Chapter 1

An Overview of the Family of Rasch Measurement Models

Benjamin D. Wright
University of Chicago

Magdalena M. C. Mok
The Hong Kong Institute of Education

The family of Rasch measurement models is a way to make sense of the world. Experience is continuous. But the moment we notice experience, it becomes discrete. We sense the fragrance of flowers. The sensation is continuous. But when we distinguish between flowers—with and without fragrance; strong from weak fragrance, fragrance we like, don't mind, or dislike, then our observations become discrete. As we notice and remember particulars, we begin the counting that can become measurement. Counting is never accidental. It is always underpinned by the intention of replication. But replication is never exact. Its approximation depends on the situation, how much we care and what we are going to do with the count. A vacationer may count seashells according to size, shape or color. But an Aboriginal would count them according to whether or not their contents were edible. Any idea that all seashells are sufficiently identical to be counted is based on a fiction that each shell makes an equal contribution to an intention—which for practical purposes we keep constant.

This is true for all counting. As soon as we start counting, we have decided on a useful identity, namely that, at least for us, the objects we count are sufficiently identical to be infinitely exchangeable.

We choose a dimension according to its utility. Then we define what is (and what isn't)—a sign of the more or less of that dimension. Then we count indicators of the dimension. To make our counting useful, we look beyond the raw objects counted to the dimension which we have decided our counts imply. We decide, discover and verify the extent to which counting these particular observations contains inference about the dimension. Our raw data take such forms as:

Yes/No
Present/Absent
Right/Wrong

Which we score as observations: $x = 0,1$.

There are situations where indications of more or less of a dimension can be introduced as categories within each observation. Counting in this way gives rise to data such as:

Frequently/Sometimes/Rarely for $x = 0, 1, 2$
Strongly Agree/Agree/Disagree/Strongly Disagree for $x = 0, 1, 2, 3$

The family of Rasch measurement models provides the means for constructing interval measures from these kinds of raw data.

All observations begin as counts. But raw counts are only indications of a possible measure. Raw counts cannot be the measures sought because in their raw state, they have little inferential value. To develop metric meaning, the counts must be incorporated into a stochastic process which constructs inferential stability. There are many examples in everyday life where raw counts are not useful for inference. Suppose we want to measure how long we can support a heavy pile of books. We may take a stop-watch to record the length of time, but the seconds counted do not "measure" our experience. The first seconds are easy and pass quickly. But the final seconds become painfully difficult and "take forever". In this situation, each raw second counted has a different experiential meaning, depending on when it occurs. As such, the "second" itself is not a useful unit of measurement for how it feels to support heavy books.

Raw counts may give the impression that they are interval (or ratio) measures of experience. But this is always an illusion. In particular, raw counts at the beginning and end of a raw score scale are problematic be-

cause while the counts necessarily terminate at "none" or "all", the measures they might imply have no boundaries. Problems with counting can also occur at other parts of the scale. The question, "How many oranges do we need to squeeze an 8 ounce glass of orange juice?" makes no sense because the answer depends on the size of the oranges. We withdraw from the concrete reality of counting real oranges and advance instead to approximating an abstract fiction of perfect ounces of weight to construct a stable answer. A pint is a pound the world around. Oranges are half juice. Therefore it takes 1 pound of oranges to produce 8 ounces of juice.

Consider observations derived from commonly used survey rating scales such as "strongly agree"/ "agree"/ "disagree"/ "strongly disagree". The assignment of the number labels (numerals) 1, 2, 3, 4 to these options does not make these numerals become equally distanced measures. But if the category labels are not equally distanced, then none of the conventional statistics we like to use, including the mean and standard deviation, provide legitimate processing for these non-interval category labels.

There is also the issue of missing data. Data may be missing because of oversight or non-compliance. It may result from incidental interference, perhaps from the physical condition of the person. If the purpose of research is to use existing information to make inferences about what is still unknown, then missing data are of the essence. It follows that a useful measurement model for constructing inference from observation must be unaffected by missing data. Further, for a measurement model to be useful, it must enable us to estimate the precision of our inference and it must provide for the detection and evaluation of discrepancy between observation and expectation.

If raw counts cannot be relied upon to serve as measures, how can we construct inferences from observations? In order for measurement to be useful for inference, it needs to be linear and reproducible.

A feel for precision can be achieved through replication. The ability to replicate is the first condition for precision. When similar results occur repeatedly, we gain confidence that the same will happen in the future. However, replication does not guarantee the accuracy of a measure. If we use a broken typewriter to measure typist speed, then no matter how many times the test is repeated, we will get the wrong measurement of the typing speed. Likewise, a testing instrument has to operate in the region of a candidate's proficiency. This is called 'targeting'. A second condition for precision is noise control. This includes using a relevant tool to carry out the observations and making sure that the observations take place under

reproducible conditions. A typewriter in good condition should be used for measuring typing speed and the room should be well-lighted, not too noisy and neither too cold nor too hot. Nevertheless, no matter how hard we try to control the intrusion of noise into the observation process, there are always factors beyond our control. The person can become careless or have had an argument with their family, which may have affected their performance. Such factors are often unknown to the person collecting the data. It is therefore important that the measurement model has indicators not only of the precision of inference but also the quality.

Measures must be as independent as possible of incidental circumstances. As long as a good typewriter is used, the measure of my typing speed must not depend on who else before me has been measured on the typewriter. And, so far as all typewriters are in good conditions, my speed should not depend on which one I use. The measured proficiency of a candidate cannot depend on who else takes part in the examination or the difficulty level of the test items. This requirement for measurement is called 'parameter separation'. This condition is met in the family of Rasch measurement models.

Thus, in order to construct inference from observation, the measurement model must: (a) produce linear measures, (b) overcome missing data, (c) give estimates of precision, (d) have devices for detecting misfit, and (e) the parameters of the object being measured and of the measurement instrument must be separable. Only the family of Rasch measurement models solve these problems.

We will begin our discussion with the simplest case, that of a dichotomous outcome. The method generalizes easily to situations with finer gradations. Imagine jumpers jumping fences. The jumpers vary in strength from weak to strong, and the fences are of various heights posing different challenges (Figure 1).

The outcomes among jumpers of varying strengths attempting fences of different heights can be summarized in a data matrix. For any jumper n, (n = 1,...,N) attempting fence i (i = 1,...,L), the outcome is either a success (denoted by x_{ni}=1) or a failure (denoted by x_{ni}=0). The attempts of jumper n against all fences tried can be represented by a response vector such as (1,1,-,0,1,...,0) where a '1' represents a successful attempt by jumper n, '0' represents a failed attempt and a '-' records that jumper n was not observed to attempt that particular fence. A raw count of successes ($\sum_i x_{ni} = R_n$) can be obtained by summing the elements of a response vector. But unless all

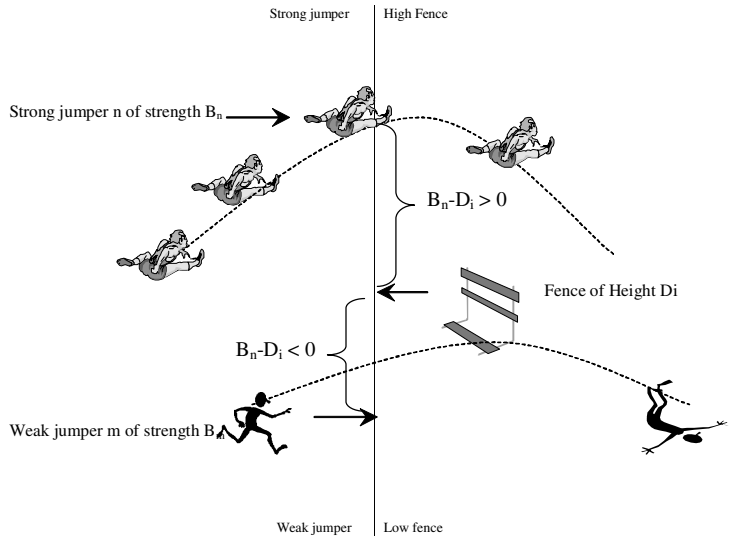


Figure 1. Jumper stronger than fence clears. Jumper weaker than fence tumbles.

jumpers have an equally fair-shot at all fences, the raw sum of successes R_n made by jumper *n* and R_m by jumper m remain incomparable. Because they do not share the same fences, there is no way one raw sum can be compared with another in order to infer that one jumper is better than the other. Similarly, for any fence *i*, the attempts made by all jumpers can be represented by a vector of 1, 0, and –'s, with a total number of successful jumpers equal to the sum of '1's ($\sum_n x_{ni} = S_n$). Again, unless all fences have been challenged by all jumpers, the raw sum of successful jumpers over two particular fences cannot be used to compare the relative difficulties of those two fences. Figure 2 shows such a data matrix.

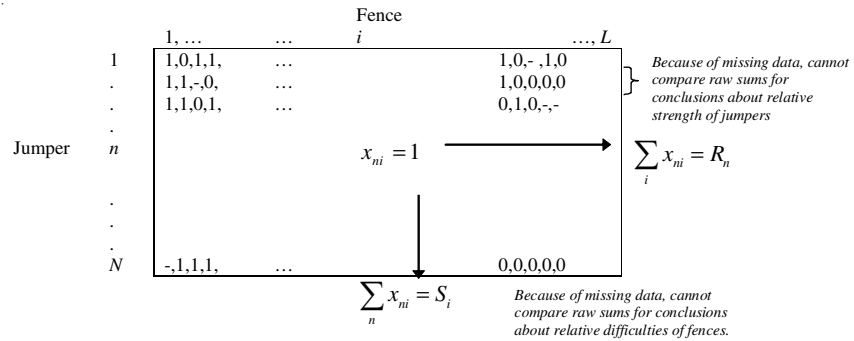


Figure 2. Observation of jumpers over fences.

While this raw data matrix is all the observation we have, as it stands, it is of limited utility. Even though it contains everything we could observe, as it stands, it doesn't help us to predict what will happen in the future. In order for it to be useful, we must build a useful expectation of whether jumper n will succeed on fence i the next time round.

To know about the strength of a jumper, we must challenge him with a fence and to find out about the height of a fence, we must challenge it with a jumper. The meaning of the observation is derived conjointly from fence and jumper, simultaneously. Nevertheless, we must back away from the mere manifestation of jumpers negotiating fences because, after all, we are not interested in the specific incidents of success and failure on this already passed occasion. Instead, we want to infer from these data, assertions of the relative strengths of jumpers and fences, expectations as to what will happen next. Our expectations must be grounded on an abstraction of this conjoint situation. Counts are concrete and limiting; expectations are abstract and liberating. The ability to expect and so to infer is the impetus of

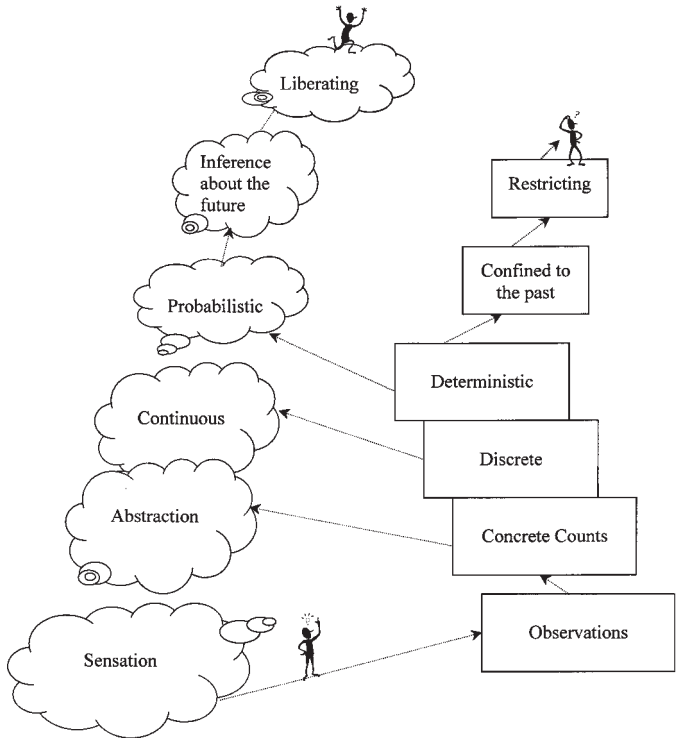


Figure 3. The spiral of inferential development.

human development, the prime tool of civilization. The transition from continuous sensation to discrete counts, and from discrete observations of current events to continuous inferences about the future underpins the evolution of our ability to survive, let alone build science (Figure 3).

For jumper and fence, the meaning of experience is created by abstracting from observations of '0's, '-'s, and '1's into expectations, P_{ni}, as in Figure 4. In this matrix of expectations there are no missing data.

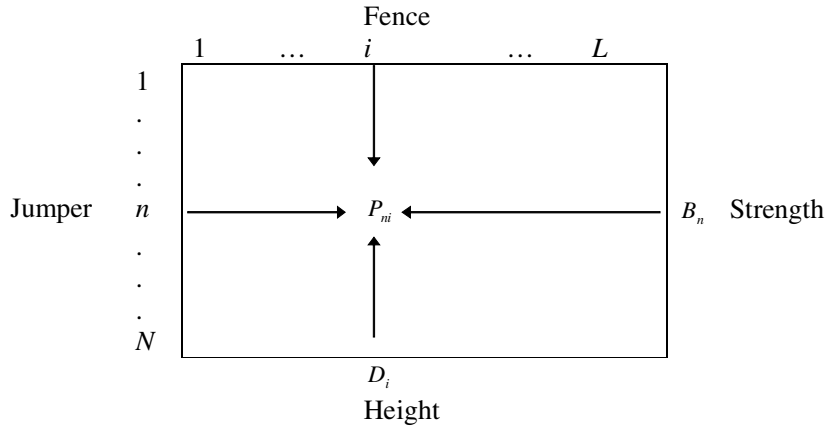


Figure 4. The stochastic interpretation of observations of jumpers over fences.

The basic Rasch model for this kind of analyses can be derived from the simplest paired comparison. Consider comparing the strengths of two jumpers or the heights of two fences (Wright and Linacre, 1995). Consider jumpers Mike and Nick of strengths B_m and B_n making a jump at the same fence i. Let x_{mi} denote the outcome of Mike's attempt at fence i and x_{ni} denote the outcome of Nick's attempt at fence i. x_{mi} can be either 0, if Mike fails to jump fence i, or 1, if he succeeds. x_{ni} is also scored 0 or 1, depending on whether Nick fails or succeeds with fence i. If Mike and Nick each makes one jump at fence i, there are 2 x 2 possible outcomes. Of these four possibilities, the two outcomes in which either both jump over the fence or both fail the fence do not contain any information regarding the relative strengths of Nick and Mike. Only the two off-diagonal outcomes are informative, because only these outcomes tell us whether Mike or Nick is a stronger jumper.

To get an idea of who is a stronger jumper, in pursuit of the necessity for replication, Mike and Nick make many attempts at fence i and the results are recorded in Table 1.

Table 1
Outcome when Mike competes with Nick on many attempts at fence i

		Nick Wins $x_{ni}=1$	Nick Fails $x_{ni}=0$
Mike Wins	$x_{mi}=1$	N_{11} = the number of times both are successful (no useful information)	N_{10} = the number of times Mike beats Nick
Mike Fails	$x_{mi}=0$	N_{01} = the number of times Nick beats Mike	N_{00} = the number of times both fail (no useful information)

Let P_{ni} be the probability of Nick jumping fence *i* and (1-P_{ni}) be the probability of Nick failing fence *i* and the same for Mike with P_{mi} and (1-P_{mi}). The probabilities of the four possible outcomes are given in Table 2.

Table 2
Probability matrix of possible outcomes

		Nick Wins $x_{ni}=1$	Nick Fails $x_{ni}=0$
Mike Wins	$x_{mi}=1$	$P_{ni}\ P_{mi}$	$P_{mi}\ (1-P_{ni})$
Mike Fails	$x_{mi}=0$	$(1-P_{mi})\ P_{ni}$	$(1-P_{mi})\ (1-P_{ni})$

Let N_{10} be the number of times Mike succeeds but Nick fails and N_{01} be the number of times Mike fails but Nick succeeds. As Nick and Mike compete on a number of occasions, it is the ratio of times N_{10}/N_{01}, rather than the difference (N_{10}–N_{01}), by which Mike beats Nick that produces a stable picture of how much Mike is better than Nick. To illustrate this point, four possible off-diagonal outcomes are presented in Table 3.

In Table 3, A, B and C describe situations where Mike is nine times better than Nick, a condition clearly stable in the ratios but unstable in the differences. Situation D, on the other hand, is clearly a case where Mike and Nick are almost identical in strength, as reflected in the ratio N_{10}/N_{01}. Their difference (N_{10}–N_{01}) gives a distorted picture because it implies that Mike is as much stronger than Nick as he was in situation A where the N_{10}/N_{01} ratio was 9. These examples demonstrate that the ratio N_{10}/N_{01}, of the off-diagonal elements contains the replication stable information about the relative strengths of Mike and Nick. Introducing a probability model for this ratio produces:

Table 3
Ratios or Differences?

		Situation A	B	C	D
Mike Succeeds Nick fails	N_{10}	9	90	9000	5004
Mike Fails Nick Succeeds	N_{01}	1	10	1000	4996
Difference	$N_{10}-N_{01}$	8	80	8000	8
Ratio	N_{10}/N_{01}	9	9	9	≅1

$$\frac{N_{10}}{N_{01}} \approx \frac{P_{mi}(1-P_{ni})}{P_{ni}(1-P_{mi})} \quad ,$$

which, if Mike and Nick have a meaningful relation, must hold for any fence *i*, that is, for all fences. To be objective and hence useful, the comparison between Mike and Nick cannot depend on which fence they compete on. Expressed mathematically this becomes:

$$\frac{P_{mi}(1-P_{ni})}{(1-P_{mi})P_{ni}} \equiv \frac{P_{mj}(1-P_{nj})}{(1-P_{mj})P_{nj}} \quad ,$$

for all *i*, *j*, or

$$\left(\frac{P_{ni}}{(1-P_{ni})}\right) \equiv \left(\frac{P_{nj}}{(1-P_{nj})}\right)\left(\frac{(1-P_{mj})}{P_{mj}}\right)\left(\frac{P_{mi}}{(1-P_{mi})}\right).$$

By the same argument, to maintain objectivity, the relation between any pair of fences *i* and *j* must hold for any arbitrary jumper *m*. Any jumper and any fence can be chosen to define the frame of reference for these comparisons. Choosing person 0 and fence 0 to be of equal strength sets P_{00} at 0.5. Mathematically, this becomes:

$$\left(\frac{P_{ni}}{(1-P_{ni})}\right) \equiv \left(\frac{P_{no}}{(1-P_{no})}\right)\left(\frac{P_{0i}}{(1-P_{0i})}\right)\left(\frac{(1-P_{00})}{P_{00}}\right) = \left(\frac{P_{no}}{(1-P_{no})}\right)\left(\frac{P_{0i}}{(1-P_{0i})}\right),$$

that is, $$\frac{P_{ni}}{(1-P_{ni})} \equiv f(n)\times g(i) = \frac{b_n}{d_i} \quad , \qquad (1)$$

where $f(n) = b_n$

and $g(i) = 1/d_i$.

Equation (1) specifies that, for measurement objectivity to be obtained, the odds of jumper n succeeding over fence i must be a product of a function of jumper strength, represented by $f(n)=b_n$, and a function of fence difficulty, represented by

$$g(i) = 1/d_i$$

and nothing else. Note that

$$b_n = \frac{P_{n0}}{(1-P_{n0})}$$

is solely a trait of jumper *n* and the metric origin, and that

$$1/d_i = \frac{P_{0i}}{(1-P_{0i})}$$

is solely a trait of fence *i* and the same metric origin. In this measurement model, the jumper parameter and the fence parameter are completely separated, making it possible to estimate jumper strength independently of fence difficulty, and to estimate fence difficulty independently of jumper strength.

The odds ratio is defined as the ratio of b_n, which takes the value between 0 and infinity, depending only on person n and the stipulated frame of reference, and d_i, which takes the value between 0 and infinity, depending only on item i and the same frame of reference. We have found a way to estimate which jumper is stronger. The next question is: "How much stronger?" But "How much" is not a ratio question; it is a difference question. Taking the logarithms of both sides of equation (1) gives:

$$\ln\left(\frac{P_{ni}}{(1-P_{ni})}\right) \equiv \ln\left(\frac{P_{n0}}{(1-P_{n0})}\right) \times \ln\left(\frac{P_{0i}}{(1-P_{0i})}\right) \equiv \ln(b_n) - \ln(d_i)\,,$$

ie, $$\ln\left(\frac{P_{ni}}{(1-P_{ni})}\right) \equiv B_n - D_i \quad , \tag{2}$$

or $$P_{ni} = \frac{\exp(B_n - D_i)}{[1+\exp(B_n - D_i)]} \quad , \tag{3}$$

where $B_n = \ln\left(\frac{P_{n0}}{1-P_{n0}}\right) = \ln(b_n)$

depends only on attributes of person *n* and the metric origin,

and $D_i = \ln\left(\frac{P_{0i}}{1-P_{0i}}\right) = \ln(d_i)$

depends only on attributes of fence *i* and the metric origin.

The Rasch model can be equally well derived by applying the same argument in the case of one jumper negotiating two fences *i* and *j*, i.e.,

$$\frac{N_{10}}{N_{01}} \approx \frac{P_{ni} \times (1-P_{nj})}{(1-P_{ni}) \times P_{nj}}$$, for all n.

Equations (2) and (3) are equivalent forms of the dichotomous Rasch model. B_n and D_i are commonly referred to as person ability and item difficulty parameters respectively. All other forms of the Rasch model can be derived from this basic form.

The Dichotomous Rasch Model

The simplest Rasch model is for dichotomies (derived in the previous section):

$$\ln\left(\frac{P_{ni}}{(1-P_{ni})}\right) \equiv B_n - D_i$$, or equivalently,

$$P_{ni} = \frac{\exp(B_n - D_i)}{[1+\exp(B_n - D_i)]}$$.

Here, P_{ni} is the probability of person *n* with ability B_n succeeding on item *i* which has difficulty level D_i. In the case of one trial, P_{ni} is the expectation (abstraction) of the observed (concretization) x_{ni}. The correspondence between abstraction and concretization is evaluated by the size of the observed discrepancy $Y_{ni} = x_{ni} - P_{ni}$ (Figure 5). A large discrepancy means that the concrete experience is not a useful example of the abstraction. A small discrepancy implies that the abstraction is robust with respect to the experience and thus by inference to similar future experiences. Since each trial is a Bernoulli experiment, the variance of the x_{ni} is given by $P_{ni}(1–P_{ni})$, so it is possible to evaluate the significance of a discrepancy by computing an approximate χ^2 with one degree of freedom for each x_{ni}:

$$z_{ni}^2 = \frac{y_{ni}^2}{P_{ni}(1-P_{ni})} \sim \chi_1^2 .$$

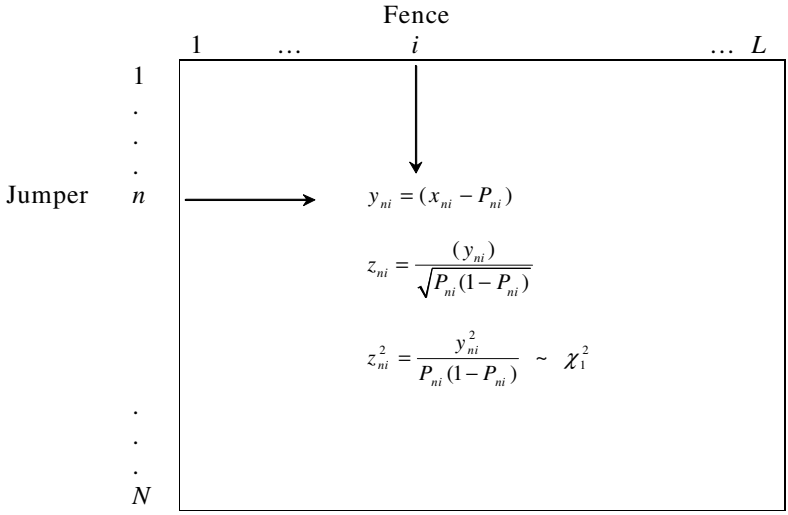


Figure 5. Verification of relation between interpretation and observation.

The discrepancy between observed, x_{ni}, and expected, P_{ni}, with expected model variance $V_x = P_{ni}(1–P_{ni})$ enables us to verify, fine-tune, and validate our measurement constructions. Each residual, in raw form $y_{ni} = (x_{ni}–P_{ni})$, or in standardized form,

$$z_{ni} = (x_{ni} - P_{ni}) \Big/ \sqrt{P_{ni}(1-P_{ni})} ,$$

shows us a piece of information about the quality of our data and the corresponding validity of our construction. A positive residual indicates that the observation is higher than that expected. A negative residual indicates that the observation is lower than expected. Large residuals raise doubts with regard to the match between data and model. We can study these standardized residuals z_{ni} one at a time. But that can be laborious. To expedite our evaluations we organize our study of residuals according to three points of view:

1. We begin with the most unexpected, that is improbable, observations to see what they suggest as to data quality and construct validity.

2. We square our residuals in raw and standardized form to calculate the outfit and infit mean squares for each person, each item and each

adjacent category threshold estimate. "Outfit mean square" is shorthand for "Out-lier sensitive mean square residual goodness of fit statistic" which is the unweighted version of the fit statistic. It measures the average mismatch between data and model. The item outfit mean square is calculated by taking the sum of squared residuals averaged over the total number of persons taking the item (Wright and Masters, 1982: 99). That is:

$$u_i = \sum_{n=1}^{N} z_{ni}^2 \Big/ N .$$

Similarly, the person outfit mean square is given by:

$$u_n = \sum_{i=1}^{L} z_{ni}^2 \Big/ L .$$

Mean square statistics are sensitive to extreme values. Wright and Masters (1982) caution researchers that outfit mean squares are exaggerated by unexpected responses made by persons to items for whom the items are either far too easy or far too difficult.

An alternative suggested by Wright and Masters (1982) to outfit mean squares is the infit mean square statistic, which stands for "information weighted mean square residual goodness of fit statistic". The outfit mean square statistic is calculated by taking the weighted average of squared residuals so that remote responses are given less weight than proximal responses. Mathematically, this is

$$v_i = \sum_{n=1}^{N} z_{ni}^2 \, W_{ni} \Big/ \sum_{n=1}^{N} W_{ni} = \sum_{n=1}^{N} y_{ni}^2 \Big/ \sum_{n=1}^{N} W_{ni} \quad ,$$

where weight W_{ni} is the variance $W_{ni} = P_{ni}(1–P_{ni})$.

3. We decompose the whole matrix of residuals into its principal components among items and among persons. This brings out whatever patterns of misfit lurk among the leftovers from our measurement construct.

Linacre (1998) highlighted several options of factor analysis for identifying multidimensionality. They are factor analysis of the observations and factor analysis of the residuals, namely, (a) the raw Rasch residuals, (b) the standardized Rasch residuals, and (c) the logit residuals. The mathematical expressions of these three residuals are presented in Table 1. On the basis of a series of simulation studies involving both orthogonal and correlated dimensions, Linacre (1998) concludes that although factor analysis of the original observations is informative of the factor structure, this method does not construct the measures of the factors. Further, principal components factor analysis of the standardized Rasch residuals is most

effective amongst the three residual factor analyses in identifying multidimensionality of the measurement instrument. It is followed by factor analysis of the raw Rasch residuals, which is only slightly inferior in effectiveness. Factor analysis of the logit residuals is the least effective in identifying multidimensionality (Linacre, 1998).

Typically these principal components identify structural differences between positive versus negative questions, feeling versus thought versus behavior questions and so on.

A stable inference is obtained when experience points repeatedly in a same direction with a same meaning. When jumper ability is stronger than fence height, we expect the jumper to make the jump most of the time. There will always be some occasions, however, when a jumper fails a jump, particularly as $(B_n – D_i)$ comes close to zero. On the other hand, when jumper ability is weaker than fence height, we expect the jumper to fail the jump most of the time, with odd occasions when he is successful, perhaps by luck. When jumper ability is equal to fence height, however, we expect the jumper to fail the jump about half the time. Thus:

$(B_n - D_i) > 0$ if and only if $P_{ni} > 0.5$,

$(B_n - D_i) = 0$ if and only if $P_{ni} = 0.5$,

$(B_n - D_i) < 0$ if and only if $P_{ni} < 0.5$.

This function is represented graphically in Figure 6 and mathematically by the Rasch measurement model:

$$(B_n - D_i) = \ln\left(\frac{P_{ni}}{(1-P_{ni})}\right), \text{ or}$$

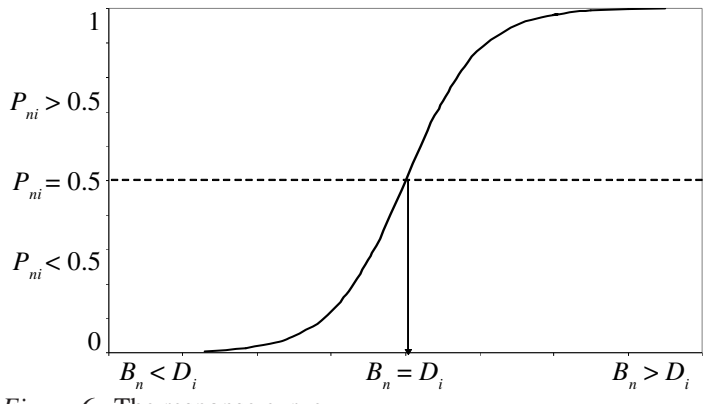


Figure 6. The response curve.

$$P_{ni} = \frac{\exp(B_n - D_i)}{[1 + \exp(B_n - D_i)]} \quad .$$

Rasch Model Overview

There are many Rasch models. We will discuss six of them here. Their relationships are shown in the flow diagram in Figure 7 on the next page.

Binomial Trials

Binomial trials (Wright and Masters, 1982: 51) are situations where several independent attempts are made at an item and the number of successes is counted. In shooting contests, instead of determining the ability of a shooter from one trial, the shooter is allowed to take several, say *m*, attempts at a target and the total number of hits, say *x*, within *m* attempts is counted. The probability of a shooter with ability B_n aiming at a target with difficulty level D_i and getting *x* hits in *m* attempts is:

$$\pi_{nix} = C_x^m \, P_{ni}^x (1 - P_{ni})^{m-x},$$

where

$$C_x^m = \frac{m!}{x!(m-x)!} \cdot$$

Substituting

$$P_{ni} = \frac{\exp(B_n - D_i)}{[1 + \exp(B_n - D_i)]}$$

from equation (3) and simplifying gives:

$$\pi_{nix} = C_x^m \left[\frac{\exp[x(B_n - D_i)]}{\left[1 + \exp(B_n - D_i)\right]^m} \right].$$

Similarly, the probability of the shooter getting (*x*–1) hits in *m* attempts is:

$$\pi_{ni\,(x-1)} = C_{x-1}^m \left[\frac{\exp[(x-1)\ (B_n - D_i)]}{\left[1 + \exp(B_n - D_i)\right]^m} \right].$$

Combining these expressions produces the ratio of probabilities of shooter *n* aiming at target *i* and making *x* hits instead of (*x*–1) hits in *m* attempts, that is, the odds for *x* rather than (*x*–1):

$$\frac{\pi_{ni\,x}}{\pi_{ni\,(x-1)}} = \left(\frac{m-x+1}{x}\right)\left(\frac{P_{ni}}{(1-P_{ni})}\right).$$

Taking the logarithm of both sides gives:

$$\ln\left(\frac{\pi_{ni\,x}}{\pi_{ni\,(x-1)}}\right) = \ln\left(\frac{P_{ni}}{1-P_{ni}}\right) - \ln\left(\frac{x}{m-x+1}\right), \text{ i.e.,}$$

$$\ln\left(\frac{\pi_{ni\,x}}{\pi_{ni\,(x-1)}}\right) = B_n - D_i - C_x,$$

where $C_x = \ln\left[\frac{m-x+1}{x}\right]$.

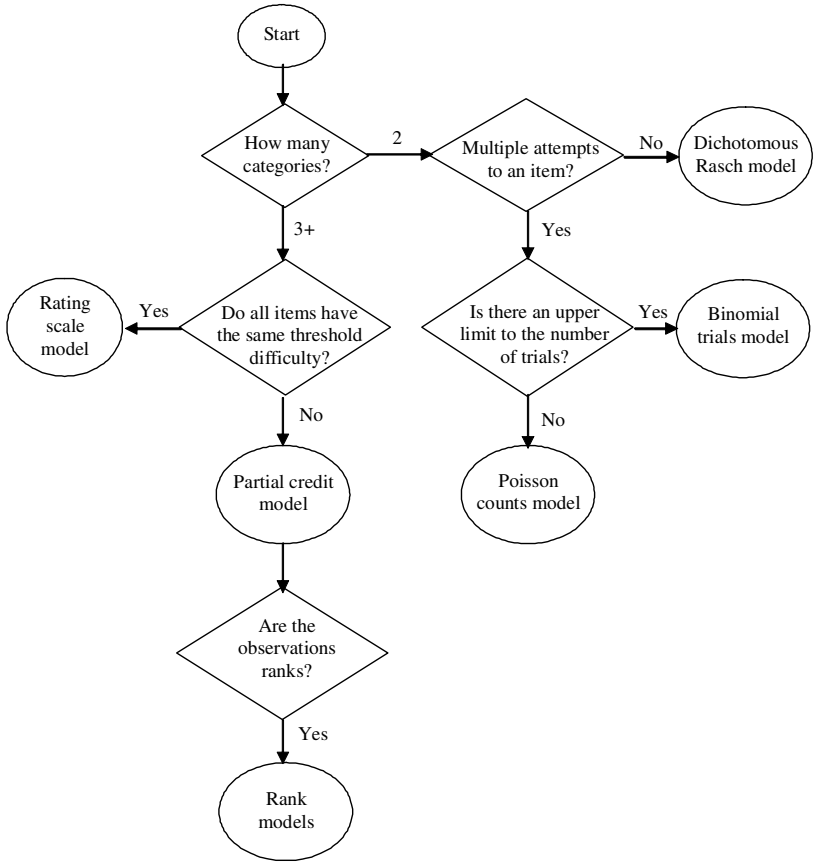


Figure 7. Six commonly encountered Rasch models.

Poisson counts

If the number of trials in the binomial model is infinite and the probability of success is small, as in the case of counting the number of customers buying a certain product at the supermarket in some given time period, such that the buying behavior of a particular customer is independent of that of previous customers, and (mP_{nix}) remains approximately constant, then a binomial distribution approaches a Poisson distribution (Wright and Masters, 1982: 52-54).

$$\pi_{nix} = C_x^m \; P_{ni}^x (1 - P_{ni})^{m-x} \quad \text{approaches}$$

$$\frac{\exp(-\lambda_{ni}) \; (\lambda_{ni})^x}{x!}, \text{ where } \lambda_{ni} = \exp(B_n - D_i),$$

ie, $$\pi_{nix} = \frac{\exp[x\,(B_n - D_i)]}{x!\exp[\exp(B_n - D_i)]} \quad ,$$

and $$\pi_{ni(x-1)} = \frac{\exp[(x-1)\,(B_n - D_i)]}{(x-1)!\exp[\exp(B_n - D_i)]} \quad ,$$

so that $$\frac{\pi_{ni\,x}}{\pi_{ni(x-1)}} = \frac{\exp(B_n - D_i)}{x} \, ,$$

or, $$\ln\left(\frac{\pi_{ni\,x}}{\pi_{ni\,(x-1)}}\right) = B_n - D_i - \ln(x) \quad .$$

Rating Scale Model

The previous models are useful for binary outcomes. However, there are often situations where outcomes can be given finer gradations than just “present/absent”, “yes/no” or “right/wrong”. Response categories in Likert questionnaires may include ordered ratings such as “Strongly Disagree/ Disagree/ Agree/ Strongly Agree”, to represent a respondent’s increasing inclination towards the concept questioned. The response rating scale, when it works, yields ordinal data which need to be transformed to an interval scale to be useful. This is achieved by the Rasch rating scale model (Andrich, 1978). The literature (e.g. Wright and Masters, 1982; Andrich, 1988) discusses many useful applications of the rating scale model, including the study of testlets made up of sets of dichotomously scored items and the analysis of partial credit test items. A testlet is a section of a test comprising a stimulus, such as a reading passage or dia-

gram, with several items referring to the stimulus. An example of stimulus and the associated items is given in Figure 8.

It is reasonable to expect responses to the items within a testlet to correlate higher with one another than with items on other testlets. As a consequence, although it is possible to score each item as either right or wrong, to take into account their testlet clustering, a score can be given to the testlet instead of to the individual items. In the above example (Figure 8), possible testlet scores are 0, 1 and 2 indicating respectively: both items wrong, either item correct, and both items correct. A score of 1 does not distinguish which item in the testlet is right.

A typical item characteristic curve is in Figure 9. Score x=2 represents a higher level of ability than score x=1, which in turn stands for more than x=0. If the ability level is between x=1 and x=2, that rules out the possibility that the ability level is x=0. Likewise, if the ability level is between x=0 and x=1, that rules out the possibility that the ability level is x=2. As a consequence, response interpretation is always between adjacent categories.

Similarly, in the case of a Likert item, such as, "What do you think

Stimulus material:

"It was Mother's Day and every street-kid was given a free phone card so that they could call home. John picked up the handset but he hesitated. What if his mother had not forgiven him? It was three years since he spoke to her, although there had not been a single night he went to sleep without thinking about her. He learned from occasional chats with his brother that she did miss him and hoped that they were friends again."

Questions to be answered using the above stimulus material.

1. Each street-kid was given a phone card
   - ☐ so that they could contact their friends.
   - ☐ so that they could talk with their family.
   - ☐ so that they learned to use the phone.

2. Why did John hesitated in ringing?

   John hesitated because
   - ☐ his mother missed him.
   - ☐ he had not talked with his mother for a long time.
   - ☐ his mother might not have forgiven him.

Figure 8. Examples of a testlet which involves a stimulus material and two test items.

of the amount of homework this term?" and the response categories are: "Too Much/ Just Right/ Too Little". If we choose between "Too Much" and "Just Right", then we have already decided that the amount of homework given is not "Too Little", but if we choose between "Just Right" and "Too Little", then we have already decided that it is not "Too Much". Thus, no matter how many categories are included in a response scale, the response decision and interpretation is always between adjacent categories. The point at which the probability of opting for the next category is equal to that for the previous one is called a threshold. There are two thresholds, represented by F_1 and F_2, in the examples, involving possible scores of 0 ("Too Much" or "Both item wrong"), 1 ("Just Right" or "Only one item correct") or 2 ("Too Little" or "Both item correct") as shown in Figure 10. If we don't like homework, then our inclination is below the

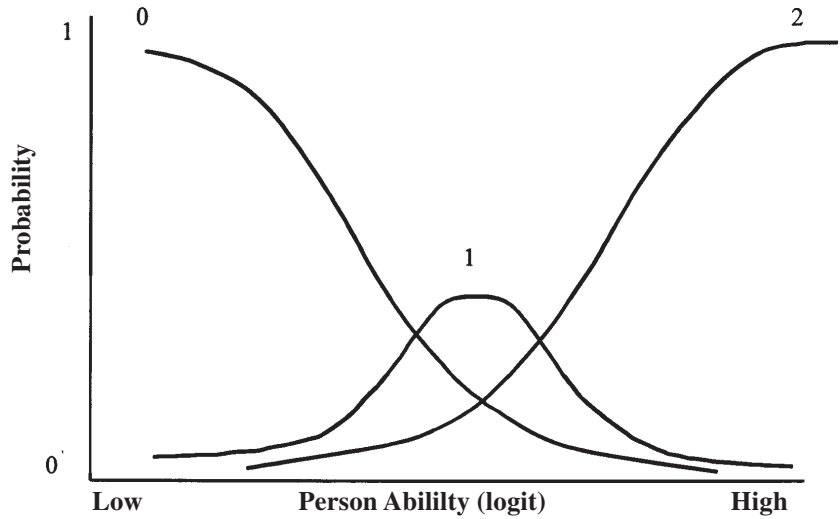


Figure 9. Item characteristic curve for the 0, 1 and 2 responses in a three-category item.

first threshold F_1, and we choose category "Too Much". But, if we enjoy doing homework, but are not crazy about it, our inclination would pass threshold F_1, but would not pass F_2, so we would choose "Just Right" over "Too Much". On the other hand, if we are good students who love homework, we would choose "Too Little". The situation is like having two dichotomous items operating simultaneously.

At the boundary of each threshold, there is a possibility of scoring either '0' or '1', depending on whether the threshold is "failed" or "passed".

This implies that there should be 2 x 2 = 4 possible outcomes in combination. But Figure 10 depicts only 3 of these 4 possible outcomes, namely, (failed, failed), (passed, failed) and (passed, passed). The outcome not included in Figure 10 is (failed, passed), which would mean that the respondent hates homework but thinks that the amount of homework is "Too Little"—an obviously illegitimate situation in real life and one that would work against the concept of an underlying continuum.

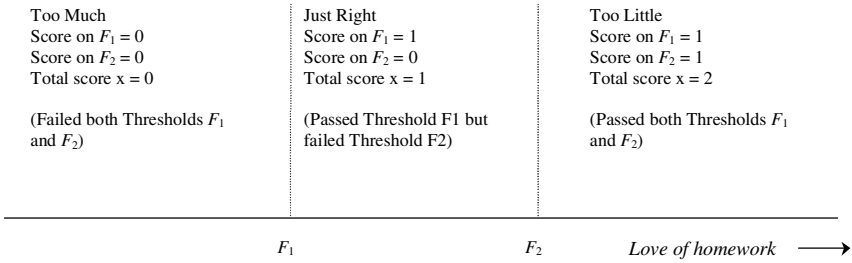
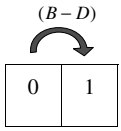


Figure 10. Interpretation of thresholds in a three-category item.

The probability of passing or failing each threshold can be described by a Rasch model. If there are only two categories, denoted by '0' and '1' respectively (Figure 10), then the probabilities of choosing category each of '1' and '0' are:



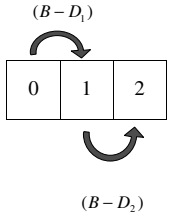
$$P_0 = 1/C_2$$

$$P_1 = \exp(B - D)/C_2$$

$$C_2 = 1 + \exp(B - D)$$

Where C_2 is the sum of the two numerators.

In the case of three categories, denoted by '0', '1' and '2', the probabilities of choosing the categories are:



$$P_0 = 1/C_3$$

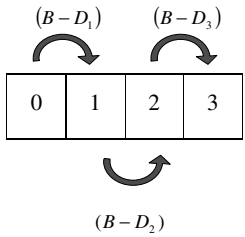
$$P_1 = \exp(B-D_1)/C_3$$

$$P_2 = \exp[(B-D_1)+(B-D_2)]/C_3$$

$$C_3 = 1+\exp(B-D_1)+\exp[(B-D_1)+(B-D_2)]$$

Where C_3 is the sum of the three numerators.

Similarly, in the case of four categories, denoted by '0', '1', '2' and '3', the probabilities of choosing the categories are:



$$P_0 = 1/C_4$$

$$P_1 = \exp(B-D_1)/C_4$$

$$P_2 = \exp[(B-D_1)+(B-D_2)]/C_4$$

$$P_3 = \exp[(B-D_1)+(B-D_2)+(B-D_3)]/C_4$$

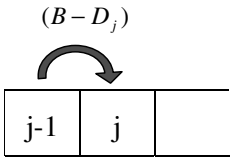
$$C_4 = 1+\exp(B-D_1)+\exp[(B-D_1)+(B-D_2)] + \exp[(B-D_1)+(B-D_2)+(B-D_3)]$$

Where C_4 is the sum of the four possible numerators.

For this final case, the log-odds of choosing a category over the previous adjacent one is given by the following computation:

$P_1/P_0 = \exp(B-D_1)$	, ie, log-odds of '1' over '0' is:	$\ln(P_1/P_0) = B-D_1$
$P_2/P_1 = \exp(B-D_2)$	, ie, log-odds of '2' over '1' is:	$\ln(P_2/P_1) = B-D_2$
$P_3/P_2 = \exp(B-D_3)$	, ie, log-odds of '3' over '2' is:	$\ln(P_3/P_2) = B-D_3$

In general, the odds of choosing a category 'j' over the previous category '*j*–1' is given by:



$$P_j / P_{j-1} = \exp(B - D_j)$$

and the corresponding log-odds is:

$\ln(P_j/P_{j-1}) = (B\text{-}D_j)$, the basic Rasch model.

The general Rasch rating scale model is given by:

$$\ln\left(\frac{P_{nix}}{P_{ni(x-1)}}\right) = B_n - D_i - F_x .$$

If there are several Likert items sharing the same response categories, it is reasonable to specify that the thresholds for all are the same and the rating scale model described above can be applied to the group of items. On the other hand, if the thresholds are not the same across all items, then a partial credit model, which will be discussed in the next section, is applicable.

Partial Credit Model

The partial credit model is similar to the rating scales model except that now each item has its own threshold parameters (Wright and Masters, 1982). This is achieved by a reparameterization:

$$F_x = F_{ix},$$

and the partial credit model becomes:

$$\ln\left(\frac{P_{nix}}{P_{ni\,x-1}}\right) \equiv B_n - D_i - F_{ix} .$$

The examples of partial credit model discussed in the literature (Wright and Masters, 1982) are achievement items where: (a) credits are given for partially correct answers, (b) there is a hierarchy of cognitive demand on respondents in each item, (c) each item requires a sequence of tasks to be completed or (d) there is a batch of ordered response items

with individual thresholds for each item. Such examples occur frequently in grading situations. For instance, a writing assignment is scored as follows:

3 points for work of a superior quality.
2 points for work of predominantly good quality.
1 point for work that is satisfactory.
0 point for work that is of poor quality.

It is clear from the marking scheme that a score of 3 represents more writing proficiency than that represented by a score of 2, which in turn represents higher proficiency than a score of 1, and so on.

Ranks Model

A Ranks Rasch model is useful when respondents are asked to rank order a group of objects instead of giving a rating to each object. Examples include a judge ordering pianists from the strongest to the weakest, or a worker sequencing jobs from most to least urgent. Before 1998, the Higher School Certificate Examination result in New South Wales Australia was reported in the form of a number which indicated the candidate's position in the list formed by the ordering from most to least able of all candidates who sat the same examination that year.

Rank order is a familiar concept. Most people have preference hierarchies for car models, living styles and personal values. Linacre (1994) highlighted the utility of ranks as a way to avoid the problem of having to define a rating scale. He also alerted researchers to the drawbacks of rank data, namely, that rank data are ipsative and that such data contain no information about the preference levels of the rankers. That is, in the case of a judge giving ranks to three objects, the ranks must be "1", "2", and "3" irrespective of what the objects are or how much they differ. For instance, Mike likes green over purple and he loves both colors. Nick also prefers green over purple but he hates either color. The ranks themselves give no information on how much Mike or Nick like these colors.

Further, the ranks assigned to a basket of objects depend on what else is in the basket. A child prefers "chocolate" (rank 1) over "strawberry" (rank 2) if these are the only ice-cream flavors available. But when mango is available, he prefers it to strawberry. So ranking becomes chocolate (1), mango (2), strawberry (3), if all three flavors are presented. We see that ranking differences are not item free and therefore ranks are not measures.

Linacre conceptualizes ranks as a special case of the rating model in which objects are given the rating corresponding to their ranking scale. With this approach the problem of tied ranks is easily solved.

Conclusion

Rasch measurement provides a complete solution to almost every measurement problem encountered in science. It is especially apt for social science, where the raw data is so unruly and so vaguely conceived. An easy way to begin Rasch measurement is to down load MINISTEP and its manual from www.winsteps.com and to run some of the included examples. As one's understanding increases, one can turn the program to 25 item by 100 respondent segments of one's own data to see how useful the program can be for one's own work. If the result is satisfying and a version of the program with greater than 25 x 100 capacity is desired, then WINSTEPS itself with capacity 10,000 x 1,000,000 can be obtained through the same home page.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, *1*, 363-378.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual works best. *Journal of Outcome Measurement*, *2*, 266-283.

Wright, B. D., and Linacre, J. M. (1995). Rasch model derived from objectivity. *Rasch measurement transactions, Part 1*, pp. 5-6. Chicago: MESA Press.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.