been in education. A current outstanding example of practical application follows to show the precision that can be achieved in measuring latent traits using the Rasch method. It is Jack Stenner's "Lexile Framework for Reading". This is an educational tool that links readers with reading materials using a common measure called a Lexile. Recognized as the most widely adopted reading

measure, a Lexile denotes both reading ability and text difficulty on the same scale. When used together, Lexile reader measures and Lexile text measures enable educators, parents and students to select reading materials that meet or else challenge a reader's unique abilities and interests.

But, overall, the biggest difference is pictures and charts versus tables of numbers. Of course we need the numbers, but it is pictures of placement on the latent trait variable, and pictured profiles of patient performance, that forcefully impart the message. The main point is that no matter where two points are place on a ruler and moved down or up, as long as the distance between the two points remains the same, it is the same distance on the ruler.

(Stenner, 1996)

The Lexile Map
How difficult is it to read?

Easier to read books

More difficult to read books

Clifford, the Small Red Puppy

Maniac Magee

Profiles in Courage

Equal Intervals

330 Lexiles

820 Lexiles

1410 Lexiles

Level of Reading
How well does the person read?

The young boy on the left has a reading level of 330 Lexiles and the book, "Clifford, the Small Red Puppy" has matching difficulty. When he becomes older he would be expected to read "Maniac Magee" which the older girl with a Lexile level of 820 is able to read. Beyond that, a college student would be comfortable reading a book such as "Profiles in Courage" at a difficulty level of 1410 Lexiles.

<u>Rasch Analysis (RA) is Important to Understand and Use for Measurement</u>
▪ In measurement, our intent is to use numbers (which are really raw scores/ratings) to indicate "*more*" or "*less*" *of the trait that is presumed to be homogeneous,* actually *an important part of investigation is to verify that the data reflect that homogeneity*.

▪ **RA is a unique approach of** *mathematical modeling* **based upon a** *latent trait* **and accomplishes stochastic (***probabilistic***)** *conjoint additivity* **(conjoint means measurement of persons and items on the same scale and additivity is the equal-interval property of the scale).**

▪ **The purposes of RA are to** *maximize the homogeneity* **of the trait and to allow greater** *reduction of redundancy* **at no sacrifice of information by decreasing items and/or scoring levels** *to yield a more valid and simple measure***. At times this requires extracting from messy data measures that conform to a homogeneous latent variable and/or identifying for removal features of the data (e.g., bad items, mis-categorization) which contradict measure homogeneity.**

▪ **RA permits** *rating of a limited set of attributes* **that are representative of the underlying trait, limited means that a small set may be sufficient.**

▪ **Whether observed or self-reported, the summed rating of the attributes represents how much of the trait has been mastered, since the raw score is the "sufficient statistic" for the Rasch measure.**

▪ **The model assumes that the probability of a given person/item interaction (in terms of rating high or low) is** *only governed by the difficulty of the item and the ability of the person***, that are determined by the item locations on the presumed latent variable along with the rating scale structure.**

▪ **Raw scores have unknown spacing between them. Rasch builds estimates of true intervals of item difficulty and person ability by** *creating linear measures.*

▪ **In this process,** *item values are calibrated and person abilities are measured on a shared continuum that accounts for the latent trait***. Should an item rating be missing, the model estimates the person's probable rating based upon the observed data without imputing the missing data.**

▪ **Concurrently, the improbability of a person's passing or failing a particular item is estimated item by item in terms of fit statistics. This is a comparison between what actually happened and what the model predicts should have happened based on the estimated measures.**

▪ *INFIT and OUTFIT* **statistics are the most widely used diagnostic Rasch fit statistics. Comparison is with an estimated value that is near to or far from the expected value. INFIT is more diagnostic when item measures are close to the person measures. OUTFIT is more diagnostic when item measures are far from the person measures. But, for long rating scales, like the FIM^{TM} instrument, this difference tends to disappear.**

▪ **The fit statistics indicate where the operator should decide whether to either delete, rescore, or reword an item. Deciding to how to select the number and cut-points of the rating categories is more complex, requiring a combination of fit, reliability and substantive meaning. Lopez (Lopez W (1996) Communication Validity and Rating Scales. Rasch Measurement Transactions 10:1 p482-3) identifies that respondents rarely make stable discriminations among more than 6 levels, especially if the categories are not clearly defined (criterion referenced). Excess categories can introduce more noise than information. Thus, Rasch analysis provides an opportunity to collapse categories in order to produce the best fit of data to the model.**

See http://www.rasch.org/rmt/rmt101k.htm

▪ The Rasch linear measures are originally expressed in log-odd units but may be rescaled to suit conventional scaling, as from 0 to 100 while still retaining conjoint additivity. The model also estimates the scoring error at each level as standard errors of the measure.

▪ Error is always greater at the upper and lower ends of a scale because the *Rasch model is not limited at the extremes, but measures from the middle of the range of values and anticipates infinity in both directions*. Measurement is better when the middle values of subjects lie close to the middle values of the measure. In other words, the true score is less certain as the limits of the scale are approached.

See http://www.rasch.org/rmt/rmt204f.htm.

▪ RA transforms ordinal scales into interval measures that may be used in parametric statistical analyses and the measures are characterized with standard errors for even more sophisticated analyses. Patient measures and calibration of individual item values are measured on the same metric and are locally independent, provided that Rasch criteria are met.

▪ Measures constructed using RA are unidimensional and have predictable hierarchies of item calibrations that span the range of difficulty within a domain of assessment.

▪ Final measures are built by the operator based upon the best judgments of:
   - spread of item values (eveness of steps)
   - reduced error of measurement (precision)
   - probability and improbability (fit) of item and person values to that expected
     from the model
   - overall reliability (noise)
   - simplicity, and
   - conformity to the nature of the clinical values that are being measured

▪ Building measures using RA requires that the data fit the model, not that the model fit the data

▪ Rasch modeling facilitates analysis of responsiveness of individual items with respect to their calibrated positions within a measure

▪ Overall, Rasch analysis provides an internally valid measure that, when developed from an appropriate sample, is independent of the particular sample to which it is applied, meaning that the findings for the sample extrapolate to its population. In building a measure, the sample must be of sufficient size, the ranges of person-ability and of item-difficulty must coincide, and relationships between items must maintain a coherent pattern from low to high.

Precision case management for rehabilitation inpatients is facilitated by having the physician manage and steer the clinical team toward specific FIM[TM] item goal values that are consistent with returning patients to live in the community, across the various impairment types.

How Rasch analysis (RA) has been used to shape the LIFEware® System of UDSMR

- **The top priority in measurement is validity. Validity is evidence that an instrument or measure is being used appropriately and measures what it is supposed to measure. While reliability tells us whether an instrument is measuring something consistently, only validity can provide us with information about what an instrument really measures.**

- **Construct validity is the degree to which an instrument accurately measures what it is supposed to measure. One method used to establish construct validity is RA. RA is based on mathematical probability. RA takes ordinal level data such as items rated 0 or 1 (for correct or incorrect, high or low), or 1, 2, 3, 4 or 5 (for several ordered categories, such as degree of impairment from "severely impaired" to "not at all impaired") to indicate levels of a response on some variable. These responses are then added across all items to give each person a total rating on that measure. This total score summarizes the responses to all the items, and a person with a higher total rating displays more of the variable assessed. Summing the ratings of individual items to give a summed rating for a person implies that the items are intended to measure a single dimension or construct (pain, for example), often referred to as unidimensional.**

- **RA is a logarithmic version of IRT (item response theory) however it is unique because it relies upon a mathematical model based on probability. Rasch analysis is the only methodology known to form a "true measure." RA requires that the data fit the model, not that the model fits the data.**

- **In addition to providing results at the person-level, RA also places each item of the measure into a hierarchical order of item difficulty. Rasch provides indicators of how well items cooperate with each other in maintaining their expected calibrated positions in the hierarchy, thus, fitting within the underlying construct being described by the measure. For example, if measuring pain, how well do the items combine as a measure of pain? Items that do not fit the unidimensional, hierarchical construct, deviate (or misfit) unacceptably by not maintaining an expected calibrated position in the hierarchy.**

- **While Rasch analysis is not reliant on sample size, it is necessary that sample size is sufficient to assure that items and categories are representative of the construct/dimension being measured. The intent of Rasch modeling is to create a measure that is "sample free," meaning being independent of the sample from which it was derived. Thus, while a minimum sample size is not a requirement to perform Rasch analysis, the interpretation is most reliable with at least 50-100 subjects. While RA can handle missing data by estimating the value of a missing item, we prefer not to use this feature because we wish to test how each item responds on each individual assessment. RA detects and excludes persons who are rated at the extremes (when all items are rated either at the highest or at the lowest response categories).**

In building a Rasch measure, it is imperative to be guided by clinical reasoning in the organization of and performance of Rasch analysis, as well as in the interpretation of the results. Rasch modeling requires satisfaction of the following attributes:

- "Raw" scores are labels, not quantities – raw scores must be transformed in order for the sum to be a "true" measure
- Unidimensionality – all items contribute to a single 'latent trait'
- Hierarchical structure – items individually represent a known amount of difficulty
- "Fit" – item values are calibrated within a range of values and must maintain close to their expected positions under low or high test conditions
- Categorical steps/levels must maintain their order
- Rasch measures are independent of the data source as long in creating the Rasch model the amount and distribution of items and levels had been representative of data to be measured
- Rasch modeling establishes "conjoint additivity" meaning that
  1. the two crucial elements of the model are "person ability" and "item difficulty" and
  2. both elements are measured on the same metric.
- Prior to applying the measure for testing data, the item difficulty values and categorical steps/levels are anchored in order to produce a stable measure

In summary: Verified measurement properties require use of Rasch analysis, which is a mathematical theory-based method for transforming raw score values into measures that satisfy the requirements for construct validity in measuring latent trait variables.

Latent trait variables refer to those qualities of human functioning and behavior that are not measured by characteristics of weight, length, etc. Yet, whole person qualities of being independent in one's personal care, emotion/ mood, social interaction, role participation, and pain (sometimes included in 'quality of life' assessments) are real, but there is a lack of appropriate metrics.

The structural requirements for creating measures include (a) a hierarchical, calibrated relationship of items to each other and (b) predictive validity with respect to ratings for individual patients, in which the total for the measure should reflect likely ratings of individual items.

Display of item values within a measure reveals what specific functions a particular patient may or may not be able to do, with respect to the items that comprise the measure. This feature enhances a clinician's ability to manage a patient in more specific ways. Such information is useful for medical decision-making as to what needs to be the next focus of treatment.Thus, functional assessment tools must have the characteristics of measurement – such as unidimensionality, equal interval and hierarchical calibration of item values – in order to accurately reflect the qualities that are inherent in analyzing

the *latent traits of human functioning and behavior*. From a larger perspective, non-linearity of raw scores (i.e., ratings that have not been "Rasched"), may give misleading information about treatment effectiveness.

Rasch Analysis- A Brief Description: Rasch analysis has been used for instrument assessments in various disciplines including health and medicine, education, business and marketing, and the social sciences. Rasch analysis takes ordinal level data such as items scored 0 or 1 (for correct or incorrect, high or low), or 1, 2, 3, 4 or 5 (for several ordered categories, such as degree of impairment from severely impaired to not at all impaired) to indicate levels of a response on some variable. These responses are then added across items to give each person a total score. This total score summarizes the responses to all the items, and a person with a higher total score displays more of the variable assessed. Summing the scores of the items to give a single score for a person implies that the items are intended to measure a single dimension or construct (pain, for example), often referred to as a unidimensional. In addition to providing results at the person-level, Rasch analysis also converts each item of the scale into hierarchical properties, or simply put, into an order of item difficulty. Rasch provides indicators of how well each item fits within the underlying construct or measure (if measuring pain for example, how well does each item measure pain), items that do not fit the unidimensional construct (straight line) are those that diverge (or misfit) unacceptably from the expected pattern. One of Rasch analysis's greatest uses is to establish construct validity of an instrument. Construct validity is evidence that an instrument accurately measures the construct of concern, or simply stated the instrument measures what is says it measures. In addition, Rasch analysis is not reliant on sample size and can handle missing data; it can also detect and account for persons scoring at the extremes (all very high or very low responses).

Further clarification of the Rasch model may be read in the chapter by A H Elhan, A A Kucukdeveci, and A Tennant entitled The Rasch Measurement Model in Advances in Rehabilitation, Research Issues in Physical & Medical Rehabilitation Medicine, Vol 19, 2010, pp 89-102, Pavia, Maugeri Foundation Books, Franco Franchignoni, editor.

http://www.udsmr.org/Documents/UDSMR_What_Is_Measurement_Article.pdf

Quote from a sign in Albert Einstein's office: "Not everything that counts can be counted, and not everything that can be counted counts"