The Use of Rasch Analysis To Produce Scale-Free Measurement of Functional Ability

Craig A. Velozo, Gary Kielhofner, Jin-Shei Lai

Key Words: independent living skills (human activities) • measurement scales

Innovative applications of Rasch analysis can lead to solutions for traditional measurement problems and can produce new assessment applications in occupational therapy and health care practice. First, Rasch analysis is a mechanism that translates scores across similar functional ability assessments, thus enabling the comparison of functional ability outcomes measured by different instruments. This will allow for the meaningful tracking of functional ability outcomes across the continuum of care. Second, once the item–difficulty order of an instrument or item bank is established by Rasch analysis, computerized adaptive testing can be used to target items to the patient's ability level, reducing assessment length by as much as one half. More importantly, Rasch analysis can provide the foundation for “equiprecise” measurement or the potential to have precise measurement across all levels of functional ability. The use of Rasch analysis to create scale-free measurement of functional ability demonstrates how this methodology can be used in practical applications of clinical and outcome assessment.


Rasch analysis has been increasingly used in the study and development of occupational therapy and health care measurement. In occupational therapy, Rasch analysis has been used in the development of the Assessment of Motor and Process Skills (Fisher, 1993), Assessment of Communication and Interaction Skills (Forsyth, Lai, & Kielhofner, in press), Work Environment Impact Scale (Corner, Kielhofner, & Lin, 1997), Volitional Questionnaire (Chern, Kielhofner, Heras, & Magalhaes, 1996), and the Pediatric Evaluation of Disability Inventory (Haley, Coster, Ludlow, Haltiwanger, & Andrellos, 1992). It has also been used to assess three major global functional status measures, the Functional Independence Measure (FIM™; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994), Patient Evaluation Conference System (PECS; Silverstein, Fisher, Kilgore, Harley, &

Craig A. Velozo, PhD, OTR, is Associate Professor, Department of Occupational Therapy, University of Illinois at Chicago, M/C 811, 1919 West Taylor Street, Chicago, Illinois 60612.

Gary Kielhofner, DPh, OTR, FAOTA, is Professor and Head, Department of Occupational Therapy, University of Illinois at Chicago, Chicago, Illinois.

Jin-Shei Lai, PhD, OTR, is Research Assistant Professor, Department of Occupational Therapy, University of Illinois at Chicago, Chicago, Illinois.

This article was accepted for publication July 2, 1998.
focuses on the test, the Rasch measurement model focuses on the test item. This paradigm shift technically frees measurement from the use of particular tests, a concept referred to as scale-free measurement (McHorney, 1997).

The purpose of this article is to demonstrate two practical examples of scale-free measurement as applied to rating scales of functional ability: (a) the translation of measures across instruments and (b) the use of item banks and computerized adaptive testing to achieve equiprecise measurement (i.e., precision in measuring persons across the entire range of abilities).

**Rasch Measurement Model**

To demonstrate the logic underlying the applications of scale-free measurement, we will use the Rasch formula. Although potentially intimidating, this formula represents the simplest mathematical formula of what is both necessary and sufficient for measurement. Hopefully, through understanding the mathematical formula underlying Rasch analysis, one can become more confident in the claims of this psychometric approach. All the applications of Rasch analysis proposed in this article are founded on this formula.

The many-faceted Rasch formula, which is appropriate for analyzing rating scales (Linacre, 1994), is

$$\log \left[ \frac{P_{nijk}}{P_{nijk-1}} \right] = B_n - D_i - C_j - F_k$$

where

- $P_{nijk}$ = probability of person $n$ being rated at step $k$ on domain $i$ by rater $j$
- $P_{nijk-1}$ = probability of person $n$ being rated at step $k-1$ on domain $i$ by rater $j$
- $B_n$ = ability of person $n$
- $D_i$ = difficulty of item $i$
- $C_j$ = severity of rater $j$
- $F_k$ = difficulty of rating step $k$ relative to step $k-1$

The left side of the equation presents the linear transformation of raw scores into interval-based measures. Taking the log of any value or arithmetic operation yields linear values or measures with equal intervals. This method of log transformation is well documented in the basic sciences as a way of creating a linear measurement between two variables. Converting ordinal scores into interval measures is a critically important aspect of Rasch analysis because all rating scales of functional ability yield raw scores, which are ordinal. The kinds of mathematical operations we routinely perform on raw scores (e.g., statistical analysis) improperly treat ordinal data as if they are interval.

Although this practice is common, it is not inconsequential. Treating ordinal data as interval data can result in incorrect conclusions about, for example, whether the outcomes of two different treatment approaches or facilities are different. (See Fisher [1994] for a more detailed explanation of the importance of interval measures.)

The left side of the equation represents, in its simplest form, the probability of a person passing a test item divided by the probability of failing a test item. This is easiest to understand when a person is tested on a dichotomous item (i.e., an item asking whether a person can or cannot perform a given functional task). However, the concept equally applies to an item that is rated with a rating scale. In this case, the person either passes or fails a given rating on the item. For example, receiving a score of 4 (minimal contact assist) on the upper-extremity dressing item on the FIM (which uses a 7-point scale) assumes that the patient has passed the item at ratings 1, 2, 3, and 4 and failed the item at ratings 5, 6, and 7.

Said another way, the equation represents the probability of receiving a particular rating (in this case 4) versus the rating below or above that rating.

The right side of the equation presents each observed rating as the probabilistic result of four interacting components: (a) the ability of a person being measured; (b) the difficulty of an item or, conversely, how much ability is required to pass the item; (c) the severity of a rater; and (d) the structure of the rating scale (Linacre, 1994). The component $B_n - D_i$ is the core of the Rasch formula and fundamental to understanding scale-free measurement. The essence of any assessment situation is that a person of a particular ability is assessed on an item that has a particular difficulty. The result of the interaction between the person’s ability and item difficulty is that the person either passes or fails that item or receives or does not receive a particular rating on that item. From the result of this basic person–instrument interaction, our goal is to solve the equation for $B_n$ (the ability of the person).

Rasch methodologies also take into account more complicated assessment situations. For example, when a therapist scores a patient, $C_j$ (rater severity) comes into play. Severity refers to a rater’s tendency to give higher or lower scores relative to other raters. The dramatic influence that raters have on test scores has been demonstrated by a number of researchers (Englehard, 1994; Fisher, 1993; Forsyth, Lai, & Kielhofner, in press; Lai, Velozo, & Linacre, 1997; Lunz & Stahl, 1993; Wilson & Wang, 1995). By having all therapists rate a set of the same patients, Rasch analysis can be used to determine each rater’s severity and remove the influence of rater severity for all future ratings of patients.

The final aspect of the Rasch formula is $F_k$, which rep-
resresents the “step” from one rating to the next. We assume that for multiple-point Likert scales, the intervals between ratings are equal across all ratings (e.g., 1–2 = 2–3 = 3–4 on a 4-point scale). It has been extensively demonstrated that although rating scales are routinely assumed to have equal intervals, in reality, the intervals between each pair of the rating points are not equivalent (Merbitz, Morris, & Grijp, 1989). Rasch analysis can display and correct for the inequalities in these intervals. These unequal intervals may have important implications for assessing patient improvement. If, for example, the actual interval between a rating of 2 and a rating of 3 is twice that of the interval between 3 and 4, a patient who improves on a functional ability item from 2 to 3 has shown twice as much improvement as a patient who has improved from 3 to 4. With conventional psychometric approaches, both patients would be assumed to have improved the same amount.

The importance of converting ordinal rating scale data to interval data, assessing rater severity, and equalizing rating steps is among the reasons why many researchers increasingly use Rasch analysis to develop instruments or surveys. Nonetheless, assessment development and improvement of individual instruments represent only the most basic application of Rasch analysis. Because the Rasch measurement model focuses on the relationship between what we are attempting to measure (i.e., patient ability) and the test item (i.e., item difficulty or amount of ability required to pass the item), it becomes possible to think beyond the bound of any single instrument to the larger problem of measuring functional ability. This inherent relationship between person ability and item difficulty has led researchers to the idea of scale-free measurement and its application in translating measures across instruments.

Translating Measures Across Instruments

The first example of scale-free measurement is the application of Rasch methodologies to translate measures across instruments. The inability to convert scores from one instrument to another is a major challenge of outcome measurement. Without this capability, patient progress cannot be monitored across the continuum of health care (acute hospitalization, subacute care, postacute care, nursing home care, outpatient services), where different instruments are typically used to measure functional ability. Furthermore, the linking of instruments could do much in resolving the inability to replicate scientific findings. The difficulty with replication may often be the result of different studies using different instruments to measure the same trait or construct (e.g., quality of life).

The most often-cited solution to this dilemma of comparing different measures is to abandon the use of multiple measures and to adopt a “gold standard” in measurement. To measure general health, the instrument of choice has been the SF-36 and, in rehabilitation, the measure of choice has been the FIM. The problem is that choosing a gold standard for measurement is fraught with flaws (Velozo, 1994). For example, for patients undergoing vocational rehabilitation, the FIM demonstrates ceiling effects that can compromise demonstrations of treatment effectiveness (Jessop, 1996).

Social science is not the first science to be challenged with the problem of scale-free measurement. Many physical science measures originally depended solely on the specific instrument used. For example, time was originally measured by different and often incompatible instruments (e.g., sundial, hourglass, water clock) (Fisher, Harvey, Taylor, Kilgore, & Kelly, 1995). Renowned social science statisticians (e.g., Thurstone, 1929) recognized that the size of a measuring unit should not be affected by what is measured or by the measuring instrument (Fisher et al., 1995). Although this ideal was recognized, the methods for realizing it were not yet available. This began to change when George Rasch (1960) formulated the idea mathematically in what he called the separability theorem. This theorem required that the data be modeled to support comparisons of subjects without regard to items and comparisons of items without regard to subjects (Fisher et al., 1995).

Scale-free measurement in the social sciences means that what we measure is independent of the instrument used for measurement. The independence of what we measure and how we measure can be explained by the Rasch formula. This scale-free principle is built into the Rasch formula in such a way that two subjects’ abilities can be compared no matter what tests are used to assess the particular subject’s ability.

Figure 1 presents scale-free measurement using the Rasch model. The core of the model, $B_n - D_i$, is used to compare the abilities of two patients, $B_1$ and $B_2$. The center line in the dashed frame represents functional ability measured by a common scale (i.e., a standard unit of functional ability). Instrument X is represented by four items of increasing difficulty ($D_{x1}, D_{x2}, D_{x3}, D_{x4}$) and Instrument Y is represented by another four items of increasing difficulty ($D_{y1}, D_{y2}, D_{y3}, D_{y4}$). Each item’s placement along a scale represents the relative difficulty of the item (items toward the bottom of the scale are easier, items toward the top of the scale are harder).

As presented in the Rasch formula, person ability ($B_n$) is determined by comparing it to item difficulty ($D_i$). This comparison is achieved from the rating data (i.e., Does the patient pass or fail the items? What ratings does the patient achieve on the items?). Given the difficulty of the two sets of items and the ability of the two patients shown, we would expect that both patients passed the two lower ability items and failed the two higher ability items or, alternatively, that they received higher ratings on the lower ability items and lower ratings on the higher ability items.

If items $D_{x1,2,3,4}$ and items $D_{y1,2,3,4}$ are placed on the
same scale or a common scale (as indicated by the center line in the dashed frame), the ability of Patient B1 can be directly compared with Patient B2. This is accomplished even though the two patients are assessed on two different instruments because the instruments measure the same trait or construct and can be linked to a single standard or common scale for measuring. Just as a mechanical clock and a quartz digital wristwatch both measure time according to a common standard of hours and minutes but in different ways, two instruments may measure the same functional ability using different methodologies but using a common scale.

The possibility of scale-free measurement of functional ability was underscored when the order of activities of daily living (ADL) mobility items (from less to more difficult) on the FIM, on the PECS, and on the LORS-III were found to be virtually identical (Linacre et al., 1994; Silverstein et al., 1991; Velozo et al., 1995). This finding gave evidence that the scales were likely measuring the same underlying functional ability, despite differences in item definitions and rating method.

The feasibility of equating instruments can be demonstrated by what Fisher (1997) referred to as “pseudo-common item equating” (p. 89). In this process, the calibrations or difficulty levels of similar items from two different instruments can be compared. If two instruments are measuring the same construct, the item difficulty order of the two instruments should be similar. Figure 2 shows a scattergram of Rasch measures of the ADL-mobility items of the FIM plotted against the ADL mobility items of the LORS-III calculated on different samples of rehabilitation patients obtained from Medirisk of Chicago, Illinois, Inc. Item calibrations are based on 10,509 rehabilitation patients for the FIM and 2,962 patients for the LORS-III. Measures for each item are presented in logits (log equivalent units), with the high negative numbers representing the scores of “easy items” and the high positive numbers representing the scores of “hard items.” As can be seen from the figure, except for the feeding/eating item, all items fall very close to the 45° diagonal line, indicating similar calibrations for similar items.

Whereas the demonstration seen in Figure 2 suggests that the translation of measures across instruments is feasible, actual instrument linking can be accomplished through common-sample equating (Fisher et al., 1995). For this method, a group of patients needs to be assessed on both instruments. Data from both instruments then are “co-calibrated” by analyzing the items from both instruments as if they were a single scale, producing item measures using the same scaling unit. The final step in linking instruments is to analyze the data from each instrument in separate Rasch analyses, anchoring items at the co-calibrated item measures. Rasch analysis provides tables that convert raw scores on each instrument to the common scale, thereby linking the instruments. Because the results of Rasch analysis are sample free, these tables can be used for all future instrument-to-instrument conversions.

Fisher et al. (1995) have demonstrated functional scale linking. Using the previously described methodology, they showed that the 13 FIM and 22 PECS motor skill items
could be scaled together in a 35-item instrument. The authors found the correlation of separate FIM and PECS measures to be .91 and with the co-calibrated values produced by Rasch analysis to be .94. Furthermore, Fisher et al. demonstrated that either instrument's ratings were easily and quickly converted into the other's via a table that used a common unit of measurement, which they referred to as the “rehabit.” This common unit of measurement (represented by the “common scale” in Figure 1) allows the translation of scores from one instrument to another.

The linking of similar instruments is well overdue in occupational therapy and in all of health care. There are obvious similarities across instruments developed to measure similar phenomenon, such as functional ability. The linking method provided by Rasch analysis would allow the continued use of existing instruments, while making patient scores obtained on those instruments comparable. This linking method would also enhance the use of existing databases because, for example, we would be able to follow patient progress across the continuum of care even when different instruments are used to measure functional ability at different points in time.

It should be noted that the linking of instruments should not be taken as a causal statistical exercise. Different functional assessments may measure different and unique aspects of function. For example, Glass (1998) showed that measures of functional capacity should not be equated to measures of functional performance. It is important that instrument linking be preceded by a strong theoretical foundation. Statistical methods used without such a foundation can lead to considerable misinference.

The linking of existing measures of functional ability would be an important step toward the standardization of how we measure functional ability. It would lead to advances in research by allowing comparisons across instruments that previously could not be legitimately made. It would also have practical significance for therapists who would be able to compare a patient's score on different instruments.

**Equiprecise Measurement**

A common critique of global functional status measures, such as the FIM, is their lack of precision for clinical assessment (Fisher, 1992). Therapists often complain that these instruments are insensitive to the actual improvements that are the result of their clinical interventions. For example, a patient who is initially rated as minimal contact assistance for dressing may show improvements as a result of treatment (e.g., may advance from not being able to get trousers over hips to now being able to accomplish this aspect of the task) but will not show improvements in the FIM rating if he or she has not achieved independence (e.g., cannot zipper trousers). To measure smaller increments of improvement, more detailed aspects of functional ability need to be assessed. The problem is that such an assessment would require an inordinate number of items. Furthermore, such detail would at times be cumbersome and unnecessary because many patients may have few problems with activities such as dressing.

A solution to this problem is the development of what Weiss (1982) calls equiprecise measurement. This equiprecise measurement refers to having high precision potential in measuring a trait or construct across the desired range of that trait or construct. For example, with considerable precision, equiprecise measurement of functional ability suggests the capability of measuring a wide range of abilities from the patient with the least function to the patient with the most function. Unlike current assessment practice, wherein the items used depend on the instrument, the selection of items to assess a given patient would be determined by the ability level of the patient being measured. For example, when measuring the physical ability of a mildly injured Olympic athlete, items would be chosen that matched the athlete's superior physical ability. Alternatively, when measuring the physical ability of a patient weakened by muscular dystrophy, a different set of items would be chosen that matched the patient's severely impaired physical ability. These two persons, although assessed on different items, would nevertheless be assessed on the same functional ability scale in the same way that persons 5 ft tall or 6 ft tall are measured on different parts of the ruler but still in feet and inches.

The ability to measure patients on different items and yet be able to compare the patients is founded on principles of the Rasch measurement model. Interpretation of the Rasch formula $B_n = D_i$ suggests that the most important information from a patient is obtained when items are matched to the difficulty level of the patient, where $B_n = D_i$. Technically, patient ability matches item difficulty when the patient has a 50% probability of passing or being successful on an item. That is, items with difficulties nearest to this 50% probability are the most sensitive to the patient's actual abilities. Conversely, items for which the patient has much lower or higher probability of passing provide less useful information about the patient's ability.

The efficiency of measuring patients with items focused at their ability level is demonstrated in Figure 3, where the ability of the patient with lower ability is represented by B1, and the ability of the patient with higher ability is represented by B2. Both patients are being measured on a scale of items of low difficulty (D1) to items of high difficulty (D10). In this figure, the ability of the less able patient is more appropriately measured with items of difficulty D1 to D5 because these items match the low abilities of Patient B2. In contrast, the ability of the more able patient is more appropriately measured with items of difficulty D6 to D10, because these items match the greater abilities of Patient B2. It would be imprudent to measure a patient of ability B1 with items of difficulty D6 to D10 (the difficult items) and equally questionable to...
measure a patient of ability B2 with items of difficulty D1 to D5 (the easy items).

In addition to efficiency, there is an increase in precision associated with using ability-matched items. Figure 4 presents a comparison of two patients of ability B1 and B2 who have similar but not identical abilities. Example 1 shows an attempt to assess these patients with items that are poorly matched to their abilities. In Example 1, the two patients’ abilities would not be differentiated; they would get the same expected score. In contrast, if these patients were assessed on items that were well matched to their ability levels, as in Example 2, they would be differentiated because they would get different expected scores.

The efficiency and precision of equiprecise measurement can also be presented in a practical example, such as attempting to compare two patients of similar “moderate” culinary abilities (e.g., the ability to make a three-course meal). If we tested the patients with an easy activity, such as making a peanut butter sandwich, the patients would not be differentiated, because both would easily perform that type of activity. Similarly, if we tested the patients with a hard item, such as preparing a 7-course meal, again they would not be differentiated, because both would be unable to perform that type of activity. The most information would be obtained when the patients attempted three-course meals of slightly different difficulty levels. For example, one patient may be capable only of frying a steak (a task requiring few steps), whereas the other patient may be capable of completing a more complex task such as making a soup (a task requiring several steps).

The question is, if the patient’s ability level before testing is not known, how can items be targeted to the patient’s ability level? The answer lies in the combination of item banking and adaptive test administration techniques. With a large pool of items calibrated according to their difficulty through Rasch analysis, modern testing techniques, such as computerized adaptive testing (CAT), can be used to quickly locate the patient’s ability level. CAT techniques provide an algorithm, or set of rules, for efficient presentation of items in a testing situation. For example, the simplest algorithm consists of first presenting a “middle-difficulty” item. If the patient passes this item, a more difficult question will be asked. If the patient fails the item, an easier item will be asked. This process continues, targeting the questions toward the patient’s ability level, where the patient has 50% probability of passing an item. This method, therefore, provides high precision while reducing the number of unnecessary test items presented to the patient. It is estimated that CAT procedures reduce testing length to one half while maintaining precision of measurement across the construct continuum (Weiss, 1982).

Equiprecise measurement is in its infancy, but it is likely to transform assessment in the not too distant future. Fisher (1997) has already argued that rehabilitation professionals need to move toward the identification of a standard unit of functional ability. The ongoing pressure for cost containment combined with the increasing need to accurately measure functional ability makes the approach of item banking for equiprecise measurement highly attractive. Moreover, the rapid emergence of new computer technologies will make the collection of data in the electronic medium increasingly feasible and cost-effective.
Conclusion

The promise of scale-free measurement in occupational therapy and health care appears inevitable. Rasch measurement is quickly becoming the standard for creating and evaluating measurement instruments. Data collection technologies, such as CAT, are highly advanced and quickly being adopted in other fields, such as medical licensing examinations and standardized educational testing (Bergstrom, 1996; Eignor, 1993). Because short-form surveys, such as the SF-36, have been shown to lack precision, the application of equiprecise measurement to generic health surveys appears imminent (Bjorner & Ware, 1998; McHorney, 1997).

The use of scale-free measurement to translate scores across assessments could be valuable in resolving the problem of monitoring patient progress across the continuum of care and facilitating a postacute health care prospective payment system (PPS). For example, three generic assessments presently dominate postacute health care: the FIM for rehabilitation, the Minimum Data Set (MDS) for skilled nursing, and the Outcomes and Assessment Information Set (OASIS) for home health. A fourth instrument, the Minimal Data Set for Post-Acute Care, is presently under development to address the needs of skilled nursing facilities, subacute rehabilitation hospitals, and long-term care hospitals (Health Care Finance Administration [HFCFA], 1998). The HCFAs (1998) goal for a “beneficiary-centered” PPS will require the linking of common data elements from these instruments. Rasch analysis would be the ideal methodology to accomplish this linking.

Whereas the linking of the above instruments is possible, such linking would do little to improve precision or efficiency of measurement in postacute care. For example, the ADL items of the FIM, MDS, and OASIS appear more similar than different (each instrument includes eating, dressing, toilet use, bowel, bladder, transfer, and locomotion items). Although the rating scales differ, none represent a major psychometric advantage in the measurement of function.

It would seem antithetical to continue to use assessments that lack the precision and efficiency of which we are now capable. An alternative to using the accepted global functional status measures is to develop an equiprecise measure of function. Rasch analysis can be used to identify functional items that have a distinct hierarchy of difficulty. Once this hierarchy is established, a computer algorithm that identifies the optimal set of items can be created to provide the most precise information for a particular patient (Bjorner & Ware, 1998). Patient ability measures could be compared even though patients answered different items because Rasch models score all items on a common metric.

Occupational therapy has a long history of concern for and expertise in functional ability assessment, and occupational therapists, along with other health professionals, have a major role in measuring functional ability. However, our contributions to the creation of widely used functional ability instruments have been less obvious. As health care is increasingly managed and standardized, certain core assessments will be required for administration to patients. To a large extent, such core assessments drive both what is considered important in health care and how health services are assessed. Consequently, the measurement tools that are used have a considerable impact in shaping health care. Through successfully competing for federal funds to develop more advanced measures of function, occupational therapists could lend their perspective and expertise to impact the future health care system. ▲

References


