Maintaining a Common Unit in Social Measurement

Steve Humphry

University of Western Australia

(Email: Stephen.humphry@uwa.edu.au)

Abstract

The purpose of this paper is to present a framework and model within which it is possible to investigate and account for the influence of empirical factors on the unit of a scale. The framework is an extension of that developed by Rasch. The concept of discrimination is explicitly defined, and it is shown that this definition implies discrimination is a scale parameter pertaining to empirical characteristics of the frame of reference for measurement; i.e. the assessment context. The implications of this definition for the interpretation of the logit are made explicit. Implications of the developments for maintaining a common unit of scale are illustrated in terms of the Western Australia Literacy and Numeracy Assessment (WALNA) program. An overview of background theoretical developments is presented. In particular, it is shown that it is possible to extend the Rasch model to incorporate a scale factor while preserving the distinguishing feature of the model; namely sufficiency. Key features of the process of applying the extended form of the model are presented, and relevant issues considered. Implications, including those for future work, are also briefly discussed.

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

Introduction

Differences between units of scales carry important implications for quantitative research if they are not accounted for within the approach to measurement adopted by the researcher. The role of the unit in social and psychological measurement is, however, generally left implicit, and although a relationship between discrimination and the unit of a scale is often noted (eg. Brink, 1971; Wood, 1978; Andrich, 1988; Embretson & Reise, 2000), the relationship has not been formulated in a manner that permits productive research into the influence of empirical factors on the unit of scale. The objective here is, accordingly, to formulate this relationship and to develop a framework which provides a basis for investigating and accounting for the influences of empirical factors on the unit of a scale. Preliminary empirical investigations are outlined which illustrate applications and implications of the approach.

In the physical sciences, the magnitude of the unit of a measuring instrument is generally defined relative to a standard, such as one of the *Système International d'Unités* (SI units). Instruments designed to measure physical quantities are deliberately constructed and used under controlled conditions in order to measure in terms of particular units. In the social sciences, it is difficult to deliberately construct instruments in such a manner as to influence the unit in terms of which measurements are made as is the case in the physical sciences. Nevertheless, key features of the empirical context, including the manner in which instruments are constructed, are also likely to influence the magnitudes of units. For this reason, the framework developed in this paper recognizes the influence of empirical factors and conditions on the magnitudes of units of scales. A key motivation is to build a foundation for acquiring knowledge regarding empirical factors which must be controlled in order to obtain invariant units of scales. The framework is developed from the Rasch model (RM) for dichotomous response data and surrounding conceptual framework, which were explicitly founded upon criteria for measurement deduced from analysis of measurement in the physical sciences (Rasch, 1960/1980, 1961, 1977). Accordingly, a central criterion is that the distinguishing property of the RM is preserved; namely, sufficient statistics (Andersen, 1977) exist for person and item parameters within any specified empirical context.

The RM shares a similar structure with Birnbaum's (1968) Two Parameter Logistic Model (2PLM). Both models contain an item parameter intended to represent the location of the item on a latent continuum. In addition, the 2PLM features a discrimination parameter for every item, the magnitudes of which are estimated. The estimation of *magnitudes* of item discrimination implies that the parameters are in some sense treated as *quantitative* terms in the 2PLM. However item discrimination parameters do not represent levels of the latent trait being measured; rather they pertain to other item factors which influence the degree of discrimination obtained between levels of the trait.

The RM is often regarded as a specific case of the 2PLM in which the discrimination parameter is uniform for all items (e.g. Maris & Bechger, 2005). This perspective implies, however, that discrimination is related only to item factors, whereas the RM requires discrimination to be uniform given all relevant empirical factors associated with the context for measurement. In this paper, discrimination parameters are formulated in quantitative terms by making explicit the connection between discrimination and the unit of a scale. The approach is developed explicitly so that discrimination may be parameterized more generally for a range of empirical factors, including conditions under which an assessment is administered, item factors, person factors, and features of a specified assessment context. The framework allows for classifications $k = 1,...,K$ of any given empirical factor, and is designed to permit evaluation of the hypothesis that any given classification has a *fixed* level of influence on discrimination. If *degree of discrimination* is viewed as a dependent variable, the classifications are somewhat analogous to levels of an empirical factor in the context of Analysis of Variance. The motivations for proceeding in this way are to obtain a fixed unit of scale and to preserve sufficiency.

Although the Rasch model has no explicit discrimination parameter, Rasch (1960/1980, p. 121) had identified what he referred to as a general form of a measuring function with the inclusion of two *constants*, commenting 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". With the constants incorporated, Rasch's model for dichotomous response data becomes

$$\Pr\{X_{in} = 1\} = \frac{\gamma \exp(\rho(\beta_n - \delta_i))}{1 + \gamma \exp(\rho(\beta_n - \delta_i))}. \qquad (1)$$

The RM is usually written as

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

which is the same model as Equation (1), with γ and ρ implicitly defined as 1. Because the term ρ can be arbitrarily specified as a choice of unit, it follows that specifying different values of ρ amounts to the choice of different units. This consequence will be exploited, shortly, in order to develop a basis for accounting for differences between units that arise from empirical factors by treating ρ as a parameter pertaining to a particular classification of an empirical factor. In so doing, the connection between ρ and the unit of a scale will be made explicit.

Multiple frames of reference

Before introducing notation that allows ρ to be used as a basis for accounting for differences between units, it is useful to consider Rasch's (1977) conceptualisation of a *Specified Frame of Reference*. Rasch (1977) defined a frame of reference $\mathbf{F} \equiv [\mathbf{O},\mathbf{A},\mathbf{X}]$ in terms of a collection of objects $\mathbf{O}$, a collection of agents $\mathbf{A}$, and the set of outcomes $\mathbf{X}$ arising from the interaction between $\mathbf{O}$ and $\mathbf{A}$. The RM finds its application within Specified Frames of Reference in which, for example, the agents may be items on an assessment or questionnaire, the objects a class of persons, and the outcomes response data arising from the interaction between persons and items under specified assessment conditions.

In this paper, the focus is on contexts in which measurements are obtained from *multiple* frames of reference constructed with the intention of measuring a *common* latent trait. In order to distinguish between separate frames of reference, a given Specified Frame of Reference must either be defined in terms of an empirical factor which can be classified $k = 1,2,..,K$, or in terms of interactions between classifications of different empirical factors. Examples of such factors are well-defined environmental conditions for an assessment, time available for completion of items and the mode of delivery of an assessment. Person and item characteristics constitute important *special cases* of empirical factors. For example, in the first of the empirical investigations presented later, persons are classified according to school year group, and in the second, items are classified according to their membership to assessment forms constructed by different item developers at different times.

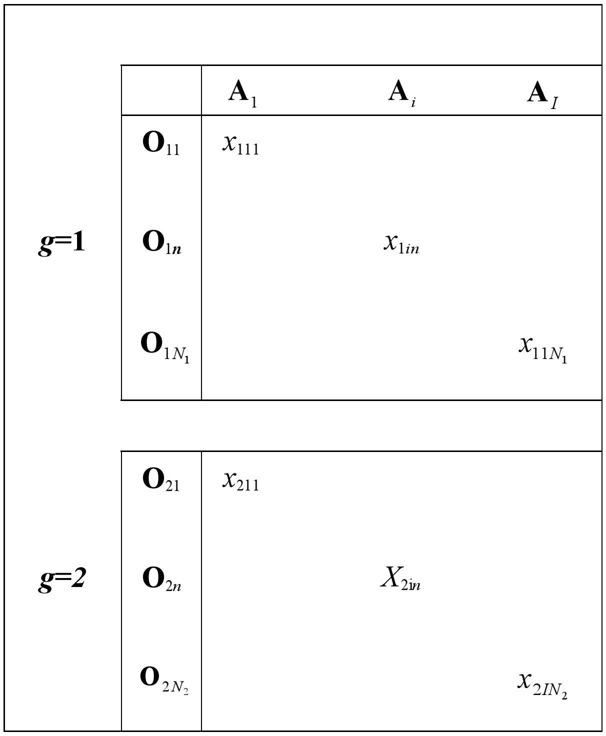
Given that person and item factors are the focus of the empirical investigations, it is instructive to consider cases in which Specified Frames of Reference are defined in terms of classifications of such factors. Accordingly, an assessment context comprising two Specified Frames of Reference, and involving two collections of collections of items, is portrayed in Figure 1. Items are classified according to an item factor with classifications $s = 1,...,S$ with $S = 2$ in this example.

	s = 1			**s = 2**		
	$\mathbf{A}_{11}$	$\mathbf{A}_{1i}$	$\mathbf{A}_{1I_1}$	$\mathbf{A}_{21}$	$\mathbf{A}_{2i}$	$\mathbf{A}_{1I_s}$
$\mathbf{O}_1$	x_{111}			x_{211}		
$\mathbf{O}_n$		x_{1in}			x_{2in}	
$\mathbf{O}_N$			x_{1I_1N}			x_{2I_2N}

Figure 1: Multiple frames of reference defined in terms of an item factor

A given frame of reference as shown in Figure 1 is denoted $\mathbf{F}_s$. The collection of items categorized in terms of classification s of the item factor contains a total of I_s items. The matrix of response data obtained within a particular frame of reference is denoted $\mathbf{X}_s$, and individual responses or outcomes are denoted x_{sin}.

In Figure 2, frames of reference are defined in terms of classifications of a person factor. A given frame is denoted $\mathbf{F}_g$ and individual responses are denoted x_{gin}. Notice the same collection of persons is contained within the two frames shown in Figure 1 and the same collection of items is contained within the two frames shown in Figure 2. As will become evident, a pair of frames must share either common persons or items in order to compare units and origins.



		$\mathbf{A}_1$	$\mathbf{A}_i$	$\mathbf{A}_I$
	$\mathbf{O}_{11}$	x_{111}		
***g*=1**	$\mathbf{O}_{1n}$		x_{1in}	
	$\mathbf{O}_{1N_1}$			x_{11N_1}
	$\mathbf{O}_{21}$	x_{211}		
***g*=2**	$\mathbf{O}_{2n}$		X_{2in}	
	$\mathbf{O}_{2N_2}$			x_{2IN_2}

Figure 2: Multiple frames of reference defined in terms of a person factor

Signifying natural units within parameters

The notation used for the parameters of models such as the RM and 2PLM implicitly takes for granted that parameters are expressed in terms of a single and common unit. In order to investigate the influence of empirical factors on the unit of a scale, it is necessary to develop a notation which recognizes differences between units and origins of scales obtained from different Specified Frames of Reference. In this paper, the term *parameter* refers, in relation to individual persons and items, to the *measure* of the level of a trait or ability *in terms of a specific unit and origin*, where the term *measure* "is reserved for the theoretical value of the object of measurement, of which the measurements … are estimates" (Andrich, 2003). Thus, it is necessary to develop notation which recognizes that measures of individual persons and items may be estimated in terms of different units.

To facilitate the exposition of notation employed in this paper, we will begin with a tangible physical analogy in which units can be clearly defined in terms of the empirical structure of measuring instruments and the specific conditions under which the instrument is used. Thus, consider a situation in which an experimenter measures the masses of a common set of objects using two instruments, one of which is designed to measure in pounds, the other of which is designed to measure in kilograms. Accordingly, let $k = 1$ represent the empirical conditions necessary in order to measure the mass of an object in pounds, including the empirical structure of the instrument itself and the conditions under which the instrument must be used. Similarly, let $k = 2$ denote conditions under which measurements are obtained in kilograms. Following from this, let ξ_{kv} be the *measure* of object *v* within a frame of reference defined in terms of the empirical factor *k*. Andrich (2003) refers to such a unit, which is *integral* to a measuring instrument, as a *natural unit*, and distinguishes this unit from an *arbitrary unit*, the size of which is unrelated to empirical factors. Hence, the first subscript in ξ_{kv} signifies that measurements are obtained in terms of a *natural unit* obtained under empirical conditions classified *k*. In the example above $\xi_{1v} \cong 2.2\xi_{2v}$; i.e. the measure of the mass of an object in pounds is approximately 2.2 times the measure of the mass of the same object in kilograms.

A collection of frames of reference which provide for measurements of the same quantitative attribute or trait is referred to as an *Extended Frame of Reference*. Due to the commonality of the quantitative trait it is, in principle, possible in such a context to express measurements in a *common unit* of the trait. That is, it is possible to express measurement in terms of an interval of a fixed size irrespective of the Specified Frame of Reference from which measurements are obtained. In order to express measurements obtained from more than one frame of reference in terms of the same unit, it is necessary to make an arbitrary choice of a common unit. In this paper, the superscript * is incorporated in parameters to signify that measurements obtained from a collection of Specified Frames of Reference are expressed in a *common arbitrary unit*. To continue with the example from the preceding paragraph, suppose that the kilogram is chosen as the arbitrary unit. It then follows that $\xi_{2v} \equiv \xi_v^*$ and $\xi_{1v} \cong 2.2\xi_v^*$.

In the present context, it is useful to allow for differences between both the units and origins associated with Specified Frames of Reference, given that different origins may result from constraints imposed during estimations. Employing the notation introduced above, let β_{kn} denote the measure of person *n*, in which the subscript *k* signifies that the magnitude of the trait is expressed in terms of a unit which depends on empirical factors associated with $\mathbf{F}_k$ relative to an origin determined by estimation constraints imposed on parameters pertaining to $\mathbf{F}_k$. Similarly, let δ_{ki} be the measure of the level of trait of item *i* expressed in terms of the same unit and origin. In the context of multiple frames of reference, and expressed in terms of this notation, the RM in the form of Equation (1a) becomes

$$\Pr\{X_{kin} = 1\} = \frac{\exp(\beta_{kn} - \delta_{ki})}{1 + \exp(\beta_{kn} - \delta_{ki})}. \qquad (2)$$

Equation (2) is referred to as the *Specified RM* (SRM). The SRM permits different locations for any given person or item within separate frames of reference when the units or origins associated with those frames differ.

As mentioned earlier, because ρ in Equation (1) can be considered an arbitrary scaling constant, it is an obvious candidate for dealing with differences between units that arise between assessment frames of reference. Accordingly, let ρ_k denote a constant which is specific to the Specified Frame of Reference defined in terms of classification k of a well-defined empirical factor. The objective is to develop an approach in which it is possible to account for differences between units of scales by estimating parameters in terms of a *common* unit and origin; i.e. the same unit and origin irrespective of the frame of reference from which estimates are obtained. As the next step toward achieving this objective, we will define the relationship between parameters expressed in natural units and parameters expressed in arbitrary units as follows:

$$\beta_{kn} \equiv \rho_k\left(\beta_n^* - c_k^*\right) \text{ and} \tag{3}$$

$$\delta_{ki} \equiv \rho_k\left(\delta_i^* - c_k^*\right) \ , \tag{4}$$

in which c_k^* is the location of an origin specific to the frame of reference $\mathbf{F}_k$, and β_n^* and δ_i^* are person and item parameters. As before, the superscript * signifies that the locations of the parameters and origin of a Specified Frame of Reference are expressed in terms of a common arbitrary unit relative to a common origin across multiple frames of reference. Stated in these terms, the model takes the form

$$\Pr\{X_{kin} = 1\} = \frac{\exp\left(\rho_k(\beta_n^* - \delta_i^*)\right)}{1 + \exp\left(\rho_k(\beta_n^* - \delta_i^*)\right)}. \tag{5}$$

in which ρ_k pertains to a particular Specified Frame of Reference whereas β_n^* and δ_i^* are *invariant* across frames of reference because they are by definition expressed in terms of a common unit and origin. Equation (5) is referred to as the *Extended Frame of Reference Model* (EFRM) because the purpose of its application is to extend the frame of reference of measurement. Of particular importance in terms of preserving sufficiency, the EFRM possesses the same fundamental structure as

Rasch's general form of a measuring function for dichotomous data, Equation (1), where $\gamma = 1$.

Discrimination and the magnitude of the natural unit

It was mentioned earlier that the relationship between discrimination and the unit of a scale is often implicitly recognised. 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 an ability estimate is "always scaled by a factor $D\bar{a}$", where "$\bar{a}$ is the common level of discrimination for all items", and *D* is a constant. In addition, the potential for person characteristics to influence discrimination and the unit of a scale has also been 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. Andrich (1988, p. 75) also discussed the possibility that person characteristics might influence discrimination, and therefore the unit of a scale, in the process of examining item scale values for Eysenck's (1958) Neuroticism questionnaire.

In order to formulate a rational definition which relates discrimination to the unit, we will firstly introduce b_k to represent the *natural unit*. This is the *interval* on the relevant latent continuum which is the unit of $\mathbf{F}_k$ when data obtained within that frame of reference accord with Equation (2), the SRM. The symbol b_k is employed in the same essential manner that, for example, the abbreviation g is used to represent a specific mass of a specific type of quantitative property; i.e. one gram of mass. There is, however, an important difference, which is that that the introduction of b_k does not imply that this unit is a *standard* relative to which measurements can be reproduced by others in other contexts; rather b_k is intended only to represent the magnitude of the unit of a latent trait associated with a Specified Frame of Reference, whatever that magnitude may be given relevant empirical factors. The *arbitrary unit* is then denoted without subscripts, as b, because its magnitude is unrelated to empirical factors.

The level of discrimination associated with a frame of reference $\mathbf{F}_k$ is defined as

$$\rho_k \equiv \frac{\mathrm{b}}{\mathrm{b}_k}. \tag{6}$$

Consider, for example, an Extended Frame of Reference within which the levels of discrimination obtained within two frames $\mathbf{F}_1$ and $\mathbf{F}_2$ are compared directly. From Equation (6), it follows that $\rho_1 / \rho_2 \equiv \mathrm{b}_2 / \mathrm{b}_1$. In general, the *ratio of levels of discrimination is inversely proportional to the ratio of the sizes of the natural units of the frames of reference*.

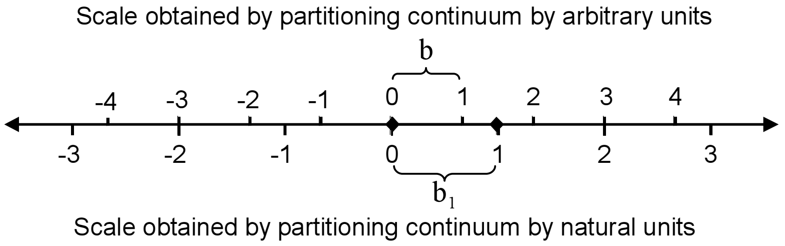

Scale obtained by partitioning continuum by arbitrary units
b
-4
-3
-2
-1
0
1
2
3
4
-3
-2
-1
0
1
2
3
b_1
Scale obtained by partitioning continuum by natural units

Figure 3: A latent continuum partitioned into natural and arbitrary units

An example is shown in Figure 3 in which the size of the natural unit of a frame $\mathbf{F}_1$ is 1.5 times the size of the arbitrary unit and hence $\rho_1 = \mathrm{b} / \mathrm{b}_1 = 0.\dot{6}$. It can be seen according to this definition of discrimination, the measure of any given interval on the common latent continuum in natural units is ρ_k times the measure of the same interval in arbitrary units. For example, let $b_k^{(k)} \equiv 1$ be the measure of the natural unit on a scale partitioned by the natural unit itself, and let b_k^* be the measure of the natural unit on a scale that is partitioned by arbitrary units. It follows that $b_k^{(k)} = \rho_k b_k^*$ and hence $\rho_k = 1/b_k^*$. Similarly, the measure in natural units of an interval between any given person *n* and item *i* is $\beta_{kn} - \delta_{ki} = \rho_k \left(\beta_n^* - \delta_i^*\right)$, consistent with Equations (3) and (4) which define the relationship between the parameters. The measure in arbitrary units of the same interval is thus $\beta_n^* - \delta_i^* = (\beta_n^{(k)} - \delta_i^{(k)}) / \rho_k$.

Importantly in terms of the present objectives, the discrimination parameter in Equation (6) is defined in quantitative terms, for it is defined as a *ratio between two intervals on a single latent continuum*. This definition therefore provides a clear justification for estimating the relative magnitudes of levels of discrimination obtained in the presence of different classifications of empirical factors. Consequently, the definition of discrimination given in Equation (6) also makes precise the sense in which item discrimination can be conceived as a quantitative term in the 2PLM. When interpreted in terms of the framework developed here, the 2PLM is a special case of Equation (5) in which an item factor has a different classification for every item. By defining discrimination explicitly, the nature of the interdependence between discrimination parameters and individual person and item parameters is also made explicit. The structure of the 2PLM implies that every item defines a separate Specified Frame of Reference each of which potentially has a different natural unit; a consequence which presents problems to be highlighted later.

Clearly, the RM is also a special case of Equation (5) in which there is an implicit requirement that the influence of any empirical factor on discrimination is uniform within the Specified Frame of Reference in which the model is applied. Accordingly, within a single frame of reference ρ_k can be arbitrarily specified as any value. As shown in the following section, however, it is possible to estimate relative levels of discrimination across frames of reference without introducing either assumptions regarding the distribution of any set of parameters or unnecessary constraints on parameters, provided a fixed level of discrimination can be obtained within each Specified Frame of Reference. The fundamental reason for this is that comparisons between discrimination parameters can be obtained from *ratios of differences* between location estimates in natural units as shown in Figure 2.

Estimating relative levels of discrimination and equating scales

Given the dependence of *relative* magnitudes of the natural units on empirical factors in terms of which Specified Frames of Reference are defined, the specific magnitudes of all natural units are determined by the choice of an arbitrary constraint. For instance, if $\rho_1 / \rho_2 = 1.2$, it is equally justifiable to specify $\rho_1 = 1.2$ and $\rho_2 = 1$ as it is to specify $\rho_1 = 2.4$ and $\rho_2 = 2$. Each constraint will result in the arbitrary unit being a different interval on the latent continuum. A simple and useful constraint which can be applied across frames of reference as a choice of the arbitrary unit is

$$\prod_k^K \rho_k = 1 . \qquad (7)$$

Values of ρ_k, $k = 1,...,K$ can be estimated given such an arbitrary constraint provided common persons or items are contained within the relevant frames of reference. In the first of the empirical investigations which follow, the influence of a person factor on discrimination is investigated where common items are administered to students in different school year groups. Thus, to focus on a case in which frames of reference are defined in terms of a person factor, from Equation (4) it can be seen that when there are two classifications *g* and *h* of the person factor,

$$\frac{\mathrm{V}\left[\delta_{g.}\right]}{\mathrm{V}\left[\delta_{g.}\right]} = \frac{\mathrm{V}\left[\rho_g \delta^*_{.}\right]}{\mathrm{V}\left[\rho_h \delta^*_{.}\right]} = \frac{\rho_g{}^2\,\mathrm{V}\left[\delta^*_{.}\right]}{\rho_h{}^2\,\mathrm{V}\left[\delta^*_{.}\right]},$$

$$\Rightarrow \sqrt{\frac{\mathrm{V}\left[\delta_{g.}\right]}{\mathrm{V}\left[\delta_{h.}\right]}} = \frac{\rho_g}{\rho_h} = \frac{\mathrm{b}_h}{\mathrm{b}_g}. \qquad (8)$$

An estimate of the ratio of levels of discrimination is therefore obtained from the ratio of the standard deviation of the item estimates obtained from one Specified Frame of Reference to the standard deviation of the estimates for the same items obtained from a second Specified Frame of Reference.

The formal symmetry of the model means it is also possible to compare levels of discrimination based on the estimated locations of common *persons* obtained within when two frames of reference contain common persons but no common items. For example, the comparison between the levels of discrimination associated with of the Specified Frames of Reference $\mathbf{F}_s$ and $\mathbf{F}_t$, defined in terms of classifications of an item factor, can be obtained as follows:

$$\sqrt{\frac{\mathrm{V}\left[\beta_{s.}\right]}{\mathrm{V}\left[\beta_{t.}\right]}} = \frac{\rho_s}{\rho_t} = \frac{\mathrm{b}_t}{\mathrm{b}_s}. \qquad (9)$$

This is analogous to taking ratios of standard deviations of the measurements of a common set of objects in two natural units of measuring instruments, such as the pound and kilogram, in the physical sciences. Such a method is unnecessary when the relative magnitudes of the units are known, but becomes necessary when the relative magnitudes are unknown as is the case in social and psychological measurement.

A minimum of three dichotomous items is required within each of two Specified Frames of Reference in order to compare levels of discrimination in this fashion since a minimum of two possible person total scores is required to obtain the variance of person estimates. Thus, it is not possible to compare levels of discrimination in this way if every item has a different associated level of discrimination as is permitted in the 2PLM. Nonetheless, as elaborated in the following section, analysis of the Maximum Likelihood (ML) estimation equation shows that, in principle, ρ_k can be estimated for a Specified Frame of Reference containing only one item *provided* another frame of reference containing several items is firstly used to obtain person estimates.

When separate constraints are imposed in separate frames as arbitrary choices of origins, it is necessary to obtain an estimate of the difference between the origins of those scales in order to equate them. When common items are contained within two frames of reference, the origins of the scales can be compared as follows:

$$\left(\bar{\delta}_{g.} / \rho_g\right) - \left(\bar{\delta}_{h.} / \rho_h\right) = \left(\bar{\delta}^*_{.} - c^*_g\right) - \left(\bar{\delta}^*_{.} - c^*_h\right)$$

$$\Rightarrow \left(\bar{\delta}_{g.} / \rho_g\right) - \left(\bar{\delta}_{h.} / \rho_h\right) = c^*_h - c^*_g . \qquad (10)$$

When common persons are contained within two frames of reference, the comparison between the origins is given by

$$\left(\bar{\beta}_{s.} / \rho_s\right) - \left(\bar{\beta}_{t.} / \rho_t\right) = c^*_t - c^*_s . \qquad (11)$$

Thus, it is possible to compare both the units and origins of any given pair of scales by applying the SRM within each context, then comparing firstly the units and secondly the origins. Once the values of ρ_k and c^*_k, $k = 1,...,K$ have been estimated given relevant constraints, person and item locations can be expressed on a common scale: that is, estimates of β^*_n , $n = 1,...,N$ and δ^*_i, $i = 1,...,I$ can be obtained. In the case of just two sets of estimates, this process specialises to the commonly employed procedure in which the mean and variance of the two sets of estimates are set at specific common values, as described for example by Embretson & Reise (2000). In terms of the framework and terminology developed here, such a procedure entails accounting for the influence of different classifications of an empirical factor on discrimination, and therefore the natural unit of a scale.

Item Characteristic Curves and sufficiency

When person and item locations are expressed in terms of a common arbitrary unit, the level of discrimination obtained in a given Specified Frame OF Reference is reflected within the *slopes* of Item Characteristic Curves (ICCs). An ICC shows the probability of the discrete outcome $X_{kin} = 1$ as a function of person location. In general, the slope of the ICC is steeper for any given probability when the parameter ρ_k assumes a greater magnitude. Figure 4 shows the ICCs of a set of items within each of two Specified Frames of Reference for which $\rho_1 / \rho_2 = 1.5$ and hence given the constraint shown in Equation (7), $\rho_1 \cong 1.22$ and $\rho_2 \cong 0.82$. These items might, for example, be a single set of items administered in the presence of two different conditions such as a systematic difference between the times available to respond to the items.

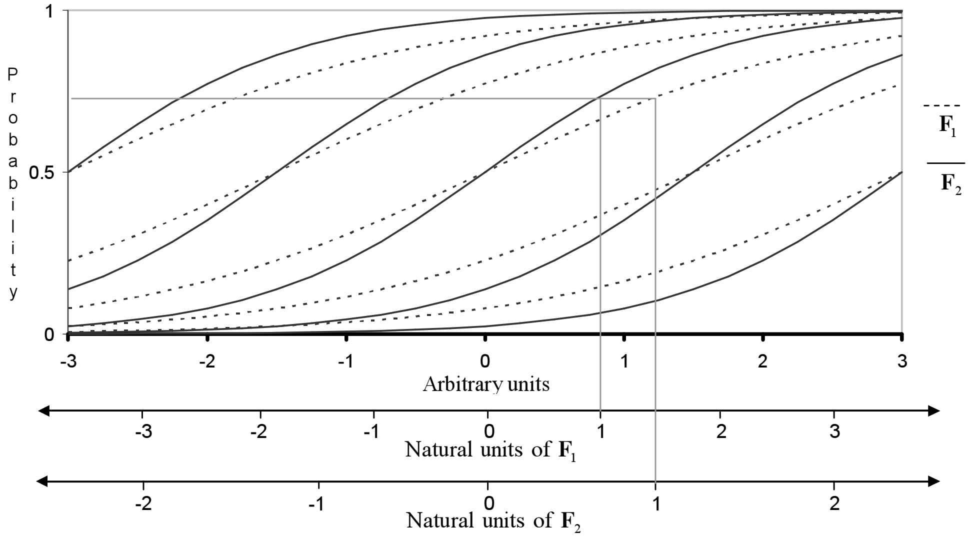


Figure 4: ICCs for items within two Specified Frames of Reference

Three sets of coordinates are provided as horizontal axes in Figure 3 to highlight the fact that each set of ICCs accords with the RM. Thus, when the ICCs for frame $\mathbf{F}_k$ are referenced to the scale partitioned into the natural units of $\mathbf{F}_k$, the ICCs accord with the RM in its standard form; i.e. they follow Equation (1a). On the other hand, when the same two sets of ICCs are referenced to the scale portioned into common arbitrary units, the ICCs accord with the EFRM, Equation (5), where $\rho_1 = 1.22$ and $\rho_2 = 0.82$. By partitioning *the same latent continuum* into units of different sizes, it is therefore possible for data to accord with the RM irrespective of any difference between the slopes of ICCs *between* the two frames of reference when all ICCs are referenced the scale portioned by common arbitrary units. With regard to preserving sufficiency, the important thing to note is that given the structure of the EFRM the slopes of ICCs *within* any particular Specified Frame of Reference remain parallel, while the slopes vary only *between* Specified Frames of Reference. Sufficient statistics are therefore preserved within any given Specified Frame of Reference because separation of the person and item parameters of the RM entails parallel ICCs (Rasch, 1961; Andersen, 1977). Thus, while it is not apparent that Rasch (1960/1980) intended ρ in Equation (1) to be treated as a parameter, it is clearly possible to do so without destroying sufficiency within a Specified Frame of Reference provided ρ_k is of fixed magnitude *within* any given Specified Frame of Reference.

On the surface, the situation presented here may seem to represent something of a paradox: sufficient statistics are preserved within a given context, yet sufficiency would be destroyed were a *single* raw score formed from data contained within the contexts in combination. However, a given raw score is a sufficient statistic for a parameter *in which a certain unit and origin are implicit.* For example, let $r_{kn} = \sum_{i=1}^{I_k} x_{kin}$, $k = 1,...,K$ be the raw scores for person n obtained in the frames of reference $\mathbf{F}_k$, $k = 1,...,K$. In Equation (2), the SRM, r_{1n} is the sufficient statistic for β_{1n} and r_{2n} is the sufficient statistic for β_{2n}. This formulation recognizes the empirical possibility that the unit will differ due to the nature of the Specified Frames of Reference and, accordingly, that it is equally justifiable to estimate or express the measure of a given person n in terms of each of these units. To preclude this possibility would seemingly imply that some particular unit is, in some sense, privileged over others. Now, the relationship between the probability of a particular response and a given difference $\beta_n^* - \delta_i^*$ is not the same irrespective of the frame of reference, as can be seen in Figure 4. However, this does not seem a satisfactory justification for privileging one unit over another, for the mapping of any given difference to the probability of particular response is equally well preserved within each context. For example, let $\delta_3^* = 0$ denote the location of the item shown in the center of the graph in Figure 4. The grey lines mapping location to probability in

Figure 4 show that, in keeping with the RM in its standard form, when $\beta_{1n} - \delta_{13} = 1$, the probability $\Pr\{X_{11n} = 1\}$ is approximately 0.73 and similarly, when $\beta_{2n} - \delta_{23} = 1$, $\Pr\{X_{21n} = 1\}$ is also approximately 0.73.

Following from this, as shown in Appendix I, *vectors* of total scores are sufficient statistics for vectors of parameters of the SRM across multiple frames of reference. This result follows naturally given the preservation of sufficiency within each frame separately since the score vector is "the natural generalization of the raw score" (Andersen, 1973, p. 73). Thus, let $\mathbf{x}_n = (x_{11n},...,x_{KI_kn})$ be the vector of individual responses for person n and let $\mathbf{r}_n = (r_{1n},...,r_{Kn})$ be the vector of total scores across Specified Frames of Reference defined in terms of classifications k of an empirical factor. It is shown in Appendix II that the conditional density function $\Lambda = \Pr\{\mathbf{x}_n \mid \mathbf{r}_n\}$ derived from the SRM does not contain the vector of person parameters $\boldsymbol{\beta}_n = (\beta_{1n},..,\beta_{Kn})$. Consequently, $\mathbf{r}_n$ is sufficient for $\boldsymbol{\beta}_n$ and, hence, by partitioning the response space in terms of $\mathbf{r}_n$, it is possible to estimate item parameters independently of person parameters. The sufficiency of such score vectors in the SRM, and therefore the EFRM, closely parallels the sufficiency of score vectors for vectors of parameters in Rasch models as detailed in Andersen (1973, 1977). An illustrative example of the case studied by Andersen is provided in Appendix III.

Also following from the result described above, by conditioning on score vectors, the person parameter β_n^* of the EFRM is eliminated, in keeping with the sufficiency of the weighted raw score as shown in Verhelst & Glas (1995). Verhelst & Glas (1995) derive Conditional Maximum Likelihood (CML) equations from a model referred to as the One Parameter Logistic Model, which contains a discrimination *index* rather than a discrimination *parameter*, but is formally identical with the 2PLM. As noted by these authors, though, the problem one faces in implementing CML estimation 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 EFRM, on the other hand, sufficiency of the weighted raw score can in principle be exploited without prior knowledge of the values of ρ_k, $k = 1,...,K$ by conditioning on score vectors due to the fact that all response vectors $\mathbf{x}_n = (x_{11n},...,x_{Kin})$, which yield a particular score vector $\mathbf{r}_n = (r_{1n},...,r_{Kn})$, necessarily also yield the same weighted raw score $W_n = \sum_{k=1}^{K} \rho_k r_{kn}$.

While person parameters have sufficient statistics in the EFRM, the resulting CML equations contain item parameters expressed in terms of natural units rather than a common unit. Given estimates of vectors of parameters, it becomes an empirical question whether the frames of reference are mutually conformable in the sense that the parameters can be decomposed into products; i.e. $\beta_{kn} = \rho_k \beta_n^*$ and $\delta_{ki} = \rho_k \delta_i^*$. This question can be investigated by estimating ρ_k, $k = 1,...,K$ and subsequently conducting appropriate tests of fit as touched on in the empirical investigations to follow. The method of estimating discrimination parameters discussed in the previous section involves forming ratios of variances of estimates obtained from application of the SRM separately within two or more frames of reference. It is also instructive to briefly consider ML estimation of discrimination parameters. Thus, for example, consider a case in which frames of reference are defined in terms of classifications of an item factor. Given conditional estimates of $\delta_{ki} = \rho_k \delta_i^*$, it is in principle possible to estimate the level of discrimination obtained in a frame of reference $\mathbf{F}_2$ relative to another $\mathbf{F}_1$, provided one of the frames of reference contains several items. Suppose then that $\mathbf{F}_1$ contains a number items. By letting $\rho_1 \equiv 1$ as an arbitrary choice of unit, estimates of person abilities in the arbitrary unit, β_n^* , $n = 1,...,N$, can be obtained from $\mathbf{F}_1$ by applying the RM within that frame of reference. Given estimates of δ_{2i}, $i = 1,..,I_2$ and ability estimates derived from $\mathbf{F}_1$, there is only one unknown in the ML solution equation for ρ_2 which is provided in Appendix I, and hence this parameter can in principle be estimated iteratively.

The question of whether a matrix of parameters can be decomposed into a vector of item and discrimination parameters is of the same fundamental nature as the question discussed by Andersen (1973, 1977) in relation to item parameters and scaling values φ_p associated with each category p in the model which appears in Appendix III. In terms of the framework developed here, frames of reference are mutually conformable when they provide for measurements of a *common* latent trait in terms of natural units whose magnitudes may differ depending on the classifications of empirical factors.

Further work is needed to explore technical issues involved with estimations and tests of fit. The key point, however, is that it is in principle possible to resolve the interdependence between discrimination parameters and individual person and item parameters provided a uniform level of discrimination can be obtained within each Specified Frame of Reference defined in terms of the interaction of persons and items under specified empirical conditions, as show in Figure 1. This consequence arises from treating ρ in Equation (1) as a parameter pertaining to a frame of reference and hence is established within the framework established by Rasch (1961, 1977). In contract, for the 2PLM the conditional expression in Appendix I cannot provide a basis for estimation of the item parameters δ_{ki}, $k = 1,..,K$, $i = 1,...,I_k$ because no data reduction is achieved by partitioning the response space according to score vectors.

Before proceeding, it is noted that the focus in this section has been principally on situations in which frames of reference are defined in terms of empirical factors other than person factors. The motivation for proceeding in this way is that is simpler in theoretical terms to deal with situations in which frames of reference are defined in terms of person factors. Supposing many persons attempt each item, estimates of the item parameters δ_{gi}, $g = 1,...,G$, $i = 1,...,I$ can be obtained conditionally within each frame of reference and the variances of the item estimates can be compared as shown in Equation (8). Unlike person estimates, item estimates have negligible error and bias associated with them. The symmetry of the model means that vectors of item totals are sufficient for person parameters when frames of reference are defined in terms of person factors. However, it is neither necessary nor practicable to exploit sufficiency of vectors of item totals in order to estimate vectors of person parameters β_{gn}, $g = 1,...,G$ independently of item parameters, just as it is not necessary or practicable to exploit sufficiency of an item total within a single frame of reference in order to estimate person parameters.

In terms of classifying empirical factors so as to define frames of reference, it is recognised that there may not be an immediately available or evident basis for deciding how to partition the assessment context, and therefore the associated data matrix, into separate frames. With regard to this issue, it is proposed that any means available should be used, including theoretical predictions and empirical observations, to formulate scientific hypotheses regarding which specific factors should constitute the basis for defining Specified Frames of Reference. In the empirical investigations which follow, naturally occurring classifications are used to define separate assessment contexts with some success, indicating that it is possible to make progress by adopting such an approach. Irrespective of whether immediate applications are obtained in any particular situation, however, the key is that the framework developed here makes it possible to *properly formulate and test scientific hypotheses regarding the influence of empirical factors on discrimination*.

An illustrative example of the influence of person factor on the natural unit

The first of the empirical illustrations of the approach developed in this paper focuses on the influence of the person factor *school year level* on discrimination, in the context of reading assessment. The level of discrimination associated with a given classification of a person factors is referred to as *Person Factor Discrimination* (PFD) associated with that classification. The data used in the investigation were collected within schools across the state of Western Australia as part of the Western Australian Literacy and Numeracy Assessment (WALNA) program in 2003. The WALNA program involves participation of approximately 25,000 students across the state in each of the schooling years 3, 5, and 7. The program includes the administration of reading, writing, mathematics, and spelling assessments 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). RUMM2020 was also used to conduct CML estimations based on the RM in the present empirical investigation, and in the investigation reported in the following section. In these investigations, the SRM was applied within separate frames of reference and

Equations (8) through (10) were used as applicable in order to estimate discrimination parameters and to compare origins.

In the WALNA program, vertical equating between the Year 5 and 7 assessments is attempted on the basis of common link items embedded within the two assessments. A striking observation was made during the original analysis of these equating data subsequent to concurrent analysis of the Year 5 and 7 data according to the RM. Inspection of the ICCs of the vertical link items for evidence of Differential Item Functioning (DIF) revealed a consistent difference between the *slopes* of the empirical curves for the different year groups. Specifically, the empirical slope was observed to be steeper for the Year 7 population than for the Year 5 population.

In light of the framework that has been developed, if the differences between the slopes reflect differences between the level of Person Factor Discrimination (PFD) associated with classifications Year 5 and 7, then a difference between the dispersions of the items would be expected when the locations of the common items are estimated separately for each group through application of the SRM. The reason for this is that the level of PFD is absorbed into estimates obtained from each Specified Frame of Reference: i.e. $\sqrt{\mathrm{V}[\delta_g]} = \rho_g \sqrt{\mathrm{V}[\delta^*]}$. Indeed, such a difference between the dispersions of the estimates was apparent, indicating different levels of PFD. Letting $g = 5$ and $g = 7$ denote person factor classifications of Year 5 and 7 respectively, and imposing the constraint $\prod_g \rho_g \equiv 1$, estimates of the levels of PFD attributable to the two groups were derived as shown in Table 2.

Table 2: PFD estimation

SD	Estimate	Parameter	Estimate
$\sqrt{\mathrm{V}\left[\hat{\delta}_5\right]}$	0.695	$\hat{\rho}_5$	0.875
$\sqrt{\mathrm{V}\left[\hat{\delta}_7\right]}$	0.905	$\hat{\rho}_7$	0.905

The difference between the spread of the locations is shown diagrammatically in Figure 4. Dotted lines are used to show the correspondence of the items. Andrich (1988, p. 75) provided a similar diagram showing a systematic difference between the dispersions of item estimates for Eysenck's (1958) Neuroticism questionnaire when data were analysed separately for male and female respondents, thus providing another empirical example indicative of differential PFD.

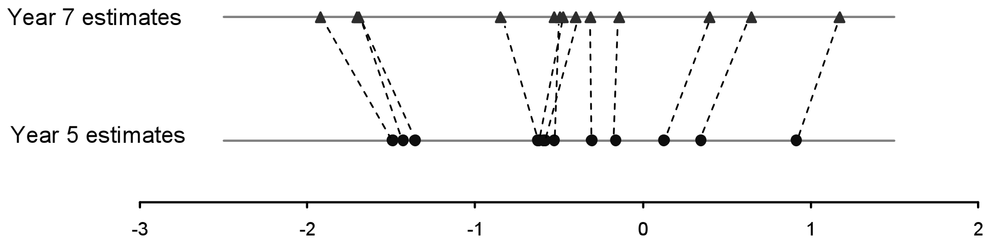


Figure 4: Estimates of common link items expressed in terms of natural units

Analysis of the WALNA data indicated that the difference of scale is attributable to differences between levels of PFD, rather than DIF in which the difficulties of the items vary systematically but for some reason other than the influence of PFD. Specifically, a fit statistic was computed based on analysis of the data using the RM and EFRM. The statistic, referred to here as a *group fit residual*, is defined as

$$Y_g = \frac{\sum_i \sum_{n \in g} \left(z_{ign}^2 - F_{ign} \right)}{\sqrt{V\left[\sum_i \sum_{n \in g} z_{ign}^2 \right]}}, \tag{13}$$

where z_{ign}^2 is the standardized residual and F_{ign} is the approximate degree of freedom per element of the data matrix, as described for example by Andrich (1988). The expected value of the statistic is 0 and, as Andrich (1988) explicates in terms of an example set of response data, a *negative* fit residual implies that response data are closer to a Guttman structure (Guttman, 1950, 1954) than expected, whereas a *positive* fit residual implies a response pattern that is more erratic than expected. In general, therefore, this type of fit statistic provides an index which is sensitive to the effects of discrimination that have not been accounted for by a given model.

The fit residuals obtained for the current empirical investigation are shown in Table 3, where the values for the EFRM were obtained by equating for the differences between the units of scales and using the EFRM to obtain expected values used for the computation of the fit residual; i.e. by computing expected values according to Equation (5). The reduction in the *difference* between the group fit residuals for the EFRM compared with the standard form of the RM indicates that PFD is the key factor underlying the difference between the scale values of the item estimates. Further detail regarding the analysis and findings can be found in Humphry (2005).

Table 3: Group fit residuals for the RM and EFRM

	Y_5	Y_7
RM: single frame of reference	1.88	-4.75
EFRM	2.04	0.20

The substantive reason for the difference between the levels of PFD is not entirely clear. However, a possibility is that guessing played a greater role in the response process among Year 5 students than among Year 7 students because the items all have a multiple choice format. The empirical plots of ICCs did not reveal obvious asymptotes above a probability of 0. However, guessing may have played a greater role generally across the range of abilities for Year 5 students due to level of familiarity and background knowledge relevant to contextual features of the reading texts. Specifically, older students with low abilities may have deliberately chosen distracters because they were more likely to have knowledge of particular facts or detail which made the distracters superficially plausible. Consequently, there may have been a tendency for distracters to appear more plausible to these Year 7 students than to the younger Year 5 students with less background knowledge and an equivalent level of reading ability. On the other hand, for Year 7 students with *high* levels of reading ability, greater background knowledge would be expected to result in a greater likelihood of responding correctly, with the distracters no longer being plausible. This is precisely what is suggested by the crossing ICCs referred to earlier in this section.

As should be clear from this brief discussion, the implied difference between the probabilities of a correct response for Year 5 and 7 students with the same ability is not without challenges in terms of substantive interpretation, as was pointed out by Andrich (1988) in the context referred to earlier involving a difference between units of frames of reference defined by gender. The point is, however, that it is by no means a given that the odds of a correct response should necessarily be the same for all persons with a common level of a trait irrespective of person factors, such as age and background knowledge, because such characteristics may play a mediating role in the manifestation of the trait. Nevertheless, in order to measure relative to a unit of fixed magnitude and to sustain sufficiency, the mapping of differences between locations and the odds of success should be uniform *within* a Specified Frame of Reference, as discussed in relation to Figure 4 earlier.

An illustrative example of the influence of item factor on the natural unit

The WALNA program also provides the context for illustrating the influence of item factors on the natural unit. The discrimination associated with a given classification of an item factor is referred to as the *Item Factor Discrimination* (IFD) associated with that classification. In 2003, common person equating was used to equate the difficulty of the 2003 Numeracy assessment with that of an assessment administered within the same program in 2000. The initial observation was that the standard deviation of the ability estimates for the 2003 population was approximately 1.2 times that of the 2000 population, based on separate analyses of the data obtained from the two assessments using the RM. The Western Australian populations in 2000 and 2003 each comprised approximately 25,000 students. Let $s = 0$ and $s = 3$ denote classification of item factors in terms of the separate 2000 and 2003 assessments, respectively. These classifications represent the fact that there were 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. As shown in Table 4, inspection of the results of the equating study revealed that for 281 *common* students involved in the relevant equating study, a difference between the standard deviations was also evident when person locations were estimated separately from the two assessments according to the SRM.

Table 4: IFD estimation

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

Thus, the ratio $\hat{\rho}_3 / \hat{\rho}_0 = 1.173$ was similar to that of the standard deviations of the ability estimates of the two populations. This evidence suggested the difference between the dispersions of estimates arises due to features of the instruments rather than the students, because common students attempted both item sets in the equating study, thus controlling for this factor.

Also relevant to this hypothesis, the two assessments were constructed by different item developers. The program effectively began in 1999, and with its evolution there have been refinements to the process of item development, trialing of items and so on, which may have contributed to there being greater discrimination among the items making up the 2003 assessment compared with those in the 2000 assessment.

A third piece of evidence also indicated that the difference between the natural units was attributable to levels of IFD. Specifically, a fit residual Y_s was computed across items contained within each of the two sets, where

$$Y_s = \frac{\sum_{i \in s}\sum_{n}\left(z_{in}^2 - F_{in}\right)}{\sqrt{V\left[\sum_{i \in s}\sum_{n} z_{ign}^2\right]}}. \tag{14}$$

The results, shown in Table 5, indicate that there was generally a lower level of IFD among the items on the 2000 assessment not accounted for in the RM, and that the effects of this differential discrimination were largely accounted for applying the EFRM.

Table 5: Item set fit residuals for common persons obtained from each of the assessments using the RM and EFRM

	Y_3	Y_0
RM: single frame of reference	-1.28	1.79
EFRM	0.32	0.57

While the results shown in Table 5 indicate that the systematic difference between levels of IFD was largely accounted for by applying the EFRM, the fit residuals for individual items in some cases suggested departure from the model. The important thing, however, is that a substantial improvement was achieved without sacrificing sufficiency, and without introducing interdependence between the IFD parameters and individual person and item parameters.

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 discrimination, including person characteristics, item characteristics, and environmental conditions. By making explicit the relationship between discrimination and the natural unit, a foundation has been established for investigating the influences of empirical factors on discrimination and the unit of a scale.

Given that assessment data are generally produced by the interaction of persons with items, it is particularly important to consider the influence of person and item factors on the level of discrimination. Accordingly, the main focus has been on developing an approach in which discrimination is parameterized to account for the influences of these factors, and the empirical investigations reinforced their importance. In addition, however, the EFRM 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 framework which has been developed provides a foundation for developing knowledge about empirical factors that must be controlled in order to obtain fixed units of scales.

Conclusion

The EFRM and associated framework represents an extension of the model and framework developed by Rasch (1960/1980). This extension makes it is possible to parameterize discrimination without destroying sufficiency within a given specified assessment context. In addition, the EFRM provides a basis for making invariant comparisons between the levels of discrimination associated with separate frames of reference, thus providing a basis for maintaining a common unit of scale across multiple Specified Frames of Reference. The EFRM therefore broadens the foundation for social and psychological measurement while preserving the distinguishing property of the class of measurement models identified by Rasch (1960/1980).

Appendix I: Conditioning on score vectors

Consider the case in which frames of reference are defined in terms of an item factor. From Equation (2), it follows that

$$\Pr\{\mathbf{x}_n \mid \beta_{kn}, \delta_{ki}\} = \prod_k \prod_i \left[\frac{\exp\left(x_{kin}(\beta_{kn} - \delta_{ki})\right)}{\left(1 + \exp\left((\beta_{kn} - \delta_{ki})\right)\right)} \right]$$

$$= \frac{\exp\left(\sum_k r_{kn}\beta_{kn}\right)\exp\left(-\sum_k \sum_i x_{kin}\delta_{ki}\right)}{\prod_k \prod_i \left(1 + \exp\left((\beta_{kn} - \delta_i)\right)\right)} , \qquad \text{(A1)}$$

Thus, it follows that

$$\Pr\{\mathbf{x}_n \mid \mathbf{r}_n ; \beta_{kn}, \delta_{ki}\} = \frac{\dfrac{\exp\left(\sum_k r_{kn}\beta_{kn}\right)\exp\left(-\sum_k \sum_i x_{kin}\delta_{ki}\right)}{\prod_k \prod_i \left(1 + \exp\left(\beta_{kn} - \delta_{ki}\right)\right)}}{\dfrac{\exp\left(\sum_k r_{kn}\beta_{kn}\right)\sum_{(\mathbf{x})|\mathbf{r}} \exp\left(-\sum_k \sum_i \delta_{ki}\right)}{\prod_k \prod_i \left(1 + \exp\left(\beta_{kn} - \delta_{ki}\right)\right)}}$$

$$= \frac{\exp\left(-\sum_k \sum_i x_{kin} \delta_{ki} \right)}{\sum_{(\mathbf{x})|\mathbf{r}} \exp\left(-\sum_k \sum_i \delta_{ki} \right)}, \qquad \text{(A2)}$$

in which the vector of person parameters $\boldsymbol{\beta}_n$ is eliminated by partitioning the response space in terms of the score vector $\mathbf{r}_n$. Consider, as a specific example, the probability of the vector of raw scores $\mathbf{x}_n = \{(1,0) \cap (1,0)\}$ conditional on the vector of total scores $\mathbf{r}_n = (1,1)$. The conditional probability is

$$\Pr\{(1,0) \cap (1,0) | (1,1); \beta_{kn}, \delta_{ki}\} = \frac{\exp(r_{1n}\beta_{1n} + r_{2n}\beta_{2n})\exp(-\delta_{11} - \delta_{21})}{\exp(r_{1n}\beta_{1n} + r_{2n}\beta_{2n})\sum_{(\mathbf{x})|\mathbf{r}} \exp\left(-\sum_k \sum_i \delta_{ki} \right)}$$

$$= \frac{\exp(-\delta_{11} - \delta_{21})}{\exp(-\delta_{11} - \delta) + \exp(-\delta_{11} - \delta_{22}) + \exp(-\delta_{12} - \delta_{21}) + \exp(-\delta_{12} - \delta_{22})}. \qquad \text{(A3)}$$

Appendix II: Maximum Likelihood solution equation for discrimination parameters in the EFRM

Let

$$\Gamma = \Pr\{\mathbf{X}; \rho_k, \beta_n^*, \delta_i^*\} = \prod_k \prod_i \prod_n \left[\frac{\exp(\rho_k x_{kin} (\beta_{kn}^* - \delta_{ki}^*))}{1 + \exp(\rho_k (\beta_{kn}^* - \delta_{ki}^*))} \right]$$

be the likelihood function for the EFRM where $\mathbf{X}$ is a data matrix contained within multiple frames of reference. Then the log-likelihood function is

$$\log \Gamma = \sum_k \sum_i \sum_n \rho_k x_{kin} (\beta_{kn}^* - \delta_{ki}^*) - \sum_k \sum_i \sum_n \log[1 + \exp(\rho_k (\beta_{kn}^* - \delta_{ki}^*))].$$

In the case in which frames of reference are defined in terms of an item factor, the Maximum Likelihood (ML) solution equations, obtained from setting the partial derivative with respect to ρ_k equal to zero, are as follows:

$$0 = \sum_i^{I_k} \sum_n x_{kin} (\beta_n^* - \delta_i^*) - \sum_i^{I_k} \sum_n \left[\frac{\exp(\rho_k (\beta_n^* - \delta_i^*))}{1 + \exp(\rho_k (\beta_n^* - \delta_i^*))} \right] (\beta_n^* - \delta_i^*), \quad k = 1, \ldots, K.$$

Let $\delta_{ki} \equiv \rho_k \delta_i^*$, such that $c_k^* = 0$, $k = 1, \ldots, K$. The solution equations are then

$$0 = \sum_i^{I_k} \sum_n x_{kin} \left(\beta_n^* - \left(\frac{\delta_{ki}}{\rho_k} \right) \right) - \sum_i^{I_k} \sum_n \left[\frac{\exp(\rho_k \beta_n^* - \delta_{ki})}{1 + \exp(\rho_k \beta_n^* - \delta_{ki})} \right] \left(\beta_n^* - \left(\frac{\delta_{ki}}{\rho_k} \right) \right). \qquad \text{(A4)}$$

for $k = 1,\ldots,K$.

Appendix II: Sufficiency of score vectors for Rasch models shown by Andersen (1973, 1977)

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

$$\Pr\{X_{in}^{(p)}\} = \frac{\exp(\theta_{np} - \psi_{ip})}{\gamma_{in}} , \qquad \text{(A5)}$$

where θ_{nq} and ψ_{iq} are person and item parameters pertaining specifically to category q of a total of Q categories, and $\gamma_{in} \equiv \sum_q^Q \exp(\theta_{nq} - \psi_{iq})$. Let the event of individual n responding in category p of item i be represented as a vector $\mathbf{X}_{in} = (x_{in}^{(1)},\ldots,x_{in}^{(q)},\ldots,x_{in}^{(Q)})$ where $x_{in}^{(p)} = 1$ and $x_{in}^{(q)} = 0$ for $q \neq p$ (Rasch, 1961; Andersen, 1973, 1977). Let $\boldsymbol{\theta}_n = \theta_{nq}$, $q = 1,..,Q$ be a vector of person parameters for person n associated with the categories. Also, let $t_{nq} = \sum_{i=1}^{I} x_{in}^{(q)}$ 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

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

$$= \frac{\exp(\theta_{n3} + \theta_{n2})\exp(-\psi_{22} - \psi_{13})}{\exp(\theta_{n3} + \theta_{n2})\exp(-\psi_{13} - \psi_{22}) + \exp(-\psi_{12} - \psi_{23})}$$

$$= \frac{\exp(-\psi_{22} - \psi_{13})}{\exp(-\psi_{13} - \psi_{22}) + \exp(-\psi_{12} - \psi_{23})} . \qquad \text{(A6)}$$

Although the model of Equation (A5) pertains to items with multiple categories, it can be seen that this conditional equation has the same basic form as Equation (A3).

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

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

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.

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.

Eysenck, W.J. (1958). A short questionnaire for the measurement of two dimensions of personality. *Journal of Applied Psychology*, 47, 14-17.

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

Maris, G. & Bechger, T.M. (2005). An introduction to the DA-T Gibbs sampler for the Two-Parameter Logistic (2PL) model and beyond. *Psicologica*, 26, 327-352.

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. (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.

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.

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