Chapter 8
Equating Measuring Instruments in the Social Sciences: Applying Measurement Principles of the Natural Sciences

David Andrich and Dragana Surla

Abstract The concept of measurement in which the magnitude of a property is quantified in a common unit relative to a specified origin is a deep abstraction. This chapter shows the application of measurement in a social science context where the motivation is transparency and equity rather than the advancement of scientific laws. However, to achieve these, the realization of measurement needs to be no less rigorous than it is in the advancement of scientific laws. Rasch measurement theory provides the basis for such rigor. The context in this chapter is competitive selection into universities in Western Australia based on a summary performance on a series of instruments which assess achievement in a range of discipline areas. Such selection tends to determine life opportunities; therefore to ensure consistency and fairness, performances on different instruments need to be transformed into measurements which are in the same, explicit unit relative to a specified origin. Because the illustrative context is complex, it is considered that the Rasch measurement theory applied in this chapter could be applied to a range of social contexts where assessments on different instruments need to be transformed to measurements in a common unit referenced to a common origin and where the focus is on making decisions at the person level.

Keywords Instrument equating · Quantification · Theory, Rasch models, Measurement

D. Andrich (✉) · D. Surla
The University of Western Australia, Perth, Australia
e-mail: david.andrich@uwa.edu.au

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_8

8.1 Introduction

Measurement, the quantification of the magnitude of some property of an object from a specified, convenient or natural *origin* in a constant *unit* of an instrument, is a deep abstraction. For example, the elementary measurement of mass using a beam balance is both simple and sophisticated. Though the balance is a relatively simple instrument, the *conceptualization* to quantify the mass of the object in a meaningful way is a remarkable abstraction. The balance involves having a three-dimensional object, potentially of any shape, volume, color, and so on, on one side, and a set of equivalent objects on the other side that balances it, mapping the count of this set on a real number line, itself partitioned into equal contiguous distances. There is little in the natural world that even approximates a real number line, which is totally abstract, where even a drawn line to represent it is ragged when looked at through a microscope.

Despite aspects of deep abstraction and scientific implications, school-children understand the concept of measurement of mass using a beam balance, and understand more generally the idea of measurement in a constant unit relative to a specified origin. The motivation for measurement of common properties was not to advance physical laws, but the fair trade of objects with properties such as mass, length, and volume. The standardization of units for this purpose is exemplified by the development of the metric system [1]. Although standardized for purposes of fair trade, the metric system was developed by scientists of the highest order. The relationship of science to measurement from this perspective is summarized by Alder:

> We often hear that science is a revolutionary force that imposes radical new ideas in human history. But science also emerges from within human history, reshaping ordinary actions, some so habitual we hardly notice them. Measurement is one of our most ordinary actions. We speak its language whenever we exchange precise information or trade objects with exactitude. This very ubiquity, however, makes measurement invisible. To do their job, standards must operate as a set of shared assumptions, the unexamined background against which we make agreements and make distinctions. So it is not surprising that we take measurement for granted and consider it banal. Yet the use a society makes of its measures expresses its sense of fair dealing. That is why the balance scale is a widespread symbol of justice [1, p. 1].

This chapter is concerned with scientific measurement in which the prime purpose is transparency and fairness, and not the advancement of laws. Nevertheless, for this important purpose, measurement needs to be as rigorous as that needed to advance quantitative, scientific laws, and needs to be of the same kind that advances such laws. Thus, although rigorous measurement will be required for the development of quantitative laws in the social sciences, there are social contexts where transparency and fairness of decisions seem sufficient to require rigorous scientific measurement. This chapter provides such an example.

8.1.1 A Complex Social Science Context

In educational assessment of proficiency, it is common to refer to the instruments of assessment as *tests*. Because the example of this chapter is illustrative of a general approach to equating, and because of the connections made with measurement in the natural sciences, they are referred to as *instruments*. The typical study and process of equating scores from different instruments assumes that they assess the same variable [24]. The context of the example of this chapter, referred to as the *frame of reference* for reasons that emerge later, is substantially more complicated than that. The context is the selection of students into universities in Western Australia based on their assessment on *instruments* from an array of *disciplines* at the end of 12 years of schooling. There are some 40 disciplines of possible study, and students must have scores on at least four, including in the discipline of English, though many have scores on five or more disciplines. Although there are prerequisites for university entry in some fields, such as engineering, to meet other policy requirements – for example, not specializing studies too early – these are minimal. In addition, where there are prerequisites, additional electives may be chosen, and these are not the same among the candidates. Each instrument's scores range from 0 to 100 and, because a summary score of each student's profile is calculated and the students are ranked for competitive entry into universities in Western Australia, the scores of all instruments need to be equated onto a measurement scale with the same unit relative to a specified origin.

Within each discipline area, the instrument scores can be considered to reflect *causal* variables in which proficiency of the student in the discipline area governs the performance on the instrument. In principle, within each instrument, different items that assess the same proficiency are exchangeable. However, the summary score across a range of disciplines, can be considered an index variable [4, 39, 40, 43]. This variable can be considered a higher order, *thick* variable, one thicker than those from each of the disciplines, that reflects a general capacity to profit from a university education as evidenced from previous relevant study.

The frame of reference is complex for the following reasons. First, the summary index variable is even more complex than the kind illustrated by Stenner, Burdick and Stone [39]. Their example defines socioeconomic status (SES) by *education, occupational prestige, income, and neighborhood.* In defining SES in this way, none of the components are exchangeable. These same components define the variable for each person. However, in the example of university selection, they are not defined by a fixed set of disciplines, and in principle, they are exchangeable. For example, two candidates may be competing for entrance to the same university study, such as law or psychology, with only one instrument score out of four or five which is common. Second, because candidates self-select the disciplines they study, and they have scores on different combinations of instruments, no pair of instruments have entirely common candidates. Third, although the instrument scores are positively correlated, the correlations among instrument scores are not homogeneous. For example, the correlation between instruments of Mathematics and Chemistry is greater than that

between either discipline and English or History. Fourth, the scores are probabilistic, not deterministic, relative to proficiency, and for their analysis a measurement theory that is inherently probabilistic is required. Finally, as indicated above, for various historical reasons, the scores have a finite range, and therefore, especially close to the higher limits of the range where competitive scores are most relevant, the relationships among the scores are not linear. How each of these complexities is accommodated in producing measurements on the same scale relevant for the purpose of university entry is the substance of this chapter. Because of the complexities of this frame of reference, it is considered that many other contexts that require equating of instruments might be accommodated by the application of the Rasch measurement theory described in this chapter.

8.1.2 Empirical Understanding and the Role of the Rasch Model for Measurement

In the natural sciences, advanced measurement is derived from the scientific, theoretical understanding of the relevant variables and their relationships [26]. Direct reading of measurements from the instrument hides the substantive theory and design that *manifests* the property measured and *controls* properties that disturb it. The beam balance exploits the effect of gravity and simply controls other factors, for example measurements in very small units of chemical properties used antique balances that were enclosed to ensure no disturbance from air movement. Appropriately anchored, and elegantly for such an elementary instrument, the balance will give the same measurement whether under the gravitational force of the Earth or the Moon and whether it is stationary or accelerating relative to either of them. From the perspective of scientific theory, the separation of the concepts of mass and gravity took the geniuses of Galileo and Newton, and at the time was controversial. The studies of mass and gravity in physics, from nuclear energy to gravity waves, continue to be advanced areas of physics. However, to ensure transparency and equity, everyday transactions also require accurate measurement of mass.

The beam balance is not only familiar, but its application is relatively tangible, thus disguising some of the sophistication in the understanding of mass and gravity. On the other hand, the now equally familiar mercury thermometer for measuring human and day temperatures, though now easy to apply, is conceptually much less tangible than the beam balance. In particular, the origin and unit are more abstract than a unit of mass, for example, a kilogram. In Celsius, at one atmosphere of pressure, 0° C is set at the freezing point of water, 100° C at the boiling point, and the unit is one 100th of this range. These end-points are chosen for convenience of relatively observable everyday phenomena. Empirical work was required to ensure that the uniform expansion of mercury in a thin tube in the range specified, also implied uniform increases of temperature.

The thermometer requires the control of the relative expansion of all of its other components. This control requires the scientific understanding of vacuums, pressure of gases, and so on. This is the reason that the construction of reliable, practical thermometers required the work of the best scientists of the nineteenth century. Indeed measurement of temperature was developing before the concept of heat as the kinetic energy of atoms was understood, meaning that the measurement of this variable and its understanding were more or less simultaneous [15]. An important inference from this example is that the reliable measurement of a variable indicates a substantial understanding of its properties, and reciprocally, that successful measurement can enhance further understanding.

In the frame of reference in the example of this chapter, the counterpart to the theoretical understandings of variables are the transparent, explicit syllabuses for each discipline for Years 11 and 12 of schooling, the teaching of these syllabuses by qualified teachers, the design of instruments to assess a range of proficiencies that reflect each syllabus, and the anonymous scoring of the performances of each candidate by two independent qualified markers. Even the studies up to Year 10, in the development of student proficiencies to enable them to study at Years 11 and 12, are relevant. Thus a great deal of professional, intellectual and empirical work goes into the production of scores on the instruments by students. The ultimate, substantive, validity of each instrument is the public outcry in the newspapers, and these days, social media, if questions stray from the syllabus.

It is stressed that such substantive, empirical work is essential and that the application of Rasch measurement theory to equate instruments and to map observed scores into measurements cannot overcome, through probabilistic modelling, any shoddy, superficial, and poorly constructed instruments that do not validly reflect candidates' relative proficiencies in the respective disciplines. The function of the application of Rasch measurement theory, where scores of all instruments are valid, is to ensure that the different instruments are transformed onto a measuring instrument with the same unit relative to the same specified origin. The reason this is necessary is that each instrument is designed to align the range of the difficulties of its items to the proficiencies of the candidates, and that relative to the higher order index variable described above, these proficiencies are not the same in each discipline. In addition, the structure, format, and scoring of different questions, which is natural to the different disciplines, is not the same across these disciplines. As a result, every instrument has its own implied relative unit, which may not even be consistent across its own continuum of proficiency. Finally, each is referenced to its own relative origin; in common parlance, instruments are more or less difficult and instrument scores are more or less skewed.

8.1.3 Descriptions of Measurement

Because of its role in science and everyday applications, it is not surprising that measurement has been studied by physical and social scientists, among many of them are Campbell [13], Duncan [16], Finkelstein [19], Fisher and Stenner [20],

Luce and Tukey [27], Mari [30], Ramsay [35], and Wright [47]. Perhaps surprisingly, physical scientists can, and most do, take the concept and need for measurement for granted. On the other hand, social scientists generally do not have that luxury. This section does not provide a review of discourses on the history, structure, function and definitions of measurement; instead, to set up the distinctive definition in Rasch measurement theory, and why this definition is most germane to the example of social science measurement of this chapter, it only summarizes briefly three common definitions of measurement.

Measurement assumes that magnitudes of properties can be mapped on a real number line. This is assumed with the measurement of mass and temperature summarized above. With this assumption, three definitions are most common. First *classical*, second *representational*, and third *additive conjoint*. They all attempt to describe measurement by formalizing the properties illustrated above with the measurement of mass and temperature.

The classical definition emphasizes measurement as the ratio of the amount of the property of an object relative to an amount defined as the unit [33]. The representational definition emphasizes that the relationship among the properties is the same as the relationship among the numbers assigned to the objects [25]. For example, for certain physical relationships among masses, concatenating two objects is equivalent to having a single object whose mass is the sum of the measures of each object. The additive conjoint definition is more abstract, requiring that the rows, columns, and cells of a two-way table with real numbers, which reflect the magnitudes of properties, can be transformed monotonically to produce an additive structure [27].

Rather than considering some of its operational aspects as a basis for formalizing a definition of measurement, Rasch arrived at his requirement from his empirical studies in reading proficiency [36]. Specifically, that within a *specified frame of reference*, and with stimuli (referred to as instruments in this chapter) and individuals characterized by real numbers,

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison... Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; ... [37, p. 332].

From the mathematical abstraction of his definition, and further derivations from them, Rasch observed that the properties of measurement that are characterized by classical, representational, and additive conjoint, are satisfied [36]. However, he considered that the requirement of *invariant comparisons* stood as a more fundamental basis for measurement than any *description* of measurement [38]. Andrich [6] makes the case that, not only is Rasch's definition compatible with the other three, but that it *explains* them. Moreover, the other definitions are deterministic, whereas Rasch's theory is set in both deterministic and probabilistic contexts, where the probability characterizes the uncertainty in the observation when one person encounters one instrument. It is not concerned with the distributional properties of populations of persons. This chapter applies the probabilistic formulation.

8.1.4 Measurement of Variables with No Physical Counterpart

Rasch's formulation of invariant comparisons, conveniently and elegantly, abstracts measurement further than typically considered in the natural sciences in that it makes no reference to any physical properties. Rasch was not the first to consider such an abstraction and the quest for invariance of comparisons. In abstracting measurement from his work on comparative judgements of pairs of a set of objects with respect to a physical property such as sound pitch or luminescence, Thurstone commented:

> One of the main requirements of a truly subjective metric is that it shall be entirely independent of all physical phenomena. In freeing ourselves completely from physical measurement, we are also free to experiment with aesthetic objects and with many other types of stimuli to which there does not correspond any known physical measurement [44, pp. l82–83].

Thurstone emphasized that the mapping of the magnitude of the property on a real number line involved very specific focus on a variable, and the control of all other variables. In the construction of instruments for measuring attitudes in terms of statements that reflected different degrees of the attitude through the opinion they expressed, he writes that

> The various opinions cannot be completely described merely as "more" or "less". They scatter in many dimensions, but the very idea of measurement implies a linear continuum of some sort, such a length, price, volume, weight and age. When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind [44, pp. 218–19].

On the requirement of invariance of comparisons, he articulates that

> If a scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale [44, p. 228].

Rasch's requirements of invariant comparisons with respect to any properties that can be characterized by real numbers, physical or not, are compatible with Thurstone's. However, the distinctive part of Rasch's formulation is that the requirements are expressed formally in mathematical terms [37, 38]. As a result, further mathematical derivations can be carried out. Among other epistemological implications of these derivations [6], it makes it possible to apply the resultant model to real data. This chapter is concerned with one such example where the measurement model applied to define a unit and origin of a standard instrument could only have been formulated through a mathematical derivation.

8.1.5 Structure of This Chapter

The rest of this chapter is structured as follows. Section 8.2 provides a summary of the polytomous Rasch model and the special case of the *Rasch distribution* which makes tangible the analogy between measurement in the natural sciences and Rasch measurement theory. The thresholds in the Rasch distribution are equidistant, and the common distance is identical in interpretation to the unit of a measuring instrument in the natural sciences. In addition, this distribution, which is the inferred distribution of replications, is a discrete analogue of the continuous Gauss distribution of uncertainty for replicated measurements. In order to satisfy the unidimensionality property of the Rasch model, and therefore optimize the relationship between the instruments in deriving the equating functions, this section also provides a rationale for editing the profiles of persons. This editing is based on obtaining relatively homogeneous profiles, that is, those for which the total score is sufficient. In a complementary fashion, the section also provides a rationale for identifying, at the person measurement stage, those profiles which can and cannot be characterized by their total score. Then, because of the frame of reference of the measurement, those profiles that cannot be summarized by their respective total scores need to be considered in terms of both a summary estimate and the properties of the profile.

Section 8.3 provides an empirical example from the frame of reference described above. Specifically, it shows the details of real data from six instruments used for university entrance examinations and the results of equating these instruments to a common instrument of the Rasch distribution, where the origin and unit are defined identically, and not merely analogously, to those of an instrument measuring physical variables and are chosen for convenience in their frame of reference. Descriptive statistics before and after equating are provided with the example, emphasizing its illustrative properties that can be transferred to other contexts. The final section is a summary.

8.2 The Rasch Model and Distribution

From three successive papers [2, 3, 37], where the theoretical characteristic of these papers is emphasized by there being no data analysis in any of them, the Rasch model for ordered categorical data can be expressed in the now familiar forms, as either a *rating* or *partial credit* parameterization [31] according to

$$\mathrm{P}\{X_{ni} = x; \beta_n, \psi_{xi}\} = [\exp(\psi_{xi} + x\beta_n)]/\gamma_{ni}, \tag{8.1}$$

where X_{ni} takes integer values $x = 0, 1, 2, \ldots, m_i$ when person n, measure β_n, is assessed with instrument i; $\psi_{xi} = -\sum_{k=0}^{x} \tau_{ki}\, k = 1, 2, \ldots m_i$, τ_{ki} are the instrument's m_i thresholds where $\tau_{0i} \equiv 0$ is introduced for notational convenience; and γ_{ni} is a normalizing factor. In common psychometric applications, the instrument in

Eq. (8.1) is an item of an instrument or questionnaire. The relevant part of the partial credit parameterization, relative to the rating parameterization, is that in the former all the thresholds of the different instruments can have different values, whereas in the latter the means of the thresholds of the instruments can be different but the deviations of the thresholds from their mean is the same across instruments.

The response structure for any one person responding to one instrument is the same in both parameterizations. However, interpreting the thresholds as *steps,* and presenting Thurstone thresholds reconstructed from the Rasch model [31, 32], is not compatible with the Rasch model [5, 34]. Moreover, although *estimates* from data for an instrument can result in reversed threshold values relative to their natural order, it is evidence of a problem with the responses produced by the instrument. Although there are multiple reasons for this conclusion, as seen later, there are two particular reasons thresholds must be in their natural order for the purposes of this chapter. First, if they were disordered, it would not be possible to define a unit of an instrument as the common, equal distance between successive thresholds, as in a measuring instrument in the physical sciences and in this chapter. Second, the distribution of uncertainty around any measurement would be bimodal, whereas all random error distributions of uncertainty, including the Gauss distribution are not only strictly unimodal, but the transition between successive probabilities is smooth. With thresholds in their correct order, the Rasch model satisfies this criterion of smooth, strict unimodality [9].

In any analysis, the parameter β is expressed in what are commonly referred to as *logits*, though the logit is not a unit of the instrument of the kind found in the natural sciences, nor of the unit as defined in this chapter [22]. Specifically, the same logit value across separate analyses of different data sets is not expressed in the same unit in the sense of the unit used in the present chapter.

At *threshold* τ_{ki}, the probabilities of responses in its two adjacent categories are equal. With a maximum score of 100 for each instrument, as in the example of this paper, there are 100 thresholds. This is an over-parameterization of the model, and with zero frequencies in the data, especially with low scores, direct estimation fails without modifying the model in some way [28, 46]. Therefore, instead of attempting to estimate all thresholds directly, they are estimated through their first four *principal components* given by

$$\begin{aligned}\psi_{xi} = \quad & -x\delta_i + \{x(m_i - x)/2\}\Delta_i + \{x(m_i - x)(2m_i - x)\}\lambda_i \\ & + \{x(m_i - x)\left(5x^2 - 5xm_i + m_i^2 + 1\right)\}\zeta_i,\end{aligned} \tag{8.2}$$

where δ, Δ, λ, ζ characterize the location, spread, skewness and kurtosis of the thresholds of the instrument. It is stressed that these parameters characterize properties *of the thresholds*, *not* the distribution of persons. It is also emphasized that the person parameter β is a scalar, and therefore is said to be *unidimensional* [7].

It is relevant to stress the implied interpretation of the Rasch model of Eq. (8.1). Namely, that given the values of the instrument parameters, then for a given value of β, Eq. (8.1) is the inferred distribution of responses as if the same person responded

to the same instrument an infinite number of times. Clearly, this is not administratively feasible, but even if it were, if the same person responded to the same instrument multiple times, there would be substantial local dependence. However, that is not the point; as part of the abstraction from the data by applying the model, the distribution is an *inferred distribution* of replicated responses. An important aspect of this distribution when applied to real data is that, by analogy to the Gauss distribution, it is a *random* distribution [9]. This implies that no unaccounted-for factors are producing systematic errors in the measurements, which is checked by both ensuring fit between the data and the model and that the thresholds estimates are in their natural order [10].

8.2.1 The Expected Value Curve and the Equating Function

The estimation of the instrument parameters is considered briefly in a later section. Here we consider the relationship between observed scores and person estimates given an instrument's values $(\delta_i, \Delta_i, \lambda_i, \zeta_i)$, which may be estimates. The estimate of $\widehat{\beta}_n$ for each person is given by a maximum likelihood estimate (MLE) individually. However, the relationship between a score x_{ni} and the estimate is analytic, and holds whether or not any person in a sample obtained such a score.

For a set of instruments for which person n has scores, $\widehat{\beta}_n$ is given by the solution to the implicit equation

$$t_n = \sum_{i_n=1}^{I_n} x_{ni} = \sum_{i_n=1}^{I_n} E[X_{ni}] = \sum_{i_n=1}^{I_n} \sum_{x=0}^{m_i} x\widehat{\Pr}\{x_{ni}\}, \tag{8.3}$$

where i_n, I_n indicate the instruments to which person n has responded and t_n is the total score on these instruments. Thus, for the person estimates β to be on the same scale, not all persons need to respond to all instruments, a key feature of the application of the model in the frame of reference of the example.

For each instrument, Eq. (8.3) specializes to Eq. (8.4), which given the observed score x_i, gives the estimate $\widehat{\beta}_{xi}$ on instrument i. Reciprocally, given any person value β, Eq. (8.4) also gives the expected value, $E[X_i|\,\beta]$, on the instrument:

$$x_i = E[X_i] = \sum_{x=0}^{m_i} x\widehat{\Pr}\{x_i\}. \tag{8.4}$$

Equation (8.4) can be solved for each total score x_i whether or not any person obtained that score, giving a unique estimate which we denote $\widehat{\beta}_{xi}$ and drop the person subscript n. In this case, though real numbers, because $x_i \in \{0, 1, 2, \ldots, m_i\}$ are discrete, the $\widehat{\beta}_{xi}$ are discrete. On the other hand, Eq. (8.4) can be solved for $E[X_i]$ given any β and is in principle continuous. Both relationships of Eq. (8.4) are applied in this chapter. This application is introduced in Fig. 8.1.

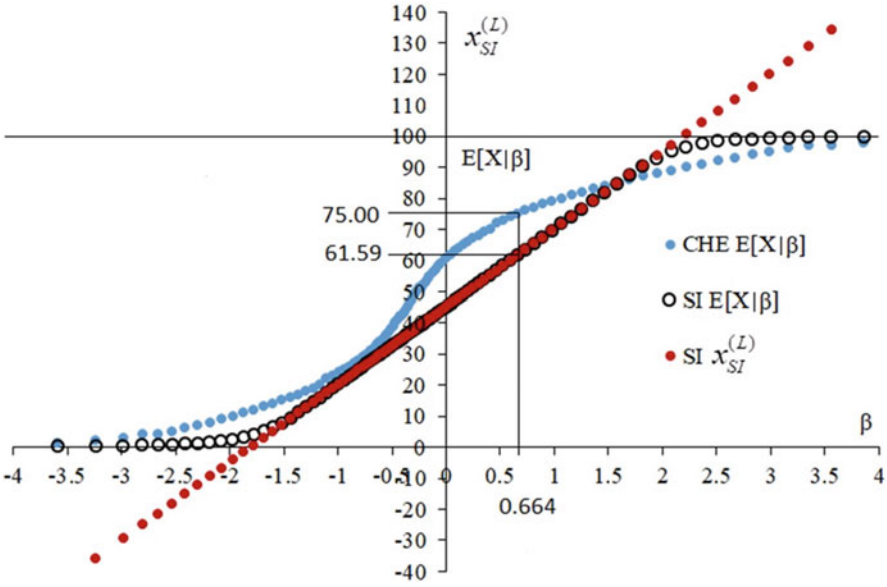

Fig. 8.1 $E[X|\beta]$, $(\delta, \Delta, \lambda, \zeta) = (-0.06325, 0.05025, 0.00024, 0.00001)$ for CHE; $E[X|\beta]$, $(\delta, \Delta) = (0.20000, 0.04000)$; and the linear measurement function $x_{SI}^{(L)}$ of the Standard Instrument (SI), $x_{SI}^{(L)} = E\left[x_{SI}^{(L)}|\beta = 0\right] + \beta_x/\Delta = 45 + 25\beta_x$

Figure 8.1 shows the relationship between the observed score x_i and person estimate $\widehat{\beta}_{xi}$, and the expected value curve $E[X_i|\beta]$ as a function of β, for one of the instruments called CHE which is analyzed in detail later in this chapter. The values of the parameters of this instrument are shown in the caption of Fig. 8.1. Because of the large value of 100 for the maximum score, the values of these parameters are small – therefore they are shown to five decimal places. Although referred to as an item characteristic curve when responses are dichotomous because it is more descriptive in its meaning, the graph of Fig. 8.1 is referred to as an *expected value curve* (EVC).

Figure 8.1 also shows a second EVC of an instrument that is linear over a substantial range of the continuum with curvature only at the extremes. This instrument has only two parameters, the first two principal components (δ_i, Δ_i) of Eq. (8.2), whose values are shown in the caption of the Figure. In anticipation of further detail below, this instrument is referred to as the Standard Instrument (SI), and is the instrument to which all instruments are equated or mapped. The SI has equidistant successive thresholds, Δ_i, the value which is identical to $\beta_x - \beta_{x-1}$ over a well-defined range. Therefore, it is analogous in interpretation to the unit of a standard measuring instrument. Moreover, again in anticipation of further elaboration, over the same range of the values, the random, uncertainty distribution is the discrete counterpart of the Gauss distribution, the distribution of random variation of replicated measurements. The SI is not simply an empirical regression equation, but

an instrument whose unit and origin are both explicit and are deliberately chosen for context relevance and convenience. Accordingly, a value on the SI is referred to as a *measurement*.

Here we note that Fig. 8.1 shows the effective mapping function of the instrument CHE onto the SI. In particular, the score $x = 75$ on CHE, $\left\{\widehat{\beta}|x = 75\right\} = 0.664$, gives the measurement, $E[X_{SI}|\beta = 0.664] = 61.591$, on the SI. Finally, Fig. 8.1 shows a full linear extrapolation of the SI beyond the minimum and maximum scores of 0 and 100, which is explained later in this chapter.

8.2.2 The Rasch Distribution of Uncertainty

We now explain in detail the SI. A special case of the Rasch model of Eq. (8.2) involves just the first two principal components, (δ_i, Δ_i), giving

$$\mathrm{P}\{X_{ni} = x; \beta_n, (\delta_i, \Delta_i)\} = [-x\delta_i + \{x(m_i - x)/2\}\Delta_i + x\beta_n]/\gamma_{ni}. \tag{8.5}$$

This distribution has a distinctive role in showing measurement of the kind found in the physical sciences. Therefore, two of its characteristic features are elaborated now: first, the presence of the explicit unit Δ relative to the specified origin δ, directly equivalent to that in the physical sciences; and second, the resultant random distribution of measurement uncertainty which is the discrete analogue of the Gauss distribution of uncertainty of replicated measurements in the physical sciences.

Because of these features, Eq. (8.5) is referred to as the *Rasch distribution*, rather than simply a model.

8.2.3 The Unit in the Rasch Distribution

The EVC of the empirical instrument in Fig. 8.1 has a non-linear relationship with the continuum β. The non-linearity results from the presence of skewness and kurtosis in the thresholds. On the other hand, because its thresholds have no skewness or kurtosis, the SI is linear over a substantial range. An excerpt of the relationship between x, $E[X_{SI}|\beta]$ and β is shown in Table 8.1, where the observed scores and expected values in the range 15–86 inclusive are highlighted in bold. Table 8.1 also shows a column referred to as $x_{SI}^{(L)}$ which is defined formally below as *measurements* on the SI.

Not only is the relationship between x, $E[X_{SI}|\beta]$ and β linear in the range shown, but it makes the unit explicit. Thus the difference between two successive values β_{x+1}, β_x is not only constant and linear with the observed score x, but the difference between them is exactly the unit $\Delta = 0.04$, that is, $\beta_{x+1} - \beta_x = \Delta$; $x = 15, 16, 17,$

Table 8.1 The relationship between x, $E[X_{SI}|\beta]$, β of the Standard Instrument of Fig. 8.1

x, $E[X\|\beta]$	β	$\beta_{x+1}-\beta_x$	$x_{SI}^{(L)}$		x, $E[X\|\beta]$	β	$\beta_{x+1}-\beta_x$	$x_{SI}^{(L)}$
0	−3.114	.	−32.9		...	...	...	...
1	−2.400	0.714	−15.0		82	1.480	0.040	82.0
2	−2.051	0.349	−6.3		83	1.520	0.040	83.0
3	−1.880	0.171	−2.0		84	1.560	0.040	84.0
4	−1.768	0.112	0.8		85	1.601	0.040	85.0
5	−1.684	0.084	2.9		86	1.641	0.040	86.0
6	−1.616	0.068	4.6		87	1.682	0.041	87.1
7	−1.557	0.059	6.1		88	1.724	0.042	88.1
8	−1.504	0.053	7.4		89	1.766	0.042	89.2
9	−1.456	0.048	8.6		90	1.810	0.044	90.3
10	−1.410	0.046	9.8		91	1.856	0.046	91.4
11	−1.366	0.044	10.9		92	1.904	0.048	92.6
12	−1.324	0.042	11.9		93	1.957	0.053	93.9
13	−1.282	0.042	13.0		94	2.016	0.059	95.4
14	−1.241	0.041	14.0		95	2.084	0.068	97.1
15	−1.201	0.040	15.0		96	2.168	0.084	99.2
16	−1.160	0.040	16.0		97	2.280	0.112	102.0
17	−1.120	0.040	17.0		98	2.451	0.171	106.3
18	−1.080	0.040	18.0		99	2.800	0.349	115.0
19	−1.040	0.040	19.0		100	3.514	0.714	132.9

$SI: \delta = 0.200;\ \ \Delta = 0.040; x_{SI}^{(L)} = E\left[x_{SI}^{(L)}|\beta = 0\right] + \beta_x/\Delta = 45 + 25\beta_x$

..., 84, 85, 86. This relationship can be shown algebraically. Thus let the measurement on the SI be notated $x_{SI}^{(L)}$ where the superscript (L) indicates that $x_{SI}^{(L)}$ is linear throughout, and not an expected value or an observed measurement which is constrained between 0 and 100. Then relative to $E[X_{SI}|\beta = 0]$,

$$\beta_x = \left\{x_{SI}^{(L)} - E[X_{SI}|\beta = 0]\right\}\Delta, \tag{8.6}$$

showing that values of β_x increase by the value of the unit Δ for each integer increase in the observed score x. In addition, relative to the origin, this relationship between an observed integer count x and the value of β is directly analogous to a measurement of an object, relative to its origin, in the unit of the instrument.

Rearranging Eq. (8.6), gives the general relationship

$$x_{SI}^{(L)} = E\left[x_{SI}^{(L)}|\beta = 0\right] + \beta/\Delta. \tag{8.7}$$

The last column of Table 8.1 shows values of $x_{SI}^{(L)}$. Note Eq. (8.7) is not estimated as a regression equation, but is expressed analytically as a relationship between the measurement $x_{SI}^{(L)}$ given the value of β. Table 8.1 shows that in the range in which the

relationship is linear between x, $E[X_{SI}]$ and β, their values are identical to three decimal places. These values are also highlighted in bold. Outside this range, for the same value of β, $x_{SI}^{(L)}$ is different from x, $E[X_{SI}]$. Table 8.1 also shows that only at the extremes of the range of the instrument, 0, 1 and 99, 100, the values of $x_{SI}^{(L)}$ show very large extrapolations. This in part is a result of the choice of the unit and origin. How this range is determined is described in the next sub-section. However, it is because of the properties of the distribution, in particular that of Eqs. (8.6) and (8.7), the values of $x_{SI}^{(L)}$ have been, and continue to be, referred to as *measurements*.

8.2.4 The Rasch Distribution of Measurement Uncertainty

As indicated above, the Rasch model distribution of Eq. (8.1) is the *inferred distribution* of replications of responses of the same person to the same instrument. This is simply a property of the probabilistic model. However, this inference holds for data only if the responses also fit the model. If they *do* fit the model, and the thresholds are ordered, then this distribution is of random uncertainty, with no evidence that any unaccounted-for factors are disturbing the responses [10]. This same inference holds for the Rasch distribution of Eq. (8.5) which has only the two parameters, the origin and the unit of the instrument specified. Of course, in this distribution of the SI, the thresholds are defined to be ordered. Because this distribution is directly analogous to the Gauss distribution of measurement uncertainty, now taken for granted in the natural sciences, we briefly review the motivation and role of the Gauss distribution in measurement.

Although the Gauss distribution is so mainstreamed that it is referred to generally as the *normal distribution*, the motivation and lengthy evolution of this distribution is generally not presented in textbooks. Besides Gauss, the derivation of the distribution exercised the best mathematicians in the late 18th and early 19th centuries, including De Moivre, Lagrange, La Place, and others [17, 41]. It was derived to account for the consistent evidence that . . .*repeated measurement of a fixed quantity by the same procedure under constant conditions*. . . did not give the same values but a distribution of values. The derivations culminated . . . *in the quadratic exponential law of Gauss* [17, p. 1]. This distribution satisfied the requirement that it characterized variation that was random, it having been realized that rather than propagating errors, random variation cancelled them. Thus the distribution is a theoretical distribution of random variation of replicated measurements, not a distribution derived to describe any particular data set. However, to the degree that any data set does conform to the Gauss distribution, to that degree it provides evidence that variation is no more than random, and therefore that relevant inferences can be drawn from the data, for example, that the mean is an ideal characterization of the object of measurement. Measurement of uncertainty continues to be a concern of natural scientists [23].

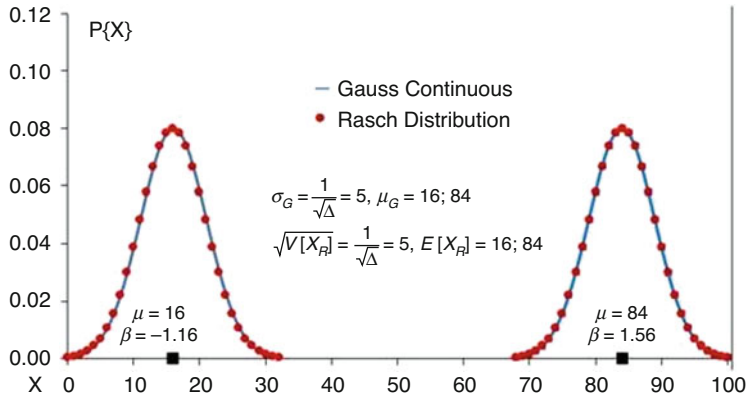

Fig. 8.2 Two Rasch distributions of the Standard Instrument where the probabilities of extreme measurements vanish, interpolated by the Gauss distribution in which $E[X_R] = \mu_G$, $V[X_R] = \sigma_G^2 = 1/\Delta$

Figure 8.2 shows the Rasch distribution for the SI of Fig. 8.1 in which $(\delta, \Delta) = (0.200, 0.040)$. The figure also shows the mean, $E[X|\beta]$ for the two measurements $(\beta_l, \beta_u) = (-1.160, 1.560)$. The Rasch distribution is clearly discrete with respect to the possible integer measurements. The discrete probabilities in Fig. 8.2 are interpolated with a continuous distribution. Perhaps unexpectedly, this is the continuous Gauss distribution. The possibility of this interpolation is no coincidence. For completeness, Eq. (8.8) shows the now common form of the Gauss distribution,

$$P\{X = x|\mu, \sigma^2\} = \left[\exp\ -(x-\mu)^2/2\sigma^2\right]/\sqrt{2\pi}\sigma, \tag{8.8}$$

where $(\mu,\ \sigma^2)$ are the mean and variance of the distribution.

In elaborating this relationship, we first note an observation made by Gauss regarding the limits of the applicability of the distribution of Eq. (8.8):

> Gauss commented that (1) (*the distribution*) cannot represent a law of error in full rigor because it assigns probabilities greater than zero to errors outside the range of possible errors, which in practice always has finite limits; that such a feature is unavoidable because one can never assign limits of error with absolute rigor; but this shortcoming is of no importance in the case of (1), because it "decreases so rapidly, when $[(x - \mu)^2/2\sigma^2]$ has acquired a considerable magnitude, that it can safely be considered as vanishing." [18, p. 2].

The values $(\beta_l, \beta_u) = (-1.160, 1.560)$ were chosen because they are just inside the limits of the range for the SI where the probabilities of the extreme measurements, 0 and 100, *vanish*. Thus they are the limits within which the Gauss distribution would be applicable. Four aspects of the relationship between the Rasch and Gauss distributions are relevant to note here. First, as evident from the first term in Eq. (8.2), just as is the Gauss distribution, the Rasch distribution is a *quadratic exponential*. Second, within the range in which the Gauss distribution is applicable,

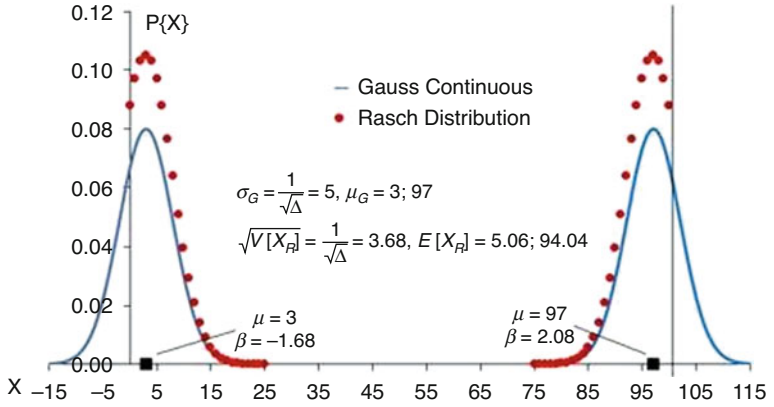

Fig. 8.3 Two Rasch distributions of the Standard Instrument where the probabilities of extreme measurements do not vanish and in which $E[X_R] \neq \mu_G$, $V[X_R] \neq \sigma_G^2 \neq 1/\Delta$

the mean of the Rasch distribution is identical to the mean of the Gauss distribution, $E[X_R] = \mu_G$, and in particular, the distribution is *symmetrical* about the mean. Third, and perhaps most surprisingly, the variance of the Gauss distribution is not only the variance of the Rasch distribution, $V[X_R] = \sigma_G^2$, but it is also the *inverse* of the unit in the Rasch distribution, $V[X_R] = \sigma_G^2 = 1/\Delta$. Finally, and importantly, it is within the range in which the Gauss distribution holds, that the relationship between the measurements β, the observed scores x, and expected values, are linear as shown in Fig. 8.1, Table 8.1 and Eq. (8.6).

For completeness, and because it is relevant to account for the finite range of the instrument, Fig. 8.3 shows the Rasch and Gauss distributions of two locations, $(\beta_l, \beta_u) = (-1.680, 2.080)$, where the limits of the instrument play a role. The figure shows the Gauss distributions as if the range of the instrument did not play a role, and the Rasch distributions which take account of the range. It is evident that the Gauss distributions are outside the range of the instrument, a feature which concerned Gauss as indicated in the above quote. The Rasch distributions of course are within the limits of the range. However, reflecting the impact of the limited range of the instrument, the mean and variance ($E[X_R]$, $V[X_R]$) of the Rasch distributions are regressed relative to their values if the range were not constrained. The property of the estimates $\widehat{\beta}_x$ is that they *undo* this regression to a large degree.

Given the properties described above, this chapter presents a justification for transforming the estimates of proficiencies β from each instrument to a measurement on the SI of the form of the Rasch distribution with the origin and unit chosen for convenience in the frame of reference. Specifically, from the estimates of the person locations β_x for each score x of an instrument, measurements in the chosen unit of the SI can be obtained from the linear extrapolation of Eq. (8.7). This linear extrapolation is shown in Fig. 8.1 and Table 8.1. It is evident that from this extrapolation, the measurements close to the limits of the range of the instrument are outside this range. However, these measurements are linear extrapolations that indicate the values that

would have been obtained had the limits of the instrument not regressed the observed scores. It is these linearized measurements to which the observed scores on all instruments are mapped in the example shown in the next section.

It would be more desirable if assessments were such that few if any person locations were affected by the limits of the range of the scores (0 and 100) with these instruments, and it is generally achieved in the frame of reference of the example described next. However, where they do occur, especially at the higher limit where competition is most relevant, then the limits of the range for scores close to the extreme need to be taken into account. It is an example where the Rasch distribution, which is discrete and is a function of both the unit and maximum score on each instrument, can be applied to advantage.

8.2.5 *Maximum Likelihood Estimates of the Instrument Parameters in the Rasch Model*

Estimates of instrument parameters ψ_{xi} of Eq. (8.2) can be derived from Eq. (8.9) below, where $(x_{ni}, x_{nj}) \mid r_n$ are the responses of person n to two instruments (i, j), and where $r_n = x_{ni} + x_{nj}$ is the sufficient statistic for β_n:

$$\Pr\{(x_{ni}, x_{nj})|r_n; \beta_n, \psi_{xi}, \psi_{xj}\} = [\exp(\psi_{xi}, \psi_{xj})]/\gamma_{ni}. \tag{8.9}$$

Because of sufficiency, Eq. (8.9) is *independent* of β. To estimate parameters $(\delta_i, \Delta_i, \lambda_i, \zeta_i)$ of multiple instruments, Eq. (8.9) is generalized over multiple instruments I. The algorithm, described in detail in Andrich and Luo [7], is implemented in the software RUMM2030Plus [12] and is used in the analysis of the example of this chapter. The only constraint required in the estimation is that $\sum_{i=1}^{I} \widehat{\delta}_i = 0$. Then δ_i is the *relative* origin of each instrument in an analysis. This origin, as shown with the SI above and illustrated with the example below, can be defined independently.

Because the coefficient of each principal component is a function of frequencies in all categories, rather than of each category, the estimation is not impeded by the presence of zero frequencies, except in very extreme cases, nor is it impeded by the structurally missing data.

8.2.6 *Profile Analysis and Editing of Profiles*

One of the complexities of the frame of reference listed above is that the latent correlation among the instruments is not 1, which implies lack of unidimensionality. The effect on the parameter estimates of the instruments when analyzed with the Rasch model is that they are all regressed to their mean [29, 42].

To account for this feature of the data from the perspective of the frame of reference, a complementary focus to unidimensionality, one which focuses on the profiles of persons, is taken [42, 45]. Unidimensional instruments, those with a latent correlation of 1, have two implications for profiles with respect to the Rasch model: first, that each person's profile is relatively homogeneous; second, that the total score is the sufficient statistic for the person parameter, with no further information in the profile. For example, a relatively high score on one instrument implies a relative high score on other instruments, where the variation among the scores of the profile on the instruments is no more than random. In contrast, and at the other extreme, if the latent correlation were 0, there is no such relationship.

Therefore, not only from the perspective of the application of the Rasch model, but also from the requirements of the frame of reference, the profiles that need to be used to obtain the equating functions between instruments are those that are relatively homogeneous. This ensures that the properties of the instruments as reflected by their parameters, for example their relative difficulties are a property of the instruments and not of the persons who happen to have scores on the instruments. In more general terms, the profiles that need to be used are from those persons who are relatively equally proficient on all instruments.

A comparison and contrast might be made with equipercentile scaling which can be applied when instruments do not have a latent correlation of 1, but are administered to the same sample of persons [24]. Here the assumption is that, although the individual persons are not expected to have homogeneous profiles, the latent distributions of the proficiencies of the *sample as a whole* are the same for the different instruments, and any differences in scores is a property of the instruments. Therefore, with this assumption, scores with the same cumulative percentage on each instrument are deemed equivalent. This method has its own problems for equating, including with zero frequencies and extreme scores, and in addition, impeding its application in the example of this chapter, the students with scores on different pairs of instruments are not common. However, when applicable, the assumption is the equivalence of the distributions on the different instruments. It is relevant to compare the assumptions made between the equipercentile and Rasch model applications to equating. In applying the former, the assumption is that the sample has the same proficiency distribution on the different instruments; in applying the latter, it is ensured that the persons whose profiles are used have equivalent proficiencies on the different instruments.

The method of obtaining the subset of profiles whose scores are homogeneous requires two successive analyses of the data. The first is simply the standard analysis of all data. Then, given the estimates of the instruments' parameters, and each person's estimate of β, the expected value, $E[X_{ni}]$, is calculated for each instrument using Eq. (8.4). A comparison is then made between the observed score x_{ni} and $E[X_{ni}]$ for each person n on each instrument i for which they have a score. This comparison is made in terms of the standardized residual,

$$z_{ni} = (x_{ni} - E[X_{ni}])/\sqrt{V[X_{ni}]}. \tag{8.10}$$

Then if the absolute value of the residual is greater than some chosen magnitude, that score in the profile is deleted, creating further missing responses. Because there are already structurally missing responses, further missing responses are no impediment to the estimation. However, it is necessary to choose the magnitude of the residual criterion judiciously. If it is too large, there will be a substantial number of heterogeneous profiles in the analysis; and if it is too small, then relative to the variation of the model, there will be insufficient variation creating a form of local dependence. This method of editing profiles is analogous to that used by Andrich, Marais and Humphry [11] and Andrich and Marais [8] in editing responses to control the bias on item parameter estimates from guessing for multiple choice items. Specifically, given each person's and each item's parameter estimates, if there is a greater probability than random that the person guessed or at least partially guessed a response on that item, whether the response is correct or not, the response is converted to missing data. This editing of responses removes the bias in the item parameter estimates due to guessing in the data.

The criterion for deleting a response chosen in the analysis of the example in this chapter is $|z_{ni}| > 0.85$. Evidence which shows this choice is reasonable is that the latent correlation between each pair of instruments, when corrected for error, is of the order of 1. This implies that, for the purpose of obtaining equating functions, the complementary properties of unidimensionality and sufficiency of the total score of the Rasch model are satisfied.

The classical definition of reliability, when applied to the Rasch model estimates, is given by

$$r_{\beta i} = \left(V\left[\widehat{\beta}_i\right] - [V[\widehat{\varepsilon}_i])/V\left[\widehat{\beta}_i\right], \tag{8.11}$$

where $V[\beta_i]$ is the estimate of the variance of the persons on instrument i, and $V[\varepsilon_i]$ is the estimate of its error variance [21]. The latent correlation between two instruments, corrected for attenuation because of error, is given by

$$\rho_{ij} = r_{ij}/\sqrt{r_{\beta i} r_{\beta j}}. \tag{8.12}$$

In the application of Eqs. (8.11) and (8.12) in the Rasch model, $\widehat{\beta}_{xi}$ is the estimate of proficiency of the person given the score x and $V[\widehat{\varepsilon}_i]$ is the mean error variance of the estimates $\widehat{\beta}_{xi}$ from the persons who have scores on both instruments. These estimates are elaborated next.

8.2.7 Person Estimates

Given estimates $(\widehat{\psi}_{xi})$ of instrument i, the maximum likelihood estimate $\widehat{\beta}_n$ for each person n from all instruments the person has responded to, is given *individually* by the solution to implicit Eq. (8.3). However, in application to the frame of reference of this chapter, it is necessary to have an estimate, $\widehat{\beta}_{ni}$, of each person on each instrument. This estimate is given by the solution to Eq. (8.4). For this estimate of each person's proficiency on each instrument, the parameters $(\widehat{\psi}_{xi})$, following the editing of the profiles in the terms described above, are used. These are the estimates based on homogeneous profiles which, because of sufficiency of the total score for these profiles, are independent of the actual distribution of the person parameters.

However, for the person estimates based on these instrument parameters, and for evidence of sufficient proficiency for selection into university studies, every score of each person, that is the full profile before editing, must be used. Finally, each estimate of each person from each instrument is transformed to a measurement on the SI in the form of the Rasch distribution described above.

Because the original profiles are used for the final estimates, the latent correlations between instruments will not be in the range 0.90–1.00. That means that there will be profiles which, when transformed to the SI, will not be homogeneous and the sum of the estimates, or their mean, will not characterize the profile fully. How the distinction between those profiles that are homogeneous, and those that are not, is dealt with in its frame of reference is described in the context of the example. The next section provides the results of analysis of data from the example.

8.3 An Illustrative Example

The data for university selection from 2018 were provided by the School Curriculum and Standards Authority of Western Australia. As indicated above, the example comes from a series of instruments used to assess the proficiency of students in a range of disciplines for university selection in Western Australia. The total number of disciplines is as large as 40. For the purpose of illustration in the example of this chapter, scores from examinations of the following six disciplines were analyzed: English (ENG), English Literature (LIT), Mathematics 1 (MA1), Mathematics 2 (MA2), Modern History (HIM), and Chemistry (CHE) which was introduced in Fig. 8.1. These disciplines were chosen because relative properties of these instruments can be anticipated, and they illustrate some complexities that are overcome.

First, one of the disciplines of English must be taken to be eligible for university entry, and either ENG and LIT are acceptable. Therefore, very few students have scores in both disciplines. On the one hand, because it is a specialized unit, students studying LIT may be expected to have greater proficiency in English, and therefore a higher mean proficiency on the SI than those studying ENG.

Second, MA1 and MA2 have a partly different relationship from that between ENG and LIT. MA2 has more challenging material than MA1. However, to study MA2 it is necessary to either study MA1 simultaneously or otherwise know its content. Therefore, it is expected that MA2 will be shown to be more difficult than MA1, and that the mean proficiency of students studying MA2 (and MA1) will be greater on the SI than that of those studying only MA1. Finally, the disciplines HIM and CHE are chosen because one is a humanities and the other a science discipline, and they are expected to show properties that are commensurate with ENG and MA1 respectively.

For purposes of efficiency of exposition, the results are not presented in the order in which they were obtained. Following the summary of the raw data, the results of the estimates of the proficiencies and their relationships among the disciplines are shown first, followed by the estimates of the instruments' parameters, the equating functions, and finally graphical presentations of the equating functions and distributions.

8.3.1 The Raw Scores on the Instruments

An excerpt of the data file used for analysis is presented in Table 8.2, illustrating the structurally missing responses in the data file and integer scores recorded for the six disciplines.

Table 8.3 shows descriptive data in the form of the number of students, the means, standard deviations and the skewness of the distribution, and the observed pairwise frequencies and correlations. First, it is evident that ENG has the greatest number of students, which is expected because an English discipline assessment is a

Table 8.2 Excerpt of the first 15 cases from the data file for analysis

ID	CHE	ENG	HIM	LIT	MA1	MA2
S00001		45				
S00002	56	56			64	
S00003	39				50	
S00004		10				
S00005	39				50	
S00006	56	56			64	
S00007	59	61	67		43	
S00008	52				79	72
S00009			46			
S00010	58	50			35	
S00011					64	57
S00012					64	
S00013	59	61	67		43	
S00014		62			40	
S00015		80				

Table 8.3 Pairwise frequencies (F), observed correlations r_{ij} between instruments, and descriptive statistics of observed scores for each instrument from the total sample of 13617

F/r_{ij}	ENG	LIT	MA1	MA2	HIM	CHE
ENG		*	2800	937	1589	3480
LIT	*		661	267	381	698
MA1	0.289	0.247		1530	237	3255
MA2	0.409	0.348	0.867		39	1221
HIM	0.581	0.635	0.434	0.566		266
CHE	0.417	0.433	0.764	0.788	0.578	
N	10,974	1463	4426	1594	2014	4973
Mean	57.96	70.89	65.14	61.24	60.06	58.50
St dev	11.38	9.96	18.65	19.43	13.18	17.31
Skew	−0.37	−0.89	−0.65	−0.50	−0.89	−0.44

Note. *Frequency of less than 20 not shown. Number of common persons above the diagonal; observed correlations below the diagonal. Correlation between ENG and LIT not shown because of the small number of common students

requirement for university entry. LIT and MA2 have the least number, reflecting their specialist status. The table shows no common students between ENG and LIT. There were seven common students but the reason for students taking both disciplines is idiosyncratic to different circumstances and the correlation between the two disciplines was −0.188. Therefore, this frequency and the correlation are not shown in Table 8.3.

Second, all but one mean is in the range between 55 and 65, showing that the difficulties of the instruments were well aligned to the expected proficiencies of the students, except for LIT which has a mean above 70. Proficient students choose LIT, but this mean seems relatively high. There is a greater relative range in the standard deviations, where the two mathematics and the science instruments show greater standard deviations than the humanities instruments and all distributions are skewed negatively. The table also shows that the number of students assessed by each instrument varies, as does the number who are assessed by any pair of instruments. Because of the different samples, the properties of the distributions such as their means cannot be compared directly. Important from the perspective of measurement on a single variable, is that the observed initial correlations among the instruments has a large variation, ranging from 0.247 to 0.867.

The data in Table 8.3 are shown graphically in Fig. 8.4. The differences in the distributions, including their negative skewness is clear. Some of the low scores might be taken as resulting from not taking the examination seriously, but these scores are included in the analysis for completeness and illustration. Not only does the distribution of LIT show a high mean, it is also very narrow. It is also clear that the distributions of CHE, MA1 and MA2 are somewhat similar as are those of ENG and HIM. The former three instruments assess mathematics and science disciplines, the latter two assess humanities. Although the distributions cannot be compared directly, it can be inferred how well the instruments align themselves to the proficiencies of the samples. This alignment is important in distinguishing validly

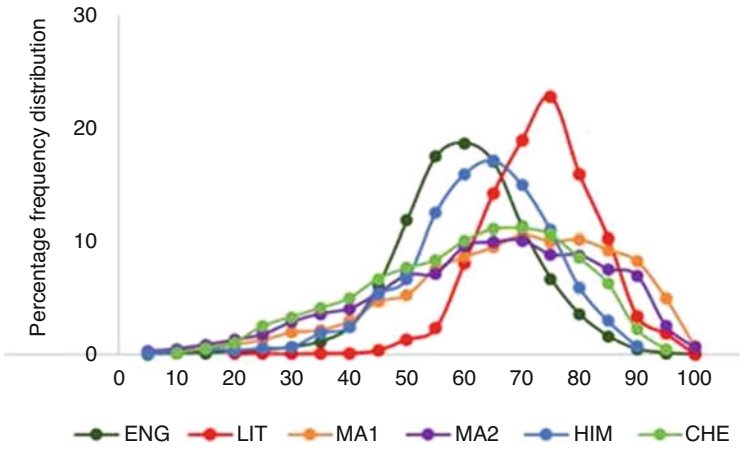

Fig. 8.4 Percentage frequency distributions of observed scores in class intervals of 5 score points

between candidates. In addition to LIT, and even though they spread the students well, the three science instruments were relatively lenient, while the two humanities instruments were better aligned. ENG, which assesses an effectively compulsory discipline in which the sample is less self-selected than the other disciplines, has noticeably the smallest value for its mode.

8.3.2 The Equating Functions

The section begins with a graphical presentation of the equating functions and the principal component estimates from which the equating functions are derived. The equating functions are obtained from an analysis in which the profiles were edited as described above. The section finishes with the latent correlations between instruments from these equating functions and the equated scores of all instruments on the SI.

Figure 8.5 shows the equating functions based on EVCs of the kind shown for CHE in Fig. 8.1. In addition, each curve has its observed means in 10 class intervals shown. It is evident that the means are very much on the curves, indicating fit to the model for the data from these instruments. An approximate Chi Statistic value which compares these observed means and their expected values across all instruments is 32.237 on 54 degrees of freedom, confirming excellent statistical fit. It is clear from Fig. 8.5 that the curves are non-linear and intersect, and require equating before comparisons can be made. This evidence that the edited data fit the model is considered sufficient for this example.

Table 8.4 shows the principal component estimates and their standard errors. Because, with a maximum score as large as 100, the values of the parameters are small in magnitude, they are shown to five decimal places. As expected, the relative

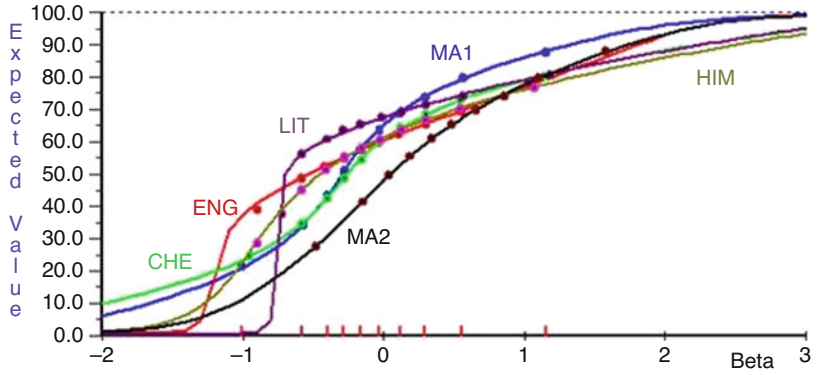

Fig. 8.5 EVCs from edited profiles with observed means in 10 class intervals

Table 8.4 Estimates of the four principal components $(\delta, \Delta, \lambda, \zeta)$ of the thresholds from a profile analysis with residuals $|z_{ni}| > 0.85$ removed

	δ(origin)	$SE(\delta)$	Δ(unit)	$SE(\Delta)$	λ(skew)	$SE(\lambda)$	ζ(kurt)	$SE(\zeta)$
ENG	−0.08586	0.00220	0.03747	0.00020	0.00055	0.00000	0.00000	0.00000
LIT	0.08213	0.00630	0.03257	0.00040	0.00110	0.00000	0.00000	0.00000
MA1	−0.23313	0.00290	0.03530	0.00020	0.00008	0.00000	0.00000	0.00000
MA2	0.19814	0.00510	0.03385	0.00020	0.00019	0.00000	0.00000	0.00000
HIM	0.10197	0.00510	0.04553	0.00040	0.00069	0.00000	0.00000	0.00000
CHE	−0.06325	0.00280	0.05025	0.00020	0.00025	0.00000	0.00000	0.00000
Mean	0.00000	0.00407	0.03916	0.00027	0.00048	0.00000	0.00000	0.00000
St Dev	0.14267	0.00150	0.00649	0.00009	0.00035	0.00000	0.00000	0.00000

difficulty of MA2, the specialist mathematics discipline, is more difficult ($\delta = 0.19814$) than MA1, the one that has the content as its prerequisite ($\delta = -0.23313$). The difficulty of LIT, the specialist English discipline, ($\delta = 0.08213$) is likewise more difficult than ENG, which is studied by most students as a required discipline, ($\delta = -0.08586$).

8.3.3 *The Equated Scores to Measurements on a Standard Instrument*

For each observed score of each person on each instrument, the proficiency estimate $\widehat{\beta}_{ni}$ is obtained from Eq. (8.4) and CHE is illustrated in Fig. 8.1. Then each estimate $\widehat{\beta}_{ni}$ is transformed further to a measurement on the SI according to Eq. (8.7) in which $E[X_{SI}|\beta_{SI} = 0] = 45$ and $\Delta = 0.04$..

For completeness, Table 8.5 shows the equivalent measurements on the SI for a series of scores on each of the instruments. Figure 8.1 illustrates this equivalence for CHE. Table 8.5 shows that a score of 50 on ENG, for example, is a measurement of

Table 8.5 Equivalent measurements on the SI for the same score on each of the instruments

Score	ENG SI	LIT SI	MA1 SI	MA2 SI	HIM SI	CHE SI
20	15.3	26.0	18.8	27.6	18.7	15.3
30	16.8	26.4	28.0	34.4	22.9	26.9
40	21.7	26.8	33.6	40.3	27.5	33.5
50	31.8	27.6	37.6	46.2	33.8	38.5
60	44.7	34.2	42.0	52.9	43.4	44.5
70	59.4	49.6	48.6	61.5	57.9	54.6
80	74.8	71.9	59.8	73.0	78.8	72.3
90	90.1	101.4	78.0	88.5	107.6	101.1
95	98.9	119.5	91.0	99.1	125.8	121.0

Table 8.6 Pairwise latent correlations and descriptive statistics for each instrument from the total sample of 13617, anchored to principal components from the analysis of edited profiles at 0.85 and transformed according to $\beta_{SI} = \beta_i/0.04 + 45$

ρ_{ij}/r_{ij}	ENG	LIT	MA1	MA2	HIM	CHE
ENG		*	0.9723	1.0160	0.9680	0.9907
LIT	*		0.9836	0.9989	0.9484	1.0092
MA1	0.2987	0.2632		1.0592	1.0389	1.0192
MA2	0.4418	0.4096	0.9845		0.9187	1.0348
HIM	0.6759	0.7483	0.4881	0.6510		1.0338
CHE	0.4640	0.5167	0.8496	0.8698	0.7344	
N	10974	1463	4426	1594	2014	4973
Mean	43.25	54.68	49.61	57.05	47.17	47.86
St Dev	14.26	18.24	16.41	17.19	15.58	17.33
Skew	0.48	0.83	0.22	0.17	0.62	0.49

Note. *Frequency of less than 20 not shown. Mean of the pairwise latent correlations for homogeneous profiles: $\overline{\rho}_{ij} = 0.9994$. Some observed correlations are slightly greater or less than 1 due to random variation around 1. Mean of the pairwise latent correlations for measurements of all data: $\overline{r}_{ij}$ =0.5997

31.8 on the SI, while the same score on MA2 is a measurement of 46.2. For scores up to 70, the greatest measurement on the SI is for MA2, the advanced mathematics discipline. For scores greater than 70, HIM (modern history) has greater measurements on the SI. This order results from it being more difficult to obtain a very high score in HIM than in MA2 for the respectively very proficient students. It is not uncommon for it to be difficult to obtain very high scores in humanities disciplines, while it is much more common to obtain very high scores in the mathematics and science disciplines. The EVCs of Fig. 8.5 reflect the relative difficulty at the higher end of the proficiency continuum.

Table 8.6 shows two sets of latent correlations and the distribution properties of measurements on the SI for scores on each instrument. Above the diagonal, it shows the latent correlations corrected for errors of measurement for the analysis of the edited profiles, which results in homogeneous scores. It will be recalled that this

analysis ensures that the equating functions are obtained from profiles that are effectively unidimensional. The average latent pairwise correlation of 0.9994 ensures that the profiles are homogeneous. Of the 5708 profiles that had at least two measurements after the profiles were edited, 24% had a standard deviation less than 5. Given the unit of the SI, $\Delta = 0.04$, which implies a variance of $\sigma^2 = 1/\Delta = 25$ for replicated measurements in the linear range between 15 and 86, where the majority of measurements are, it would be expected from a Gauss distribution that some 32% would have a standard deviation less than 5. Thus if anything, the profiles are slightly more homogeneous than under total randomness.

Below the diagonal, Table 8.6 shows the latent correlations between the instruments for measurements of all profiles of all persons. Clearly, with all profiles measured, the latent correlations are not homogeneous, and of course not close to 1. As expected, with a mean of 0.5997, they mirror the observed correlations between the raw scores shown in Table 8.3. Such correlations are reflected in non-homogeneous profiles. Of the 7334 profiles with two or more scores, 55.83% have a standard deviation greater than 5 which indicates that they deviate from their respective means by approximately five score points. For example, a profile with two scores and a standard deviation of 5, has measurements of 48.90 and 58.90. 22.08% have a standard deviation greater than 10, which indicates scores that essentially deviate 10 points in either direction from their respective means. For example, a profile with two measurements and a standard deviation of 10.00 has measurements of 69.30 and 89.30.

The profiles with a standard deviation greater than 10 are not characterized by their total scores. This seems very relevant in the system of selection, which is based primarily on the total score. Depending on which course of study the student is planning, those profiles with a marginal case for selection on the basis of their total score, would need to be considered individually. For example a relatively low score in the required discipline of English and relatively high scores in the mathematics and science disciplines, may not preclude a student marginal for selection on the basis of the total score, being selected for an engineering course. It is stressed that in formal university entry, at least four measurements, including one of ENG or LIT, are required to meet the eligibility requirements for university entry. Then the application is based on the mean of the four highest scaled measurements. In the illustrative example of this chapter, where the maximum number of measurements is only six, very few people have the minimum four and therefore the means and rankings from these illustrative data are not useful to study.

Recognizing that many profiles are not characterized by their total scores, Table 8.6 shows the proficiencies of all students on all instruments. The measurements on MA2 have the greatest mean (57.05), followed by LIT (54.68). It will be recalled that these are more advanced specialist disciplines in mathematics and English respectively, and therefore it is expected that their relative means will be the greatest. That they are, also confirms the success of the equating. It is noticeable, however, that the mean of LIT, which was 70 from the raw scores and the largest, is no longer the largest. ENG, the discipline taken by many students because English proficiency is required for university entry, has the smallest mean among this group of students.

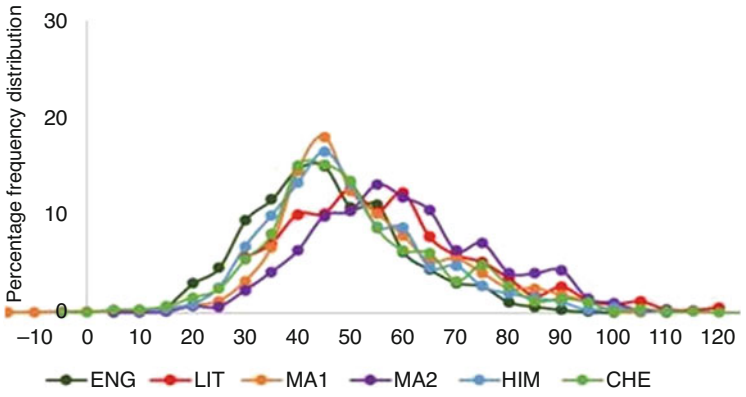

Fig. 8.6 Distributions of equated scores to a standard instrument in class intervals of 5 score points

Figure 8.6 shows the distributions of these measurements in class intervals of five score points. It reflects the distributional statistics in Table 8.6, with the standard deviations relatively homogenous. On the other hand, unlike the original scores, which show a negative skew, the measurements on the SI all show a positive skew. It is noticeable that only the distributions of the two mathematics disciplines, MA1 and MA2, have a skew value clearly less than 0.5, suggesting they are the only ones with clearly normal distributions, while the others, and in particular LIT, show deviation from normality. A final point to observe is that the distributions of the two specialist disciplines, LIT and MA2, show similar features, while the distributions of the other disciplines are also similar.

8.4 Summary and Discussion

This chapter began by observing two contrasting but interrelated aspects of the idea of measurement. First, that the historical motivation for physical measurement, the standardization of a unit and origin within a frame of reference, was fairness of transactions of goods for trade in everyday applications. The aim of such measurement was to ensure that comparisons among objects, for the relevant property, were invariant with respect to which specific instrument was employed. Second, that although in its elementary form, measurement is understood by school children, the mapping of the magnitude of a property of an object onto a number line partitioned into equal units, is a deep abstraction, and that the advanced application of measurement has evolved in conjunction with the remarkable advancement of the quantitative, natural sciences. Although many scientific measurements remain in the realm of scientific theory, many have also become mainstreamed and are applied in everyday applications. Measurement of temperature is an example, where although integral to the theory of thermodynamics, thermometers for measuring human temperatures in the case of potential illness have become indispensable.

The success of quantitative laws in explaining phenomena in the natural sciences provides an aspiration for quantitative laws in the social sciences. The coevolution of measurement and quantitative laws of the natural sciences implies that any quantitative laws of the social sciences will evolve in conjunction with social measurement, where the concept of measurement transcends the natural and social sciences. However, there are frames of reference in the social sciences where the historical motivation for measurement in the natural sciences prevails – that of fairness of transactions. By analogy to the use of rigorous measurement of temperature in routine applications, measurement in the social sciences may require measurement which is just as rigorous as is the measurement of variables that will produce scientific laws. This chapter is concerned with such an example.

The frame of reference of this chapter is the competitive selection of students into universities in Western Australia based on their assessments in a range of disciplines in their Year 12 studies. The assessments are not of the form of intelligence or aptitude tests, but based on explicit syllabuses that the students have been taught. The students may choose to pursue their university studies in a wide array of courses based on their Year 12 studies in a similarly wide array of disciplines. In general, satisfactory performance in the discipline of English is required for selection into all university courses. Many courses, such as law, psychology, and economics, have no prerequisite studies, while some courses such as engineering, natural sciences, and mathematics, may require studies in those disciplines, but permit electives which may vary among students. The selection is based primarily on the mean of four performances, from the disciplines they have studied, which give the greatest mean. The implication of these features is that, in the interest of fairness, the selection of students for various university courses needs to be *invariant* with respect to which disciplines they studied. This requires rigorous measurement, with the same unit and origin, of each discipline.

It is recognized that, because different students may have studied different disciplines, their summary scores do not reflect a degree of proficiency on the same substantive, content variable. Instead, the variable is an abstracted index variable of causal variables in which the proficiency of a student in a discipline governs their performance on the relevant assessment instrument. The variable is inferred to be one of a capacity to succeed and benefit from university studies based on previous studies. The example is described in some detail with the expectation that it has properties that can be transferred to other social sciences cases where rigorous measurement is required.

In part because of the advancement of the natural sciences with the coevolution of quantitative laws and measurement, social scientists have studied and attempted to define measurement generally from the perspective of advancing quantitative laws in the social sciences. One of these is the work of Rasch, termed in this chapter, *Rasch measurement theory*. The principle on which Rasch's theory is based is the requirement of invariant comparisons of objects with respect to instruments (and vice versa) within a specified frame of reference. The consequences of this definition have been shown to lead to quantitative relationships which are entirely compatible with other definitions of measurement and with the laws of the natural sciences.

Rasch's formulation has at least three distinctive elements compared to other definitions of measurement. First, it is relevant in both probabilistic and deterministic frameworks. The probabilistic contexts immediately provide evidence of statistical variation that may or may not be random. Second, because Rasch formulated his requirement of invariant comparisons mathematically, rather than merely descriptively, further derivations of the model for measurement are possible. Third, the requirement of invariant comparisons is relevant for quantitative laws in general, and not only measurement; therefore, it seems a more fundamental basis for understanding measurement than simply describing measurement. All three features of Rasch measurement theory are applied explicitly in the example of this chapter.

Rasch's probabilistic formulation is used in this chapter. It specifies the probability that a person will obtain a score on the instrument as a function of the person's hypothesized, scalar, proficiency and the threshold parameters of the instrument which reflect its relative difficulty (origin) and their tendency to spread and skew the responses. The thresholds are points of equal probability of two adjacent scores.

Then, given estimates of its parameters, the estimate of every person's proficiency on each instrument is transformed to an expected value on a SI (standard instrument) that is termed a *measurement*. The procedure was introduced and summarized in Fig. 8.1. The distribution of the SI has been derived from Rasch's original formulation and is unlikely to have been formulated in any other way. The origin and unit of the SI are as explicit as they are in measuring instruments in the natural sciences, and are chosen for convenience. In particular, over a substantial and defined range of the instrument, the difference between two successive measurements is the unit. In addition, in this range, the distribution of inferred replicated measurements of each person to each instrument is a discrete analogue of the continuous Gauss distribution in which the variance on the SI is the inverse of the unit. It was recalled that the Gauss distribution was derived to characterize random variation of replicated measurements of the same object with the same instrument. Finally, the region in which the relationship between the measurement on the SI and student proficiency is not linear is near scores of 0 and 100 where the limits of the instrument interfere with the random variation from replications. This is a region in which the Rasch distribution of the SI is applicable, but in which the Gauss distribution is not.

The origin and unit of the SI in the example were chosen to minimize the number of measurements in the region in which the proficiency and the expected value of the SI are not linear. The advantage of this choice is that the variance of the SI for any person reflects no more than random variation. Therefore, if the observed variance of the SI scores on a profile is greater than random variation, the total score does not summarize the profile. In this case, not only should the magnitude of the mean measurement of a profile be considered for selection, but the profile should also be studied for evidence of specific capabilities that might be relevant for the choice of further studies.

The motivation for transforming the scores of each instrument to a measurement on the SI is that of invariance of comparisons – that comparisons between students for competitive selection is invariant with respect to the subset of disciplines that they have studied from a wider set of relevant disciplines. This motivation, rather

than that of advancing quantitative laws in the social sciences, is identical to the original motivation of much of measurement of physical variables. Importantly, although the motivation for invariant comparisons in this case is clearly that of fairness of selection, it is the same motivation that led to Rasch's measurement theory of invariance and which is relevant for understanding and constructing measurements which can lead to quantitative laws in the social sciences.

It was also indicated that, because the frame of reference was complex, the example can be taken as illustrative and that the approach taken to equating can be generalized to other frames of reference. One of these is person-centred outcomes in health assessment. As is evident throughout the example, the concern is with personal profiles and selection of individuals; therefore the example can be considered to be person-centred in its concern, an approach exemplified in Cano, Pendrill, Melin, and Fisher [14]. In the case of the assessment of an individual on multiple instruments in the health outcomes area, the principles, first that a summary score characterizes a higher order variable which is primarily an index variable, and second, that there will be individuals whose profiles are summarized by the total score and others which are not, is readily applicable.

In summary, it is stressed that, analogous to the Gauss distribution, the Rasch distribution of the SI employed in this chapter was derived from theoretical considerations, and not to describe any particular data set. Just as the Gauss distribution sets up a criterion that the variance of real or inferred replications is no more than random, and therefore that the mean can be used as a summary measure for the replications, the Rasch distribution sets up the criterion that the distribution of inferred replications from an instrument is no more than random, that the total score is sufficient to characterize the profile, and that the mean can be used to summarize the profile. By implication, comparisons which are invariant with respect to the instruments across profiles can be made. In short, it is a criterion for measurement. In conjunction with stressing that the Rasch distribution is derived as a criterion for measurement and not to describe any data set, it is stressed that the observed scores from instruments can only be transformed successfully to measurements if the data themselves permit such a transformation. To ensure such a possibility, extensive substantive empirical and theoretical work and understanding is required.

Acknowledgments The School Curriculum and Standards Authority, Department of Education, Western Australia, gave permission for the use of the data analyzed in the example. The research was supported in part by grants from the Australian Research Council.

References

1. K. Alder, *The Measure of All Things: The Seven-Year Odyssey and Hidden Error That Transformed the World* (Free Press, 2002)
2. E.B. Andersen, Sufficient statistics and latent trait models. Psychometrika **42**, 69–81 (1977)
3. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**(4), 561–574 (1978)

4. D. Andrich, A structure of index and causal variables. Rasch Measur. Trans. **28**(3), 1475–1477 (2014)
5. D. Andrich, The problem with the step metaphor for polytomous models for ordinal assessments. Educ. Meas. Issues Pract. **34**(2), 8–14 (2015)
6. D. Andrich, Chapter 7: Perceived health and adaptation in chronic disease: Stakes and future challenge, in *Advances in Social Measurement: A Rasch Measurement Theory*, ed. by F. Guillemin, A. Leplège, S. Briançon, E. Spitz, J. Coste, (CRCS Press/Taylor and Francis, 2018), pp. 66–91
7. D. Andrich, G. Luo, Conditional estimation in the Rasch model for ordered response categories using principal components. J. Appl. Meas. **4**, 205–221 (2003)
8. D. Andrich, I. Marais, Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. Appl. Psychol. Meas. **38**(6), 432–449 (2014)
9. D. Andrich, P. Pedler, Modelling ordinal assessments: Fit is not sufficient. Commun. Stat. **48**(12), 2932–2947 (2019a)
10. D. Andrich, P. Pedler, A law of ordinal random error: The Rasch measurement model and random error distributions of ordinal assessments. Measurement **131**, 771–781 (2019b)
11. D. Andrich, I. Marais, S. Humphry, Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. J. Educ. Behav. Stat. **37**(9), 417–442 (2012)
12. D. Andrich, B.S. Sheridan, G. Luo, *RUMM2030Plus: Rasch Unidimensional Models for Measurement* (RUMM Laboratory, Perth, 2020)
13. N.R. Campbell, *Physics: The Elements* (Cambridge University Press, 1920)
14. S.J. Cano, L.R. Pendrill, J. Melin, W.P. Fisher Jr., Towards consensus measurement standards for patient-centered outcomes. Measurement **141**, 62–69 (2019)
15. M. De Podesta, Absolute zero, in *Nothing*, ed. by J. Web, (New Scientist, 2013), pp. 164–173
16. O.D. Duncan, Rasch measurement in survey research: Further examples and discussion, in *Surveying Subjective Phenomena*, ed. by C. F. Turner, E. Martin, vol. 2, (Russell Sage Foundation, 1984), pp. 367–403
17. C. Eisenhart, Law of error I: Development of the concept, in *Encyclopedia of Statistical Sciences*, ed. by S. Kotz, N. L. Johnson, vol. 4, (Wiley, 1983a), pp. 530–547
18. C. Eisenhart, Law of error II: Development of the concept, in *Encyclopedia of Statistical Sciences*, ed. by S. Kotz, N. L. Johnson, vol. 4, (Wiley, 1983b), pp. 547–562
19. L. Finkelstein, Well-defined measurement – an analysis of challenges. Measurement **42**(9), 1270–1277 (2009)
20. W.P. Fisher Jr., A.J. Stenner, Theory-based metrological traceability in education: A reading measurement network. Measurement **92**, 489–496 (2016)
21. H. Gulliksen, *Theory of Mental Scales* (Wiley, 1950)
22. S. Humphry, D. Andrich, Understanding the unit implicit in the Rasch model. J. Appl. Meas. **9**, 249–264 (2008)
23. Bureau Internationale des Poids et Mesures: Joint Committee for Guides in Metrology (JCGM/WG 1). (2008). Evaluation of measurement data--Guide to the expression of uncertainty in measurement. Sevres, France: International Bureau of Weights and Measures--BIPM. www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.
24. M.J. Kolen, R.L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd edn. (Springer, 2004)
25. D.H. Krantz, R.D. Luce, P. Suppes, A. Tversky, *Foundations of Measurement*, vol 1 (Academic, 1971)
26. T.S. Kuhn, The function of measurement in modern physical science. Isis **52**, 161–190 (1961)
27. R.D. Luce, J.W. Tukey, Simultaneous conjoint measurement: A new type of fundamental measurement. J. Math. Psychol. **1**, 1–27 (1964)
28. G. Luo, D. Andrich, Estimating parameters in the Rasch model in the presence of null categories. J. Appl. Meas. **6**(2), 128–146 (2005)

29. I. Marais, D. Andrich, Formalising dimension and response violations of local independence in the unidimensional Rasch model. J. Appl. Meas. **9**(3), 200–215 (2008)
30. L. Mari, Epistemology of measurement. Measurement **34**, 17–30 (2003)
31. G.N. Masters, A Rasch model for partial credit scoring. Psychometrika **47**, 149–174 (1982)
32. G.N. Masters, B.D. Wright, The partial credit model, in *Handbook of Item Response Theory*, ed. by W. J. van der Linden, R. K. Hambleton, (Springer, 1997), pp. 101–121
33. J. Michell, Measurement: A beginner's guide. J. Appl. Meas. **4**, 298–308 (2003)
34. R. Ostini, M.L. Nering, *Polytomous Item Response Models*, Sage University Paper Series on Quantitative Applications in the Social Sciences (07-144) (Sage Publications, 2006)
35. J.O. Ramsay, in *Review of Foundations of Measurement, Vol. I*, ed. by D. H. Krantz, R. D. Luce, P. Suppes, A. Tverskey, vol. 40, (Psychometrika, 1975), pp. 257–262
36. G. Rasch, *Probabilistic Models for some Intelligence and Attainment Tests* (Danish Institute for Educational Research, Copenhagen, 1960). Expanded edition (1980) with foreword and afterword by B. D. Wright. The University of Chicago Press. Reprinted (1993): MESA Press
37. G. Rasch, On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*, ed. by J. Neyman, (University of California Press, 1961), pp. 321–334
38. G. Rasch, On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. Dan. Yearb. Philos. **14**, 58–94 (1977)
39. A.J. Stenner, D. Burdick, M.H. Stone, Formative and reflective models: Can a Rasch analysis tell the difference? Rasch Measur. Trans. **22**, 1059–1060 (2008)
40. A.J. Stenner, W.P. Fisher, M.H. Stone, D.S. Burdick, Causal Rasch models. Front. Psychol. **4**, 536–557 (2013)
41. S.M. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900* (The Belknap Press of Harvard University Press, 1986)
42. D. Surla, Application of the Rasch Model of Modern Test Theory to Equate Multiple Tests Using Their Total Scores. Unpublished PhD dissertation, The Univesity of Western Australia (2020)
43. L. Tesio, Items and variables, thinner and thicker variables: Gradients, not dichotomies. Rasch Meas. Trans. **28**(3), 1477–1479 (2014)
44. L.L. Thurstone, *The Measurement of Values* (University of Chicago Press, 1959)
45. J. Tognolini, D. Andrich, Analysis of profiles of students applying for entrance to universities. Appl. Meas. Educ. **9**(4), 323–353 (1996)
46. M. Wilson, G.N. Masters, The partial credit model and null categories. Psychometrika **58**, 87–99 (1993)
47. B.D. Wright, A history of social science measurement. Educ. Meas. Issues Pract. **16**(4), 33–45 (1997)