Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 57, No. 8, December 15, 2007, pp 1358–1362
DOI 10.1002/art.23108
© 2007, American College of Rheumatology

SPECIAL ARTICLE

The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper?

ALAN TENNANT[1] AND PHILIP G. CONAGHAN[2]

Introduction

When evaluating outcome tools, rheumatology researchers are familiar with the traditional standards of measurement science: validity, reliability, and responsiveness (1). In these approaches construct validity may have been supported through factor analytic techniques that confirmed the presence of 1 or more valid unidimensional scales (a scale measuring a single construct). Reliability is usually reported as Cronbach's alpha (2). Attention has also been given to the concept of responsiveness, the ability of the scale to detect change through measures such as the effect size or standardized response mean (3).

Modern psychometric approaches have been adopted to supplement this traditional approach (4). These new approaches have been mostly associated with the application of the Rasch measurement model (5), usually referred to as Rasch analysis. Although Rasch has been widely used in education for the last 40 years, over the last decade a wide variety of applications of Rasch analysis have been published in the health sciences. These range from early work on widely used patient-reported scales such as the Health Assessment Questionnaire (6) through the revision of observer-evaluated scoring systems such as Larsen radiographic scoring (7) to the recent introduction of the concept of item banking (8). However, understanding of this analytical technique and its applications remains somewhat limited. This has not been helped by the use of different software packages producing different Rasch-related statistics. This article seeks to provide an update on the use of Rasch analysis; explain what it is, why it should be used, when to apply it; and provide guidance on what the current state-of-the-art analysis should comprise, and consequently what should be expected in a paper that uses Rasch analysis.

[1]Alan Tennant, BA, PhD: University of Leeds, Leeds, UK; [2]Philip G. Conaghan, MBBS, PhD, FRACP, FRCP: University of Leeds, Leeds Teaching Hospitals Trust, and Northwest Primary Care Trust, Leeds, UK.

Address correspondence to Alan Tennant, BA, PhD, Academic Unit of Musculoskeletal Disease, Faculty of Medicine and Health, The University of Leeds, 36 Clarendon Road, Leeds, LS2 9NZ, UK. E-mail: a.tennant@leeds.ac.uk.

Submitted for publication November 7, 2006; accepted in revised form July 3, 2007.

What Is Rasch Analysis?

Rasch analysis is the formal testing of an outcome scale against a mathematical measurement model developed by the Danish mathematician Georg Rasch (5). The model operationalizes the formal axioms that underpin measurement (9). These axioms, of additive conjoint measurement, are the rules for making measurements and will determine if ordinal or interval scales have been constructed (10,11). Rasch analysis provides the opportunity to examine to what extent the responses from a scale approach the pattern required to satisfy the axioms, and thus construct measurement.

The Rasch model shows what should be expected in responses to items if interval scale measurement is to be achieved. Dichotomous (5) and polytomous (12) versions of the model are available, the latter dealing with scales whose items have multiple response categories (e.g., 1, 2, and 3). The response patterns achieved from a set of items in a questionnaire that are intended to be summed together are tested against what is expected by the model, which turns out to be a probabilistic form of Guttman scaling (13). Guttman scaling is a deterministic pattern that expects a strict hierarchical ordering of items (e.g., from low to high levels of activity limitation) such that if (in the dichotomous case) a patient has affirmed an item representing a task of average difficulty, then all the items below that task on the scale (i.e., easier tasks) should also be affirmed. The Rasch model relaxes this to say that if a harder task is affirmed, then there is a high probability that easier tasks will also be affirmed.

The model assumes that the probability of a given respondent affirming an item is a logistic function of the relative distance between the item location and the respondent location on a linear scale. In other words, the probability that a person will affirm an item is a logistic

function of the difference between the person's level of, for example, pain and the level of pain expressed by the item, and only a function of that difference.

Why Should Rasch Analysis Be Used?

Patient-reported outcomes are frequently used for evaluation in clinical trials, where valid change scores and access to parametric statistics are required (14). Rasch analysis can support this process by providing a transformation of an ordinal score into a linear, interval-level variable, given fit of data to Rasch model expectations.

Consequently, Rasch analysis allows for a unified approach to several measurement issues, all of which are required for the validity of the transformation to interval scaling: testing the internal construct validity of the scale for unidimensionality, required for a valid summed raw (ordinal) score; testing the invariance of items (that is, the ratio of difficulties between any pair of items remains constant across the ability levels of respondents), required for interval-level scaling; appropriate category ordering (whether or not the category ordering of polytomous items is working as expected); and differential item functioning (DIF; whether bias exists for an item among subgroups in the sample).

When Should Rasch Analysis Be Used?

Rasch analysis is used when a set of questionnaire items (or items from an administered scale) are intended to be summed together to provide a total score (which may include several subscale totals, as well as an overall score). There are different occasions when Rasch analysis would be applied. First, it would be applied in the development of a new scale, where it is possible to design the item set to fit the model expectations from the outset (15). Items should be selected that are free of DIF, fit model expectations, and demonstrate unidimensionality (16).

Second, Rasch analysis would be used in reviewing the psychometric properties of existing ordinal scales. For example, where studies have found that existing category structures for items (e.g., a 1–6 response structure) do not work as intended, it is necessary to collapse categories into fewer options (17,18). Also, the unidimensionality of some existing scales has been called into question, resulting in the necessity of item deletion (19).

Third, Rasch analysis would be applied in examining hypotheses about the dimensional structure of ordinal scales. Sometimes second-order factors or higher-order constructs are proposed (that is, when several sets of items from different subscales are added together), and these can be tested by fit to the Rasch model (20). For example, can we develop a higher-order construct of health status from 2 subscales of impairment (e.g., pain) and activity limitation?

Fourth, Rasch analysis would be used in constructing item banks as the basis of computer adaptive testing. New developments are emerging in medical outcome measurement (21). With calibrated item banks (where the difficulty of items has been previously established on a single metric), it is possible to use computer algorithms to present items to patients in such a way that their level on the construct to be measured can be determined by just a few questions. This is called computer adaptive testing (22).

Finally, Rasch analysis would be applied whenever change scores need to be calculated from ordinal scales. The data must be shown to meet model expectations so that an interval (logit-based) estimate can be derived.

What Software Is Available to Perform Rasch Analysis?

In the medical outcomes literature, most Rasch analysis is undertaken with proprietary software. The most commonly used packages are WINSTEPS (23), RUMM2020 (24), and ConQuest (25), but many more are available (26). Each reports findings in a slightly different way, although the basic premise is to test whether the response pattern observed in the data matches the theoretical pattern expected by the model (i.e., the probabilistic form of Guttman scaling). This difference (between observed and expected) is at the heart of the statistics used to test if the data fit the model.

What to Look for in a Good Paper

Although different software can give rise to different reporting characteristics, there are some common fundamental aspects to the Rasch approach that should be reported in any analysis. These are 1) the model chosen; 2) where polytomous, the appropriate ordering of categories and any necessary rescoring; 3) fit of items and persons to the model (including any relevant summary statistics) and justification for fit levels chosen, strategy for improving fit (e.g., item deletion) and subsequent fit statistics; 4) test of the assumption of the local independence of items, including response dependency and unidimensionality; 5) the presence of DIF and any action taken to adjust; 6) the targeting of the scale; and 7) the person separation reliability.

Which derivation of the model is used? The first step in the process of Rasch analysis is a decision on which mathematical derivation of the Rasch model should be chosen. When items in a scale have only 2 response options, the dichotomous model is chosen. When items have 3 or more options, then the Rasch model takes a slightly different form, either what is commonly known as the Andrich Rating Scale Model (12) or the Masters Partial Credit Model (27). These derivations use the same Rasch model, but they differ slightly in their mathematics and are named after the people who developed them. The principal difference between the 2 is that the former expects the distance between thresholds to be equal across items. A threshold is the probabilistic midpoint (i.e., 50/50) between any 2 adjacent categories. This means that the metric distance between, for example, the thresholds separating categories 1 and 2 and that separating categories 2 and 3 is the same across all items. In RUMM2020 there is a likelihood ratio statistic to help decide which polytomous version of the model to use. Whatever version of the model is chosen needs to be reported, and if polytomous, the reason for the choice should also be reported.

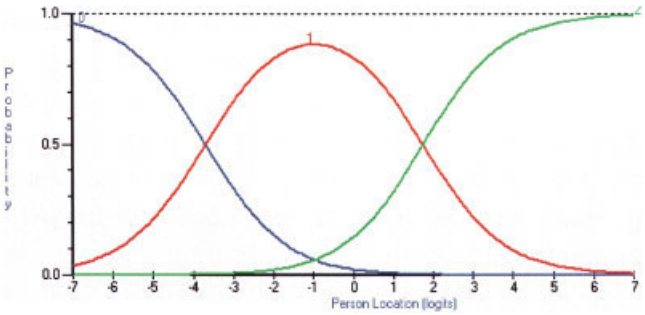

Figure 1. Ordered category responses representing an increase in trait value (logits) for each category. Note that each response option (0, 1, and 2) has a distance across the trait when it is the most probable response.

The threshold ordering of polytomous items. The second step for assessing polytomous scales, when using the partial credit Rasch model, is to examine their category structure. Whether the responses to the items are consistent with the metric estimate of the underlying construct is indicated by an ordered set of response thresholds for each of the items (Figure 1). When disordered thresholds occur, some analysts will rescore items by collapsing categories, and others will cite the increase in average measure to indicate that the categories are working properly. Again, whatever approach is used should be made explicit, along with any remedial action taken.

Tests of fit to the model. Given the emphasis on the difference between observed response and that expected by the model, it comes as no surprise that many of the fit statistics are chi-square based. In WINSTEPS these are called INFIT and OUTFIT statistics, whereas in RUMM2020 they are labeled as chi-square statistics. WINSTEPS also has some standardized fit statistics (reported as ZSTD), and RUMM2020 has a residual statistic, which is the standardized sum of all differences between observed and expected values summed over all persons. Each statistic gives slightly different information about the difference between the observed and expected response, similar to viewing something from a different angle. For example, INFIT takes particular note of the difference between observed and expected response for those items that have a difficulty level near the person's ability level. OUTFIT includes the differences for all items, irrespective of how far away the item difficulty is from the person's ability. Thus the former is a weighted fit statistic in that it gives greater weight to responses to items close to the person's ability level. In RUMM2020 the chi-square statistic compares the difference in observed values with expected values across groups representing different ability levels (called class intervals) across the trait to be measured (e.g., pain). Consequently, for a given item, several chi-squares are computed (the number of groups depend on sample size), and then these chi-square values are summed to give the overall chi-square for the item, with degrees of freedom being the number of groups minus 1. If the value is less than 0.05 (or a Bonferroni-adjusted value [28]) then the item is deemed to misfit model expectation. None of the fit statistics are directly comparable across the software packages, although in theory, because of their standardized nature across all persons, the OUTFIT ZSTD of WINSTEPS and the residual statistic in RUMM2020 are very similar. RUMM2020 also reports an item-trait interaction chi-square, reflecting the property of invariance across the trait. This sums the chi-squares for individual items (as described above) across all items. A significant chi-square indicates that the hierarchical ordering of the items varies across the trait, compromising the required property of invariance. A wide variety of texts are available to help the reader understand fit and the other relevant topics discussed in this article (18,26,29,30). The choice of fit statistic and the justification for the choice should always be included in the methods of any report.

In addition to item fit, examination of person fit is important. A few respondents with bizarre response patterns (identified by high positive residuals) may seriously affect fit at the item level. Such aberrant response patterns may be due to unrecorded comorbidity or, for example, respondents with cognitive deficits. Therefore, where some respondents misfit in this way, removal from the analysis may make a significant difference to a scale's internal construct validity, while at the same time raising questions about the external construct validity of the scale with the particular patient group. Consequently, some summary of person fit should be reported.

Testing for differential item functioning. DIF, or item bias, can also affect fit to the model. This occurs when different groups within the sample (e.g., younger and older persons) respond in a different manner to an individual item, despite equal levels of the underlying characteristic being measured. Therefore, this does not preclude a different score between younger and older persons, but rather indicates that, given the same level of, for example, pain, the expected score on any item should be the same, irrespective of age. Two types of DIF may be identified. One is where the group shows a consistent systematic difference in their responses to an item, across the whole range of the attribute being measured, which is referred to as uniform DIF (31). When there is nonuniformity in the differences between the groups (e.g., differences vary across levels of the attribute), then this is referred to as nonuniform DIF, as is demonstrated in Figure 2, which shows the difference in response to a dichotomous variable across countries. The

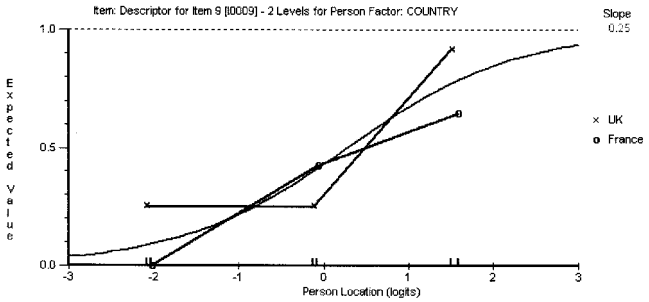

Figure 2. Nonuniform differential item functioning. Note that respondents in France have a higher probability of affirming this item when they are located in the middle of the construct, but a lower probability at the margins.

analysis of DIF has been widely used to examine cross-cultural validity, and readers can find an explanation of the approach, including the analysis of variance–based statistical analysis used in RUMM2020, in several recent reports (17,32,33). At the very least, expect to see DIF for age and sex reported.

Targeting of persons and items, and sample size. In Rasch software the scale is always centered on zero logits, representing the item of average difficulty for the scale. Given this is an interval scale variable, it can be rescored to a convenient value and range (e.g., a value of 5, with a range of 0–10). Comparison of the mean location score obtained for persons with that of the value of zero set for the items provides an indication of how well targeted the items are for people in the sample. For a well-targeted measure (not too easy, not too hard), the mean location for persons would also be around the value of zero. A positive mean value for persons would indicate that the sample as a whole was located at a higher level (e.g., of pain) than the average of the scale, while a negative value would suggest the opposite. Some indication of the quality of targeting should always be reported.

Reliability. Both WINSTEPS and RUMM2020 have a measure of reliability, although again they differ slightly in their interpretation. In RUMM2020 an estimate of the internal consistency reliability of the scale is available as a person separation index. This is equivalent to Cronbach's alpha (2), only using the logit value (linear person estimate) as opposed to the raw score in the same formulae. It is interpreted in a similar manner, that is, a minimum value of 0.7 is required for group use and 0.85 for individual use. In WINSTEPS an item separation ratio is reported, and the equivalent values of this would be 1.5 and 2.5, respectively. Whichever package is used, these values should be given and explained in the Methods section.

Response dependency. The assumption of local independence that underpins the Rasch models implies that once you have extracted the Rasch factor, that is, the main scale, there should be no leftover patterns in the residuals. A breach in the assumption of local independence of items can be found in 2 ways, through response dependency and multidimensionality. Response dependency is where items are linked in some way, such that the response on one item will determine the response on another. The classic example of this is where several walking items are included in the same scale. If a person can walk several miles without difficulty, then that person must be able to walk 1 mile, or any lesser distance, without difficulty. Such sets of items inflate classic reliability and affect parameter estimation in Rasch analysis. They can be identified through the residual correlation matrix and dealt with by combining the items into a subtest, which, in the example above, would be the equivalent of making one walking item with response options that relate to how far a person can walk. Expect to see that this issue has been dealt with, if only to report that no response dependency was found.

Unidimensionality. The Rasch model is a unidimensional measurement model, therefore the assumption is that the items summed together form a unidimensional scale. There are various ways to test this assumption, and these can be thought of as a series of indicators to support the assumption. Rasch programs usually provide a principal components analysis of the residuals. This allows for a test of the local independence of the items (34). This test implies that once the Rasch factor has been taken into account, there should be no further associations between the items other than random associations. The absence of any meaningful pattern in the residuals will also be deemed to support the assumption of unidimensionality. A test for this has been proposed by Smith (35). This test takes the patterning of items in the residuals, examining the correlation between items and the first residual factor, and uses these patterns to define 2 subsets of items (i.e., the positively and negatively correlated items). These 2 sets of items are then used to make separate person estimates, and, using an independent *t*-test for the difference in these estimates for each person, the percentage of such tests outside the range −1.96 to 1.96 should not exceed 5%. A confidence interval for a binomial test of proportions is calculated for the observed number of significant tests, and this value should overlap the 5% expected value for the scale to be unidimensional. Given that the differences in estimates derived from the 2 subsets of items are normally distributed, this approach is robust enough to detect multidimensionality (36) and appears to give a test of strict unidimensionality, as opposed to essential dimensionality (37). In the latter case a dominant factor occurs, and although other factors exist, they are not deemed to compromise measurement. However, it is our experience that a *t*-value that exceeds 1.96 usually equates to a substantial difference in the person estimate derived from different sets of items. This raises concerns about the validity of such approaches when scales are to be used at the individual person level (for example, with a clinical cut point) or where item banks are to be established, given that in these circumstances estimates may be made from just a few items using a computer adaptive testing approach (22). We argue that the observed differences in person estimates associated with a significant number of *t*-tests cannot be sustained under these circumstances. Therefore, any article should report on the unidimensionality of the scale, including one or more of the tests above, preferably appropriate to the use to be made of the scale.

Conclusion

The Rasch measurement model is now firmly established as the standard for modern psychometric evaluations of outcome scales. Whether constructing a new scale or reviewing and revising existing scales, performing Rasch analysis provides a powerful tool for bringing together key issues such as unidimensionality, category ordering, and DIF within the framework of measurement science. It is important to remember that all ordinal scales are nonlinear, and the raw score remains so even when data fit the Rasch model. The curve is always a sigmoid, and thus the values at the margins cover a wider part of the underlying

trait than those at the center. However, fit to the model ensures that a logit person estimate can be exported for parametric analysis. Defining measurement, the Rasch model thus provides a template for the appropriate pattern of responses if a unidimensional scale is to be constructed and a linear conversion of an ordinal score is required.

AUTHOR CONTRIBUTIONS

Dr. Tennant had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study design. Tennant, Conaghan.
Acquisition of data. Tennant.
Analysis and interpretation of data. Tennant, Conaghan.
Manuscript preparation. Tennant, Conaghan.
Statistical analysis. Tennant.

REFERENCES

1. Nunally JC. Psychometric theory. New York: McGraw-Hill; 1978.
2. Cronbach IJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–333.
3. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27 Suppl 3:S178–89.
4. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004;7 Suppl 1:S22–6.
5. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago; 1960.
6. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheumatol 1996;35:574–8.
7. Wolfe F, van der Heijde DM, Larsen A. Assessing radiographic status of rheumatoid arthritis: introduction of a short erosion scale. J Rheumatol 2000;27:2090–9.
8. Kopec JA, Sayre EC, Davis AM, Badley EM, Abrahamowicz M, Sherlock L, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. Health Qual Life Outcomes 2006;4:33.
9. Perline R, Wright BD, Wainer H. The Rasch model as additive conjoint measurement. Appl Psychol Meas 1979;3:237–56.
10. Karabatos G. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. J Appl Meas 2001;2:389–423.
11. Luce RD, Tukey JW. Simultaneous conjoint measurement: a new type of fundamental measurement. J Math Psychol 1964; 1:1–27.
12. Andrich D. Rating formulation for ordered response categories. Psychometrika 1978;43:561–73.
13. Guttman LA. The basis for Scalogram analysis. In: Stouffer SA, Guttman LA, Suchman FA, Lazarsfeld PF, Star SA, Clausen JA, editors. Studies in social psychology in World War II. IV. Measurement and prediction. Princeton: Princeton University; 1950. p. 60–90.
14. Svensson, E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 2001;33:47–8.
15. Gilworth G, Chamberlain MA, Bhakta B, Haskard D, Silman A, Tennant A. The development of the BD-QoL: a quality of life measure specific to Behçet's disease. J Rheumatol 2004; 31:931–7.
16. Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000;2:199–218.
17. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 2004;42 Suppl 1:37–48.
18. Andrich D. Rasch models for measurement. Newbury Park (CA): Sage; 1988.
19. Pallant JF, Tennant A. Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. BMC Psychiatry 2006; 6:28.
20. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46:1–18.
21. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res 2003;12:485–501.
22. Gershon RC. Computer adaptive testing. J Appl Meas 2005;6: 109–27.
23. Linacre JM. WINSTEPS Rasch measurement computer program: version 3.64.1. Chicago: Winsteps.com; 2007.
24. Andrich D, Lyne A, Sheridan B, Luo G. RUMM 2020. Perth: RUMM Laboratory; 2003.
25. Wu ML, Adams RJ, Wilson MR. ACER ConQuest. Melbourne: Acer Press; 1998.
26. Fischer GH, Molenaar IW, editors. Rasch models: foundations, recent developments, and applications. New York: Springer; 1995.
27. Masters G. A Rasch model for partial credit scoring. Psychometrika 1982;47:149–74.
28. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.
29. Smith EV Jr, Smith RM. Introduction to Rasch measurement. Maple Grove (MN): JAM Press; 2004.
30. Wilson M. Constructing measures. Mahwah (NJ): Lawrence Erlbaum; 2005.
31. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Stat Med 2000; 19:1651–83.
32. Hagquist C, Andrich D. Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. Pers Indiv Differ 2004;36:955–68.
33. Lawton G, Lundgren-Nilsson A, Biering-Sorensen F, Tesio L, Slade A, Penta M, et al. Cross-cultural validity of FIM in spinal cord injury. Spinal Cord 2006;44:746–52.
34. Wright BD. Local dependency, correlations and principal components. Rasch Meas Trans 1996;10:509–11.
35. Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3:205–31.
36. Tennant A, Pallant JF. Unidimensionality matters. Rasch Meas Trans 2006;20:1048–51.
37. Stout WF. A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. Psychometrika 1990;55:293–325.