# Chapter 29
# Review of Principles of Test Analysis Using Rasch Measurement Theory


The case for applying the Rasch model arises from the requirement of invariance of comparisons within a specified frame of reference, of a particular property of objects (individuals) relative to the stimuli (instruments) which manifest that property, and vice versa. This requirement, and meeting it, can be said to lead to a Rasch Measurement Theory (RMT). This theory might be compared and contrasted to Item Response Theory (IRT) and to Classical Test Theory (CTT) in different ways. In this book, RMT is presented as an elaboration of many of the principles, explicit or implicit, in CTT, but also different in specific ways. It is also presented as different from IRT which is based primarily on principles of modelling responses rather than a priori requirements.

Following a brief review of the principles of RMT, this chapter summarizes the approach within RMT to consideration of item and threshold locations and tests of statistical fit between responses and the Rasch model. In doing so, we stress a point made by Duncan (1984):

> The Rasch model … does not revoke the criteria scientists normally cite in deciding whether right variables have been measured. (pp. 398–399)

In parallel, in applying Rasch measurement theory, scientists must also not revoke criteria they normally cite in applying statistical and empirical methods and principles of measurement that must apply. In summary, the fit to the Rasch model is taken as a necessary, but not sufficient, condition, to achieve measurement.

## Invariance of Comparisons and RMT

An excerpt of Rasch' specification of the requirement for invariant comparisons is shown below:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; …



D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_29

> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; …
>
> (Rasch, 1961, p. 332)

Table 29.1 shows a frame of reference of a class of stimuli comprising an instrument and a class of persons (individuals) to be compared using the instrument. In social science assessment, the *stimuli* above are generally referred to as *items*.

In addition to a well-defined class of persons which is to respond to a well-defined class of items that form an instrument, the frame of reference includes the specifications of the relevant conditions for the administration of the items to the persons, for example, the time available for responding to the instrument, and so on.

The well-defined class of items includes all the features that make the set of items valid for assessing the intended variable. In the assessment of variables in education, psychology and other social sciences, this means that the items are relatively homogeneous in content, but have meaningful differences in difficulty or intensity. Rasch summarized this relationship as follows:

> Altogether these experiences – limited as they are to intelligence tests and attainment tests – suggests that once items have been constructed with an eye to uniformity of content, but variance in difficulty – which may even cover *complexity* – then there is a fair chance that they on the whole fit well into the model of simple conformity. (Rasch, 1960, p. 125)

Then, the *comparisons* referred to are with respect to a variable characterized by a real number for persons and a vector of real numbers for items, with the number of elements in the vector depending on the number of ordered categories for an item. With dichotomous items, there is only one value for each item. The variable indicates more or less of some property, also referred to as a latent trait or construct, and the comparisons are with respect to this property. In Table 29.1, the value of an item, which characterizes its relative difficulty, is designated $\delta_i$ and the value of a person, which characterize his or her relative proficiency, is characterized by $\beta_n$.

**Table 29.1** A frame of reference

| Response Random variable | | | | | | Stimulus $A_i$ Value $\delta_i$ | | |
|---|---|---|---|---|---|---|---|---|
| $X_{ni} = x_{ni}$ | $X_{ni}$ | $A_1$ | $A_2$ | $A_3$ | … | $A_i$ | … | $A_I$ |
| | $O_1$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | | $x_{1i}$ | | $x_{1I}$ |
| | $O_2$ | $x_{21}$ | $x_{22}$ | $x_{23}$ | | $x_{2i}$ | | $x_{2I}$ |
| | $O_3$ | $x_{31}$ | $x_{32}$ | $x_{33}$ | | $x_{3i}$ | | $x_{3I}$ |
| | ⋮ | | | | | | | |
| Person $O_n$ | $O_n$ | $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | | $x_{ni}$ | | $x_{nI}$ |
| Value $\beta_n$ | ⋮ | | | | | | | |
| | $O_N$ | $x_{N1}$ | $x_{N2}$ | $x_{N3}$ | | $x_{Ni}$ | | $x_{NI}$ |

Rasch's specification of invariance as a requirement in a probabilistic model results in the class of Rasch models, and only this class of models. These models are characterized by sufficient statistics for the parameters. In the case of a unidimensional model for responses in ordered categories (polytomous responses), the general form of the model can be written as

$$P_{nix} = P\{X_{ni} = x; m_i, \beta_n, (\delta_i)\} = [\exp(\psi_{xi} + x\beta_n)]/\gamma_{ni}, \tag{29.1}$$

where $x \in \{0, 1, 2, \ldots, m_i\}$ is an integer variable for the $m_i+1$ successive categories, $\beta_n$ is the location of person $n$, $\psi_{xi} = -\sum_{k=0}^{x} \delta_{ik}$, $(\delta_i) = \delta_{ik}, k = 0, 1, 2, \ldots, m_i$ is a vector of $m_i$ thresholds of item $i$ where, for notational convenience, $\delta_{i0} \equiv 0$, and $\gamma_{ni} = \sum_{x=0}^{m_i} [\exp(\psi_{xi} + x\beta_n)]$ is the normalizing factor. It is simply the sum of the numerators and ensures that the sum of the probabilities is 1.

The category coefficients $\psi_{xi}$ can be reparameterized so that the threshold parameters are deviations from the location of the items giving

$$\delta_{ik} = \delta_i + \tau_{ik}; \tau_{ik} = \delta_{ik} - \delta_i, \tag{29.2}$$

where $\sum_{k=0}^{m_i} \tau_{ik} = 0$ and where for notational convenience again, $\tau_{i0} \equiv 0$.

Then, the model takes the form

$$P_{nix} = P\{X_{ni} = x; m_i, \beta_n, \delta_i, (\tau_i)\} = \exp[\kappa_{xi} + x(\beta_n - \delta_i)]/\gamma_{ni}, \tag{29.3}$$

where $\kappa_{xi} = -\sum_{k=0}^{x} \tau_{ik}$ and where in proficiency assessment, $\delta_i$ can be interpreted as the overall *difficulty* of item $i$. In general terms, $\delta_i$ is referred to as the *location* of the item on the variable or the continuum.

In the case of a dichotomous item, where $m_i = 1$, there is only the one threshold $\delta_{i1} = \delta_i$ and the model specializes to

$$P_{nix} = P\{X_{ni} = x; \beta_n, \delta_i\} = \exp[x(\beta_n - \delta_i)]/\gamma_{ni}. \tag{29.4}$$

We note that the integer scoring for the polytomous case does not arise from *equidistant* successive thresholds. Instead, it arises from a common discrimination at all thresholds.

## Total Score as the Sufficient Statistic

A characteristic feature of the requirement of invariance is that the total score of a person, on the items that the person has responded to, is the sufficient statistic for the estimate of the person parameter. Sufficiency has two implications. First, that the person estimate can be characterized by a single parameter, the estimate of which, $\beta$, is a linearization of the total score. It also means that if the items conform to the

model, that is, they conform to a probabilistic Guttman structure reviewed below, and there is no information in the pattern of responses. Second, it implies that the item parameters can be estimated independently of the person parameters, and therefore independently of any person distribution. For example, unlike the estimation with many other models, there is no need to assume the person distribution is normal. Of course, as indicated below, for other properties of the theory and for inferences that can be made, in any data set analysed the locations of the persons and items need to be reasonably well aligned. Achieving such an alignment is part of the specification and articulation of the frame of reference, and part of applying usual criteria for statistical, empirical and measurement principles.

## Dichotomous Items: The Probabilistic Guttman Structure

In the earlier chapters, we considered in some detail the Guttman structure on manifest responses in which the relative difficulties of items and persons define the structure. It will be recalled that in the Guttman structure, if person $n$ has a greater score than a second person $l$, then person $n$ will have positive responses on all the items on which person $l$ has positive responses, and in addition, positive responses to the next most difficult items up to $n$'s total score. This structure, which is deterministic, leads into the implication of relative difficulty of items in the Rasch models.

One of the consequences of the Rasch model is that, when the items have a range of difficulties, the responses form a probabilistic, rather than a deterministic, Guttman structure. In the case of dichotomous items, let person $n$ with location $\beta_n$ have a probability $p_{ni}$ of providing a positive response to item $i$ with location $\delta_i$. Then, the probabilistic Guttman structure has the following implications:

(i) If for a second item $j$, $\delta_j < \delta_i$, then $p_{nj} > p_{ni}$. That is, the same person will have a greater probability of a positive response to the item with the lower location.
(ii) If for a second person $l$, $\beta_l < \beta_n$, then $p_{li} < p_{ni}$. That is, the person with the lower location will have a smaller probability of a positive response to the same item.

We return to the further implication of this structure for the relationship among items.

## Reasons for Multiple Items in Instruments

Most instruments in the social sciences are composed of multiple items. Responding to multiple items provides a kind of replication of responses of each person, and replications contribute to both precision of person estimates and to the validity of the instrument in assessing the required variable (e.g. proficiency or attitude).

In principle, the effect of multiple items applied to each individual is equivalent *statistically* to estimating the parameter of each individual from multiple replications of responses to a single item. Of course, it would be pointless substantively and statistically to ask a person to respond to the same item on multiple occasions, and therefore different items assessing the same variable are used. However, in terms of precision, the effect of having more than one item is equivalent to having that many replications. By analogy, the precision of the estimate of the mean of a distribution increases with the number of independent replications, commonly referred to as the sample size. Precision can also be understood as being potentially enhanced because with more items, there are more potential score points. For example, with just one dichotomous item, persons can be placed into just two categories; and with 10 dichotomous items, persons can be placed potentially into 11 categories. It needs to be appreciated that the precision is increased only if the items are operating as required, in particular, that they are operating independently. For example, if two dichotomous items have identical responses for all persons, then the persons would be placed into just two categories and either one of the items would be redundant relative to the other.

Validity is also enhanced with an increase in the number of items because each item assesses a somewhat different aspect of the same variable, or assesses the same variable in a slightly different though still relevant way, or some of both. Assessment restricted to only one way with one item may provide information that is too narrow for the kinds of decisions that need to be made from the assessments. Of course, there are situations where one response to one item can be decisive, and situations where multiple items do not enhance the validity of the assessment.

## Evidence from the Location and Thresholds of Items

### *Construction of Items*

The construction of items, both in content and in response format, needs to be carried out carefully in relation to the variable to be assessed in the frame of reference. Here, we note the quote from Duncan (1984) above as particularly relevant. Poorly constructed items cannot be saved by a statistical analysis using the Rasch model. Indeed, the model will just expose the problems with the items. In the construction of these items, understanding the features that make different items more or less difficult and constructing items accordingly is a central element.

However, the starting point for many analyses of instruments according to the Rasch model is an existing instrument which was refined on the basis of CTT. One rationale for doing so is that in both CTT and RMT, the total score characterizes a person—in CTT by definition and in RMT as a consequence of the model.

Although many instruments constructed with CTT analyses are likely to show general fit to the Rasch model, potentially they can also show deviations of one kind

or another. One of the reasons for both possibilities (fit and misfit) is that the locations of items and thresholds, central to the Rasch model, do not exist in the basic true score equation of CTT.

In addition to the role of the total score, another common condition between CTT and RMT is that items and thresholds discriminate equivalently. The location of a dichotomous item can be considered its threshold. Then, the discrimination in RMT is characterized by the slope of the item characteristic curve (ICC) in dichotomous items and the latent threshold characteristic curves (TCC) with ordered category items, which is a kind of average discrimination at all the thresholds among all the items. It is not formalized in this way in CTT, but the assumption that the items have a common latent correlation is equivalent to having the same discrimination.

Therefore, generally, and especially with dichotomous items, if items of an instrument are selected based on CTT, in which items with low discrimination are modified or eliminated, then they are likely to fit the Rasch model. However, because items are selected based on having a discrimination which is greater than some *minimum*, some items may have a very high discrimination compared to the majority of retained items. In this case, these highly discriminating items are likely to show misfit in a Rasch model analysis by over-discriminating relative to the *average* discrimination of the remaining items. These over-discriminating items in turn might induce some of those with minimum discrimination to under-discriminate relative to the *average* discrimination. Essentially, the RMT criterion is somewhat tighter, and symmetrical, compared to the CTT criterion which is asymmetrical in studying the discrimination of items.

### Implications of Item Locations—Dichotomous Items

One of the specific implications of the probabilistic Guttman structure of the Rasch model, and an advantage of RMT over CTT, is the relationship among the locations of the items. Statistically, the relative locations of the items are a function of the number of persons who score positively on the items. Wherever possible, this relationship needs to be more than merely a reflection of the relative frequency of positive responses. That is, the relative frequency should reflect a substantive relationship among items in which different items have an inherently and theoretically greater demand of the variable than other items.

In a well-designed test, the items with different locations, *difficulties* in proficiency assessment, should be generated with a hypothesis regarding at least the rank ordering of the locations. Although item locations are not part of CTT, experienced item writers in proficiency assessment nevertheless construct instruments with a range of item difficulties. Not only do these relative difficulties reflect an inherent hierarchy of proficiencies, but having a range of difficulties is sound assessment practice, with the easy items being presented earlier in the test.

An example might be helpful. Figure 29.1 shows an example of three items from a test administered by the State Department of Education in Western Australia as part

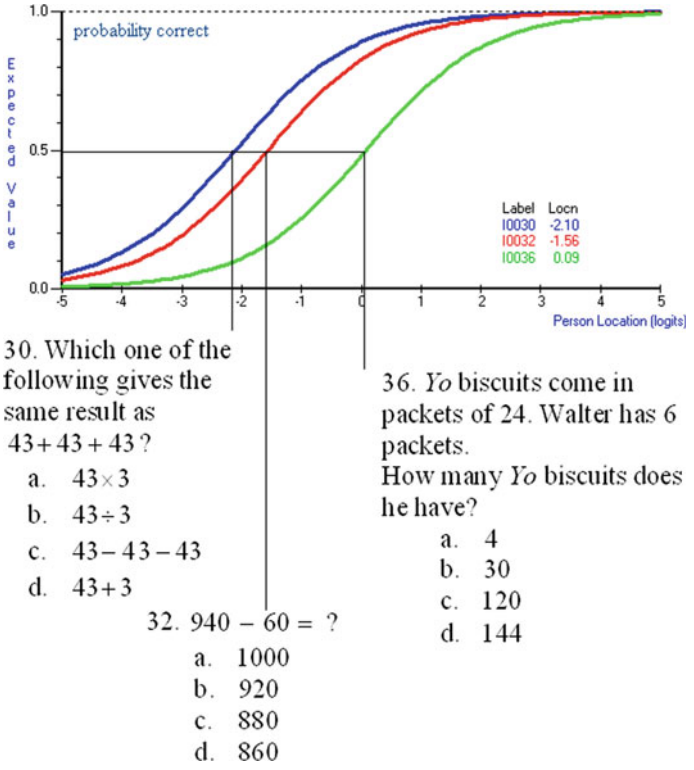


**Fig. 29.1** Three arithmetic items of different difficulties with an inherent hierarchy of proficiency

of its Monitoring Standards Program in government schools in 2005.[1] The reason that the probabilistic Guttman structure would be present with these three items is that there is an order of proficiency, with success on either of items 36 or 32 implying having achieved the proficiency inherent in item 30. Thus, item 36, which requires proficiency in division, implies having already understood the concept of subtraction which is taught earlier, and item 32 which requires proficiency in subtraction implies having already understood the concept of addition tested in item 30 which is taught earlier again. Ideally, as indicated above, items with different locations (difficulties) would be generated deliberately based on the understanding of the variable of assessment. In the case of proficiency assessment in schools, this understanding would include the curriculum and relevant syllabuses. Such an examination of the order of the item difficulties can also be conducted post hoc rather than the order hypothesized in advance. A post hoc examination is better than no examination at all. If the ordering of the items is very different from that predicted or explained post

[1] Reproduced with permission from the School Curriculum and Standards Authority for assessments originally developed for the Department of Education, Western Australia.

hoc, then some substantive explanation based on a qualitative analysis is required for the unexpected results.

### Ordered Category Items and Implications of Threshold Order

Ordered category items have an average difficulty of their thresholds, $\delta_i$ in Eq. (29.3), and these may be ordered in the same way that dichotomous items are ordered. Sometimes, however, in the context in which they are used, they are not ordered and there is not a substantive basis for the ordering. Often in Likert-style questionnaires this is the case. However, the hypothesis of the order of the location of thresholds *within items* with ordered categories is built into the format of the items and into the structure of the Rasch model.

An example in the application of a proficiency assessment, this time in health outcomes, is illustrated below. The example is the assessment of muscle tone shown in Andrich (2011) where the items are different functions of different limbs. In the illustrated example, assessments were eight ratings (items) of the parts of the lower limbs (hip adduction, knee extension, knee flexion and foot plantar flexion) for each side. The total score was taken as a summary of the muscle tone of the lower limbs for each person.

The assessment design was in ordered categories with the format shown in Table 29.2. In this example, it might *not be expected* that the average difficulties of the thresholds will be different. This indeed proved to be the case. However, there is an expected ordering of the threshold estimates. Importantly, from the point of view of understanding what it means to have more muscle tone, the hypothesis of threshold order was *not* confirmed between thresholds 3 and 4. Figure 29.2 shows the small reversal for item 5. That there is an anomaly between these two thresholds was confirmed by all items showing the same small reversal. It is evident that the proficiencies of *catch* and of *normal tone* were not distinguished in the assessments. In this case, all aspects of the assessment, from the definitions of *catch* and *normal tone* to the interpretation and implementation by the assessors (clinicians), need to be examined.

In developing items which have ordered categories, the structure of the ordering of the categories and the check on the empirical ordering needs to be as rigorous as that of the content of the items. Clearly, the ordering in the example is based on a presumed understanding of different levels of muscle tone and muscle rigidity. The

**Table 29.2** Format of the assessment of muscle tone

| Limb rigid (minimal movement) | Increased tone (restricting movement) | Increased tone (easily flexed) | Catch | Normal tone |
|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 |
| | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\delta_4$ Thresholds |

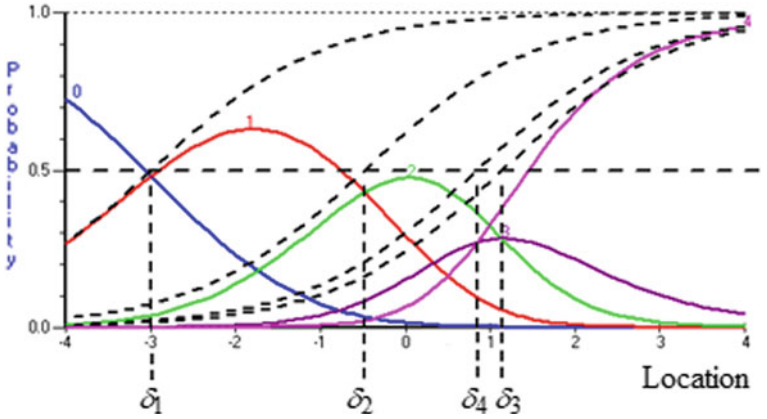


**Fig. 29.2** The category characteristic curves in the assessment of muscle tone

ordered categories are intended to reflect what it means to have more of the property (in this case muscle tone), and if the categories are not working as intended, then it is a reflection on a lack of some aspect of this understanding.

To be more explicit, when there are disordered thresholds, as in the example, a qualitative explanation regarding the operation of the format needs to be sought and hypothesized, and ideally tested empirically. Although *how* the thresholds are disordered can give a clue as to the possible explanation, *why* the threshold estimates are disordered cannot be explained solely from the statistical analysis. It must be recognized that the probability of a response in any category is a function of all thresholds, and therefore the responses in all categories affects all threshold estimates.

One way of overcoming disordered threshold estimates in many cases is to collapse adjacent categories, that is, score two adjacent categories the same way and reduce the maximum score of the item accordingly. However, simply collapsing categories to overcome the disordered thresholds may not replicate in another sample. Therefore, collapsing categories, and thereby obtaining ordered threshold estimates, should be taken as generating a hypothesis regarding a clarification and possible redefinition of the category. The hypotheses generated might then lead to (i) better distinguishing in the definition of the categories or (ii) if it is considered that there may be too many categories, to redefining the categories into a smaller number of categories. This new category system then needs to be checked empirically. The evidence regarding the ordering of thresholds within ordered category items is based on an a priori structure. This evidence is different from the evidence that comes from the statistical tests of fit.

## Assessing the Fit Between Responses and the Rasch Model

The statistical evidence of fit is directed at checking if the responses are consistent with each other as summarized by the Rasch model.

### *Meaning of Fit to the Rasch Model*

The property of invariance and sufficiency of the total score only holds for the data if the observed responses fit the model. We stress that the criterion of invariance, a criterion that provides measurements, comes before any data are collected—we take it as a *requirement* of the responses.

Constructing items for a frame of reference that fit the Rasch model is not an end in itself—it results from the intention to have invariance of comparisons and indirectly, to have the total score characterize each person. In addition, fitting the Rasch model is not sufficient to establish the validity of an instrument. As indicated above, the items need to be substantively valid, and in the case of items with ordered categories, the threshold estimates need to take on their natural order.

To say that responses *fit the Rasch model* is shorthand for two implications about the responses. First, it is shorthand for saying that the *responses provide invariant comparisons and that their total scores characterize the persons*. Second, it also implies that the items *work together and reinforce the evidence from each other*. No item can fit the Rasch model on its own. An item fits the model to the degree it is working consistently with the other items analysed in the data set.

Thus, *fit the Rasch model* is shorthand for *responses of the items were consistent with each other as expressed by the Rasch model to provide invariant comparisons which ensure that total scores characterize the persons*.

However, even this expanded statement needs to be qualified to indicate that fit to the model is at a particular level of precision. The precision in this case refers primarily to the number of persons and the spread of the persons. However, it is also affected by the number of items, the number of categories and the alignment of persons to the thresholds. Thus, (i) the greater the number of persons, (ii) the greater the number of items, (iii) the greater the number of categories and (iv) the better the alignment between the persons and items, the greater the potential precision. Then, the greater the precision of estimates, the greater the power of the test of fit, that is, the more likely it is that deviations from the model will be identified.

With respect to the sample size, with a very small sample any set of data will fit statistically; on the other, because no model can describe a particular data set to an infinitely high level of precision, a large enough sample can always be found to show that the data do not fit the model.

Because of this effect of the sample size on the power of the test of fit, if the sample is very large it might be useful, for this general check of fit, to reduce the sample size. Perhaps a sample size of approximately a factor of 10–20 persons for

each threshold for the set of items will suffice. Thus, with 15 items each with 3 categories (2 thresholds each), and therefore a total of 30 thresholds, a sample of the order of 15 * 30 = 450 persons can be used. However, the full sample should be used to establish the parameter estimates.

It is necessary to have a reasonable spread of persons relative to the thresholds of the items to achieve power in the test of fit. This ensures that there are opportunities for improbable responses to occur. If all persons had a similar location, and these were well aligned to the items, then responses do not have a range of probabilities. The extreme case of similar probabilities of responses is when persons are all similar in value and well aligned to the difficulty of a dichotomous item. Then, the probabilities of both responses are close to 0.5, and it is impossible to decide if a response is unlikely, and therefore misfits in some sense.

With respect to the spread of the persons and the test of fit, the important index is that of person separation. In the case that the person distribution is well aligned to the threshold distribution, it is analogous in construction to the CTT index of reliability. In the case that its assumptions are met, traditional true score reliability of CTT is well estimated by coefficient alpha, and if the persons are well aligned to the thresholds of the items, then the person separation index from the Rasch model as defined in this book is similar in value to the value of coefficient alpha. When they are not aligned and there are floor and ceiling effects rendering a violation of the assumptions of CTT, then these reliability indices diverge. These indices increase systematically with the number of thresholds within an item as well as with an increase in the number of items, providing the items and thresholds are functioning as required. With well-aligned thresholds and persons including a spread of items that covers the range of the person locations, it is possible to obtain a value greater than 0.80 for these reliability indices with 20 or so thresholds (e.g. 20 dichotomously scored items or 5 items with 5 ordered categories and 4 thresholds each). A value of the order of 0.80 for this index gives excellent power for the test of fit. A value of the order of 0.5 gives very weak power in detecting misfit.

As indicated above, no item can fit the Rasch model on its own. Even two items cannot be assessed for fit. It is necessary to have at least three items to test the fit.

Therefore, *items fit the Rasch model* is also shorthand for *the items' responses operate consistently with each other in reflecting a single variable as summarized by the Rasch model*.

Complementary-wise, *this item does not fit the Rasch model* is shorthand for *this item's responses do not operate consistently with the responses of the other items in reflecting a single variable as summarized by the Rasch model nor invariantly across the continuum*.

## *Identifying Misfitting Items*

It is expected that in carrying out an analysis with the Rasch model, there are theoretically and empirically hypothesized reasons as to why these items are placed in an instrument and why they should be operating together to assess a single variable.

Then, the purpose of analysing a set of data with the Rasch model is to check if the items do fit the model, recognizing that fitting the model is shorthand for checking for the invariance of comparisons, the sufficiency of the total score and for items operating with each other as summarized by the Rasch model. If the data do fit the model, and there is other evidence of the validity of the instrument's assessment of the intended variable, then all the benefits of measurement follow. For example, the Rasch estimates are linear, subsets of items will estimate the same parameter for each person and so on.

The decision that an item does not fit well and that it needs further consideration needs to be made on the basis of multiple pieces of evidence, statistical and graphical, and not merely on the basis of one statistic. Because there should be substantive and theoretical reasons for the inclusion of every item in an instrument, two consequences follow. First, it should be expected that only a few items will not fit the model, that is, do not operate consistently with the majority, perhaps something of the order of 10–15%. If many more items than this proportion misfit, it suggests an immediate examination of the construction, theory and administration of the items, taking into account which of the items do not fit, and the clues that the fit statistics give to the sources of the misfit. All features of the assessment need to be considered in deciding the sources of the problems, and these sources are outside the statistics themselves. For example, evidence of multidimensionality, response dependence, discrimination and so on, which can be statistical manifestations of poor instructions, poorly constructed items, inconsistent items, unclear marking keys and so on, needs to be examined from the perspective of the intended and required assessment.

Second, if only a few items misfit, then an attempt needs to be made to explain qualitatively the reasons for misfit of each of these misfitting items. The Rasch model analysis simply indicates that an item is not working consistently with the majority of other items, but the statistics cannot reveal the substantive reason why the item misfits. Much can be learned from items that operate differently from the original expectation that they will fit.

## *Dealing with Misfitting Items*

As indicated above, every item deemed to misfit relative to the operation of the majority does so for one or more reasons, and these need to be identified. Sometimes this might be challenging; and if it cannot be used in the particular context, there may be no option but to simply discard the item. For example, in a linking design where two groups of persons are administered different items with some common items, it is important that the common items do not show differential item functioning (DIF) among the groups of persons, which is a form of misfit. If one of these common items does show misfit in a particular data set, the item may need to be eliminated from the particular application.

However, the source of the DIF or other misfit should be studied and understood and the understanding used to ensure that such sources of DIF are controlled in

future test designs. The attempted explanation of misfit should be a hypothesis for future empirical testing. Dealing with misfit statistically is no match to anticipating problems and removing them in the design of the items and their administration.

In contrast to understanding, the source of the misfit, deleting items routinely and justifying the deletion of items only on the grounds of statistical misfit, is not consistent with either RMT or with sound instrument design and measurement practice.

Deleting many items, for example, 20% of the items or more lends itself to two inferential problems. First, it capitalizes on chance, and second it can distort the assessment so that the intended variable is not assessed. First, because of capitalizing on chance, the same item parameter estimates are unlikely to be found in another sample of responses and it is unlikely that the responses will fit the model. Second, by deleting many items from the large pool of items, which means retaining just a small subset of items that do operate consistently with each other, it is likely that many aspects of the variable intended to be assessed will not be assessed, thus distorting the original, intended variable of assessment.

Although the Rasch model can be used as a criterion to which data should fit to provide invariant comparisons, the responses must nevertheless subscribe to other substantive and methodological principles of scientific research, in particular, principles of statistical inference and reference to the substantive variable to be assessed. Not capitalizing on chance and not distorting the original substantive variable are two of these principles. Here again, the quote from Duncan (1984) at the beginning of the chapter is particularly apt.

Rasch (1960) set the precedent for careful analysis and test construction following his first two applications of the dichotomous model of Eq. (29.4). He derived the model from his analysis of reading tests, and then applied it to two sets of data he had at hand. One was from the Raven's test of progressive matrices, a well-known non-verbal intelligence test; the second was from a general test of intelligence. The former fitted the model to a very satisfactory degree of precision, but the latter did not. However, in this second case, rather than discarding items, or complicating the model, Rasch discerned from a study of the items that they appeared to fall into four different classes. When he showed these results, and their implications for *not* using a single summary score, to the original users of the test, they decided to reconstruct the original test into four new tests with each test having only one kind of the original kinds of items, which were intended to fit the dichotomous Rasch model. In particular, the total score on each new test would retain all the information in characterizing a person on that test. Thus, with all four tests, each person would be characterized by four proficiencies, not just one. Of course, within each test, the multiple items assess their own, finer aspect relative to the original test. No doubt, the performances on the four tests were correlated, but that is a different matter.

This reconstruction of the original test meant that all four aspects of that test continued to be assessed, but each aspect was assessed by its own test. Had Rasch proceeded to delete many items based on fit statistics alone, he is not only likely to have finished with a subset of items that might not have shown invariant properties in another sample, but also to have ended up with only one of the four original aspects being assessed. This outcome is most likely to have been the case had one

of the aspects had more items than each of the other aspects. In such a situation, this majority of items would define the most common variable and the items from the other three aspects would have shown misfit relative to this majority. Clearly, a selection of the majority that fit is likely to be from one aspect, which would have violated the substantive validity of the intended assessment.

Thus, fit statistics, which need to be considered in conjunction with each other, can only point to where there is an internal inconsistency with respect to the assessment of a single variable. They cannot explain the substantive reason for that internal inconsistency. Sometimes this inconsistency might be statistically significant but the item may be retained because the information it provides outweighs the effect of the statistical misfit. However, in each case that an item is deemed to misfit (taking account of the sample size, any misalignment with persons, and the like), a substantive explanation needs to be at least hypothesized as to why it misfits. Then, whether it is modified, retained or discarded, a substantive justification, outside the statistical analysis, needs to be provided. To the degree that the item misfits, to that degree it detracts from invariance of comparisons and from the total score being a sufficient statistic for the person estimate.

## *Separating the Scale Construction and Person Measurement Stages*

The separation of the person parameters from the item parameters in estimation in the Rasch model is an important mathematical and statistical property that provides the property of invariance of comparisons. However, it can be, and even needs to be, seen also as an empirical and experimental characteristic. In principle, the instrument construction stage should be conceptually, and often empirically, separated from the person measurement stage. The construction itself may require more than one iteration. Often, and often unfortunately, they are carried out from the same set of data.

Sampling of persons to help construct an instrument may be different from the sampling of persons who are to be assessed. To provide the evidence to check the operation of items, it is necessary to have persons whose responses contribute to relevant information. Thus, in the study of items in a test of proficiency, it is important to have more and less proficient persons so that the persons in this sample contribute the same information across the relevant range of the difficulties of the items. Thus, ideally, one would try to obtain persons across the whole required range of proficiency and as close as possible to being uniformly spread. Unless deliberately selected this way, samples are more likely to have a unimodal rather than a uniform distribution. If they have a unimodal distribution, such as the normal, they provide much more information about items in around the mean of the persons than at the tails of the distribution.

When the range of the proficiencies of the variable to be assessed is relatively large, another challenge, but one which can also be exploited, presents itself. Because it is pointless to administer items that are very easy to students who are very proficient, and items that are difficult to students who are not very proficient, some kind of tailored or adaptive testing and linking design needs to be constructed. Then, the less proficient persons are administered the less difficult items, the moderately proficient the moderately difficult ones with some overlap, and the highly proficient the most difficult items, again with some overlap. Then, the common items must be checked for DIF among the proficiency groups. This was the original challenge that Rasch met and which led to his studies reported in Rasch (1960). With modern computerized administration of tests, this becomes a very viable approach.

In addition, if one needs to check for DIF with respect to some grouping criterion, say language background, then for the stage of scale construction there should be a similar number of persons in the sample in each group. This similarity of numbers is required because the information is in the sample, and in principle, the persons who might have smaller numbers in the population should not contribute less information to the checking of DIF. Then, for the person measurement stage, the item values can be anchored to those estimated from similar sample sizes, and the responses of all persons assessed.

## Summary

In summary, because the Rasch model arises from a requirement that is independent of any particular data set, and in particular, it is not applied simply to model a data set, misfit of the data to the model implies that the data do not meet the requirement. This requirement is that the responses to the items, within a defined frame of reference, provide a particular kind of invariance. The responses will provide the invariance only if the data fit the model.

The challenge in social measurement is to construct items of instruments which do fit the model. However, fitting the model is not a sufficient condition to ensure that the instrument assesses the intended variable. Therefore, any statistical misfit needs to be considered in conjunction with the substantive variable intended to be assessed. Because every item is chosen for the reason that it assesses the relevant variable, every item deemed to misfit needs to be treated as an anomaly that needs to be explained in terms of the construction or administration, or some other feature of the item, perhaps in relation to the other items such as local dependence.

In ordered category items, the empirical ordering of the categories is assessed by the ordering of threshold estimates, and not by any statistical test of fit. However, once again, if the empirical ordering is not consistent with the intended ordering, it is an anomaly that needs to be explained.

Finally, the statistical evidence of misfit using probabilities such as those associated with the chi-square statistic needs to be understood as providing evidence to consider in assessing an instrument, not for mechanistic interpretation. For example, a chi-square probability of 0.01 for an item might imply different considerations in

different circumstances. Thus, if all the other items which have a greater probability than 0.01 jump to have values greater than 0.11, then this item and the other items which have a lower probability might be studied more closely. On the other hand, if the probabilities of the other items which have a greater probability than 0.01 increase smoothly, and the greatest jump to the next lowest probability for a chi-square for an item is say 0.005, then that item might receive less consideration as showing misfit. In every case, where some concerns with misfit are identified, the structure, format and content of the item relative to the other items and the persons to whom the instrument was administered, should be considered and the item not deleted simply, and mechanistically, on statistical grounds.

Gigerenzer (1993) laments the mechanistic application of significance testing in which some hybrid logic of Neyman–Pearson and Fisher, with which neither would have agreed, is prevalent in the social sciences.

> Statistical reasoning is an art and so demands both mathematical knowledge and informed judgment. When it is mechanized, as with the institutionalized hybrid logic, it becomes ritual, not reasoning. (Gigerenzer, 1993, p. 335).

Thus, use all the evidence and do not use significance tests or any other criterion of fit in a mechanistic way.

## Exercises

*Exercise 8: Writing up a Rasch model analysis* in Appendix C.

## References

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585.

Duncan, O. D. (1984). Rasch measurement further examples and discussion. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2). New York: Russell Sage Foundation.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, New Jersey: L. Erlbaum Associates.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by B. D. Wright (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceeding of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in D. J. Bartholomew (Ed.), *Measurement: Sage benchmarks in social research methods* (Vol. I, pp. 319–334, 2006). London: Sage Publications.