Title
Scale Alignment in the Between-Item Multidimensional Partial Credit Model

Permalink
https://escholarship.org/uc/item/0rr99001

Journal
Applied Psychological Measurement, 45(4)

ISSN
0146-6216

Authors
Feuerstahler, Leah
Wilson, Mark

Publication Date
2021-06-01

DOI
10.1177/01466216211013103

Scale Alignment in the Between-Items Multidimensional Partial Credit Model

Leah Feuerstahler

Fordham University

Mark Wilson

University of California, Berkeley

Corresponding author: Leah Feuerstahler

441 E Fordham Road

Bronx, NY 10458

(718) 817-3788

lfeuerstahler@fordham.edu

Abstract

In between-items multidimensional item response models, it is often desirable to compare individual latent trait estimates across dimensions. These comparisons are only justified if the model dimensions are scaled relative to each other. Traditionally, this scaling is done using approaches such as standardization—fixing the latent mean and standard deviation to 0 and 1 for all dimensions. However, approaches such as standardization do not guarantee that Rasch model properties hold across dimensions. Specifically, for between-items multidimensional Rasch family models, the unique ordering of items holds within dimensions, but not across dimensions. Previously, Feuerstahler and Wilson (2019) described the concept of *scale alignment*, which aims to enforce the unique ordering of items across dimensions by linearly transforming item parameters within dimensions. In this paper, we extend the concept of scale alignment to the between-items multidimensional partial credit model and to models fit using incomplete data. We illustrate this method in the context of the Kindergarten Individual Development Survey (KIDS), a multidimensional survey of kindergarten readiness used in the state of Illinois. We also present simulation results that demonstrate the effectiveness of scale alignment in the context of polytomous item response models and missing data.

Scale Alignment in the Between-Items Multidimensional Partial Credit Model

Between-items multidimensional item response models are used to simultaneously model a small number of related constructs. For example, a four-dimensional model could be specified for a test that includes four items representing each algebra, geometry, statistics, and written mathematics. In contrast to unidimensional tests for which one latent trait estimate is reported per examinee, multidimensional models result in a set of estimates that can be reported for each individual. Several empirical studies have found that the resulting estimates are more reliable than estimates from consecutive unidimensional models (Adams, Wilson, & Wang, 1997; Baghaei, 2012; Briggs & Wilson, 2003; Cheng, Wang, & Ho, 2009; Wang, 1996; Wang, Chen, & Cheng, 2004).

When reporting estimates from multidimensional models, it is often desirable to compare an individual's estimate on one dimension directly to his estimate on another dimension, especially if the dimensions are closely related or if there is thought to be a higher-order dimension. For example, consider a mathematics test that provides separate scores for algebra, geometry, statistics, and written mathematics. If the estimates for a given person equal (1.0, 1.0, 1.0, 1.0), then one might interpret this to mean that the person has "the same" proficiency on all the dimensions, that the person scored equally well on each of these areas of mathematics. However, this interpretation is not necessarily appropriate if the scores are based on multidimensional models. For example, a common constraint used to statistically identify such models is that the mean of the item difficulties equals 0.0 for every dimension. Consider then a case in which the items on one dimension have been designed to be much harder than those on the other dimensions. Then the estimate of 1.0 on that dimension would constitute a higher relative estimate than for the others. Thus, across-dimension comparisons are justified only when the dimensions have been scaled with reference to other dimensions.

One classical approach to this problem is standardization. If the latent variables for each dimension are standardized to have mean 0 and variance 1, then direct norm-referenced comparisons may be made across dimensions. This is, of course, subject to the assumption that a standard deviation on one dimension is the equivalent of a standard deviation on every other dimension. Although standardized scaling may be justifiable in the context of non-Rasch item response models, Rasch family models imply stronger measurement properties, such as the unique ordering of items, that do not necessarily hold across dimensions, regardless of whether those dimensions are standardized. Because these traditional scaling methods do not preserve the ordering of items across dimensions, the dimensions are not scaled relative to each other in terms of the strongest properties of the Rasch model.

A formal method to scale dimensions relative to each other was first described by Feuerstahler and Wilson (2019), although several authors have applied earlier approaches to scale alignment (e.g., Morell, Collier, Black, & Wilson, 2017; Osborne, Henderson, MacPherson, Szu, Wild, & Yao, 2016; Yao, Wilson, Henderson, & Osborne, 2015). Feuerstahler and Wilson's proposed scale alignment methods aimed to preserve the unique item order across dimensions for the between-items multidimensional Rasch model. To do this, they imagined projecting the dimensions from the multidimensional model onto a single reference dimension (see e.g., Ackerman, 1992) which represents a composite of the individual dimensions. On this reference dimension, a unique ordering of items is sensible for the same reason that any set of items that reflect a single dimension can be ordered. Scale alignment methods work by linearly transforming the multidimensional model such that the unique ordering of items is preserved across dimensions in the original multidimensional model. Feuerstahler and Wilson operationalized this idea by defining aligned dimensions as those for which the same observed sufficient statistics imply the same parameter estimates, regardless of dimension. It should be noted that this definition is an ideal, as it is unlikely to observe the exact same sets of sufficient statistics across dimensions. Methods to better achieve aligned dimensions were proposed by Feuerstahler and Wilson. However, their work was limited to binary data sets with no missing data. The purpose of this paper is to extend the definition of aligned scales to polytomous item responses and incomplete data sets, and to developed generalized alignment methods that account for these factors.

The remainder of the paper is organized as follows. First, we present the between-items multidimensional extension of the partial credit model (PCM; Masters, 1982) and describe how linear transformations of the latent variable affect item parameter estimates. Then, we extend the definition of scale alignment to the between-items multidimensional PCM and describe how previously developed scale alignment methods can be modified for use with this model and how scale alignment is affected by missing data. Finally, we illustrate scale alignment in terms of a large-scale rating scale instrument and provide simulation evidence that each of the scale alignment methods can lead to item response models that better meet the definition of aligned dimensions.

Partial Credit Model

Feuerstahler and Wilson (2019) defined Rasch model dimensions to be aligned if the same sufficient statistics imply the same parameter estimates, regardless of dimension. Under the Rasch model for binary item responses, the proportion of correct responses for item i belonging to dimension d, $\hat{p}_{i(d)}$, is a sufficient statistic for the item difficulty parameter $\delta_{i(d)}$. Feuerstahler and Wilson used the fact that these sufficient statistics are monotonically related to item difficulty estimates to develop alignment methods. Scale alignment methods, therefore, linearly transform dimensions such that the rank-order correlation between $\hat{p}_{i(d)}$ and $\hat{\delta}_{i(d)}$ is maximized when these quantities are not separated by dimension.

To extend scale alignment methods to polytomous data, we consider the between-items multidimensional extension of the partial credit model (PCM; Masters, 1982). This is the most general Rasch family model for responses that occur in more than two ordered categories. A general form of the between-items multidimensional partial credit model for item i with response categories $x = 0, \dots m_i$ can be written

$$\ln\left[\frac{P(X = x|\theta)}{P(X = x - 1|\theta)}\right] = \sum_{k=0}^{x} \alpha_d \left(\theta_d - \xi_{i(d)k}\right), \tag{1}$$

where by definition,

$$P(X = 0|\theta_d) = \frac{1}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} (\, \alpha_d (\theta_d - \xi_{i(d)k}))}, \tag{2}$$

$\exp \sum_{k=0}^{0} (\, \alpha_d (\theta_d - \xi_{i(d)k})) = 1$, θ_d is the latent ability on dimension d, $\xi_{i(d)k}$ is the item step parameter for step k of item i on dimension d, and α_d is the steepness of the response curve for all items on dimension d. The α_d parameter is traditionally set equal to 1 for all dimensions and is often omitted from expressions of the PCM. If the model is specified with all α_d = 1, the latent variance for each dimension is estimated. An equivalent choice is to set all latent variances equal to 1 and to estimate α_d (cf. the relationship between the Rasch model and the one-parameter logistic model, Verhelst & Glas, 1995). We include the α_d term here because it is necessary to include this term when linearly transforming the PCM dimensions, as is done in scale alignment. An equivalent expression of the PCM may be written as a function of $\delta_{i(d)}$, the average item step parameter, and $\tau_{i(d)k}$, the deviation from the average step parameter for step k such that $\sum_{k=1}^{m} \tau_{i(d)k} = 0$. The two PCM parameterizations are related as follows:

$$\delta_{i(d)} = \frac{1}{m_i}\sum_{k=1}^{m_i} \xi_{i(d)k}\,, \tag{3}$$

$$\tau_{i(d)k} = \xi_{i(d)k} - \delta_{i(d)}, \tag{4}$$

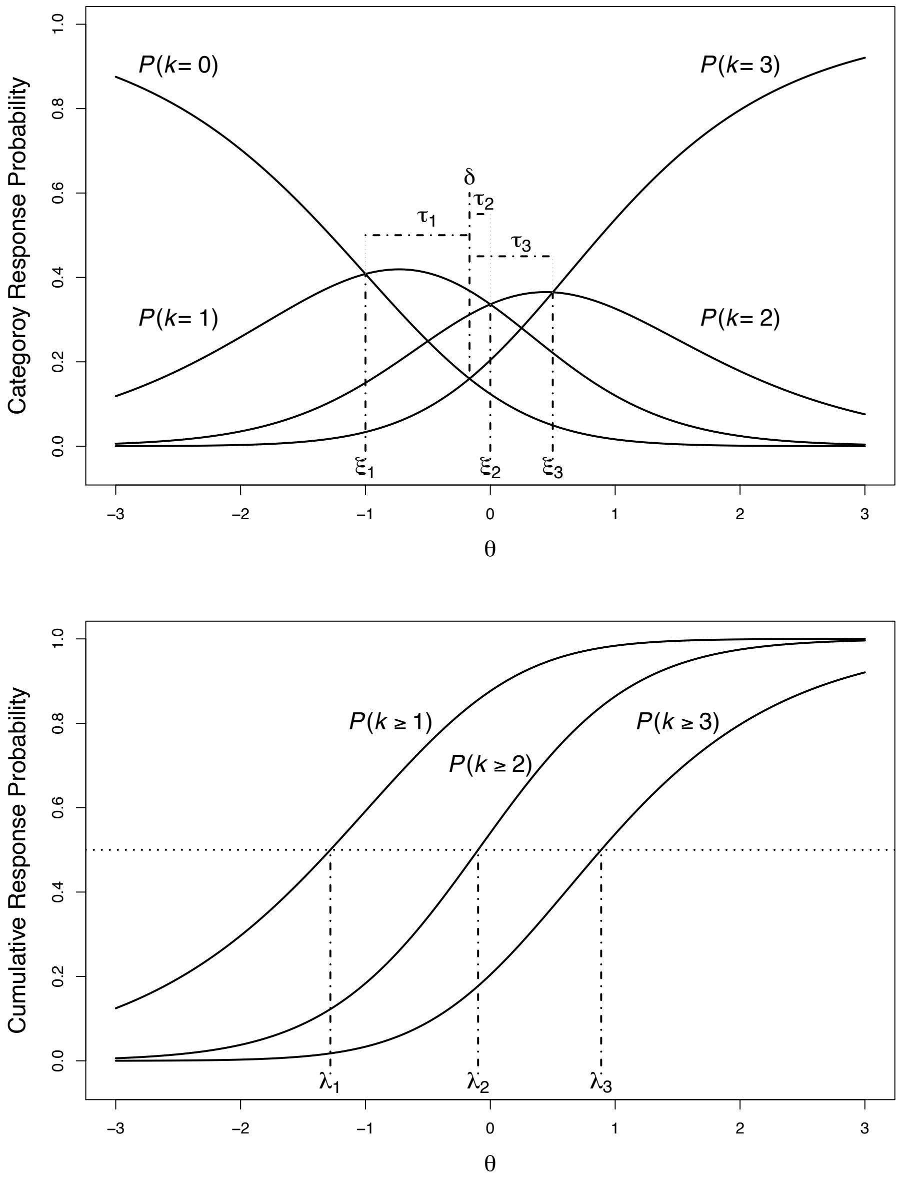
and

$$\xi_{i(d)k} = \delta_{i(d)} + \tau_{i(d)k}. \tag{5}$$

The $\xi_{i(d)k}$ parameter signifies the θ value at which the probability of responding in category $k - 1$ equals the probability of responding in category k. The $\delta_{i(d)}$ parameter can be interpreted either as the average of $\boldsymbol{\xi}_{i(d)} = (\xi_{i(d)1}, \xi_{i(d)2}, \dots \xi_{i(d)m})'$ parameters, or as the θ value at which the probability of responding in the lowest category equals the probability of responding in the highest category. Importantly, the PCM does not impose any order restrictions on $\boldsymbol{\xi}_{i(d)}$. In contrast to the formal PCM model parameters, the Thurstone threshold (Masters, 1988), denoted $\lambda_{i(d)k}$ for category $k = 1, \dots, m$, equals the θ value at which the probability of responding in category k or higher equals .5. Thurstone thresholds reflect *cumulative* response probabilities and are necessarily ordered. For this reason, many researchers often prefer to interpret the Thurstone thresholds instead of other model parameters. Thurstone thresholds are computable from the formal PCM parameters $\boldsymbol{\xi}_{i(d)}$ or $\delta_{i(d)}$ and $\boldsymbol{\tau}_{\mathrm{i}(d)} = (\tau_{i(d)1}, \tau_{i(d)2}, \dots \tau_{i(d)m})'$ using Newton-Raphson iteration or other numerical methods. An illustration of the roles of each parameter type is shown in Figure 1. See Wu, Tam, and Jen (2016) for further comparisons of these different parameterizations of the PCM.

Figure 1

Roles of Item Parameters in the Partial Credit Model



Alignment Methods

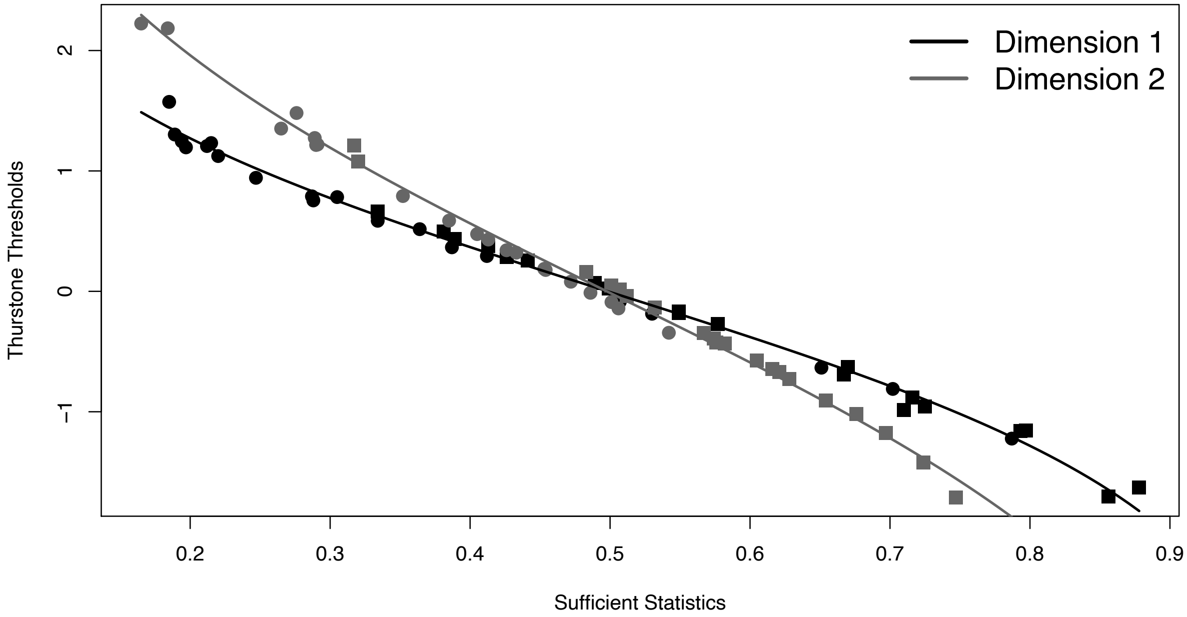
Next, we describe how scale alignment can be applied to item response models for responses that are scored polytomously. Under the PCM, the proportions of responses that achieve at least each score category $k = 1, \ldots, m$, $\hat{p}_{i(d)k}$, are sufficient statistics for the item parameters $\boldsymbol{\xi}_{i(d)}$, assuming that α_d is fixed and not estimated (Masters, 1982). Unlike for the simple Rasch model, however, the sufficient statistics need not be monotonically related to the item parameters $\xi_{i(d)k}$ or $\delta_{i(d)}$ and $\tau_{i(d)k}$. Therefore, it is not straightforward to define aligned scales in terms of a monotonic relationship between sufficient statistics and estimated item parameters, as was done previously with binary item response models. One solution is to define scale alignment for polytomously scored items in terms of parameters that are monotonically related to the sufficient statistics. As such, even though $\hat{p}_{i(d)k}$ is a sufficient statistic for $\xi_{i(d)k}$, there is not a simple monotonic relationship between these two quantities. There is similarly no monotonic relationship between the sufficient statistics and $\tau_{i(d)}$ or $\delta_{i(d)k}$. Instead, because Thurstone thresholds are necessarily ordered, we might expect that there is a monotonic relationship between $\hat{p}_{i(d)k}$ and $\lambda_{i(d)k}$. If this relationship is indeed monotonic (or nearly monotonic), it may be a useful criterion for scale alignment. Although this choice does not directly utilize the sufficiency relationship between $\hat{p}_{i(d)k}$ and $\xi_{i(d)k}$, it may be the nearest analog to the implementation of scale alignment used by Feuerstahler and Wilson (2019), which relied heavily on the monotonicity relationship between sufficient statistics and parameter estimates.

To investigate the relationship between sufficient statistics and Thurstone thresholds, we simulated data with $N = 1{,}000$ examinees from the between-items multidimensional PCM with two dimensions. Data-generating θ values were generated from a multivariate normal distribution, with correlation .5 between dimensions, a mean of 0 for both dimensions, and dimension standard deviations equaled 1 and 2. The dataset included 40 items wherein 20 were generated from the first θ dimension, 20 were generated from the second θ dimension, and each item had three response categories. These data were then fit to the the between-items multidimensional PCM using the TAM package (Robitzsch, Kiefer, & Wu, 2018) for R (R Core Team, 2019) and fixing $\alpha_1 = \alpha_2 = 1$. Figure 2 displays a scatter plot of the item sufficient statistics and the estimated Thurstone thresholds. In this figure, points corresponding to the first dimension are in black, and points corresponding to the second dimension are in gray. Moreover, circles represent the thresholds between response categories 0 and 1, and squares represent the thresholds between response categories 1 and 2. The best-fitting logistic curve is drawn through the points for each dimension. This figure suggests that, within dimensions, there is a strong nonlinear correlation between item sufficient statistics and Thurstone thresholds. For the first dimension, the absolute value of Kendall's rank-order correlation (Kendall, 1970) between these two quantities equals .974, and for the second dimension, this correlation equals .951. If dimension membership is ignored, this correlation drops to .928. Although this relationship is not exact, this example clearly shows that the relationship between sufficient statistics and Thurstone thresholds within dimensions is stronger than across dimensions. The aim of scale alignment is to linearly transform one or both of these dimensions such that all points lie on the same curve.

In theory, aligned dimensions are those for which the same sufficient statistics imply the same parameter estimates regardless of dimension. However, for most data sets, every unique sufficient statistic will not occur for every dimension. Instead, the degree to which dimensions are aligned can be evaluated in terms of the absolute rank-order correlation between sufficient statistics and Thurstone thresholds. If this correlation equals 1, then the dimensions can be considered perfectly aligned. Because any linear transformation of Rasch family models are admissible, we propose that researchers seek a linear transformation that maximizes the absolute rank-order correlation between sufficient statistics and Thurstone thresholds.

Figure 2

Scatterplot of the Relationship Between Sufficient Statistics and Thurstone Thresholds



Note. Circles correspond to the thresholds between categories 0 and 1, and squares correspond to the thresholds between categories 1 and 2.

Linear Transformations of PCM Model Parameters

Under the PCM, the θ_d metric for each dimension is determined only up to linear transformations. These linear transformations affect the scaling of the dimensions but do not affect model-data fit. Suppose that we wish to transform trait estimates from the original metric θ_d to another metric $\tilde{\theta}_d$ such that

$$\tilde{\theta}_d = r_d \theta_d + s_d, \quad (6)$$

where r_d is a multiplicative scaling constant and s_d is a shift constant. The model that results from this transformation follows the form of the PCM with

$$\tilde{\alpha}_d = \frac{\alpha_d}{r_d}, \tag{7}$$

$$\tilde{\delta}_{i(d)} = r_d \delta_{i(d)} + s_d, \tag{8}$$

$$\tilde{\tau}_{i(d)k} = r_d \tau_{i(d)k}, \tag{9}$$

and

$$\tilde{\xi}_{i(d)k} = r_d \xi_{i(d)k} + s_d. \tag{10}$$

These transformations are adapted from those that previously have been derived for the generalized partial credit model (Haberman, 2009). Therefore, the goal of any scale alignment method is to find r_d and s_d that improves alignment. In most cases, this goal may be able to be satisfied by a number of distinct linear transformations. We see lack of uniqueness as a feature, rather than a shortcoming of the concept of scale alignment. In this respect, it is helpful to compare scale alignment to factor rotation. In the same way that many factor rotation methods aim to better achieve the ideal of simple structure, scale alignment aims to better achieve the ideal relationship between sufficient statistics and item parameter estimates.

Delta Dimensional Alignment

Delta dimensional alignment (DDA; Schwartz & Ayers, 2012; Yamada, Draney, Karelitz, Moore, & Wilson, 2006) was the first method proposed for scale alignment in between-items multidimensional IRT models. This method is based on the concept of a reference dimension (Ackerman, 1992), a single dimension that represents the composite of the suite of dimensions. In practice, DDA works by fitting both a unidimensional and a multidimensional model. For binary responses, this results in two item difficulty estimates for each item—one from the multidimensional model and one from the unidimensional model. Note that the estimates from the unidimensional model (which represents the reference dimension) are required to be aligned, since all unidimensional models are automatically aligned. However, the estimates from the multidimensional model are not necessarily aligned across dimensions. Using these two sets of estimates, each dimension is transformed such that the within-dimension mean and variance of the item difficulties from the multidimensional model are equal to the within-dimension mean and variance of those item difficulties from the unidimensional model.

When dealing with polytomous responses, however, there are several quantities that serve the role of the item difficulties in dichotomous models and may be eligible for use in the DDA method. For a dimension with n items, these candidate parameters include $\boldsymbol{\delta}_d$, $\boldsymbol{\xi}_d$, and $\boldsymbol{\lambda}_d$, where bold notation indicates all parameters of the given type belonging to the given dimension. Let $\boldsymbol{\eta}_d$ indicate one of these types of parameter vectors (either $\boldsymbol{\delta}_d$, $\boldsymbol{\xi}_d$, or $\boldsymbol{\lambda}_d$). Then, the DDA method then defines the linear transformation parameters (see Equations (6)— (10)) as follows:

$$\hat{r}_d = \frac{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{U}d})}{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{M}d})} \tag{11}$$

and

$$\hat{s}_d = \text{mn}(\widehat{\boldsymbol{\eta}}_{\mathcal{U}d}) - \hat{r}_d\text{mn}(\widehat{\boldsymbol{\eta}}_{\mathcal{M}d}), \tag{12}$$

where $\mathcal{M}$ indicates parameters from the multidimensional model, and $\mathcal{U}$ indicates parameters from the unidimensional model. If dimension 1 is set as an unchanged reference dimension ($\hat{r}_1 = 1$ and $\hat{s}_1 = 0$), then

$$\hat{r}_d = \frac{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{U}d})\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{M}1})}{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{M}d})\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{U}1})} \tag{13}$$

and

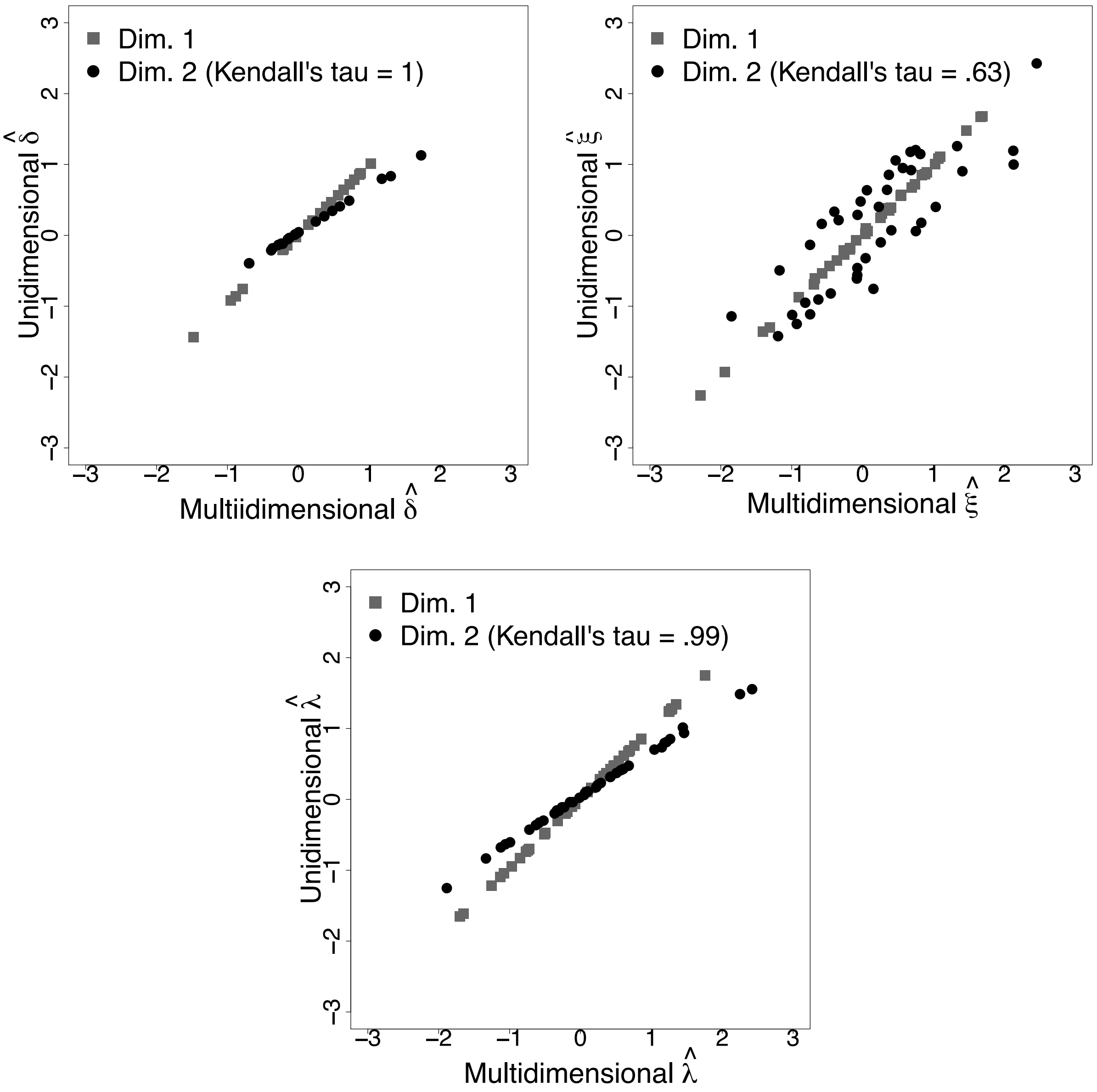
$$\hat{s}_d = \frac{\text{mn}(\widehat{\boldsymbol{\eta}}_{\mathcal{U}d}) - \text{mn}(\widehat{\boldsymbol{\eta}}_{\mathcal{U}1}) + \frac{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{U}1})}{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{M}1})}\text{mn}(\widehat{\boldsymbol{\eta}}_{\mathcal{M}1}) - \frac{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{U}d})}{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{M}d})}\,\text{mn}(\widehat{\boldsymbol{\eta}}_{\mathcal{M}d})}{\frac{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{U}1})}{\text{sd}\,(\widehat{\boldsymbol{\eta}}_{\mathcal{M}1})}}. \tag{14}$$

The above results leave open the question of which parameters—$\boldsymbol{\delta}_d$, $\boldsymbol{\xi}_d$, or $\boldsymbol{\lambda}_d$—should be used with DDA. The logic of DDA implies that, within dimensions, there should be a strong linear relationship between multidimensional parameter estimates and unidimensional parameter estimates. To investigate which parameters might be appropriate for DDA, we again consider the simulated two-dimensional data set presented in Figure 1. We fit these data to a unidimensional model again using the TAM package and compared these estimates to the multidimensional parameter estimates obtained earlier. Scatterplots of multidimensional and unidimensional estimates of these three parameter types are shown in Figure 3.

As Figure 3 shows, there exists a strong linear relationship between $\hat{\boldsymbol{\delta}}_{\mathcal{M}d}$ and $\hat{\boldsymbol{\delta}}_{\mathcal{U}d}$, and between $\hat{\boldsymbol{\lambda}}_{\mathcal{M}d}$ and $\hat{\boldsymbol{\lambda}}_{\mathcal{U}d}$. However, there is a weaker relationship between $\hat{\boldsymbol{\xi}}_{\mathcal{M}d}$ and $\hat{\boldsymbol{\xi}}_{\mathcal{U}d}$. Note that because the first dimension is the reference dimension, these relationships are perfectly linear. Further investigation (not reported here) shows that the strength of the relationship between $\hat{\boldsymbol{\xi}}_{\mathcal{M}d}$ and $\hat{\boldsymbol{\xi}}_{\mathcal{U}d}$ depends on the correlation between dimensions, whereas the relationships between the other sets of parameters are not sensitive to dimension correlations. This suggests that the degree of model misspecification when fitting the unidimensional model to multidimensional data affects the spacing of the item steps, which are represented by the $\boldsymbol{\xi}$ or $\boldsymbol{\tau}$ parameters. As a result, unless the correlation between $\hat{\boldsymbol{\xi}}_{\mathcal{M}d}$ and $\hat{\boldsymbol{\xi}}_{\mathcal{U}d}$ is close to 1 (as will happen as the correlation between dimensions approaches 1), DDA computed from the $\hat{\boldsymbol{\xi}}_d$ parameters may not align dimensions reliably. For this reason, we will only further investigate DDA implemented with $\hat{\boldsymbol{\delta}}_d$ and $\hat{\boldsymbol{\lambda}}_d$.

Figure 3

Scatterplots of Unidimensional and Multidimensional Item Parameter Estimates



Note. Data were generated according to a two-dimensional model and fit to both unidimensional and two-dimensional models.

Logistic Regression Alignment

Another viable scale alignment method for polytomous item responses is logistic regression alignment (LRA), first described by Feuerstahler and Wilson (2019). For dichotomously scored item responses, the LRA methods works by fitting a logistic regression curve between the sufficient statistics and the item difficulties for each dimension. When applied to the PCM, a logistic regression curve can be fit between the sufficient statistics and the Thurstone thresholds for each dimension. Specifically, for each dimension,

$$\text{logit}(Pr\left(y_{ij} \geq k \middle| \lambda_{i(d)k}\right) = \hat{\gamma}_{0d} + \hat{\gamma}_{1d}\lambda_{1(d)k}, \tag{15}$$

where $\hat{\gamma}_{0d}$ and $\hat{\gamma}_{1d}$ are the estimated intercept and slope for the logistic regression curve. After the logistic regression is fit to each dimension, one dimension (here, the first dimension) is chosen as the fixed reference dimension. The non-fixed dimensions are then transformed such that the fitted logistic regression curves are identical for each dimension. Specifically,

$$\hat{r}_d = \frac{\hat{\gamma}_{1d}}{\hat{\gamma}_{11}} \tag{16}$$

and

$$\hat{s}_d = \frac{\hat{\gamma}_{0d} - \hat{\gamma}_{01}}{\hat{\gamma}_{11}}. \tag{17}$$

Previous comparisons of DDA to LRA for binary response data (Feuerstahler & Wilson, 2019) demonstrated that both methods are broadly effective for scale alignment and that no one method systematically leads to greater alignment (as operationalized by the absolute rank-order correlation between sufficient statistics and transformed item difficulties). The LRA method may have a slight advantage over DDA in that it is usually more computationally efficient to compute a series of logistic regressions than to fit the unidimensional PCM. However, these computational differences are relatively small, and we recommend applying both methods and retaining the solution that leads to the highest absolute rank-order correlation.

Missing Data

The above presentation of scale alignment methods assumes complete data sets, which is unrealistic for most real testing applications. Complete data is required because the sufficient statistics for the Rasch model and the PCM are only truly sufficient insofar as there are no missing data. Scale alignment is certainly possible in the context of missing data, but first we must consider the relationship between similar quantities and the estimated parameters. In the presence of missing data, the relationship between the data and the parameter estimates depends on how the estimation algorithm handles missing responses. Multidimensional item response models are often fit using the marginal maximum likelihood (MML; Bock & Lieberman, 1970; Bock & Aitkin, 1981) algorithm. Within the EM algorithm, rather than maximizing the marginal likelihood of the full data, the marginal likelihood is maximized with respect to expected response counts. The expected response counts, which are obtained in the "E" step of the EM algorithm, take into account that different numbers of respondents are expected to exist at each quadrature point. In the full-information maximum likelihood application of the EM algorithm to IRT item calibration (Bock, Gibbons, & Muraki, 1988), if some data are missing, the "E" step is computed from the complete data part of the likelihood. This solution is valid under ignorable missing data mechanisms[1]. Because in the "E" step, the expected counts of examinees and correct responses are computed from the available data, it makes sense to consider modified

[1] Under nonignorable missing data mechanisms, treating missing responses as ignorable can lead to bias in both the item parameter estimates and the proportion of examinees in each response category (Rose, von Davier, & Xu, 2010). Methods that appropriately account for nonignorable missing data patterns (see Rose et al.) are beyond the scope of this paper.

sufficient statistics (that are no longer truly sufficient statistics) as the corresponding quantity computed from the available data. With these considerations in mind, we suggest treating the proportion of *complete* responses to that occur at or below each category $k = 1, \dots m_i$ as the $p_{i(d)k}$ values to evaluate the success of scale alignment. We believe that this solution is appropriate in the context of scale alignment because (a) most estimation algorithms assume that data are missing at random such that the proportion of complete responses is a good estimate of the "true" proportion if all responses were observed, and (b) the relationship between sufficient statistics and Thurstone thresholds is often not perfectly monotonic, even for complete data sets.

Illustration

Next, we illustrate scale alignment or the Kindergarten Individual Development Survey (KIDS), a multidimensional survey of kindergarten readiness used in the state of Illinois. On this instrument, teachers or caregivers rate each child on 38 measures (items) belonging to 5 domains (dimensions): *Approaches to Learning and Self-Regulation* (ATL-REG; 4 measures), *Social and Emotional Development* (SED; 5 measures), *Language and Literacy Development* (LLD; 10 measures)*, Math* (COG; 10 measures), and *Physical Development and Health* (PDH; 9 measures). This instrument also includes measures of the domains *History and Social Science*, *Visual and Performing Arts*, and (for children whose first language is not English) *English Language Development*—which will not be considered here. Ratings on the KIDS instrument were collected from ratings of 59,429 kindergarten-age children in the state of Illinois in the spring of 2015. In this data set, 30% of the possible ratings were missing due to intentional omission by the rater (e.g., the rater indicated that they were unable to rate the child on a measure), and item-level missingness rates ranged from 21% to 51%. On average, children were rated on 26 of the 38 measures, and only 26% of children were rated on all measures. These data were fit to the between-items multidimensional PCM using the EM algorithm as implemented in ConQuest (Adams, Wu, & Wilson, 2015) and by fixing person means equal to 0.

Before alignment, the fitted model resulted in an absolute rank-order correlation of .924 between the item sufficient statistics and the estimated Thurstone thresholds (ignoring dimension membership). The estimated dimension variances equaled 7.90, 13.58, 8.12, 9.49, and 5.89 for the ATL-REG, SED, LLD, COG, and PDH domains. We then applied the three scale alignment methods (DDA on $\widehat{\boldsymbol{\delta}}$, DDA on $\widehat{\boldsymbol{\lambda}}$, and LRA) to the fitted model, setting ATL-REG to be the unchanged reference dimension for each. Results of this model fitting are presented in Table 1. For this model, LRA and DDA $\widehat{\boldsymbol{\lambda}}$ increased the absolute rank-order correlation to .934 and .937, but the greatest increase occurring for the DDA method on $\widehat{\boldsymbol{\delta}}$ with an absolute rank-order correlation of .971. This would lead us to retain the DDA on $\widehat{\boldsymbol{\delta}}$ solution and to transform the model parameters using the first set of $\hat{r}$ and $\hat{s}$ values given in Table 1, following the parameter transformations presented in Equations (7) to (9). This solution suggests that the SED, LLD, and COG dimensions should have smaller latent variances than ATL-REG, and that PDH should have greater latent variance than ATL-REG.

Table 1

Results of three scale alignment methods on KIDS rating scale data.

		ATL-REG	SED	LLD	COG	PDH
DDA $\hat{\boldsymbol{\delta}}$	$\hat{r}$	1.000	0.832	0.973	0.912	1.128
	$\hat{s}$	0.000	-0.005	0.006	0.009	0.008
DDA $\hat{\boldsymbol{\lambda}}$	$\hat{r}$	1.000	1.097	1.071	1.034	1.148
	$\hat{s}$	0.000	0.092	0.098	0.071	0.103
LRA	$\hat{r}$	1.000	1.095	1.081	1.050	1.170
	$\hat{s}$	0.000	0.072	0.100	0.129	0.164

To illustrate the effect of scale alignment on an individual trait estimate vector, consider a child who receives a “2” on all 38 measures. In the original model, this student’s weighted likelihood (WLE; Warm, 1989) latent trait estimates equal -0.69, -1.12, -1.50, -1.02, and -1.11 for the five domains. On the aligned model, this student’s WLE latent trait estimates equal -0.69, -0.93, -1.46, -0.92, and -1.25. These differences in estimates are subtle, and it is not self-evident that the estimates from the aligned model are more appropriate comparisons than those from the unaligned model. However, these differences do reflect a scaling that more closely aligns the item parameter interpretations across dimensions, as indicated by the absolute rank-order correlation between Thurstone thresholds and the modified sufficient statistics.

Simulation

Finally, we briefly report the results of a small-scale simulation study to evaluate the effectiveness of scale alignment for polytomous Rasch family models. For the Rasch model for dichotomously scored items, Feuerstahler and Wilson (2019) found that both DDA and LRA were broadly effective at alignment and found little evidence that one method systematically outperforms the other. The goal of these simulations is to determine whether one of the three alignment methods consistently outperforms the other methods in the context of the between-items multidimensional PCM with varying proportions of missing data.

Previously, Feuerstahler and Wilson (2019) found that differences in the latent trait variance across dimensions leads to the largest alignment adjustments (and greatest increase in the absolute rank-order correlation between sufficient statistics and item parameters). For this reason, we will induce misaligned dimensions by simulating latent trait values from a multivariate normal distribution with possibly different variances per dimension. In the current set of simulations, all data sets were simulated from a between-items two-dimensional PCM with either 5 or 20 items per dimension. All simulated data sets include $N = 500$ subjects whose two-dimension latent trait values was generated from a multivariate normal distribution with a mean vector of 0 and a correlation of .5 between dimensions. For the first dimension, the latent trait standard deviation equaled 1, and for the second dimension, the latent trait standard deviation equaled .5, 1, or 2. Data-generated PCM item parameters were generated using the catR (Magis & Barrada, 2017) packages for R. The genPolyMatrix function in this package generates $\xi_{i(d)k}$ parameters from independent standard normal distributions. The number of response categories per item was set equal to either 3, 5, or 7. Missing data were simulated by deleting responses to either 0% or a randomly chosen 25% or 50% of the complete simulated data matrix. Finally, each of these conditions (2 tests lengths × 3 dimension two standard deviations × 3 response category conditions × 3 missingess conditions) was fully crossed for 54 conditions and replicated 100 times for a total of 5400 conditions. Data sets were then fit to the correctly specified between-items multidimensional PCM with TAM and by fixing person means equal to

0 and aligned using the three alignment methods described earlier. Scale alignment was performed using our scaleAlign package for R, which we will make available soon on CRAN.

The median rank-order correlations between sufficient statistics and Thurstone thresholds for each condition are displayed before and after each alignment method in Table 2. This table shows that DDA on $\hat{\boldsymbol{\lambda}}$ and LRA always lead to higher median rank-order correlations than DDA on $\hat{\boldsymbol{\delta}}$. In fact, DDA on $\hat{\boldsymbol{\delta}}$ often led to a *decrease* in rank-order correlations. The improvement in rank-order correlations for DDA on $\hat{\boldsymbol{\lambda}}$ and LRA is most pronounced when the standard deviation of θ_2 does not equal 1. This is to be expected because scale alignment is intended to account for differences in latent distributions, and there is no difference in latent distributions when standard deviation of θ_2 equals 1. In cases for which the standard deviation of θ_2 does not equal 1, the median increase in rank-order correlation equals .038 for DDA on $\hat{\boldsymbol{\lambda}}$ and .040 for LRA.

Comparing the three scale alignment methods within each replication of each condition, the highest absolute rank-order correlation is observed for DDA on $\hat{\boldsymbol{\delta}}$ in 9.6% of simulations, for DDA on $\hat{\boldsymbol{\lambda}}$ in 44.8% of simulations, and for LRA in 45.6% of simulations. The vast majority of the conditions for which DDA on $\hat{\boldsymbol{\delta}}$ is selected (i.e., has the highest rank-order correlation for a particular replication) are those in which the standard deviation of θ_2 equals 1, when scale alignment is not theoretically necessary. Effects for the other manipulated factors are relatively small, considering effects both within and across replications. DDA on $\hat{\boldsymbol{\lambda}}$ tends to be selected most often and leads to higher median rank-order correlations for models with 3 response categories, and LRA tends to be selected most often for models with either 5 or 7 response categories for all items. Test length and the proportion of missing data does not appear to affect the frequency with which scale alignment method was selected. However, the absolute value of rank-order correlations are generally higher for shorter tests and conditions with lesser proportions of missing data, particularly when the standard deviation of θ_2 does not equal 1.

These results demonstrate that for a particular replication, each scale alignment method sometimes leads to the greatest improvement in dimension alignment, and the most appropriate method depends on a number of factors that are typically not known in advance of alignment. Fortunately, in practice it is always possible to apply all three alignment methods and to retain the solution that leads to the maximum absolute rank-order correlation between sufficient statistics and Thurstone thresholds. Therefore, we recommend applying all scale alignment methods to real data sets and retaining the best performing solution.

Table 2

Median Rank-Order Correlations Before and After Alignment

5 items/dim.		0% missing				25% missing				50% missing			
sd(θ_2)	#cats	Unaligned	DDA $\boldsymbol{\delta}$	DDA $\boldsymbol{\lambda}$	LRA	Unaligned	DDA $\boldsymbol{\delta}$	DDA $\boldsymbol{\lambda}$	LRA	Unaligned	DDA $\boldsymbol{\delta}$	DDA $\boldsymbol{\lambda}$	LRA
.5	3	.945	.926	**.958**	**.958**	.937	.926	**.958**	**.958**	.937	.916	**.958**	**.958**
	5	.914	.941	.954	**.955**	.915	.938	**.954**	**.954**	.910	.937	**.951**	**.951**
	7	.894	.905	.954	**.955**	.889	.899	.948	**.950**	.886	.899	.943	**.944**
1	3	.968	.964	**.968**	**.968**	.968	.958	**.968**	**.968**	.968	.958	**.968**	**.968**
	5	.958	.950	**.960**	**.960**	.954	.945	**.956**	**.956**	.946	.944	**.951**	**.951**
	7	.943	.944	.943	**.946**	.936	.937	.937	**.939**	.924	.923	.925	**.929**
1.5	3	.923	.863	**.968**	**.968**	.926	.863	.958	**.968**	.916	.874	**.958**	**.958**
	5	.919	.899	.959	**.963**	.910	.892	.949	**.952**	.903	.887	.944	**.945**
	7	.889	.867	.926	**.938**	.882	.860	.921	**.929**	.871	.853	.906	**.914**

20 items/dim.		0% missing				25% missing				50% missing			
sd(θ_2)	#cats	Unaligned	DDA $\boldsymbol{\delta}$	DDA $\boldsymbol{\lambda}$	LRA	Unaligned	DDA $\boldsymbol{\delta}$	DDA $\boldsymbol{\lambda}$	LRA	Unaligned	DDA $\boldsymbol{\delta}$	DDA $\boldsymbol{\lambda}$	LRA
.5	3	.947	.942	**.963**	**.963**	.945	.940	**.960**	.959	.941	.936	**.956**	**.956**
	5	.921	.929	**.954**	**.954**	.919	.927	**.952**	.951	.914	.922	**.946**	**.946**
	7	.908	.918	**.944**	**.944**	.902	.912	.940	**.941**	.903	.911	**.936**	**.936**
1	3	.965	.965	**.965**	.964	.959	.959	**.960**	.959	.953	.953	**.954**	**.954**
	5	.955	.955	**.956**	**.956**	.951	.951	**.952**	**.952**	.945	.944	**.945**	**.945**
	7	.947	.947	**.949**	**.949**	.939	.939	.940	**.941**	.932	.932	.933	**.934**
1.5	3	.925	.895	**.965**	.964	.915	.891	**.956**	**.956**	.908	.883	**.946**	.945
	5	.890	.891	.952	**.954**	.887	.884	.944	**.945**	.880	.879	**.933**	**.933**
	7	.876	.867	.925	**.931**	.869	.861	.927	**.930**	.859	.857	.916	**.917**

Note. Bolded values indicate the highest median rank-order correlations within each condition. Medians are taken across 100 replications.

Discussion

In this paper, we describe how the concept of scale alignment for multidimensional Rasch models for binary item responses might be extended to polytomous item responses and incomplete data sets. In this paper, we define aligned scales as those for which the same item sufficient statistics imply the same Thurstone thresholds. To this end, we describe three scale alignment methods delta-dimensional alignment (DDA) on the average item step parameters $\boldsymbol{\delta}$, DDA on the Thurstone thresholds $\boldsymbol{\lambda}$, and logistic regression alignment (LRA). We recommend evaluating the success of scale alignment methods by calculating the absolute rank-order correlation between sufficient statistics and Thurstone thresholds before and after applying these alignment methods. Evidence from analyzing real and simulated data suggest that these three alignment methods are generally effective, and that which method is most effective varies across data sets. In practice, we recommend applying every scale alignment method and retaining the solution that leads to the maximum rank-order correlation between sufficient statistics and Thurstone thresholds.

This paper also discusses how missing data might affect scale alignment. Specifically, we recommend computing modified sufficient statistics for the PCM as the proportion of *complete* responses, effectively treating missingness as ignorable. In practice, missing responses are often nonignorable. More sophisticated modeling techniques may be used in these situations, but these methods are situation-specific, and the corresponding extension to scale alignment methods is a topic for future study. The methods described above might also be used to align dimensions for which different items have different numbers of observed response categories. However, researchers may want to account for the different numbers of categories by differentially weighting the alignment parameters (e.g., $\boldsymbol{\lambda}$ or $\boldsymbol{\delta}$). Further research is needed to determine the most appropriate and effective weighting scheme when items have different numbers of response categories.

Scale alignment is premised on the fact that all scalings of different dimensions in multidimensional item response models are equally admissible. In this way, we believe that scale alignment is conceptually similar to factor rotation in exploratory factor analysis (EFA). In EFA, an infinite number of factor rotations are admissible, but some rotations lead to more interpretable structures than others. Similarly, we view scale alignment as a way to improve across-dimension interpretability of individual trait estimates by ensuring that the ordering of items has a consistent meaning across dimensions. Overall, we present scale alignment as a concept that we believe is useful for comparing estimates across dimensions more meaningfully than if the dimensions are not aligned.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Adams, R. J., Wilson, M., & Wang, W. -C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1-23. https://doi.org/10.1177/0146621697211001

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised Item Response Modelling Software* [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment validation: An empirical example. *Electronic Journal of Research in Educational Psychology*, *10*(1), 233-252.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*(4), 443-459. https://doi.org/10.1007/BF02293801

Bock, D. R., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280. https://doi.org/10.1177/014662168801200305

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179-197. https://doi.org/10.1007/BF02291262

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87–100.

Cheng, Y. -Y., Wang, W. -C., & Ho, Y. -H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth-fidelity dilemma. *Educational and Psychological Measurement*, *69*(3), 369-388. https://doi.org/10.1177/0013164408323241

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement*, *39*(1), 59–84. https://doi.org/10.1111/j.1745-3984.2002.tb01135.x

Feuerstahler, L. M., & Wilson, M. (2019). Scale alignment in between-item multidimensional Rasch models. *Journal of Educational Measurement*, *56*(2), 280-301. https://doi.org/10.1111/jedm.12209

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations*. *Research Report RR-09-40*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x

Kendall, M. G. (1970). *Rank correlation methods*. Griffin.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, *Code Snippets*, *76*(1), 1-19. http://dx.doi.org/10.18637/jss.v076.c01

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education*, *1*(4), 27 9–297. https://doi.org/10.1207/s15324818ame0104_2

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024-1048. https://doi.org/10.1002/tea.21397

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. -Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821-846. https://doi.org/10.1002/tea.21316

Patz, R. J., & Junker, B. W. (1999) Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366. https://doi.org/10.3102/10769986024004342

R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Available from http://www.R-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2018). TAM: Test analysis modules. R package version 3.0-21. https://CRAN.R-project.org/package=TAM

Schwartz, R., & Ayers, E. (2011). Delta dimensional alignment: Comparing performances across dimensions of the learning progression for assessing modeling and statistical reasoning. Berkeley, CA: Unpublished manuscript, University of California, Berkeley.

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.). *Rasch Models* (pp. 215-237). Springer.

Wang, W. -C. (1996). *Implementation and application of the multidimensional random coefficients multinomial logit* [Unpublished doctoral dissertation]. University of California, Berkeley.

Wang, W. -C., Chen, P. -S., & Cheng, Y. -Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116-136. https://doi.org/10.1037/1082-989X.9.1.116

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wu, M. L., Tam, H. -P., & Jen, T. -H. (2016). Partial credit model. In M. L. Wu, H. -P. Tam, & T. -H. Jen (Eds.), *Educational Measurement for Applied Researchers* (pp. 159–185). Springer.

Yamada, H., Draney, K., Karelitz, T., Moore, S., & Wilson, M. (2006). *Comparison of dimension-aligning techniques in a multidimensional IRT context*. Presented at the 13th International Objective Measurement Workshop, Berkeley, CA.

Yao, S.-Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the function of content and argumentation items in a science test: A multidimensional approach. *Journal of Applied Measurement*, *16*(2), 171–192.