Maintaining a common unit in social measurement

Stephen Humphry
Murdoch University, Western Australia

Mailing address
Stephen Humphry
Murdoch University
Murdoch 6150
Western Australia

Acknowledgements
The paper is based on the author's Ph.D. thesis completed at Murdoch University, Western Australia, with Prof David Andrich principal supervisor and Assoc Prof Guanzhong Luo co-supervisor. The author wishes to acknowledge the extensive input and assistance of David Andrich in shaping the final draft of the manuscript, as well as his guidance, input and vision throughout the process of undertaking the research. Permission by the Department of Education & Training, Western Australia is acknowledged for the use of the empirical data featured within the paper. Support within the Department has also greatly benefited the research. The work for the Report was supported in part by an Australian Research Council grant with the Australian National Ministerial Council on Employment, Education, Training and Youth Affairs (MCEETYA) Performance Measurement and Reporting Task Force, UNESCO's International Institute for Educational Planning (IEP), and the Australian Council for Educational Research (ACER) as Industry Partners*.

*Report No. 6 ARC Linkage Grant LP0454080: Maintaining Invariant Scales in State, National and International Level Assessments. D Andrich and G Luo Chief Investigators, Murdoch University

Maintaining a common unit in social measurement

Abstract

The unidimensional Rasch model for measurement arises from the requirement of invariant comparisons within a specified frame of reference defined by a class of persons responding to a class of items in a well-defined response context. The defining characteristic of the Rasch model is the existence of sufficient statistics for its parameters. The response context includes empirical factors which have material relevance to the responses, such as characteristics of classes of persons and items and the circumstances under which the persons engage with the items. Responses which fit the Rasch model within a frame of reference are said to be conformable. Implicit within a set of conformable responses is a natural unit of measurement. This paper takes the perspective of classical measurement theory to derive from first principles the concepts of a natural and arbitrary unit of measurement and the relationship between them in the Rasch model. The Rasch model is consequently expressed in a form that can be applied across two or more frames of reference. It is shown that if the only difference between frames of reference is that empirical factors result in a unique natural unit for each frame of reference while measuring the same trait, then it is possible to preserve the property of invariant comparisons and to express measurements from all frames in an arbitrary unit. In this case the sufficient statistic for person parameters across frames of reference is a vector rather than a scalar. The natural unit of each frame of reference is related to the traditional concept of discrimination in item response models, and the paper shows the relationship of the *two parameter logistic model* to the *Rasch model* when the latter is expressed in a form that can be applied across more than one frame of reference. The paper also makes explicit the relationship between the unit and discrimination. An illustrative empirical example of the application of the model across two frames of reference is provided.

Keywords: Rasch model, sufficient statistic, sufficiency, discrimination, unit, scale, Two Parameter Logistic Model

Introduction

Rasch's (1961) undimensional model for measurement arises from the requirement of invariant comparisons within a specified frame of reference. A frame of reference is defined by a class of persons responding to a class of items in a well-defined assessment context. Rasch articulated the requirement of invariant comparisons as follows:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared …. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion (Rasch, 1961, p. 332).

This requirement is embodied within the formal structure and properties of the model itself and is therefore met when data conform to the model (Rasch, 1960, 1961, 1977). In a probabilistic response framework, the requirement of invariant comparison necessarily gives rise to models with sufficient statistics for person and item parameters (e.g. Andersen, 1977; Andrich, 2003; Fischer, 1994) which in turn permits the separation of person and item parameters in estimations.

The existence of sufficient statistics for person and item *parameters* entails that the comparison between two items is made in terms of the same *unit* independent of the particular persons instrumental for the comparison and of other items within the class which are compared. Accordingly, when data conform to the model within a frame of reference, a particular unit is *implicit within the parameters*. The purpose of this paper is to extend the Rasch model (RM) so that it can be applied across different frames of reference where (i) data conform to the model within each frame and (ii) only a different unit of measurement distinguishes the different frames. This objective is achieved by making the units within each frame of reference explicit and identifying these with a scale parameter which can be applied across the frames of reference. As a consequence, traditional methods of scale equating are integrated into the Rasch model and framework in a manner that preserves sufficiency.

The paper is structured as follows. In the next section we introduce Rasch's (1977) concept of the frame of reference and, following Rasch, explicate the property of invariance in the Rasch model (RM) in terms of a pair of items. Following this, we make explicit two arbitrary constants in the RM in the context of a single frame of reference and describe the sense in which one of these constants is an arbitrary scale parameter. We then introduce a simple example in which estimates of the same difference are obtained within *two* frames of reference defined in terms of assessors with different levels of experience. This example is used to illustrate basic principles throughout much of the remainder of the paper. Using this example, we first make explicit the sense in which the same interval may have two measurements in terms of two different units when the RM is applied in each of two frames of reference. We next explicitly identify the scale parameter as a ratio between two units and show that doing so allows estimates to be expressed in the same arbitrary unit irrespective of the frame of reference in which response data are collected. The scale parameter is also explicitly referenced to a frame of reference.

These developments enable expression of the RM in a form applicable across more than one frame of reference each of which has its own implicit unit. We demonstrate that sufficiency holds in the relevant form of the RM, continuing to focus, for the purpose of exposition, on the simplest extended case of two items within two frames. Consistent with results described by Andersen (1973, 1977), it is shown that score vectors rather than scalars are sufficient for person parameters when more than one frame of reference is considered. After demonstrating the preservation of sufficiency in the extended case, we consider the resolution of matrices of item estimates into vectors of scale parameters and item parameters, in order to express the item parameters in an arbitrary unit that is independent of the frame of reference. The model is also characterized graphically in terms of the same example.

Following this exposition in the simplest case, we proceed to more general cases involving other kinds of differences between frames of reference. This leads to a discussion of the key similarities and differences between the extended form of the RM and other models in item response theory, as well as the connection between the scale parameter and the traditional concept of discrimination. An empirical example illustrates the application of the developments.

A single frame of reference for measurement

Table 1 shows an example of a frame of reference where x_{ni} is a discrete response of person *n* to item *i* from the class of persons $\mathbf{O}_n$, $n = 1,...,N$ and the class of items $\mathbf{A}_i$, $i = 1,...,I$.

Table 1: A Rasch model frame of reference

	$\mathbf{A}_1$	$\mathbf{A}_i$	$\mathbf{A}_I$
$\mathbf{O}_1$	x_{11}		
$\mathbf{O}_n$		x_{ni}	
$\mathbf{O}_N$			x_{NI}

In this paper we confine attention to the case in which the response x_{ni} is dichotomous. The principles are however generalizable to the polytomous Rasch model (Rasch, 1961; Andersen,1977; Andrich, 1978; Wright and Masters, 1982).

The usual expression of the RM for dichotomous responses is

$$\Pr\{X_{ni} = 1\} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \qquad (1)$$

where $X_{ni} = x_{ni}$, $x_{ni} \in \{0,1\}$ is a Bernoulli random variable, and β_n and δ_i represent the locations of person *n* and item *i* on a latent continuum.

For purposes of exposition of the distinguishing characteristic of the model, Rasch concentrated on the details of the comparison of just one pair of items *i* and *j*. For the same purpose, we will also concentrate on the comparison of just two items, and similarly, in extending the model we will concentrate on just two frames of reference. Again, the principles can be generalised across more than one frame of reference. We begin from first principles and re-examine relevant implications, usually left implicit, of the comparison of a pair of items within a single frame of reference. This re-

examination of first principles will enable us to extend the RM to a form which can be applied across more than one frame of reference.

Let $r_n = x_{ni} + x_{nj}$. Then from Equation (1),

$$\Pr\{(X_{ni} = 0; X_{nj} = 1) \mid r_n = 1\} = \exp(\delta_j - \delta_i)/\gamma_{ij}, \quad (2)$$

and its complement is

$$\Pr\{(X_{ni} = 1, X_{nj} = 0) \mid r_n = 1\} = 1/\gamma_{ij}, \quad (3)$$

where $\gamma_{ij} = 1 + \exp(\delta_j - \delta_i)$. Equation (2) is independent of the person parameter β_n, reflecting the sufficiency of the total score r_n for this parameter (Rasch, 1966).

From Equation (2), the natural logarithm ratio

$$\ln\left(\frac{\Pr\{(X_{ni} = 1, X_{nj} = 0) \mid r_n = 1\}}{\Pr\{(X_{ni} = 0, X_{nj} = 1) \mid r_n = 1\}}\right) = \delta_j - \delta_i \quad (4)$$

gives the difference of the item parameters explicitly and independently of the person parameter.

Let F_{ij} be the frequency of the response patterns $(X_{ni} = 1, X_{nj} = 0)$ and let F_{ji} be the frequency of response patterns $(X_{ni} = 0, X_{nj} = 1)$. From Equation (4),

$$\ln(F_{ij}/F_{ji}) = \hat{\delta}_{ji} \quad (5)$$

is an estimate of the difference $\delta_{ji} = \delta_j - \delta_i$. As a particular example, suppose that $F_{ij} + F_{ji} = 100$ and that $F_{ij} = 77$ and $F_{ji} = 23$. Then

$$\hat{\delta}_{ji} = \ln(F_{ij}/F_{ji}) = \ln(77/23) = 1.21 . \quad (6)$$

An arbitrary additive constant and arbitrary origin

Consistent with Equation (6), it is only possible to estimate the difference δ_{ji} and not each individual parameter. In order to obtain a unique estimate of each item parameter in general, the arbitrary constraint

$$\sum_{i=1}^{I} \hat{\delta}_i = 0 \qquad (7)$$

is usually imposed which results in an arbitrary choice of the origin. However, for any constant ς and locations $\delta_j^{'} = \delta_j + \varsigma$, $\delta_i^{'} = \delta_i + \varsigma$, the contrast $\delta_j^{'} - \delta_i^{'} = \delta_j - \delta_i$ is independent of ς. Consequently, an arbitrary constraint

$$\sum_{i=1}^{I} \hat{\delta}_i^{'} = \varsigma \qquad (8)$$

when applied to the same set of items as Equation (7) results in a change of origin such that $\delta_i^{'} = \delta_i + \varsigma$, $i = 1,..., I$. The arbitrariness of ς and the constraint of Equation (7) are well known and Equation (8) is imposed explicitly in solution algorithms and used in linking designs when not all persons respond to all items.

An arbitrary multiplicative constant

Similarly, an arbitrary multiplicative constant is also implicit in Equation (1). The implications of this constant in the Rasch model are generally recognized only implicitly (e.g. Embretson & Reise, 2000; Wood, 1978). For the purpose of this paper it is necessary to make these implications explicit.

Rasch identified a general form of a *measuring function*, as defined in terms of invariant comparisons, with the inclusion of two constants, remarking that any values could be chosen for these constants such that person and item locations vary within an interval which “may for some reason be deemed convenient” (Rasch, 1960, p. 121). One of these constants is a multiplicative constant and, with this constant the general form of the dichotomous RM is

$$\Pr\{X_{ni}=1\} = \frac{\exp\rho(\beta_n/\rho-\delta_i/\rho)}{1+\exp\rho(\beta_n/\rho-\delta_i/\rho)} = \frac{\exp\rho(\beta_n^*-\delta_i^*)}{1+\exp\rho(\beta_n^*-\delta_i^*)} \quad (9)$$

where $\beta_n^* = \beta_n/\rho$, $\delta_i^* = \delta_i/\rho$ and ρ is the multiplicative constant. Clearly, there is no change of probability between Equations (1) and (9) for given $\beta_n, n=1,...,N$ and $\delta_i, i=1,...I$, indicating the value of ρ is arbitrary. The constant ρ is a scale parameter in the sense that it only enlarges or shrinks the person-item distance $\beta_n-\delta_i$ to $(\beta_n-\delta_i)/\rho$ (Andrich, 1995; Luo, 1998). Generally, this scale parameter is taken implicitly to be $\rho\equiv 1$ as in Equation (1). However, it can be specified to be any value. When included as a *variable* parameter in an item response theory model, ρ is generally referred to as a discrimination parameter. We return to the relationship between the scale parameter and discrimination later in the paper.

Using the full Equation (9) in estimating the difference between the two item parameters, it follows that

$$\ln(F_{ij}/F_{ji}) = \rho(\hat{\delta}_{ji}) \quad (10)$$

which remains independent of the person parameters. Given the left side of Equation (10) is a single estimated real number and the right side has two factors, it is necessary to specify ρ arbitrarily, with the consequence being an arbitrary choice of the unit of a metric.

Comparing estimates across two frames of reference

In order to explicate the sense in which a unit is implicit within response data of a frame of reference which conform to the RM, we now consider the simplest possible case of the *comparison* between two *units*. This case involves two frames of reference in which all empirical factors but one are kept constant in the generation of responses.

Let $\mathbf{F}_s$ denote the frame of reference defined in terms of classification *s* of an empirical factor and let x_{sni} denote the response associated with the interaction of person *n* and item *i* in frame *s*. Table 2 shows two frames of reference *s* =1 and *s* = 2 in which the persons and items are identical, and all factors but one empirical factor are kept constant.

Such a situation might arise in psychological and educational assessment where the same performance, under the same conditions, is assessed with respect to the same criteria by different assessors (psychologists or teachers) who have different levels of experience. In order to make this empirical factor tangible and relevant, suppose that one of the assessors is very experienced in the relevant field and that the other is a novice, and each judges whether the performance of a number of persons meets criteria specified in two assessment items. The level of experience of the assessor is then an empirical factor, and each of the assessors represents a different classification of this factor. This elementary situation is later generalized to situations in which different frames of reference may also have different items or different persons.

Table 2: Two frames of reference for which only one factor varies

	***s* = 1**			***s* = 2**		
	$\mathbf{A}_1$	$\mathbf{A}_i$	$\mathbf{A}_I$	$\mathbf{A}_1$	$\mathbf{A}_i$	$\mathbf{A}_I$
$\mathbf{O}_1$	x_{111}			x_{211}		
$\mathbf{O}_n$		x_{1ni}			x_{2ni}	
$\mathbf{O}_N$			x_{1NI}			x_{2NI}

Table 3 shows two different sets of frequencies which might arise from the two different assessors assessing the same performances on two items. Let F_{sij} and F_{sji} be the frequencies of response patterns $(X_{sni}=1, X_{snj}=0)$ and $(X_{sni}=0, X_{snj}=1)$ respectively, and let δ_{sji} be the difference between the scale locations of the items in the frame of reference $\mathbf{F}_s$.

As evident in Table 3, the data from the two different frames of reference yield different estimates of the difference between two items; namely $\hat{\delta}_{1ji} = 1.21$ and $\hat{\delta}_{2ji} = 0.80$. We suppose, for the purpose of exposition, that this difference arises only from a difference in the level of precision of measurement of the trait by each assessor and not a substantive difference in the nature of the trait that is measured.

Table 3: Relative frequencies obtained in the assessment of the same performances on two items

	s = 1 (Expert)	s = 2 (Novice)
F_{sij}	77	69
F_{sji}	23	31
$F_{sij} + F_{sji}$	100	100
$\ln(F_{sij} / F_{sji}) = \hat{\delta}_{sji}$	1.21	0.80

In general terms, differences between locations among frames of reference are usually taken to imply differential item functioning. A particular case is that in which the responses conform to the RM within each frame of reference but the relative locations of the items differ within the different frames of reference. The form of the Rasch model applicable in such circumstances is generally termed a *mixed Rasch model* (Rost et al, 1997).

This paper is concerned with an even more particular case in which differences between item locations are systematic and reflect only a difference between the units of metrics implicit within response data contained in different frames of reference. Particular conditions must be met in order to infer such a difference between units. Specifically, when the responses within each frame of reference conform to the RM and, in addition, the same ratio between estimates is preserved among all items across the two frames of reference, then the two frames of reference will be said to show a difference in their *natural units*.

The natural unit

In order to explicate the concept of a natural unit, we begin by letting roman non-italicized characters represent intervals on a linear continuum. Thus, let d_{ji} be the interval between items *j* and *i* on the latent continuum. It is stressed that d_{ji} is not a number; rather it is an *interval* which represents the magnitude of the difference between the items on a latent continuum. It is also stressed that the interval d_{ji} is by definition *invariant* across frames of reference; i.e. it is an interval of a trait that has a fixed magnitude irrespective of the frame of reference in which response data are obtained. Such an interval is shown on a continuum in Figure 1.

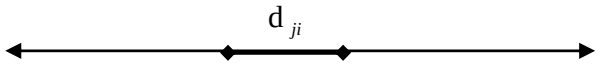

Figure 1: The interval between items *j* and *i* shown on a latent continuum

Consistent with the classical theory of measurement (Michell, 1986, 1999), a *measurement* of d_{ji} is given by the ratio of d_{ji} to another interval taken to be a unit. In Figure 2, the interval d_{ji} is augmented by two intervals b_1 and b_2 which are taken to be natural units of two different frames of reference. Specifically, in the example shown in Table 3, let $\hat{\delta}_{1ji} = 1.21$ and $\hat{\delta}_{2ji} = 0.80$ be measurements of d_{ji} in units b_1 and b_2 respectively. That is, $\hat{\delta}_{1ji} = \mathrm{d}_{ji} / \mathrm{b}_1$ and $\hat{\delta}_{2ji} = \mathrm{d}_{ji} / \mathrm{b}_2$. Thus the difference between these measurements implies the interval d_{ji} is measured relative to different units. This is illustrated in Figure 2.

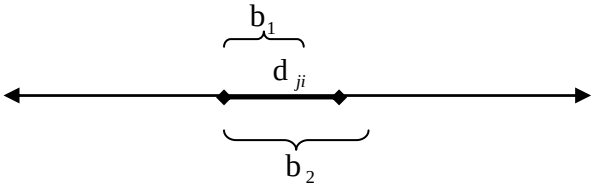

Figure 2: Units implied by different measurements of a common interval

In turn, the difference between the measurements of the same interval implies that two *metrics* are formed by partitioning the same latent continuum into two natural units. Figure 3 augments Figure 2 showing these metrics.

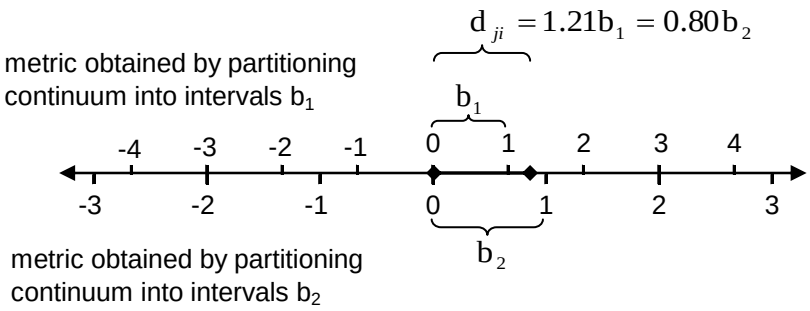

Figure 3: A continuum partitioned into two natural units

In general, the measure of d_{ji} obtained in the frame of reference $\mathbf{F}_s$ is defined as the ratio of d_{ji} to the natural unit of the frame; i.e.

$$\delta_{sji} = \frac{d_{ji}}{b_s}. \qquad (11)$$

From Equation (11) we obtain the usual expression of the measurement of the interval d_{ji} in the unit b_s; i.e.

$$d_{ji} = \delta_{sji} b_s. \qquad (11a)$$

Thus, in Table 3 and as shown in Figure 3, two measurements of the interval d_{ji} were obtained in two different units; these measurements being $d_{ji} = 1.21b_1$ and $d_{ji} = 0.80b_2$. That is, the magnitude of the interval is expressed in terms of specified units pertaining to a well-defined empirical factor, which is the level of experience of the assessor. This way of expressing magnitudes is the rule rather than the exception in the natural sciences, where it is rare not to specify the unit in which the measurement of an object's property is taken. In the social sciences, analyses tend to have their own arbitrary unit with the dependence of the unit on the empirical context left implicit.

The distinction between the natural and arbitrary unit and its implications for the relative precision of measurements are examined from an alternative perspective in Andrich (2003).

The scale parameter and the arbitrary unit

Because the frames of reference differ only in terms of their natural units and not in terms of the trait itself, the measurements can be expressed in terms of a common arbitrary unit. This implies that the frames can be unified to form a single higher-order frame of reference for measurement.

Thus suppose in our example b_2 is chosen as the arbitrary unit. Let b_* denote an arbitrary unit, such that in this example $\mathrm{b}_* = \mathrm{b}_2$. Also, let δ^*_{ji} be the measure of the interval d_{ji} in terms of the arbitrary unit; i.e.

$$\mathrm{d}_{ji} = \delta^*_{ji}\mathrm{b}_* . \tag{12}$$

It follows from Equations (11a) and (12) that $\mathrm{d}_{ji} = \delta^*_{ji}\mathrm{b}_* = \delta_{1ij}\mathrm{b}_1$, and therefore that $\delta_{1ji} = (\mathrm{b}_* / \mathrm{b}_1)\delta^*_{ji}$.

Letting $\rho_1 = \mathrm{b}_* / \mathrm{b}_1$ gives $\delta_{1ji} = \rho_1\delta^*_{ji}$. Thus, in general let

$$\rho_s \equiv \frac{\mathrm{b}_*}{\mathrm{b}_s}, \tag{13}$$

from which it follows generally that

$$\delta_{sji} = \rho_s\delta^*_{ji} , \tag{14}$$

showing that ρ_s is the scaling factor referred to in relation to Equations (9) and (10). However, it is now referenced to the frame of reference $\mathbf{F}_s$ rather than being left as a general arbitrary multiplicative constant ρ. That is, *in the context of two frames of reference we have identified the arbitrary multiplicative constant in Equations (9) and (10) as the ratio of an arbitrary unit to the natural unit of a frame of reference.*

From Equation (14), the ratio of two scale parameters is given by

$$\frac{\delta_{1ji}}{\delta_{2ji}} = \frac{\rho_1}{\rho_2}, \tag{15}$$

which implies the ratio should be the same, independent of the particular contrast between two items *j* and *i* instrumental for the comparison between the scaling constants ρ_1 and ρ_2. The requirement of invariant comparison between scale parameters directly parallels the requirement for invariant comparison between item parameters *within* a frame of reference articulated by Rasch (1961), as cited in the introduction.

In the case of the data shown in Table 3, the ratio between the scale parameters was estimated as $\hat{\delta}_{1ji} / \hat{\delta}_{2ji} = 1.51$. It follows from Equation (13) that if b_2 is chosen as the arbitrary unit, then the scale parameter of frame $\mathbf{F}_2$ is $\rho_2 \equiv \mathrm{b}_* / \mathrm{b}_2 \equiv 1$. Using Equation (14), from which $\hat{\delta}^*_{ji} = \delta_{sji} / \rho_s$, we now obtain estimates for all parameters shown in Table 4.

Table 4: Parameter estimates from the data of each assessor

	s = 1 (Expert)	*s* = 2 (Novice)
$\hat{\delta}_{sji} = \hat{\rho}_s \hat{\delta}^*_{ji}$	1.21	0.80
$\hat{\rho}_s$	1.51	1
$\hat{\delta}^*_{ji}$	0.80	0.80

The difference between the estimates shown in the first row of Table 4 implies the natural unit of the measurement obtained from the expert's judgments is smaller than the natural unit of the measurement obtained from the novice's judgments. Hence the expert is able to discriminate between performances to a higher degree. That is, the difference implies that a greater level of precision is obtained by the expert judge than by the novice judge.

In the classical theory, measurement is "defined as the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute"

(Michell, 1997, p. 358). Accordingly, an estimate of ρ_s is also a measurement because it is a ratio between two intervals on a latent trait, hence a ratio between two magnitudes of a quantitative attribute. Clearly, though, this measurement does not represent the magnitude of the latent trait attributable to either a person or item being assessed. In the present example, an estimate of ρ_s can be interpreted as a measurement of the relative degree of discrimination between levels of performance by assessor *s*. More generally, in keeping with the traditional concept of discrimination in item response theory, an estimate of ρ_s is, in a well-defined sense, a *measure* of the degree of precision or discrimination obtained within the frame of reference $\mathbf{F}_s$.

Thus we have identified ρ_s both as a scaling constant associated with a frame of reference and a measure of the degree of discrimination obtained in the presence of the relevant classification of an empirical factor in terms of which the frame is defined. Accordingly, the terms *discrimination parameter* and *scale parameter* are used interchangeably to follow in this paper, with the former emphasising the empirical character of the parameter and the latter emphasising the inherent connection between the parameter and the natural unit.

Identifying the scale parameter in the Rasch model

Making explicit the frame of reference *s* by its subscript, the Rasch model is

$$\Pr\{X_{sni}=1\}=\frac{\exp(\beta_{sn}-\delta_{si})}{1+\exp(\beta_{sn}-\delta_{si})}. \qquad (16)$$

Equation (16) is referred to as the *specified Rasch model* because parameters are specified in relation to a frame of reference. This model gives different locations for any given person or item within separate frames of reference when the units associated with those frames differ.

Now in the same way that d_{ji} represents the interval between items *j* and *i* on a latent continuum, let d_{ni} be the interval between person *n* and item *i* on the continuum. Also, let $\beta_{sn} - \delta_{si} \equiv \mathrm{d}_{ni} / \mathrm{b}_s$ be the measure of this difference in the unit b_s such that

$$\mathrm{d}_{ni} \equiv (\beta_{sn} - \delta_{si})\,\mathrm{b}_s . \tag{17}$$

Next, let $\beta_n^* - \delta_i^* = \mathrm{d}_{ni} / \mathrm{b}_*$ be the measurement of the interval between person *n* and item *i* expressed in an arbitrary unit b_* such that

$$\mathrm{d}_{ni} \equiv (\beta_n^* - \delta_i^*)\,\mathrm{b}_* . \tag{18}$$

From Equations (17) and (18) and noting from Equation (13) that $\mathrm{b}_s \equiv \mathrm{b}_* \rho_s^{-1}$, it follows that

$$\beta_{sn} - \delta_{si} \equiv \rho_s \left(\beta_n^* - \delta_i^*\right). \tag{19}$$

Finally, substituting Equation (19) into Equation (16) gives

$$\Pr\{X_{sni} = 1\} = \frac{\exp\left(\rho_s (\beta_n^* - \delta_i^*)\right)}{1 + \exp\left(\rho_s (\beta_n^* - \delta_i^*)\right)}, \tag{20}$$

in which $\beta_n^* - \delta_i^*$ is the measure of the interval between person *n* and item *i* in the arbitrary unit b_* and ρ_s is the scale parameter of the frame of reference *s*. *Equation (19) is identical with Equation (9) except that we have once again identified the general arbitrary multiplicative constant* ρ *in Equation (9) within a single frame of reference as a scale parameter* ρ_s *of each frame of reference s.*

Equation (19) is referred to as the *frame of reference Rasch model* (FRM) because it makes explicit the frame of reference for measurement in terms of the scaling constant associated with the frame of reference.

Sufficiency

We now demonstrate that *sufficiency*, which is the defining feature of Rasch models, holds in the FRM. Let $r_{sn} = \sum_{i \in s} x_{sni}$ be the person raw score in frame *s* and let $P_{xsni} = \exp(x_{sni}(\beta_{sn} - \delta_{si}))/\gamma_{sni}$, consistent with Equations (2) and (3), where again $x_{sni} \in \{0,1\}$. Consider the patterns of responses resulting in the vector of raw scores $(r_s = 1, r_t = 1)$ for frames of reference *s* and *t*. The patterns of possible responses and (unconditional) probabilities of the responses for which $r_{sn} = 1$ and $r_{tn} = 1$ are shown in Table 5.

Table 5: Patterns of responses for two items in each of two frames of reference *s* and *t* and their respective probabilities (*P*)

$\mathbf{F}_s$			$\mathbf{F}_t$		
$x_{sni}(P_{xsni})$	$x_{snj}(P_{xsnj})$	r_{sn}	$x_{tni}(P_{xtni})$	$x_{tnj}(P_{xtnj})$	r_{tn}
$1\left(\frac{e^{\beta_{sn}-\delta_{si}}}{\gamma_{sni}}\right)$	$0\left(\frac{1}{\gamma_{snj}}\right)$	1	$1\left(\frac{e^{\beta_{tn}-\delta_{ti}}}{\gamma_{tni}}\right)$	$0\left(\frac{1}{\gamma_{tnj}}\right)$	1
$0\left(\frac{1}{\gamma_{sni}}\right)$	$1\left(\frac{e^{\beta_{sn}-\delta_{sj}}}{\gamma_{snj}}\right)$	1	$0\left(\frac{1}{\gamma_{tni}}\right)$	$1\left(\frac{e^{\beta_{tn}-\delta_{tj}}}{\gamma_{tnj}}\right)$	1

In parallel with Equation (1) for a single frame of reference, in order to demonstrate sufficiency we derive the conditional probability $\Pr\{(1,0);(1,0) \big| (r_{sn} = 1, r_{tn} = 1)\}$, where the pairs of ordered pairs $(x_{sni}, x_{snj});(x_{tni}, x_{tnj})$ refer to the possible responses in each frame of reference to the two items, and the ordered pair (r_{sn}, r_{tn}) refers to the total scores of the person on the items within each frame. In order to demonstrate sufficiency of statistics for the person parameters, it is necessary to prove that this probability is independent of the person parameters; i.e. that the conditional probability $\Pr\{(1,0);(1,0)\} \big| (r_{sn} = 1, r_{tn} = 1)$ is independent of β_{sn}, β_{tn}.

Proof. First,

$$\Pr\{(r_{sn} = 1, r_{tn} = 1)\} = \Pr\{(1,0);(1,0)\} + \Pr\{(1,0);(0,1)\} + \Pr\{(0,1);(1,0)\} + \Pr\{(0,1);(0,1)\},$$

that is,

$$\begin{aligned}\Pr\{(r_{sn}=1,r_{tn}=1)\} = \{&(e^{\beta_{sn}-\delta_{si}}.1.e^{\beta_{tn}-\delta_{ti}}.1)+(e^{\beta_{sn}-\delta_{si}}.1.1.e^{\beta_{tn}-\delta_{tj}})+\\ &(1.e^{\beta_{sn}-\delta_{sj}}.e^{\beta_{tn}-\delta_{ti}}.1)+(1.e^{\beta_{sn}-\delta_{sj}}.1.e^{\beta_{tn}-\delta_{tj}})\}/\gamma_{sni}\gamma_{snj}\gamma_{tni}\gamma_{tnj}\end{aligned} \quad (21)$$

which on simplification gives

$$\Pr\{(r_{sn}=1,r_{tn}=1)\} = e^{\beta_{sn}+\beta_{tn}}\{e^{-\delta_{si}-\delta_{ti}}+e^{-\delta_{si}-\delta_{tj}}+e^{-\delta_{sj}-\delta_{ti}}+e^{-\delta_{sj}-\delta_{tj}}\}/\gamma_{sni}\gamma_{snj}\gamma_{tni}\gamma_{tnj}\,.$$

Therefore, from

$$\Pr\{(1,0);(1,0)\big|(r_{sn}=1,r_{tn}=1)\} = \frac{\Pr\{(1,0);(1,0)\}}{\Pr\{(r_{sn}=1,r_{tn}=1)\}}, \quad (22)$$

$$\begin{aligned}\Pr\{(1,0);(1,0)\big|(r_{sn}=1,r_{tn}=1)\} &= \frac{e^{\beta_{sn}+\beta_{tn}}(e^{-\delta_{si}-\delta_{ti}})}{e^{\beta_{sn}+\beta_{tn}}(e^{-\delta_{si}-\delta_{ti}}+e^{-\delta_{si}-\delta_{tj}}+e^{-\delta_{sj}-\delta_{ti}}+e^{-\delta_{sj}-\delta_{ti}})},\\ &= \frac{e^{-\delta_{si}-\delta_{ti}}}{(e^{-\delta_{si}-\delta_{ti}}+e^{-\delta_{si}-\delta_{tj}}+e^{-\delta_{sj}-\delta_{ti}}+e^{-\delta_{sj}-\delta_{ti}})},\end{aligned} \quad (23)$$

which is independent of the vector (β_{sn},β_{tn}) of person parameters.

This completes the proof.

A generalization of Equation (23), which is relevant to the illustrative empirical application of the model and is discussed further in the paper, is provided in Appendix I.

The score vector across the frames of reference is “the natural generalization of the raw score” (Andersen, 1973, p. 73) within each frame of reference. Andersen’s statement regarding the sufficiency of score vectors pertained to the elimination of a vector of person parameters in the general polytomous RM where it is assumed that the responses *m+1* categories may be characterised by *m* person and item parameters. Because of its relevance to this aspect of the paper, a full description of the vector form of the polytomous Rasch model is provided in Appendix II.

The resolution of parameters into products

In the paper referred to above, Andersen (1973, p. 43) investigated the "dimensionality of the parameters" by examining whether it was possible to resolve the matrix of *category by item* parameter estimates into the product of a vector of *item parameters* and a vector of *category coefficients*. He conducted a likelihood ratio test to evaluate the parameter reduction achieved by doing so.

Analogously, the specific question in more general cases of the example above involving several items is whether it is possible to resolve the matrix of item estimates $\hat{\delta}_{si}$, $s = 1,...,S$, $i = 1,...,I$ into the product of a vector of item parameters δ_i^*, $i = 1,...,I$ and a vector of scale parameters ρ_s, $s = 1,...,S$. This resolution for the illustrative example of Table 3 is shown in Equation (25).

Due to the possibility that different *traits* are measured within different frames of reference, it is an empirical question whether the frames of reference are mutually conformable in the sense that item parameters in natural units can be decomposed into the product of a vector of item parameters in an arbitrary unit and a vector of scale parameters. If this is the case, then the differences between the measurements in the different frames of reference can be considered to result only from a difference in the natural unit of measurement and not from differences in substantive differences in the trait measured; that is, the differences result only from differences in the level of precision of measurement in different frames of reference.

It was noted earlier that it is only possible to estimate differences between item locations and that an arbitrary constraint such as that shown in Equation (8) must be imposed in order to obtain estimates for each individual item. Similarly, it is only possible to estimate ratios such as ρ_1 / ρ_2 as shown in Equation (15), and so it is necessary to impose an arbitrary constraint in order to obtain estimates of each scale parameter separately. In the example of Table 2, the constraint $\rho_2 = 1$ was imposed by choosing the natural unit of the frame of reference $\mathbf{F}_2$ as the arbitrary unit. More generally, a multiplicative constraint analogous to the additive constraint of Equation (8), giving an arbitrary origin, may be imposed. Because the scale parameter is a

multiplicative constant, an appropriate constraint is that the geometric mean is equal to 1; i.e.

$$\prod_s \hat{\rho}_s \equiv 1. \qquad (24)$$

This constraint results in an arbitrary unit which differs from each of the natural units of the frames of reference, as shown in Figure 4 below. The constraint is employed in the empirical investigation to follow.

In the example of Table 3, if the constraint shown in Equation (24) is imposed instead of choosing $\rho_2 = 1$, then $\hat{\rho}_1\hat{\rho}_2 \equiv 1$. Since $\hat{\rho}_1 / \hat{\rho}_2 = \hat{\delta}_{1ji} / \hat{\delta}_{2ji} = 1.51$ as shown in Table 4, with these constraints imposed we obtain estimates $\rho_1 = 1.51^{0.5} = 1.23$ and $\rho_1 = 1.51^{-0.5} = 0.81$. It then follows that $\delta^*_{ji} = \delta_{1ji} / \rho_1 = 1.21/1.23 = 0.98$ and, identically, $\delta^*_{ji} = \delta_{2ji} / \rho_2 = 0.80/0.81 = 0.98$. Thus, completing the example for the data in Table 3, with $\sum_i \hat{\delta}_{si} \equiv 0$ and the constraint of Equation (24)

$$\hat{\Delta}_{\delta\rho} = \begin{bmatrix} \hat{\delta}_{11} & \hat{\delta}_{12} \\ \hat{\delta}_{21} & \hat{\delta}_{22} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} \begin{bmatrix} \delta_1^* & \delta_2^* \end{bmatrix} = \begin{bmatrix} 1.23 \\ 0.81 \end{bmatrix} \begin{bmatrix} -0.49 & 0.49 \end{bmatrix}. \qquad (25)$$

The question of whether it is possible to resolve parameters into products in this manner can be investigated by estimating each parameter set and subsequently conducting appropriate tests of fit which may involve the person parameters. The tests of fit need to be sensitive to differences between scale parameters across different frames of reference as reflected in the Item Characteristic Curves (ICCs). An example of such a test is outlined in the empirical investigations which follow.

Graphical characterization of the FRM

Given both person and item locations have been expressed in terms of a common arbitrary unit, the scale parameter of a frame of reference is reflected within the slopes of ICCs. Consider an extension of the previous example in which each person's performance is assessed by the two different assessors, under the same conditions, on several items rather than just two items. Figure 4 shows the ICCs of a set of items

within each of two specified frames of reference for which $\rho_1 / \rho_2 = 1.51$ and so, given the constraint shown in Equation (24), $\rho_1 \cong 1.23$ and $\rho_2 \cong 0.81$.

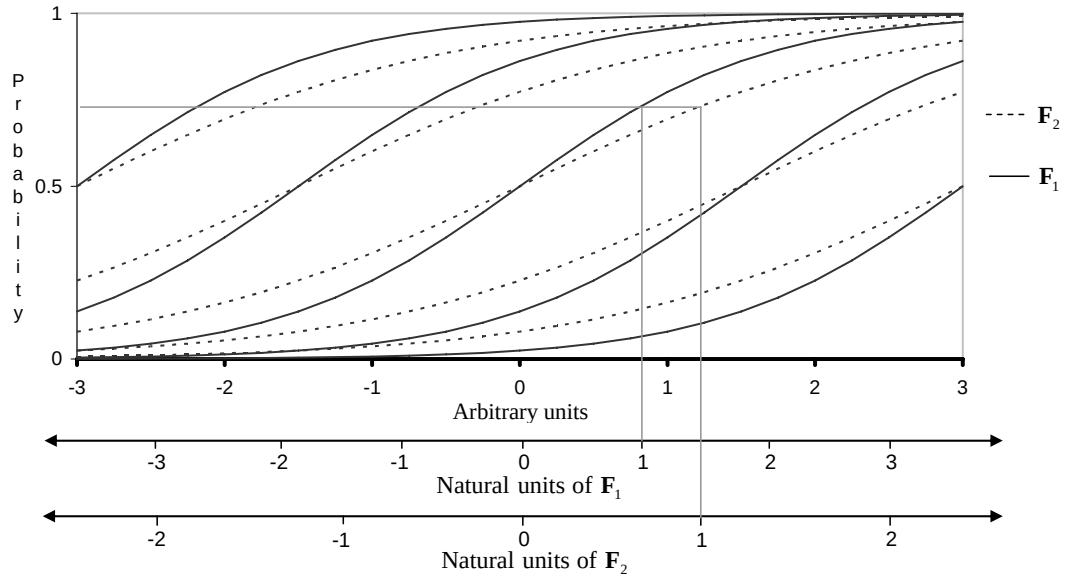

Figure 4: ICCs for items within two Specified Frames of Reference

Three sets of coordinates are provided as horizontal axes in Figure 4 to highlight that each set of ICCs accords with the RM. Thus, when the ICCs for frame $\mathbf{F}_s$ are referenced to the scale partitioned into the natural units of $\mathbf{F}_s$, the ICCs accord with the RM; i.e. they follow Equation (1). Accordingly, the grey lines in Figure 4 show that $\Pr\{X_{sni} = 1\} \cong 0.73$ when $\beta_{sn} - \delta_{si} = 1$ within each frame of reference *s* = 1 and *s* = 2. On the other hand, when the same two sets of ICCs are referenced to the scale partitioned into common arbitrary units, the ICCs accord with the FRM, Equation (20), where $\rho_1 = 1.23$ and $\rho_2 = 0.81$. As mentioned earlier, it can be seen that with the constraint of Equation (24) the size of the arbitrary unit b_* differs from the size of both natural units b_1 and b_2.

Multiple frames of reference

So far attention has been focused on the case of two frames of reference in which the identities of persons and items were the same and all empirical factors but one were held constant. We now consider more general cases, one of which also anticipates the illustrative empirical application. The focus remains on contexts in which

measurements are obtained from separate frames of reference constructed with the intention of measuring a *common* latent trait.

A collection of four frames of reference is shown in Table 6. A given frame of reference is denoted $\mathbf{F}_{gs}$. The matrix of response data obtained within a particular frame of reference consists of individual responses denoted x_{gsni}. Pairs of frames share common elements. For example the same collection of persons $\mathbf{O}_1$ is contained within frames $\mathbf{F}_{11}$ and $\mathbf{F}_{12}$, the same set of items $\mathbf{A}_1$ is contained within $\mathbf{F}_{11}$ and $\mathbf{F}_{21}$, and so on. In order to compare units and origins, a pair of frames must share either common persons or common items.

Table 6: Multiple frames of reference

		***s* = 1**			***s* = 2**		
		$\mathbf{A}_{11}$	$\mathbf{A}_{1i}$	$\mathbf{A}_{1I_1}$	$\mathbf{A}_{21}$	$\mathbf{A}_{2i}$	$\mathbf{A}_{1I_s}$
	O_{11}	x_{1111}			x_{1211}		
***g*=1**	O_{1n}		x_{11ni}			X_{12ni}	
	O_{1N_1}			$x_{11N_1I_1}$			$x_{12N_1I_2}$
	O_{21}	x_{2111}			x_{2211}		
g=2	O_{2n}		x_{21ni}			x_{22in}	
	$\mathbf{O}_{2N_2}$			$x_{21N_2I_1}$			$x_{22N_2I_2}$

Empirical factors affecting responses arising from the interaction of persons and items can be classified in order to distinguish different frames of reference. Sometimes these are empirical factors which can be controlled experimentally such as the experience of assessors, and sometimes they are factors such as gender which cannot.

In the illustrative empirical example below, items are classified according to their membership to assessment forms constructed by different item developers at different times. It is shown that there is a difference between the natural units of their frames of reference, and therefore a difference between the scale parameters.

Similarly, it is possible for differences between natural units to arise where the items are identical but persons are different. Evidence for such a case in which persons were classified according to year group of schooling is presented in Humphry (2005).

Similarities and contrasts with other item response models

Having now considered more general cases in which response data are collected in a collection of frames, it is useful to consider the similarities and contrasts between the FRM and other models, and particularly Birnbaum's (1968) two parameter logistic model (2PLM).

The key conceptual difference between the 2PLM and the FRM is that the scale parameter in the latter arises from explicitly formulating the relationship between the natural units of different frames of reference in a manner which preserves sufficiency, rather than from the goal of describing data. This is compatible with the conceptual and paradigm difference between the RM and the 2PLM in the case of a single frame of reference described in Andrich (2004); namely, the case made by Rasch for the class of models that now bears his name was not that it described any particular data set, but that it characterised invariance of comparisons.

In order to preserve invariance, the parameter ρ_s of the FRM is required to be fixed within a frame of reference, but its magnitude may vary between frames of reference as shown in Figure 4. If frames of reference are defined in terms of item factors *alone* and, in addition, *each* item is treated as representing a different classification *s* of an empirical factor, then the structure of the FRM is formally equivalent to the 2PLM. The 2PLM has the form

$$\Pr\{X_{ni}=1\}=\frac{\exp\left(\alpha_i(\beta_n-\delta_i)\right)}{1+\exp\left(\alpha_i(\beta_n-\delta_i)\right)}. \qquad (26)$$

where α_i is generally referred to as the "discriminating power" of item *i* (e.g. Lord, 1980). This parameter may vary for each item.

While the formal structure of the FRM specializes to that of the 2PLM in the case described, it is impossible to conditionally estimate parameters because in that specific case a person's response vector is identical to the person's score vector; i.e. $(x_{sn1}, x_{2n1}, ..., x_{Sn1}) \equiv (r_{sn}, r_{2n}, ..., r_{Sn})$. Consequently, no data reduction is achieved by partitioning the response space in terms of score vectors and, hence, no information can be obtained from relative frequencies of different response patterns in a conditional likelihood expression of the form shown in Equation (23) or, more generally, Appendix I. The implication is that the separation of parameters demonstrated above is impossible, and so sufficiency cannot be exploited in order to make invariant comparisons between items. Rather, comparisons depend on the person parameters and therefore also the distribution of person parameters. The complementary implication is that for estimation, the FRM requires at least two items per frame of reference where there are common items across the frames. This is consistent with requiring two points to define an interval. If general tests of fit such as those used in the empirical investigations are used, more than two items are necessary.

The similarities and contrasts between the FRM and 2PLM also become evident by considering a related item response model introduced by Verhelst & Glas. Verhelst & Glas (1995) derive conditional maximum likelihood (CML) equations from a model referred to as the *one parameter logistic model* (OPLM), which contains a discrimination *index* rather than a discrimination *parameter*, but is formally identical with the 2PLM. The authors note, however, the problem one faces in implementing CML estimation in the 2PLM is that the values of discrimination parameters are unknown, meaning the weighted raw score "is not a mere statistic, and hence it is impossible to use CML as an estimation method" (Verhelst & Glas, 1995, p. 217). In the FRM, in contrast it is possible to exploit sufficiency of the weighted raw score without prior knowledge of the values of ρ_s, $s = 1, ..., S$ by conditioning on score vectors due to the fact that all response vectors $\mathbf{x}_n = (x_{11n}, ..., x_{Sni})$, which yield a particular score vector $\mathbf{r}_n = (r_{1n}, ..., r_{Sn})$, necessarily also yield the same weighted raw

score $W_n = \sum_{s=1}^{S} \rho_s r_{sn}$. In the FRM, the vector $(\beta_{1n},...,\beta_{sn},...,\beta_{Sn})$ is by definition identical to $(\rho_1\beta_n^*,...,\rho_s\beta_n^*,...,\rho_S\beta_n^*)$. Hence, by eliminating the vector of person parameters through conditioning on score vectors, the person parameter β_n^* is necessarily eliminated. This is directly analogous to the elimination of the person parameter noted by Verhelst & Glas, except that it is made explicit in the FRM that the person parameter being eliminated is expressed in an arbitrary unit rather than a natural unit.

The relationship between discrimination and the unit of a metric is often implicitly recognised both in relation to the RM and other models such as the 2PLM (e.g. Andrich, 1988; Baker, 1983, 1984; Brink, 1971; Lord, 1980; Stocking & Lord, 1983; Wood, 1978). Embretson & Reise (2000, p. 129) note, for example, that in "the simple Rasch model, the same log odds may be predicted from infinitely many combinations of trait level and item difficulty" through arbitrary specification of different values of ρ in Equation (1). In a similar vein, Wood (1978, p. 29) noted that in the RM an ability estimate is "always scaled by a factor", this being "the common level of discrimination for all items".

In addition, consistent with the broader conceptualisation of discrimination in the FRM, the potential for person characteristics to influence discrimination and the unit of a scale is occasionally recognised. For example, in developing methods of assessing person fit to Rasch models, Klauer (1995, p. 100) presents a model containing a parameter which "regulates the overall level of the item discrimination operating for the examinee", noting the implications of this parameter for the variance of item estimates. Lord (1980, p. 35) clearly acknowledged the potential for person characteristics to influence level of discrimination as did Andrich (1988, p. 75).

Generalization of estimations across two frames of reference with the same persons and different items

In Table 2, the two frames of reference which had the same persons and items were distinguished by an empirical factor. We now consider a generalization of Table 2; specifically the case involving the combination $g = 1$ and $s = 1, 2$ in Table 6. Here, two frames of reference have the same persons and *different* items, where the items

are distinguished by an empirical factor. This is the case applicable to the empirical example to follow.

The generalization of the conditional Equation (23) in the case of two frames of reference with two different items in each is

$$\Pr\{(1,0);(1,0)\,|\,(r_{sn}=1,r_{tn}=1)\}=\frac{e^{\beta_{sn}+\beta_{tn}}(e^{-\delta_{si}-\delta_{tk}})}{e^{\beta_{sn}+\beta_{tn}}(e^{-\delta_{si}-\delta_{tk}}+e^{-\delta_{si}-\delta_{tl}}+e^{-\delta_{sj}-\delta_{tk}}+e^{-\delta_{sj}-\delta_{tl}})} \quad (27)$$
$$=\frac{e^{-\delta_{si}-\delta_{tk}}}{(e^{-\delta_{si}-\delta_{tk}}+e^{-\delta_{si}-\delta_{tl}}+e^{-\delta_{sj}-\delta_{tk}}+e^{-\delta_{sj}-\delta_{tl}})},$$

where items *i* and *j* belong to frame of reference *s*, and items *k* and *l* belong to frame of reference *t*. Clearly, Equation (27) is independent of the vector of person parameters (β_{sn},β_{tn}). The conditioning out of the person parameter in the case of frames of reference containing common persons but different items is shown in general in Appendix I. From this generalization of Equation (27) across more than four items, maximum likelihood estimates of the item parameters can be derived.

We have not yet implemented the conditional maximum likelihood estimation procedure for item estimates based on data from two or more frames to enable its use. It is however possible to apply the theory developed using existing software to estimate the relative magnitudes of the scale parameters and hence to estimate the item and person parameters in an arbitrary unit that is commensurate across the frames of reference.

In the example used earlier involving two items in each of two frames, information regarding the ratio of scale parameters was obtained from item estimates of the common items in the natural unit of each frame of reference, as shown in Equation (15). In the empirical example which follows, only *persons* are common across the two frames and so, analogously, information regarding the ratio of scale parameters derives from the person estimates in the natural unit of each frame. Provided the person estimates are expressed in terms of the same origin,

$$\beta_{sn}=\rho_s\beta_n^* \quad (28)$$

and so

$$\frac{\beta_{1n}}{\beta_{2n}} = \frac{\rho_1}{\rho_2}, \tag{29}$$

in which the person parameter in the arbitrary unit, β_n^*, is eliminated. Thus, as for the comparison between scale parameters based on items shown in Equation (15), the ratio between the scale parameters ρ_1 and ρ_2 should be independent of the particular person parameters instrumental for the comparison, in keeping with the requirement for invariant comparisons articulated by Rasch (1961).

From Equation (28) it follows that

$$V[\beta_s] = \rho_s^2 V[\beta^*]. \tag{30}$$

The item and person parameter estimates of the parameters β_{sn}, δ_{si} for each frame of reference are routinely available from software which implements estimations for the Rasch model where, within each frame of reference, the person parameters can be estimated from conditionally estimated item parameter estimates. These estimates $\hat{\beta}_{sn}$ are expressed in the natural unit of each frame of reference *s* and so the scale parameter is implicit within the estimates. From Equation (30)

$$\frac{\sqrt{V[\hat{\beta}_s]}}{\sqrt{V[\hat{\beta}_t]}} = \frac{\hat{\rho}_s}{\hat{\rho}_t}, \tag{31}$$

which gives an estimate of the relative scale values ρ_s, ρ_t with the constraint on the product of the parameters shown in Equation (24). It is necessary to have a reasonable number of well targeted items in order to achieve reasonably precise estimates of the variances to subsequently obtain estimates of the scale parameters using Equation (31).

From Equation (14) $\delta_i^* \equiv \delta_{si} / \rho_s$ and therefore given estimates $\hat{\rho}_s$ and $\hat{\delta}_{si}$, estimates $\hat{\delta}_i^*$ in an arbitrary unit common across frames of reference can also be obtained. Similarly, two sets of person parameter estimates in the arbitrary unit are also available once scale parameters have been estimated. To obtain a single estimate for each person given the set of estimates of item parameters $\delta_i^*, i = 1,...I$ and scale parameters $\rho_s, s = 1,...,S$, the person parameters $\beta_n^*, n = 1,...,N$ can be estimated by direct maximum likelihood estimation. It is readily shown that the solution equation for person estimates is

$$0 = \sum_s \sum_i \rho_s x_{sni} - \sum_s \sum_i \rho_s \left[\frac{\exp\left(\rho_s (\beta_n^* - \delta_i^*)\right)}{1 + \exp\left(\rho_s (\beta_n^* - \delta_i^*)\right)} \right] \quad n = 1,...,N \,. \qquad (32)$$

A weighted likelihood solution equation can be used instead of Equation (32) in order to minimize bias of the estimates (Samejima, 1993; Wang & Wang, 2001; Warm, 1989).

The estimates of the relative scale values, and therefore units, from Equation (31) are straightforward and consistent with traditional methods of equating for differences between units of scales. However, it is stressed that this paper shows *these estimates also arise from the requirements of invariance of comparisons for person and item parameters in the Rasch model within a frame of reference, with the complementary requirement of invariant comparison between scale parameters across frames represented by Equations (15) and (31).* Thus, the developments integrate traditional methods of scale equating into the conceptual framework of the Rasch model by making explicit the scale parameter in a manner that preserves sufficiency.

Although it is a substantial condition that the differences between the item locations across frames of reference must result only from different units, permitting differences between units nevertheless makes it possible to relax considerably what would otherwise be regarded as lack of conformity between data and the Rasch model. The Rasch model is generally applied within a single frame of reference with a single implied arbitrary unit. Thus, in the case where there are substantial differences between two sets of parameter estimates which reflect only differences between natural units, relevant tests of fit will indicate the data do not fit the Rasch

model within a single frame of reference when differences between natural units are not taken into account. As shown above, however, the same data do not necessarily violate the requirement of invariant comparisons in the Rasch model when differences between the natural units of different frames of reference *are* taken into account. As expected, if the differences between measurements result only from differences in the level of precision of measurement, then the differences do not violate the requirement of invariance of comparisons in the Rasch model provided person and item parameters are expressed in an arbitrary unit. These developments are now illustrated with an empirical investigation.

An illustrative example of different natural units between two classes of items

The data used in the empirical investigation were collected as part of the Western Australian Literacy and Numeracy Assessment (WALNA) population testing program in 2003. The program includes the administration of reading, writing, mathematics, and spelling assessments in years 3, 5, and 7 by classroom teachers based on detailed administrative instructions. The software used in the WALNA program for analyzing data is RUMM2020 (Andrich, Sheridan, & Luo, 1997-2005) which implements pairwise CML estimation (Andrich & Luo, 2003).

In 2003, common person equating was used to equate the difficulty of the 2003 and 2000 numeracy assessments. A group of 281 students completed both the 2000 and the 2003 tests which comprised 28 and 31 items respectively. Let $s = 0$ and $s = 3$ denote classification of items in terms of the separate 2000 and 2003 assessments, respectively. These classifications represent empirical differences such as differences between item developers, while other factors remained constant, such as the outcomes framework which provided the basis for constructing items. Table 7 shows the standard deviations for the subgroup of 281 common students analyzed separately within each frame of reference.

These standard deviations for the common students confirmed an initial observation from separate analyses of the *population* data in which the standard deviation of the 2003 *population* was approximately 1.2 times that of the 2000 population

Table 7: Estimates of the relative scale values of two sets of items from common persons

SD	Estimate	Parameter	Estimate
$\sqrt{V\left[\hat{\beta}_3\right]}$	1.044	$\hat{\rho}_3$	1.083
$\sqrt{V\left[\hat{\beta}_0\right]}$	0.890	$\hat{\rho}_0$	0.923

In order to limit the effects of measurement error on the estimated standard deviation of the person locations shown in Table 7, the mean squared standard error of each set of estimates was subtracted from the variance of the estimates (Andrich, 1982). The disattentuated correlation between the two sets of person estimates was 0.89, indicating a reasonably close concordance of the underlying trait measured by the two sets of items. Thus, the evidence suggests the difference between the dispersions of person estimates arises from different units of scale between items on the two assessments. The two assessments were constructed by different item developers and there were refinements to the process of item development, trials of items and so forth, which may have contributed to the greater scale value, and therefore greater precision, for items on the 2003 assessment compared with those on the 2000 assessment.

As indicated earlier, tests of fit should also be used which are sensitive to scale parameters, and hence to differences between the slopes of ICCs. Accordingly, an index of fit was computed based on analysis of the data using the RM and FRM. This index is defined as

$$Y_s = \frac{\sum_{i\in s}\sum_{n}\left(z_{ni}^2 - F_{sni}\right)}{\sqrt{V\left[\sum_{i\in s}\sum_{n} z_{sni}^2\right]}}, \qquad (33)$$

where $z_{sni} = \left((x_{sni} - E[X_{sni}])/\sqrt{V[X_{sni}]}\right.$ is the standardized residual of the response of person *n* to item *i* in the frame of reference *s*, and F_{sni} is the approximate degree of freedom per element of the data matrix as described for example in Andrich (1988).

The expected value of Y_s is 0 and a *negative* value implies that response data are closer to a Guttman structure (Guttman, 1950, 1954) than expected, that is, the item set *over discriminates*, whereas a *positive* value implies a response pattern that is more erratic than expected, that is the item set *under discriminates*. In general, therefore, this type of fit statistic provides an index which is sensitive to differences between the empirical slopes of ICCs not accounted for by a given model. It is stressed that, in this paper, this statistic is used as a *relative index* of fit and not referenced to a formal statistical distribution.

The values of Y_s of Equation (33) for each frame of reference and for analyses according to both the RM and the FRM are shown in Table 8. These show two features. First when the RM is applied, Y_3 is negative and Y_0 is positive, indicating that items in $\mathbf{F}_3$ tend to over discriminate and those in $\mathbf{F}_0$ tend to under discriminate relative to one another. This is consistent with the conclusion from the relative values of the standard deviations shown in Table 7. Second, Table 8 shows that using the FRM, the absolute value of the residuals is smaller in both sets of items than when the RM is used. This indicates that the different levels of discrimination between the frames of reference, which are not accounted for in the RM, are largely accounted for in the FRM.

Table 8: Item fit residuals for two items sets from common persons obtained using the RM and FRM

	Y_3	Y_0
RM (single frame of reference)	-1.28	1.79
FRM	0.32	0.57

Although the results shown in Table 8 indicate that the systematic difference between the scale parameter values across the two frames of reference was largely accounted for by applying the FRM, the fit residuals for individual items in some cases suggested departure from the model. These indicate departures from the FRM in the same way that differential levels of discrimination within a single frame of reference indicate departure from the RM. The key point, however, is that a substantial

improvement in fit of data to the model was achieved by accounting for differences in units without sacrificing sufficiency.

Even more importantly, in the applied situation, equating for differences between the natural units of the Year 2000 and Year 2003 assessments using the FRM enabled expression of person locations in an arbitrary unit, making the estimates considerably more comparable. This carried important implications for reporting percentages of students achieving criterion-referenced standards (Humphry, 2005).

Discussion

In order for measurements obtained from different specified frames of reference to be comparable, they must either be estimated relative to a common unit and origin, or expressed in such terms by accounting for differences between the units and origins of different scales. Various empirical factors have the potential to influence levels of discrimination, and therefore precision, including person characteristics, item characteristics, and environmental conditions. In this paper, we made explicit the natural unit of each frame of reference, and subsequently defined the scale parameter as the ratio between an arbitrary unit and natural unit. This enabled us to establish principles for investigating and accounting for the influence of empirical factors on the natural unit of a frame of reference, while preserving statistical sufficiency.

Given that assessment data are generally produced by the interaction of persons with items, it is particularly important to consider the influences on persons and items of a frame of reference on the level of precision, or discrimination. In addition, however, the FRM can also be used to investigate the influence of any key empirical factor on discrimination through experimental manipulation of the relevant factor, such as an environmental assessment condition, combined with control of common elements, such as the type of assessment items.

The empirical investigation presented above was chosen to highlight key points. Importantly, it showed that it was possible to *improve fit of data to the frame of reference Rasch model without sacrificing statistical sufficiency*. In particular, the empirical investigation examined the influence of different test developers at different

times on the natural unit, and therefore the level of precision, of two different frames of reference containing sets of items constructed by the different developers.

The example also illustrates the key difference between the FRM and 2PLM which has been described earlier in the comparison between the models. In particular, in the 2PLM each item has a different discrimination whereas in the FRM discrimination is parameterized in relation to an empirical factor; in this case the item developers and time of item development. Thus in the 2PLM the differences between discriminations of items would be considered properties integral to each item, rather than properties of an empirical factor manifested in particular circumstances in a set of items. In the FRM, empirical factors may also manifest *differently in different circumstances for the same items*.

Because this paper is concerned with fundamental principles and their illustration, the simple method of comparing standard deviations of estimates was used to estimate scale parameters. It is worth noting, however, that various authors following Stocking & Lord (1983) have provided evidence, in the context of item response theory, that the "mean and sigma" equating method is inferior to methods such as characteristic curve methods because the latter utilize more available information. There are various possible approaches to deriving more refined estimates, including a combination of conditional maximum likelihood and direct maximum likelihood. Further work is required to investigate these technical issues and perhaps this paper will stimulate such research.

Conclusion

The FRM represents an extension of the model and framework developed by Rasch (1960, 1961, 1977). This extension makes it is possible to parameterize discrimination, without destroying sufficiency, in assessment contexts involving two or more frames of reference defined in terms of empirical factors or conditions. In addition, the FRM provides a basis for making invariant comparisons between scale parameters of different frames of reference. Through these extensions of the model and framework developed by Rasch, principles are established for maintaining a common unit of scale across multiple frames of reference. Consequently, traditional methods of scale equating are also integrated into the Rasch model and framework.

The FRM therefore broadens the foundation for social and psychological measurement while preserving the distinguishing property of the class of measurement models identified by Rasch.

Appendix I: General conditional likelihood expression for the FRM in the case that different frames of reference have different items.

From Equation (16), the likelihood of $\mathbf{x}_n = (x_{sn1},...,x_{sni},...,x_{SNI})$ is given by

$$\Pr\{\mathbf{x}_n \mid \beta_{sn}, \delta_{si}\} = \prod_s \prod_i \left[\frac{\exp\left(x_{sni}(\beta_{sn} - \delta_{si})\right)}{\left(1 + \exp\left((\beta_{sn} - \delta_{si})\right)\right)} \right]$$

$$= \frac{\exp\left(\sum_s r_{sn} \beta_{sn} \right) \exp\left(- \sum_s \sum_i x_{sni} \delta_{si} \right)}{\prod_s \prod_i \left(1 + \exp\left((\beta_{sn} - \delta_{si})\right)\right)}, \qquad \text{(A1)}$$

where $\mathbf{x}_n$ is the vector of responses of person *n* across items *i* characterised by their membership to frame of reference *s*, and the items may be different in different frames of reference and there may be different numbers of items in each frame of reference.

Thus, it follows that

$$\Pr\{\mathbf{x}_n \mid \mathbf{r}_n; \beta_{sn}, \delta_{si}\} = \frac{\dfrac{\exp\left(\sum_s r_{sn} \beta_{sn} \right) \exp\left(- \sum_s \sum_i x_{sni} \delta_{si} \right)}{\prod_s \prod_i \left(1 + \exp\left(\beta_{kn} - \delta_{ki}\right)\right)}}{\dfrac{\exp\left(\sum_s r_{sn} \beta_{kn} \right) \sum_{(\mathbf{x})\mid\mathbf{r}} \exp\left(- \sum_s \sum_i \delta_{si} \right)}{\prod_s \prod_i \left(1 + \exp\left(\beta_{sn} - \delta_{si}\right)\right)}}$$

$$= \frac{\exp\left(- \sum_s \sum_i x_{sni} \delta_{si} \right)}{\sum_{(\mathbf{x})\mid\mathbf{r}} \exp\left(- \sum_s \sum_i \delta_{si} \right)}, \qquad \text{(A2)}$$

where $r_{ns} = \sum_{i \in s} x_{sni}$, and in which the vector of person parameters is eliminated by partitioning the response space in terms of the score vector $\mathbf{r}_n = (r_{n1},...,r_{nS})$

Appendix II: The polytomous Rasch model

Andersen studied a form of Rasch's (1961) class of models for multiple response categories, in which the probability that person *n* responds in category *p* on item *i* is

$$\Pr\{X_{ni}^{(p)}\} = \frac{\exp(\theta_{np} - \psi_{ip})}{\gamma_{ni}} \ , \qquad \text{(A3)}$$

where θ_{nq} and ψ_{iq} are person and item parameters pertaining specifically to category *q* of a total of *Q* categories, and $\gamma_{ni} \equiv \sum_q^Q \exp(\theta_{nq} - \psi_{iq})$. Let the event of individual *n* responding in category *p* of item *i* be represented as a vector $\mathbf{x}_{ni} = (x_{ni}^{(1)},...,x_{ni}^{(q)},...,x_{ni}^{(Q)})$ where $x_{ni}^{(p)} = 1$ and $x_{ni}^{(q)} = 0$ for $q \neq p$ (Rasch, 1961; Andersen, 1973, 1977). Let θ_{nq}, $q = 1,..,Q$ be a vector of person parameters for person *n* associated with the categories. Also, let $t_{nq} = \sum_{i=1}^{I} x_{ni}^{(q)}$ such that $\mathbf{t}_n = (t_{n1},..,t_{nq},..t_{nQ})$ is a vector of category scores for person *n*. Andersen (1973) showed in general that this score vector is sufficient for the vector of person parameters. Consider, for example, the conditional probability that person *n* responds in category 2 of item 1 given one response in category 2 and one response in category 3 across two items; i.e. given $t_{n1} = 0$, $t_{n2} = 1$, and $t_{n3} = 1$. This conditional probability is

$$\begin{aligned}
\Pr\{X_{n1}^{(3)} \,\big|\, \mathbf{t}_n = (0,1,1); \theta_{nq}, \psi_{iq}\} &= \frac{\dfrac{\exp(\theta_{n3} - \psi_{13})}{\gamma_{1n}} \dfrac{\exp(\theta_{n2} - \psi_{22})}{\gamma_{2n}}}{\dfrac{\exp(\theta_{n3} - \psi_{13})}{\gamma_{1n}} \dfrac{\exp(\theta_{n2} - \psi_{22})}{\gamma_{2n}} + \dfrac{\exp(\theta_{n2} - \psi_{12})}{\gamma_{1n}} \dfrac{\exp(\theta_{n3} - \psi_{23})}{\gamma_{2n}}} \\
&= \frac{\exp(\theta_{n3} + \theta_{n2}) \exp(-\psi_{22} - \psi_{13})}{\exp(\theta_{n3} + \theta_{n2}) + \exp(-\psi_{13} - \psi_{22}) + \exp(-\psi_{12} - \psi_{23})} \\
&= \frac{\exp(-\psi_{22} - \psi_{13})}{\exp(-\psi_{13} - \psi_{22}) + \exp(-\psi_{12} - \psi_{23})} . \qquad \text{(A4)}
\end{aligned}$$

Although the model of Equation (A3) pertains to items with multiple categories, it can be seen that this conditional equation has the same basic form as Equation (23).

References

Andersen, E.B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 26, 31-44.

Andersen, E.B. (1977). Sufficient statistics and latent trait models, *Psychometrika*, 42, 69-81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 357-74.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, 9, 95-104.

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage Publications.

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct-responses and pairwise preferences. *Applied Psychological Measurement, 19*(3), 269-290.

Andrich, D. (2003). On the distribution of measurements in units that are not arbitrary. *Social Science Information*, 42, 557-589.

Andrich, D. (2004). Controversy and the Rasch Model: a characteristic of incompatible paradigms? *Medical Care*, 42, 7-16.

Andrich, D. & Luo, G. (2003). Conditional Pairwise estimation in the Rasch model for ordered response categories using principle components. *Journal of Applied Measurement*, 4, 205-221.

Andrich, D., Sheridan, B. & Luo, G. (1997-2005). *RUMM2020*. RUMM Laboratory, Perth, Australia.

Baker, F. (1983). Comparison of ability metrics obtained under two latent trait theory procedures. *Applied Psychological Measurement*, 7(1), 97-110.

Baker, F. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement*, 8(3), 261-271.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Brink, N.E. (1971). Effect of item discrimination in the Rasch model. *Proceedings of the Annual Convention of the American Psychological Association*, 6(1), pp. 101-102.

Embretson, S. and Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fischer, G. H. (1994). Derivations of the Rasch model. In Fischer, G. H. and Molenaar, I.W. (1995) (Eds.) Rasch models: foundations, recent developments, and applications. New York: Springer. (pp. 15 – 38)

Guttman, L. (1950). The problem of attitude and opinion measurements. In S.A. Stouffler et al. (Eds.), *Measurement and prediction*. New York: Wiley.

Guttman, L. (1954). The principal components of scalable attitudes. In P.F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. New York: Free Press.

Humphry, S.M. (2005). *Maintaining a common arbitrary unit in social measurement*. Ph.D. Thesis: http://wwwlib.murdoch.edu.au/adt/browse/view/adt-MU20050830.95143

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Luo, G. (1998). A general formulation for unidimensional unfolding and pairwise preference models: Making explicit the latitude of acceptance. *Journal of Mathematical Psychology*, *42*, 400 – 417.

Michell, J. (1986). Measurement Scales and Statistics: A Clash of Paradigms. *Psychological Bulletin*, 100, 398-407.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, 355-383.

Michell, J. (1999). *Measurement in Psychology*. Cambridge: Cambridge University Press.

Klauer, K.C. (1995). The assessment of person fit. In G.H. Fischer and I.W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments and applications* (pp. 97-110). New York: Springer-Verlag.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321-334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*. Berkeley: University of Chicago Press, 1980.

Rasch, G. (1966). An individualistic approach to item analysis. In P.F. Lazarsfeld and N.W. Henry, (Eds.). *Readings in Mathematical Social Science (*pp.89-108). Chicago: Science Research Associates.

Rasch, G. (1977). On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, 14, 58-93.

Rost, J., Carstensen, C. & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost and R. Langeheine (Eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences.* Waxmann Verlag GMBH: Münster and New York, pp. 324 - 332.

Samejima, F. (1993). An approximation of the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, 58 (1), 119-138.

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201-210.

Verhelst, N.D. & Glas, C.A.A. (1995). The One Parameter Logistic Model. In G.H. Fischer and I.W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments and applications* (pp. 215-237). New York: Springer-Verlag.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, 25(4), 317-331

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427-450.

Wood, R. (1978). Fitting the Rasch model – A heady tale. *British Journal of Mathematical and Statistical Psychology*, 31, 27-32.

Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.