PSYCHOMETRIKA—VOL. 81, NO. 3, 650–673
SEPTEMBER 2016
DOI: 10.1007/s11336-015-9469-6

LINKING ITEM RESPONSE MODEL PARAMETERS

WIM J. VAN DER LINDEN AND MICHELLE D. BARRETT

CTB/McGRAW-HILL EDUCATION

With a few exceptions, the problem of linking item response model parameters from different item calibrations has been conceptualized as an instance of the problem of equating test scores on different test forms. This paper argues, however, that the use of item response models does not require any *test score equating*. Instead, it involves the necessity of *parameter linking* due to a fundamental problem inherent in the formal nature of these models—their general lack of identifiability. More specifically, item response model parameters need to be linked to adjust for the different effects of the identifiability restrictions used in separate item calibrations. Our main theorems characterize the formal nature of these linking functions for monotone, continuous response models, derive their specific shapes for different parameterizations of the 3PL model, and show how to identify them from the parameter values of the common items or persons in different linking designs.

Key words: 3PL response model, item calibration, linking design, linking function, parameter identifiability.

1. Introduction

The literature on item response model parameter linking tends to conceptualize the problem of linking the parameters from different calibrations as a step in the process of test score equating. For instance, for the well-known dichotomous logistic response models, Kolen and Brennan (2004, p. 156) treat the linking problem as the second step in a three-step process consisting of (i) estimating the item parameters in the response model for a new test form, (ii) scaling the parameters back to a base scale using a linear transformation, and (iii) if number-correct scoring is used, number-correct scores on the new form are converted to number-correct scores on an old form and then to scale scores. References to the problem addressed in this paper as an equating problem are also found, for instance, in Dorans, Pommerich, and Holland (2007), Holland and Rubin (1982), and von Davier (2011).

The main model considered in this paper is the fixed-effects three-parameter logistic (3PL) response model, which explains the probability of a correct response $U_{pi} = 1$ for a test taker p on item i with ability $\theta_p \in \mathbb{R}$ as

$$\Pr\{U_{pi} = 1; \theta_p\} \equiv p(\theta_p; a_i, b_i, c_i) \equiv c_i + (1 - c_i)\Psi[a_i(\theta_p - b_i)], \tag{1}$$

with

$$\Psi[a_i(\theta_p - b_i)] \equiv \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}. \tag{2}$$

where $b_i \in \mathbb{R}$ and $a_i > 0$ are parameters for the difficulty and discriminating power of item i, respectively, and $c_i \in (0, 1]$ represents the height of the lower asymptote to the probability for

Correspondence should be made to Wim J. van der Linden, CTB/McGraw-Hill Education, 20 Ryan Ranch Road, Monterey, CA 93940, USA. Email: wjvdlinden@outlook.com

the item. Let θ^* and θ denote the abilities of arbitrary test takers on an old and new test form calibrated under this model. The linear transformation used in the second step above is

$$\theta^* = u\theta + v, \tag{3}$$

with parameters u and v to be derived from response data for the two forms.

The standard argument for the claim of linearity of the transformation in the literature relies on the notion of indeterminacy of the scale of the (estimated) θ scores (e.g., Kim, Harris, & Kolen, 2010, p. 264; Lord, 1980, sect. 3.5). More precisely, it points at an arbitrary zero and unit for these scores, which manifest themselves by the fact that we can always transform θ as in (3), provided the two remaining parameters in (2) are replaced by

$$a_i^* = a_i/u \tag{4}$$

and

$$b_i^* = ub_i + v. \tag{5}$$

for all i.

Popular methods to estimate the parameters u and v are the mean/sigma method (Macro, 1977), the mean/mean method (Loyd & Hoover, 1980), and the methods based on the entire response functions for the common items in the two test forms by Haebara (1980) and Stocking and Lord (1983). The first two methods are based on a choice from the following relationships between the parameter values in the two calibrations:

$$\begin{aligned} u &= \frac{\mu(a)}{\mu(a^*)},\ \mu(a^*) > 0, && (6)\\ &= \frac{\sigma(b^*)}{\sigma(b)},\ \sigma(b) > 0, && (7)\\ &= \frac{\sigma(\theta^*)}{\sigma(\theta)}, \sigma(\theta) > 0. && (8)\end{aligned}$$

and

$$\begin{aligned} v &= \mu(b^*) - u\mu(b) && (9)\\ &= \mu(\theta^*) - u\mu(\theta), && (10)\end{aligned}$$

with $\mu(\cdot)$ and $\sigma(\cdot)$ denoting means and standard deviations. These methods are applied substituting the means and/or variances of the parameter estimates for the common item and persons in the linking study in these expressions. The *BILOG* computer program (Mislevy & Bock, 1990) included a version of these methods with the arithmetic means of the b parameters in (9) but the geometric instead of the arithmetic means of the a parameters in (6); for a generalization of this log-mean/mean procedure, see Haberman (2009). The Stocking-Lord method finds estimates of u and v minimizing the squared difference between the sums of the response functions for common items in the two test forms. For the three-parameter logistic (3PL) model in (1), the criterion to be minimized is

$$\left[\sum_i p(\theta; a_i^*, b_i^*, c_i^*) - \sum_i p(\theta; a_i/u, ub_i + v, c_i)\right]^2, \tag{11}$$

for a selection of θ values and with estimates substituted for the item parameters. For further details on these methods, see Kolen and Brennan (2004, sects. 6.2–6.3).

From a practical point of view, this treatment of the linking of response model parameters as a step in test score equating seems to make sense. IRT parameter linking does have some history in the context of IRT observed-score equating (Lord, 1980). And the fact that the necessary data for the estimation of the linear transformation in (3) are collected using the same type of sampling designs as used in plain observed-score equating (equivalent-groups designs; anchor-item designs; etc.) seems to lend additional support to the treatment of IRT parameter linking in this context.

From a more theoretical perspective, however, objections to this point of view are possible. First of all, a characteristic feature of all item response models is the presence of separate parameters for the effects of the properties of the items and the test takers' abilities on the response probabilities. A naïve observer may note that the item parameters already adjust the probabilities for the differences between the items in the test forms and wonder where the necessity of the additional linking does come from.

A more fundamental puzzle is the assumed linearity of the linking transformation. The argument of the logistic function in (2) is definitely nonlinear (products of a_i with θ_p and b_i). So it would be wrong to use this aspect of the parameter structure to motivate the shape of the transformation. Further, although the notion of a measurement scale for an ability with an indeterminate zero and unit has a long tradition in the behavioral and social sciences, enforced by classical publications such as Stevens (1946), its use in the current context focuses our attention exclusively on the scale of the θ parameter that is measured. But the model in (1)–(2) has four parameters for each response probability. Why should we be interested in the a_i and b_i parameters only because of features of the scale of θ? And how about the c_i parameters? If their scale is determinate, how come the Haebara and Stocking-Lord methods, which are sensitive to estimation error in these parameters, have been claimed to outperform the mean/mean and mean/sigma methods (e.g., Baker & Al-Karni, 1991)? But if their scale is indeterminate, how could we ever motivate the popularity of the last two methods in the linking literature, which ignore the c_i parameter completely?

Similar questions arise if we reparameterize the model. For example, for computational reasons (use of the Gibbs sampler), it has become convenient for Bayesian estimation to reparameterize the argument of the logistic function in (2) as $\alpha_i \vartheta_p + \beta_i$ (Albert, 1992). But does the scale of ϑ for this version of the model still have an indeterminate zero and unit? And is its linking transformation still linear?

The view of parameter linking in this paper is solely as a fundamental problem due to a formal feature of item response models—their general lack of identifiability. The notion of model identifiability seems akin to the one of indeterminacies of scales in Stevens' (1946) classification of measurement scales. But, unlike Stevens' classification, which just consists of a set of definitions of different levels for the scale of the parameter we try to measure and then lets us wonder how to establish the nature of the scale in a specific measurement situation, it implies a formal criterion that can be applied directly to the measurement model that is used. Loosely speaking, the criterion requires us to check if each possible distribution of the response data implies a unique set of values for *all* parameters in the model.

If a models lacks identifiability, the problem can be resolved by adjusting its numbers of equations and/or parameters, which in the current context of IRT parameter estimation with fixed numbers of item and person parameters leads to the necessity of extra restrictions on them. The well-known practice of setting the mean and standard deviation of the ability parameters in a maximum marginal likelihood (MML) calibration of the items equal to

$$\mu_\theta = 0, \sigma_\theta = 1 \tag{12}$$

is an example of the use of such restrictions.

However, even if the problem of identifiability is resolved, a new problem arises. The general effect of the use of identifiability restrictions is different values for the parameters of the same items and test takers in different calibrations. Hence, these parameters can only be compared if we know the function that maps the set of their values for one calibration onto those for the other. Once all parameters are linked, the response model automatically adjusts for any relevant differences between items or test takers, and future estimates of any ability parameter for given item parameters (or reversely) are always directly comparable. Thus, linking functions are not necessary to correct for arbitrary units and zeroes of the θ parameters but, more generally, *to adjust for the different effects of the identifiability restrictions used in separate calibrations*.

Observe that the differential effect also arises if we use (12) for two separate calibrations. Using superscripts to index different groups of test takers, we then have

$$\mu_\theta^{(1)} = 0, \quad \sigma_\theta^{(1)} = 1 \tag{13}$$

and

$$\mu_\theta^{(2)} = 0, \quad \sigma_\theta^{(2)} = 1. \tag{14}$$

The prevalent practice of not indexing different groups of test takers in (12) may easily lead to the erroneous belief that these restrictions always have the same effect. However, as explained in more detail below, each of these two sets of restrictions yields a different intersection with the model equations and hence different identified values for all model parameters.

As just noted, linking functions are thus mathematical functions that map the set of values for the item and ability parameters in the response model for one calibration onto those for another that are necessary because of its lack of identifiability. The main theorems in this paper characterize these functions for the general class of monotone, continuous item response models, and derive their specific shapes for different parameterizations of the 3PL model with the fixed ability parameters in (1)–(2). In addition, they show how to identify the linking functions from the parameter values of common items or persons for different linking designs. As the current focus is only on the mathematical definition of linking functions, we treat all item and ability parameters as known and postpone the treatment of the statistical problem of estimating such functions. Before presenting the theorems, a few notions from the literature on model and parameter identifiability necessary to understand the nature of these functions are reviewed.

2. Observational Equivalence and Identifiability

We restrict the review of the problem of identifiability to the class of models that serve as parametric probability functions for the distribution of (discrete) random variables. Except for an occasional definition (e.g., Casella & Berger, 2002, sect. 11.2), the problem does not have much of a history in textbooks on statistics. All families of distributions typically discussed in these texts have probability functions with standard parameters that are identifiable. But the problem does have an active history of research in econometrics, mainly because of its tradition of modeling these standard parameters as functions of quantities of substantive interest, as well as in generalized latent variable modeling. Classical papers in the econometric literature discussing parameter identifiability include Koopmans (1949), Fisher (1961; 1965), Rothenberg (1971), Richmond (1974), and Gabrielsen (1978). Bekker, Merckens, and Wansbeek (1994) offer an enlightening analysis of the problem of identifiability in structural equation modeling. For an introduction from the perspective of generalized latent variable modeling, see Skrondal and Rabe-Hesketh (2004, chap. 5).

The problem of identifiability does arise in IRT because of its similar attempt to explain standard parameters of response distributions as functions of item and person parameters. Relevant papers addressing the problem for a variety of response models include Bechger, Verhelst, et al. (2001), Bechger, Verstralen, et al. (2002), Fischer (2004), Maris (2002), Maris and Bechger (2004; 2009), Reiersøl (1950), Revuelta (2009), San Martín, Gonzáles, and Tuerlinckx (2009), San Martín, Jara, et al. (2011), Tsai (2000), and Volodin and Adams (2002). The problem of linking parameters estimated under different identifiability restrictions seems to be restricted mainly to item response theory, however. At least, these authors are not aware of any other field where parameters estimates are linked as frequently and routinely as in educational and psychological testing; the only exception known to them is Luijben's (1991) treatment of the equivalence of two differently restricted versions of an unidentifiable structural equations model (also addressed by Bekker et al., 1994, chap. 7). For the treatment of parameter linking for a response-time model with item and person parameters from the perspective of model identifiability, see van der Linden (2010).

The following definitions, which can be found throughout the literature just referred to, are for a model of a random vector with a multidimensional parameter space:

Definition 1. Two points in a parameter space are observationally equivalent if they imply the same joint distribution of the random vector.

Definition 2. A parameter is identifiable if for any of its points there is no other point that is observationally equivalent.

More formally, let $\mathbf{x}$ denote the random vector that is considered and $f(\mathbf{x};\boldsymbol{\pi})$ its probability function, which is assumed to have vector-valued parameter $\boldsymbol{\pi}$. Then, $\boldsymbol{\pi}$ is identifiable if for any pair $\boldsymbol{\pi}_0 \neq \boldsymbol{\pi}_1$, it holds that $f(\mathbf{x};\boldsymbol{\pi}_0) \neq f(\mathbf{x};\boldsymbol{\pi}_1)$ for all $\mathbf{x}$. Observe that lack of identifiability may be due to some of the components of $\boldsymbol{\pi}$ only. As $\mathbf{x}$ typically represents observed data, it is common to refer to a parameter as being "identifiable from the data." The use of this phrase emphasizes the practical meaning of parameter identifiability: If different values of $\boldsymbol{\pi}$ imply the same probability distribution, it becomes impossible to use observed data to distinguish between them, let alone infer a "true" value of $\boldsymbol{\pi}$. Consequently, if $\boldsymbol{\pi}$ is not identifiable, then for some of its values, the likelihood function $f(\boldsymbol{\pi};\mathbf{x})$ associated with the observations does not allow us to discriminate between them. Indeed, if a parameter lacks identifiability, it does not have a consistent estimator (Gabrielsen, 1978).

Definition 2 immediately suggests possible refinements of the criterion of identifiability, such as local identifiability of $\boldsymbol{\pi}$ at $\boldsymbol{\pi}_0$ (i.e., $\boldsymbol{\pi}$ is identifiable in a neighborhood of $\boldsymbol{\pi}_0$) or identifiability from a restricted set of values of x. But more important to our current goals is a discussion of a slightly generalized version of a theorem in Bartels (1985):

Theorem 1. *If $\boldsymbol{\pi}$ and $\boldsymbol{\varphi}$ are the parameters in alternate versions of a model of a given random variable that have a bijective relationship, then $\boldsymbol{\pi}$ is identifiable if and only if $\boldsymbol{\varphi}$ is.*

The theorem is immediately obvious if we realize that, in the current context, parameters serve as quantities that index individual members of families of probability distributions. As the relation between $\boldsymbol{\pi}$ and $\boldsymbol{\varphi}$ is bijective (one-to-one and onto), their role as index is entirely exchangeable.

The theorem explains why we can reparameterize a model (replace its structure with one set of parameters by a structure with another set) without losing its identifiability, provided the two sets of parameters have a bijective relationship. An example is the slope-intercept parameterization $\alpha_i\vartheta_p + \beta_i$ of the logistic model referred to earlier. In order to give the response function this parameter structure, we have to substitute

$$\begin{aligned} \theta_p &= \theta(\vartheta_p) = \vartheta_p; \\ a_i &= a(\alpha_i) = \alpha_i; \\ b_i &= b(\alpha_i, \beta_i) = -\beta_i/\alpha_i, \end{aligned} \tag{15}$$

$\vartheta_p, \beta_i \in \mathbb{R}$ and $\alpha_i > 0$, into (2). The relation between the two alternative sets of parameters is invertible, and thus bijective.

A special instance of the function $\boldsymbol{\varphi} = \varphi(\boldsymbol{\pi})$ in Theorem 1 is the vector function

$$\boldsymbol{\varphi} = (\varphi_1(\pi_1), \ldots, \varphi_d(\pi_d)), \tag{16}$$

with $\varphi_1, \ldots, \varphi_d$ being scalar-valued, bijective functions of the d different components of $\boldsymbol{\pi}$. An example of this *componentwise* type of bijective function is the well-known reparameterization of the Rasch (1960) model,

$$p(\vartheta_p; \beta_i) \equiv \frac{\vartheta_p}{\vartheta_p + \beta_i}, \tag{17}$$

$\vartheta_p, \beta_i > 0$, which, maintaining our current notation, follows from (2) with $a_i = 1$ upon substitution of

$$\begin{aligned} \theta_p &= \theta(\vartheta_p) = \ln \vartheta_p; \\ b_i &= b(\beta_i) = \ln \beta_i, \ \vartheta_p, \beta_i > 0. \end{aligned} \tag{18}$$

The critical difference between the two types of reparameterization resides thus in the fact that *each* component in (18) is a bijective function of its counterpart as well, whereas the component for b_i in (15) is not.

Reparameterization is a useful tool if we have to prove identifiability of a model. It allows us to replace the set of model equations with the current parameters by an equivalent set for which the proof is simpler. We will use this trick to prove some of our later results. Also, the problem of parameter linking will appear to be one in which we have to derive and estimate functions as in (16).

As already noted, the typical solution to an identifiability problem for an item response model is the introduction of extra restrictions on its parameters. Their effect is a reduction of the parameter space to one that uniquely represents each possible member of the family of response distributions posited by the model. However, such restrictions generally have a differential impact on the parameter space, lead to different unique representations, and consequently leave us with a linking problem. Because our focus is mainly on this problem, we only highlight the nature of the identifiability problem for the 3PL model in (1)–(2), presenting a few cases in which different sets of parameters in the 3PL model in (1)–(2) clearly show lack of identifiability (including the commonly believed to be invariant c_i parameters). It is not our intention to provide a solution for it. In fact, as follows from Theorem 3 below, in order to derive a linking function for a monotone continuous response model, it is not necessary to know the identifiability restrictions actually used in the different calibrations at all; only their differential impact on the item and ability parameters values counts.

3. 3PL Model

The distributions addressed by the 3PL model in (1) are for the dichotomous responses $U_{pi} = 0, 1$ by test takers $p = 1, \ldots, P$ on items $i = 1, \ldots, I$. The distributions are Bernoulli with probability functions

$$f(u_{pi}; \pi_{pi}) = \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}}, \ p = 1, \ldots, P; \ i = 1, \ldots, I, \tag{19}$$

which have success parameters $\pi_{pi} \in [0, 1]$ representing the probability of a correct response by each test taker on each item.

For different values of its parameter π_{pi}, each of the probability functions in (19) yields a different distribution; hence these parameters are identifiable. Observe that if these probability functions are reparameterized by substituting $\pi_{pi} = 1 - \eta_{pi}$, Theorem 1 guarantees that η_{pi} is also identifiable. We will use this feature frequently.

Making the usual assumption of independence within and between test takers, the probability function of the joint distribution of a complete response matrix, $\mathbf{U} = (U_{pi})$, is the product of $P \times I$ of these Bernoulli distributions,

$$f(\mathbf{u}; \boldsymbol{\pi}) = \prod_p \prod_i \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}} \tag{20}$$

with parameter vector $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1I}, \ldots, \pi_{P1}, \ldots, \pi_{PI})$. Clearly, as each of its components is identifiable, so is $\boldsymbol{\pi}$.

The 3PL model specifies each π_{pi} as a function of the parameters $(\theta_p, a_i, b_i, c_i)$ for the effects of the test taker's ability and the properties of the item on it. Rather than a direct probability function for a response distribution, the model is thus a (second-level) mathematical model in the form of a system of $P \times I$ nonlinear equations

$$\pi_{pi} = c_i + (1 - c_i)\Psi[a_i(\theta_p - b_i)], \ p = 1, \ldots, P; \ i = 1, \ldots, I, \tag{21}$$

one for each of the success parameters.

3.1. Lack of Identifiability

The following three cases illustrate the lack of identifiability of the 3PL model:

Theorem 2. *The system of equations for the 3PL model in (21) is not identifiable in the following cases: (i) c_i known for all i; (ii) $a_i = a \in \mathbb{R}^+$, for all i; and (iii) $\theta_p = \theta \in \mathbb{R}$ for all p.*

Proof. (i) This case basically amounts to the 2PL model in (2). The fact that different values of a_i, b_i, and θ_p parameters do not need to imply different values for π_{pi} follows from (3)–(5), and is well known. (ii) Lack of identifiability of the b_i, c_i, and θ_p parameters for this case was established by Maris (2002) and later re-analyzed in Maris and Bechger (2009). Without loss of generality, his proof sets $a = 1$ and reformulates (1)–(2) as

$$\pi_{pi} = \frac{\exp(\theta_p) + c_i \exp(b_i)}{\exp(\theta_p) + \exp(b_i)}, \tag{22}$$

which, upon substitution of $\theta_p = \ln \vartheta_p$, $b_i = \ln \beta_i$, and $c_i = \ln \delta_i / \beta_i$, leads to

$$\pi_{pi} = \frac{\vartheta_p + \delta_i}{\vartheta_p + \beta_i}. \tag{23}$$

Adding a constant to ϑ_p for all p and subtracting the same constant from β_i and δ_i for all i gives different sets of values for these parameters with the same π_{pi}. (iii) Let $\pi_{pi} = 1 - \eta_{pi}$ and

$c_i = 1 - \gamma_i$. Theorem 1 implies that π_{pi} and c_i are identifiable if and only if η_{pi} and γ_i are. From (1)–(2), using $\theta_p = \theta$ for all p,

$$\begin{aligned} \eta_i &= \frac{\gamma_i}{[1 + \exp(a_i(\theta - b_i))]} \\ &= \frac{\gamma_i^*}{\kappa[1 + \exp(a_i(\theta - b_i))]}, \end{aligned} \tag{24}$$

for all i and any $\kappa \in (0, \gamma_i]$, where $\gamma_i^* = \kappa\gamma_i$. In order to have b_i absorb κ, we need to substitute b_i^* for b_i, where b_i^* is the solution of

$$1 + \exp(a_i(\theta - b_i^*)) = \kappa[1 + \exp(a_i(\theta - b_i))], \tag{25}$$

which is

$$b_i^* = \theta - \ln[\kappa[1 + \exp(a_i(\theta - b_i))] - 1]/a_i. \tag{26}$$

For the relation in (26) to hold, it is necessary that $\kappa[1 + \exp(a_i(\theta - b_i))] - 1 > 0$; or,

$$\begin{aligned} \kappa &> \frac{1}{1 + \exp(a_i(\theta - b_i))} \\ &= 1 - \Psi_i, \end{aligned} \tag{27}$$

with Ψ the logistic function in (2). Alternatively, the change in the γ_i parameters can be traded off by

$$a_i^* = \ln[\kappa[1 + \exp(a_i(\theta - b_i))] - 1]/(\theta - b_i) > 0, \tag{28}$$

or

$$\theta^* = b_i + \ln[\kappa[1 + \exp(a_i(\theta - b_i))] - 1]/a_i \tag{29}$$

where the nonnegativity requirement for a_i^* implies both $\kappa > 2(1 - \Psi)$ and $\theta > b_i$. □

The status of the c_i parameters in the version of the 3PL model with all parameters free is still unknown. But the last two cases are already enough to illustrate the problematic role of these parameters, which has largely been ignored in the literature. Lord (1980, pp. 36, 184–185) even claims that the c_i parameters are actually identifiable. As already noted, an exception was Maris (2002), who introduced the second case above.

For the third case, Fig. 1 illustrates how dramatic the tradeoff between the γ_i and a_i, b_i and θ parameters can be. It displays each of the tradeoffs as a function of κ for an item with $a_i = 1.0$, $b_i = -.5$ and common ability parameter $\theta = 0$. For these parameter values, $1 - \Psi = .378$; hence, the range of admissible values for κ is (.378, 1] for the functions in (26) and (29) but (.755, 1] for the one in (28). Observe that $\kappa = 1$ means no change in the γ_i parameter; for smaller values of κ, γ_i decreases in size (and c_i thus becomes larger). The change in the b_i and θ parameters as a function of the decrease in γ_i is remarkable, especially closer to their vertical asymptote at $\kappa = .378$. On the other hand, the a_i parameter appears to be quite robust across its range of admissible values for κ. The fact that the tradeoff between the c_i and a_i parameters seems much less dramatic than for the other two parameters might go against our intuition, but is entirely due to the nonnegativity requirement for the a_i parameters. If we did admit negative values, this parameter could go down, for example, all the way to $a_i^* = -10$ for $\kappa = .38$ (just above the vertical asymptote).

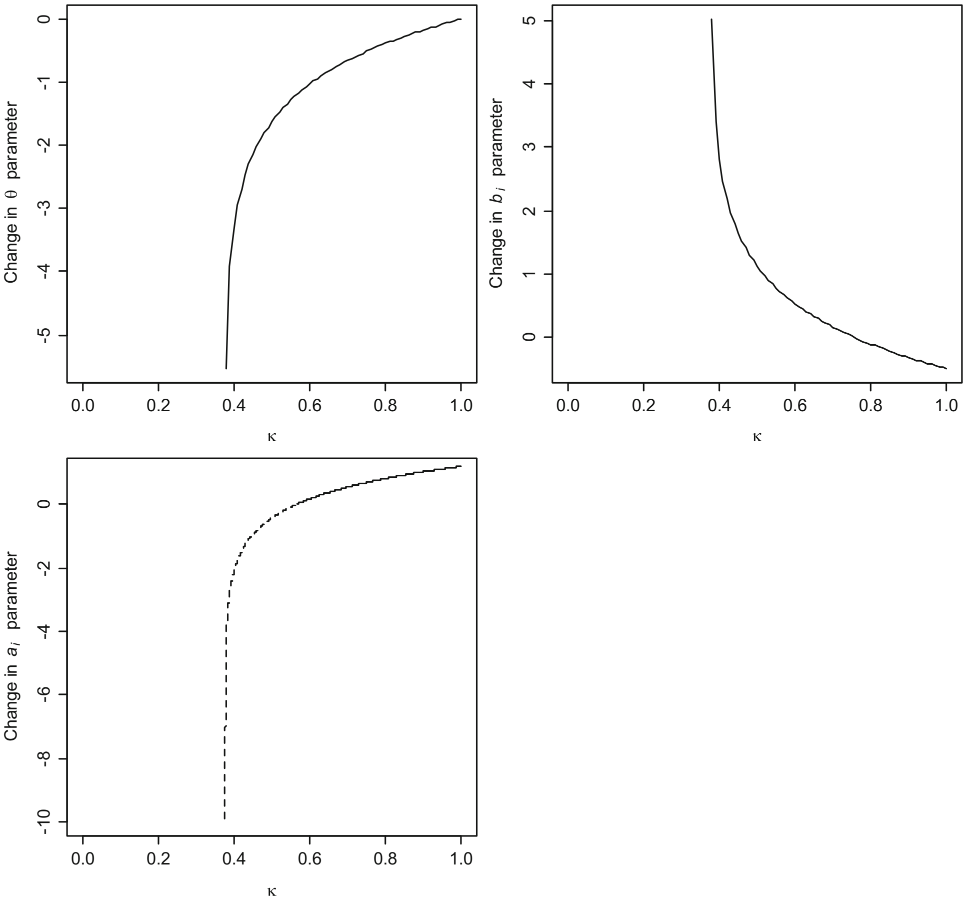

FIGURE 1.
Change in θ, b_i, and a_i parameters compensating the change in $\gamma_i = 1 - c_{\mathrm{i}}$ by a factor k, for an item with $a_i = 1$ and $b_i = -0.5$ and the ability parameter fixed at $\theta = 0$. *Dashed line* represents negative values for the a_i parameter.

The tradeoffs in (3)–(5) imply lack of identifiability of all parameters in the 2PL model and 1PL/Rasch model. Wood (1978) analyzed a special version of the Rasch model for equivalent items with success probability

$$\pi_{pi} \equiv \frac{\exp(\theta_p)}{1 + \exp(\theta_p)}, \tag{30}$$

for all i. This "0PL model" is just a reparameterization of the Bernoulli probability function in (19). It is thus fully identifiable (although unlikely to show any satisfactory fit to the items a real-world testing program).

The literature offers only a few examples of sets of identifiability restrictions for special cases of the 3PL model in (1)–(2) for which sufficiency has formally been proven. For instance, for the 1PL/Rasch model, as is well known, it is sufficient to set the difficulty parameter of one item or the ability parameter of one test taker equal to a known constant. Equivalently, we could impose a linear constraint on a subset of these parameters (e.g., constrain their mean to a known value). As shown by San Martín et al. (2009) for the 1PL-G model (i.e., 1PL/Rasch model extended with a guessing parameter for each item), it is sufficient to fix the difficulty and guessing parameters of one item to known constants. Recently, the same authors have shown that, contrary to what one might have expected intuitively, it is not sufficient to fix all three parameters of an arbitrary item to known constants to make the (fixed-effects) 3PL model in (21) identifiable (San Martín, Gonzáles, & Tuerlinckx, 2015). In fact, we are not aware of any existing proof of a set of identifiability restrictions sufficient for it. In the absence of such proofs (but the presence of large amounts of response data), the practical solution in the testing industry has been to circumvent the problem using a two-stage calibration procedure. In its first stage, the ability parameters are temporarily treated as a random sample from a population distribution, which allows for marginalization of the likelihood for the fixed-effects version of the model with respect to an assumed ability distribution and maximum marginal likelihood (MML) estimation of all item parameters. The second stage consists of subsequent maximum likelihood or Bayesian (e.g., EAP) estimation of the individual ability parameters assuming the item parameters have been estimated from enough response data to treat them as known. Typically, the ability distribution in the first stage is taken to be the standard normal, which implies the adoption of (12) as *de facto* standard set of identifiability restrictions for the 3PL model in the field of educational testing. Although its effectiveness has been confirmed in the daily practice of item calibration (e.g., convergence of parameter estimates, which would have been problematic with lack of identifiability) as well as through numerous parameter recovery studies, we still await a formal proof of its sufficiency. As for the second stage, if the item parameters can be treated as known, (21) defines known monotonic relationships between each of the ability parameters θ_p and their success parameters π_{pi}, and the former are therefore identified as well.

For all practical purposes, the use of (12) in this two-stage detour thus restricts the parameters in the specification of the 3PL model in (21) to fixed values (to which the MML estimators of the item parameters and the subsequent estimates of the ability parameters converge with the sample size and test length, respectively). The reverse does not necessarily hold, though; identifiability of a fixed-effect specification does not automatically imply the same for its random-effect specification (San Martín, Rolin, et al., 2013).

4. Linking Functions

The main goal of this section is to define the problem of linking IRT parameters from different calibration studies and derive the specific linking functions necessary for the 3PL model. More specifically, we address the problem of post hoc linking; that is, mapping the parameters from one study onto the values they would have had if they had been included in another *after* both calibrations have been conducted. This type of linking is common in the testing industry. As both the item and ability parameters are to be linked simultaneously, the linking automatically is for the fixed-effects specification of the 3PL model. In principle, it is possible to avoid the problem by concurrent recalibration of the response data collected in different studies, capitalizing on the presence of common items or test takers in them and imposing constraints on the parameters. For this approach, it has even been proposed to tentatively impose constraints, e.g., linear constraints as in (3)–(5) and check on their appropriateness using Lagrange multiplier tests (von Davier & von Davier, 2011). But such strategies are not always practical for testing programs that have to link their parameters continuously across multiple test administrations.

Our current treatment of the linking problem only deals with its mathematical aspects at the level of the model parameters, without any bothering about the fact that these parameters are unknown. In order to actually use them in practice, linking functions need to be estimated from response data. But we are only able to find defendable estimators and evaluate their statistical quality once we have an explicit definition of their estimand.

We begin with considering the more general case of a parametric response model used to calibrate the responses for a set of P test takers on I items with the vector of success probabilities

$\boldsymbol{\pi} = (\pi_{pi})$ in (20). Let $f(\cdot)$ be the response function specified by the model, $\boldsymbol{\xi}_{pi}$ its vector of parameters for the combination of test taker p and item i, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_{pi})$ the vector of parameters for all test takers and items. The choice of model amounts to the adoption of a system of $P \times I$ equations $\pi_{pi} = f(\boldsymbol{\xi}_{pi})$. As the probabilities π_{pi} are identified and thus have fixed values for all combinations of p and i, each of the equations introduces a level surface (contour) in the domain of f, which is the subset of all values of $\boldsymbol{\xi}$ for which $f(\boldsymbol{\xi}_{pi}) = \pi_{pi}$ is true. The solution set for the system of equations is the intersection of all $P \times I$ surfaces. As the system lacks identifiability of the model parameters, the set consists of more than one point. Identifiability restrictions are extra equations added to the system. The intersection of their solution sets with the set for the system reduces the latter to a unique point, whose coordinates are the true values of the item and test taker parameters for the calibration.

Now, suppose we have conducted two separate calibration studies that had both unique and common test takers and/or items in them. Both studies are assumed to have used appropriate sets of identifiability restrictions. Obviously, the use of different restrictions implies different intersections of their solution sets with those determined by the two systems of model equations, and hence different true values for the common parameters in the two calibrations. But different true values can also arise if formally identical sets of identifiability restrictions have been imposed on the two calibrations. The presence of unique test takers and/or items in the calibrations implies different vectors of success probabilities $\boldsymbol{\pi}^* \neq \boldsymbol{\pi}$ for them and thus different solutions sets for their model equations. Consequently, their intersections with the solution set for the identifiability restrictions generally differ, and the same items or test takers assigned to the two calibrations can therefore have different true parameter values. The critical factor is the scope of the restrictions. For example, if they fix the values of some of the common parameters to the same known constants in the two calibrations, obviously their impact on them is identical. But if they include unique parameters and leave the common parameters free, they yield different true values for the latter—an observation confirmed by the educational testing industry, where invariably different parameter values are found for common parameters in separate calibrations with large samples of test takers for the restrictions in (13) and (14), as well as by our example later in this paper. In fact, if this differential effect did not exist, we would not have to link any parameters.

Consider a hypothetical combination of a test taker and item assigned to two of these calibration studies with identified parameters. Let $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}$ denote the vectors with the unique true values for the combination in the two studies (where the indices have been omitted for notational convenience, as well as to emphasize the hypothetical nature of the combination). For example, for the 3PL model, $\boldsymbol{\xi}^* = (\theta^*, a^*, b^*, c^*)$ and $\boldsymbol{\xi} = (\theta, a, b, c)$. The question of how to map $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ onto one another is the topic of this section. Observe that, although different, both $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}$ are associated with the *same* success probability π for the combination of test taker and item. This fact is key in our derivation of the mapping below.

Our first theorem is for a general response model that specifies success probability π for each combination of a test taker and item only as a monotone continuous function of their parameters, where the monotonicity is taken to mean that π is strictly increasing or decreasing in each of the components of $\boldsymbol{\xi}$ with all other components fixed at any of their admissible values. We then present our results for the 3PL model. The version of this model for the regular parameterization in (1)–(2) is both continuous and monotone in each of its parameters, provided we exclude the case of $\theta = b$ for the a parameter; the version with the slope-intercept parameterization is monotone in each of its parameters without any further restriction.

Again, except for an illustrative example below, the current paper only deals with the mathematical aspects of linking functions; the problem of how to actually estimate them and evaluate their estimation error deserves separate treatment.

Theorem 3. *Assume a response model with a fixed parameter structure that (i) specifies π as a monotone continuous function of its parameters and (ii) has been used in two separate calibration studies with identified parameters $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$. Then $\boldsymbol{\xi}^*$ is linked to $\boldsymbol{\xi}$ by a vector function*

$$\boldsymbol{\xi}^*{=}\varphi(\boldsymbol{\xi}){=}\ (\varphi_1(\xi_1), \ldots, \varphi_d(\xi_d)) \tag{31}$$

with components $\varphi_1, \ldots, \varphi_d$ that are both monotone and continuous.

Proof. Let ξ^* and ξ be an arbitrary pair of corresponding components of $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}$**.** Fixing all other components, the monotonicity of the model implies the existence of monotone functions $\pi{=}f(\xi^*)$ and $\pi{=}g(\xi)$. Hence, there also exists a function $\xi^* = f^{-1}(g(\xi)) = \varphi(\xi)$. Being a composite of continuous functions, φ is continuous. Further, as both f and g are bijective, φ is bijective as well. Suppose that φ is not monotone. It then has an interior point ξ_0 in its domain with a local optimum. But this implies the existence of points $\xi' < \xi_0 < \xi''$ with $\varphi(\xi') = \varphi(\xi'')$, which contradicts the fact that φ is bijective. Thus, φ is monotone. □

The feature of monotonicity should not come as a surprise. If it did not hold, the two sets of the identifiability restrictions would imply a different order of some of the parameters, for instance, a reversal of the difficulties of two items, which is impossible without violating the requirement of observational equivalence. Observe, however, that $\varphi(\boldsymbol{\xi})$ is only required to be *componentwise* monotone; it does not need to hold that all components be increasing or all of them be decreasing.

It is thus possible to view the impact of the use of different sets of identifiability restrictions as a componentwise reparameterization of the response model, which leaves the structure of the model intact. However, unlike the earlier case of a known function in (16) applied to unknown parameter values, we now have to address the reverse problem: This time the two sets of parameter values are given, and we have to find the componentwise bijective function that maps them onto one another. Observe again that the specific identifiability restrictions used in the calibrations need not be known at all; neither do we need to assume anything about the statistical estimation method through which the restrictions might have been imposed. Only their impact on the item and test taker parameters counts.

For the 3PL model, the linking function $\boldsymbol{\varphi}{=}\ (\varphi_\theta(\theta), \varphi_a(a), \varphi_b(b), \varphi_c(c))$ has to be derived from (1)–(2) for two arbitrary sets of values $(\theta^*, a^*, b^*, c^*)$ and (θ, a, b, c). However, it is simpler to use the first equation in (24), and find φ_θ, φ_a, φ_b, and φ_γ as the solution of

$$\frac{\varphi_\gamma(\gamma)}{1+\exp[\varphi_a(a)(\varphi_\theta(\theta)-\varphi_b(b))]} = \frac{\gamma}{1+\exp[a(\theta-b)]}, \tag{32}$$

with additional back transformation of φ_γ to φ_c. The required linking function is thus the solution of a functional equation in four unknowns (for relevant theory of functional equations, see, for instance, Sahoo & Kannappan, 2011, or Small, 2007). The next theorem shows the solution:

Theorem 4. *Given the conditions in Theorem 3, the linking function for the 3PL model is*

$$\varphi_a(a) = u^{-1}a, \tag{33}$$

$$\varphi_b(b) = ub + v, \tag{34}$$

$$\varphi_c(c) = c, \tag{35}$$

and

$$\varphi_\theta(\theta) = u\theta + v, \tag{36}$$

with

$$u \equiv \frac{\varphi_\theta(\theta) - \varphi_b(b)}{\theta - b},\ \theta \neq b, \tag{37}$$

and

$$v = \varphi_b(b) - ub = \varphi_\theta(\theta) - u\theta. \tag{38}$$

Proof. From (32),

$$\varphi_\gamma(\gamma) = \frac{1 + \exp[\varphi_a(a)(\varphi_\theta(\theta) - \varphi_b(b))]}{1 + \exp[a(\theta - b)]}\gamma. \tag{39}$$

As this function is monotone in γ, φ_γ is a monotone component of $\boldsymbol{\varphi}$, and therefore

$$\frac{1 + \exp[\varphi_a(a)(\varphi_\theta(\theta) - \varphi_b(b))]}{1 + \exp[a(\theta - b)]} = \kappa > 0, \tag{40}$$

is a constant independent of γ. Thus, $\varphi_\gamma(\gamma) = \kappa\gamma$. However, since φ_γ is a monotone mapping from [0, 1] onto itself, $\kappa = 1$ and (35) follows. We now have to find φ_θ, φ_a, and φ_b as the solution of (40) for $\kappa = 1$; that is,

$$\varphi_a(a)[\varphi_\theta(\theta) - \varphi_b(b)] = a(\theta - b). \tag{41}$$

Rewriting the equation,

$$\varphi_a(a) = \frac{\theta - b}{\varphi_\theta(\theta) - \varphi_b(b)}a, \tag{42}$$

with $\varphi_\theta(\theta) \neq \varphi_b(b)$. But, as φ_a is a monotone component of $\boldsymbol{\varphi}$ as well,

$$\frac{\varphi_\theta(\theta) - \varphi_b(b)}{\theta - b} = \text{const}, \tag{43}$$

which is our key equation. First, (33) follows directly from (43) along with the definition of its constant in (37). Further, (43) shows that $\varphi_\theta(x) - \varphi_b(x)$ is equal to a constant times $\theta - b$. Substituting $x = \theta = b$ yields $\varphi_\theta(x) - \varphi_b(x) = 0$, and it thus holds that $\varphi_\theta = \varphi_b = \varphi$. Observe also that (43) implies a constant difference quotient for φ. Hence, φ is linear, and (34) and (36) hold. Finally, (38) follows from (34)–(36). □

Although the proof did not make any assumptions as to the general shape of the functions that map the values of the a_i, b_i, and θ_p parameters in a new calibration onto those in an earlier calibration, they appear to be linear, just as currently assumed in the literature; see our review in (3)–(5). However, a new result is the definition of linking parameter u, and consequently of v. Unlike (6)–(8), (37) defines it as the ratio of the differences between the test taker's ability and the difficulty of the item in the two calibrations. The reason for the difference between these new and old definitions may be the failure in the current literature to distinguish between the formal definitions of u and v and their solutions from the system of linking equations implied by the choice of linking design. As demonstrated in the next section, separating the two does give us large flexibility to derive alternative solutions for u and v from alternative designs. Another new result is the derivation of the identity function for the c_i parameters. We will further reflect on its practical implications in the last section of this article.

In addition, it is important to note the different nature of these functions for the four different types of parameters. The one for the c_i parameters is an identity function, which does not involve either of the linking parameters u and v. On the other hand, the function for the a_i parameters involves one linking parameter, u, whereas those for the b_i and θ_p parameters depend both on u and v. Thus, unlike the c_i parameters, the latter can be linked only when the numerical values of the linking parameters are known. This obvious point takes us to another identifiability requirement, namely for the system of linking equations to be derived from (33)–(38) for the specific design adopted in the linking study.

5. Identification of Linking Parameters

Basically, a linking design is a combination of two calibration designs with common items and/or test takers. Once it has been selected, (33)–(38) can be used to derive a new system of equations of the unknown linking parameters u and v in the parameter values for the common items or test takers in the two calibrations. Of course, u and v have unique values only when the system is identified. The problem of linking item response model parameters thus involves three different types of identifiability requirements, two for the system of the model equations in (21) associated with the two calibrations and another for the system of linking equations that is used. At this stage, the former have already been met through the adoption of extra restrictions in the two calibrations. The latter, although sometimes critical (e.g., Theorem 7 below), involves only an appropriate choice of linking design; no additional restrictions are necessary.

We illustrate the process for three minimal designs. In doing so, $p = 1, \dots, n$ and $i = 1, \dots, m$ are now used as indices for the common persons and items in the design, respectively, while $t = 1, 2$ will be used to denote the two calibrations. Thus, (a_{i_t}, b_{i_t}) and θ_{p_t} are the true values of the pertinent parameters for item i and the parameter for test taker p in the tth calibration, respectively. Because the linking function for the c_i parameters is already know, these parameters can further be ignored.

5.1. *One Common Item*

Linking parameters u and v are already identified if the two calibrations have one common item, $i = 1$. The system of linking equations then follows from (33) and (38) as

$$u = \frac{a_{1_1}}{a_{1_2}},\ a_{1_2} > 0, \tag{44}$$

$$v = b_{1_2} - ub_{1_1}. \tag{45}$$

5.2. *Two Common Items*

For a pair of common items $i = 1, 2$, we could use (44)–(45) for either of them. But now an alternative is available in the form of the substitution of their two sets of parameter values (b_{1_1}, b_{2_1}) and (b_{1_2}, b_{2_2}), $b_{1_1} \neq b_{2_1}$, into (38). Elimination of v then gives

$$u = \frac{b_{1_2} - b_{2_2}}{b_{1_1} - b_{2_1}}, \tag{46}$$

whereupon v is equal to

$$v = b_{i_2} - ub_{i_1}, \quad i = 1, 2. \tag{47}$$

This simple system of equations gives unique values for the linking parameters once their item parameters are identified. A similar property does not hold for the next type of design.

5.3. *One Common Test Taker*

It appears to be impossible to derive a system of equations from (33) and (38) for a common test taker with $(\theta_{1_1}, \theta_{1_2})$ from which u and v are identifiable.

5.4. Two Common Test Takers

For a pair of test takers with $\theta_{1_1} \neq \theta_{2_1}$, u and v can be obtained as

$$u = \frac{\theta_{1_2} - \theta_{2_2}}{\theta_{1_1} - \theta_{2_1}}, \tag{48}$$

and

$$v = \theta_{p_2} - u\theta_{p_1}, \quad p = 1, 2. \tag{49}$$

Because of their practical importance, we document these results as a theorem:

Theorem 5. *For the 3PL model with standard parameterization, linking parameters u and v are already identifiable for a common-item design with at least one common item and a common-person design with at least two common test takers.*

As a typical linking study has more than these minimal numbers of items or test takers, we easily have multiple systems of linking equations, each returning the same unique values for the linking parameters u and v. At the current level of true parameter values, any choice from them would thus suffice. However, in practice, except when some of common item or ability parameters were fixed at known constants, in which case we can just substitute these constants into (44)–(49), all parameters are estimated. A first suggestion of how to combine estimates of u and v from multiple systems of linking equations as in (44)–(49) is offered in the empirical example below.

5.5. Slope-Intercept Parameterization

Earlier we wondered what the impact of a change of parameter structure on the linking function would be. Theorem 6 highlights the impact for the slope-intercept parameterization for the 3PL model.

Theorem 6. *Given the conditions in Theorem 3, the linking function for the slope-intercept parameterization $\alpha\vartheta + \beta$ of the 3PL model is*

$$\varphi_\alpha(\alpha) = (\alpha - u)/v, \quad v > 0, \tag{50}$$

$$\varphi_\beta(\beta) = \beta + u, \tag{51}$$

$$\varphi_c(c) = c \tag{52}$$

and

$$\varphi_\vartheta(\vartheta) = (\vartheta - u)/w, \quad w \neq 0, \tag{53}$$

with

$$u = \varphi_\beta(0), v = \varphi_\vartheta(1), \textit{ and } w = \varphi_\alpha(1). \tag{54}$$

Proof. The component for the linking of c does not change because these parameters are left untouched by the reparameterization. But we now have to find φ_ξ, φ_α, and φ_β as the solution of

$$\varphi_\alpha(\alpha)\varphi_\vartheta(\vartheta) + \varphi_\beta(\beta) = \alpha\vartheta + \beta, \tag{55}$$

or, equivalently,

$$\varphi_\beta(\beta) = \beta + \alpha\vartheta - \varphi_\alpha(\alpha)\varphi_\vartheta(\vartheta). \tag{56}$$

Substituting $\beta = 0$ and using the definition of u in (54) yields

$$u = \alpha\vartheta - \varphi_\alpha(\alpha)\varphi_\vartheta(\vartheta). \tag{57}$$

Hence, (51) follows from (56)–(57). Likewise, substituting $\vartheta = 1$ into (57), we obtain (50) with v given by (54), while substitution of $\alpha = 1$ leads to (53). □

Although still linear, the linking functions in (50)–(54) differ considerably from those for the regular parameterization of the 3PL model; in fact, they now appear to have three rather than two unknown parameters. More importantly, an attempt to derive an identified system of linking equations from them appears to run into practical problems. Still assuming all model parameter to be known, (54) suggests selecting a common item with $\beta_{i_1} = 0$ and $\alpha_{i_1} = 1$ and common test taker with $\vartheta_{p_1} = 1$ for the first calibration and equating the three linking parameters to their values in the second calibration, that is, setting $u = \beta_{i_2}$, $w = \alpha_{i_2}$, and $v = \vartheta_{p_2}$. However, the presence of items and test takers with such exact parameter values in a linking study is highly unlikely.

An alternative seems to solve (50)–(53) for u, v, and w, obtaining them as

$$u = \beta_{i_2} - \beta_{i_1} \tag{58}$$

$$v = \frac{\alpha_{i_1} - u}{\alpha_{i_2}}, \quad \alpha_{i_2} \neq 0, \tag{59}$$

and

$$w = \frac{\vartheta_{p_1} - u}{\vartheta_{p_2}}, \quad \vartheta_{p_2} \neq 0. \tag{60}$$

for an arbitrary p and i. However, we now need a linking design with both at least one common item for (58)–(59) and one common test taker for (60), a new requirement entirely due to the change in parameter structure created by (15). Hence the following theorem:

Theorem 7. *For the 3PL model with slope-intercept parameterization, linking parameters u, v, and w are only identifiable for designs with both common test takers and common items.*

The result in this theorem has a major practical implication. If one believes that IRT models always have a scale for the person parameter with an arbitrary unit and zero, as the literature referenced in our introductory section appears to do, it may seem natural to adopt the same linking functions as for the model with the standard parameterization in (33)–(38). This choice is incorrect; the proper functions are those in (58)–(60). However, we do not expect the latter to be practised regularly as they require a linking design with test takers responding twice but independently to the same items—an assumption unlikely ever to be met because of memory effects. An alternative would be to transform the estimated slope and intercept parameters back to the standard parameters and link the latter, but then we miss the covariance matrices for the estimators of the model parameters necessary to evaluate the standard errors of linking as in the example in the next section.

6. Illustrative Example

Although the focus of this paper was not yet on a statistical treatment of the linking problem, a small example might already illustrate some of the practical consequences of our theoretical results. The example is for a common-item design for the 3PL model in its standard parameterization. The design allows us to use (44)–(45) to estimate linking parameters u and v in (33)–(38) for the two calibrations.

Response data were generated for two test forms each existing of 20 unique and 20 common items. All common items had $c_i = .25$, while their a_i and b_i parameters were chosen to represent one of the possible combinations of $b_i = -2(.5)2$ with $a_i = .5, 1.5$, using $b_i = 0$ twice to get

TABLE 1.
Generating and estimated parameter values for the common items.

Common	Generating values			Calibration 1			Calibration 2		
item	a_i	b_i	c_i	$\widehat{a_i}$	$\widehat{b_i}$	$\widehat{c_i}$	$\widehat{a_i}$	$\widehat{b_i}$	$\widehat{c_i}$
1	1.500	−2.000	0.250	2.612	−0.843	0.213	2.162	−1.725	0.191
2	1.500	−1.500	0.250	2.707	−0.558	0.234	2.334	−1.333	0.212
3	1.500	−1.000	0.250	2.612	−0.297	0.232	2.358	−0.959	0.290
4	1.500	−0.500	0.250	2.751	−0.008	0.241	2.125	−0.728	0.228
5	1.500	0.000	0.250	2.830	0.283	0.258	2.140	−0.364	0.238
6	1.500	0.000	0.250	2.722	0.296	0.264	2.283	−0.338	0.248
7	1.500	0.500	0.250	2.596	0.536	0.241	2.013	−0.046	0.233
8	1.500	1.000	0.250	2.506	0.773	0.231	2.183	0.330	0.253
9	1.500	1.500	0.250	3.150	1.090	0.249	2.478	0.663	0.257
10	1.500	2.000	0.250	2.605	1.331	0.246	2.176	1.004	0.249
11	0.500	−2.000	0.250	1.022	−0.652	0.295	0.745	−1.518	0.284
12	0.500	−1.500	0.250	0.897	−0.523	0.245	0.685	−1.486	0.256
13	0.500	−1.000	0.250	0.860	−0.510	0.171	0.698	−1.071	0.234
14	0.500	−0.500	0.250	0.854	−0.209	0.181	0.715	−0.857	0.187
15	0.500	0.000	0.250	0.938	0.311	0.246	0.770	−0.447	0.225
16	0.500	0.000	0.250	0.904	0.205	0.233	0.715	−0.455	0.228
17	0.500	0.500	0.250	0.971	0.497	0.244	0.647	−0.276	0.173
18	0.500	1.000	0.250	0.924	0.841	0.248	0.762	0.248	0.227
19	0.500	1.500	0.250	0.946	1.087	0.245	0.655	0.562	0.229
20	0.500	2.000	0.250	0.826	1.352	0.225	0.647	0.860	0.216

a total of 20 items. All unique items had $c_i = .25$ as well, but their a_i and b_i parameters were randomly sampled from $U(.5, 2)$ and $N(0, 1)$, respectively. The first calibration had ability parameters for 10,000 test takers sampled from $N(-.5, 2)$; for the second calibration, the parameters for 10,000 test takers were sampled from $N(.5, 1.5)$. The item parameters for the two test forms were estimated separately using the *MIRT Scaling Program*, version 1.0 (Glas, 2010), with MML estimation with $\theta \sim N(0, 1)$ in both runs.

Table 1 contains the generating and estimated values of the common item parameters. Observe that even though the response data were generated for exactly the same sets of parameters for the common items, the parameters of the common items in the two calibrations were estimated to be quite different. As argued in our earlier discussion of (13)–(14), the reason is the different effect of the $\theta \sim N(0, 1)$ restriction on all parameter values in the presence of test takers with different abilities in the two calibrations.

Table 2 summarizes the covariance matrices for the MML estimators for each of the common items produced by the scaling program. The data in this table are needed to evaluate the estimates of the parameters for the linking function between the two calibrations. Observe that the variances for the item parameter estimates are as expected for datasets of the current size and generating parameter values.

We know that the c_i parameters in the two calibrations are linked by the identity transformation. Thus, for example, if we needed to know the value of the c_i parameter that an arbitrary items in the first calibration would have had in the second calibration, we could just use $\widehat{c_i}$ obtained in the first calibration as its estimate. (For the common items, it makes more sense to pool their two estimates, though.) For the other parameters, we need to know the linking parameters u and v, which can be estimated simply by plugging $\widehat{a_i}$ and $\widehat{b_i}$ for each common item into (44)–(45).

TABLE 2.
Estimated (co)variances for the estimators of the common item parameters.

Common item	Calibration 1						Calibration 2					
	$\widehat{\sigma}^2_{a_i}$	$\widehat{\sigma}^2_{b_i}$	$\widehat{\sigma}^2_{c_i}$	$\widehat{\sigma}_{a_i b_i}$	$\widehat{\sigma}_{a_i c_i}$	$\widehat{\sigma}_{b_i c_i}$	$\widehat{\sigma}^2_{a_i}$	$\widehat{\sigma}^2_{b_i}$	$\widehat{\sigma}^2_{c_i}$	$\widehat{\sigma}_{a_i b_i}$	$\widehat{\sigma}_{a_i c_i}$	$\widehat{\sigma}_{b_i c_i}$
1	0.023	0.001	0.002	0.001	0.006	0.001	0.019	0.004	0.012	0.002	0.012	0.005
2	0.020	0.001	0.001	0.001	0.003	0.001	0.017	0.002	0.003	0.000	0.006	0.001
3	0.015	0.001	0.000	0.001	0.002	0.000	0.015	0.001	0.001	0.000	0.003	0.001
4	0.015	0.001	0.000	0.002	0.001	0.000	0.009	0.001	0.001	0.001	0.002	0.001
5	0.018	0.001	0.000	0.004	0.001	0.000	0.010	0.001	0.000	0.002	0.002	0.001
6	0.017	0.001	0.000	0.004	0.001	0.000	0.011	0.001	0.000	0.001	0.001	0.000
7	0.015	0.002	0.000	0.004	0.001	0.000	0.009	0.001	0.000	0.002	0.001	0.001
8	0.016	0.003	0.000	0.006	0.001	0.000	0.014	0.002	0.000	0.004	0.001	0.000
9	0.044	0.006	0.000	0.016	0.001	0.000	0.020	0.003	0.000	0.007	0.001	0.000
10	0.033	0.009	0.000	0.016	0.001	0.000	0.022	0.006	0.000	0.011	0.001	0.001
11	0.008	0.036	0.006	0.015	0.006	0.014	0.009	0.215	0.031	0.039	0.015	0.081
12	0.007	0.053	0.007	0.018	0.006	0.019	0.009	0.292	0.036	0.047	0.017	0.103
13	0.007	0.055	0.008	0.017	0.007	0.021	0.008	0.180	0.021	0.034	0.012	0.061
14	0.006	0.048	0.005	0.016	0.005	0.016	0.007	0.125	0.015	0.027	0.009	0.043
15	0.008	0.032	0.002	0.014	0.003	0.008	0.007	0.073	0.007	0.020	0.006	0.022
16	0.007	0.036	0.002	0.014	0.004	0.009	0.007	0.109	0.009	0.026	0.008	0.031
17	0.007	0.027	0.001	0.013	0.003	0.006	0.007	0.152	0.011	0.031	0.008	0.040
18	0.008	0.036	0.001	0.016	0.003	0.006	0.008	0.066	0.004	0.021	0.005	0.015
19	0.009	0.035	0.001	0.017	0.002	0.005	0.009	0.136	0.005	0.033	0.006	0.025
20	0.010	0.062	0.001	0.023	0.003	0.008	0.010	0.139	0.004	0.034	0.006	0.023

The results are shown in Table 3. Although each of these 20 estimates reveals the same trend, they show considerable random variation. In order to evaluate the variation, Table 3 also gives the estimated standard errors for each $\widehat{u}_i$ and $\widehat{v}_i$, which were derived from the (co)variances for a_i and b_i in Table 2 using the (first-order) multivariate delta method (e.g., Casella & Berger, 2002, sect. 5.5.4). As the size of these standard errors indicate, we should have expected a considerable amount of variation indeed.

Obviously, as each $\widehat{u}_i$ and $\widehat{v}_i$ is an estimate of the same u and v, respectively, rather than using them individually, they should be combined into overall estimates. A natural suggestion is to use their precision-weighted average, with the inverse of their squared standard errors, $\sigma_{u_i}^{-2}$ and $\sigma_{v_i}^{-2}$, as measure of precision. The estimator of u is then

$$\widehat{u} = \left(\sum_{i=1}^{20} \widehat{\sigma}_{u_i}^{-2} \widehat{u}_i \right) \Big/ \left(\sum_{i=1}^{20} \widehat{\sigma}_{u_i}^{-2} \right), \tag{61}$$

with estimated standard error

$$\widehat{\sigma}_u = \left(\sum_{i=1}^{20} \widehat{\sigma}_{u_i}^{-2} \right)^{-1/2}, \tag{62}$$

with similar expressions for the estimator of v.

Table 4 shows these overall estimates along with those for the mean/mean and mean/sigma methods. The former were obtained by plugging the estimates of the a_i and b_i parameters into (6) and (9); the latter by plugging the estimates of the b_i parameters into (7) and (9). The standard errors

TABLE 3.
Linking parameter and their standard errors estimated for each common item.

Common item	$\widehat{u}_i$	$\widehat{\sigma}_{u_i}$	$\widehat{v}_i$	$\widehat{\sigma}_{v_i}$
1	1.208	0.105	−0.707	0.104
2	1.160	0.088	−0.686	0.069
3	1.108	0.077	−0.630	0.045
4	1.295	0.082	−0.718	0.047
5	1.322	0.088	−0.738	0.074
6	1.192	0.079	−0.691	0.069
7	1.290	0.086	−0.737	0.106
8	1.148	0.084	−0.557	0.134
9	1.271	0.111	−0.723	0.228
10	1.197	0.117	−0.589	0.287
11	1.372	0.210	−0.624	0.413
12	1.309	0.218	−0.801	0.516
13	1.232	0.194	−0.443	0.425
14	1.194	0.178	−0.607	0.406
15	1.218	0.171	−0.826	0.396
16	1.264	0.191	−0.714	0.444
17	1.501	0.238	−1.022	0.573
18	1.213	0.184	−0.772	0.490
19	1.444	0.256	−1.008	0.724
20	1.277	0.247	−0.866	0.811

TABLE 4.
Overall estimates of linking parameters and their standard errors.

Method	$\widehat{u}$	$\widehat{\sigma}_u$	$\widehat{v}$	$\widehat{\sigma}_v$
Precision-weighted average	1.226	0.026	−0.684	0.023
Mean/mean	1.237	0.027	−0.706	0.078
Mean/sigma	1.197	0.118	−0.696	0.084

for these two methods were calculated from the (co)variances for the item parameter estimates in Table 2 using the same the multivariate delta method. The differences between the results for all three methods were generally substantial, with the precision-weighted method being uniformly best. Especially the results for the v parameter are revealing. Whereas the precision-weighted method produced an acceptable low standard error for it, the other two methods lagged behind considerably. In more practical terms, these results suggest that, even with 20 common items, these two methods are likely to seriously misspecify the location of the parameters mapped from one calibration onto the values they would have obtained in another. The extremely large errors for the mean/sigma method are assumed to be due to its ignoring of the unique information in the estimates of the a_i parameters.

It is also interesting to inspect how these overall estimates of the standard errors behave as a function of the number of common items. The curves in Fig. 2 were obtained by adding the common items to the linking design, one at a time beginning with the first item in Table 3. The precision-weighted method produced results that were generally substantially better and never worse than those for the mean/mean and mean-sigma method. In fact, it already reached stability for both estimates after some five common items, whereas there still was considerable room for the other two methods to converge. Also, observe the lack of monotonicity in the curves for the

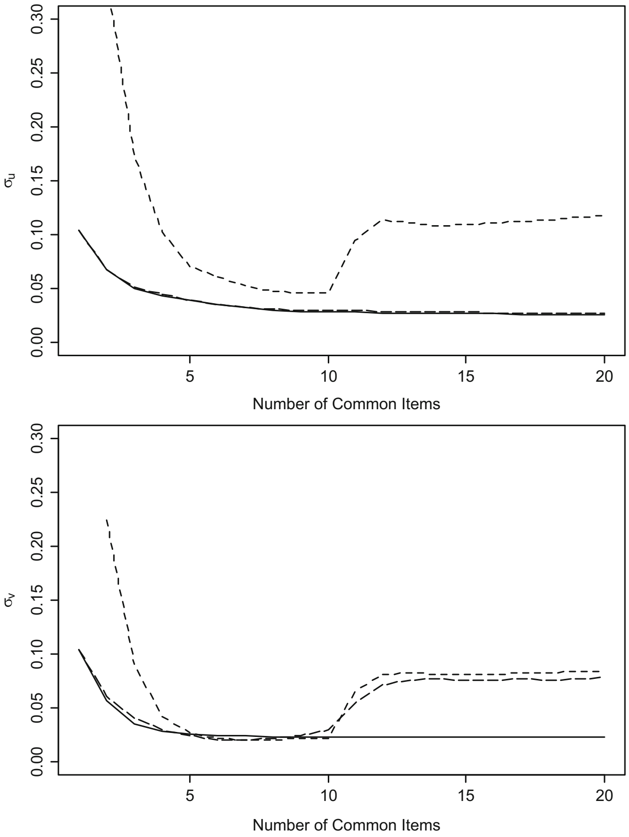

FIGURE 2.
Estimated standard errors for linking parameters u and v for the precision-weighted (*solid*), mean/mean (*longdash*), and mean/sigma methods (*shortdash*) as a function of the number of common items in the linking design.

standard errors of u and v for the mean/sigma method and in the one for the standard error of v for the mean/mean method. Whereas one would expect a decrease of them with the extra information in each common item added to the linking design, these methods actually showed a considerable increase for the eleventh and twelfth item. Finally, use of the precision-weighted method with the first ten items in Fig. 2 would yield linking estimates already superior to those for all 20 items for the mean-mean and mean-sigma methods.

The results in Fig. 2 highlight the importance of the relative precision of the linking parameter estimates contributed by each individual item to the linking design. Surprisingly, if we had added the items to the linking design in a different order, different results would have been found. Figure 3 illustrates the linking errors for the same total set of common items as in Fig. 2, but now added by increasing item difficulty rather than increasing item discrimination as in Table 2. Note that, of course, the overall error associated with all 20 common items remains the same. But the final result is now reached along different trajectories for all three methods. The difference between the results in Figs. 2 and 3 suggests further research on the use of optimal design principles to find the best possible subset of linking items for the linking design from the larger set of candidate items typically available in practical situations.

This example was only to explore what might be possible once the problem of parameter linking in IRT has been provided with a solid formal foundation. Research on the new estimation

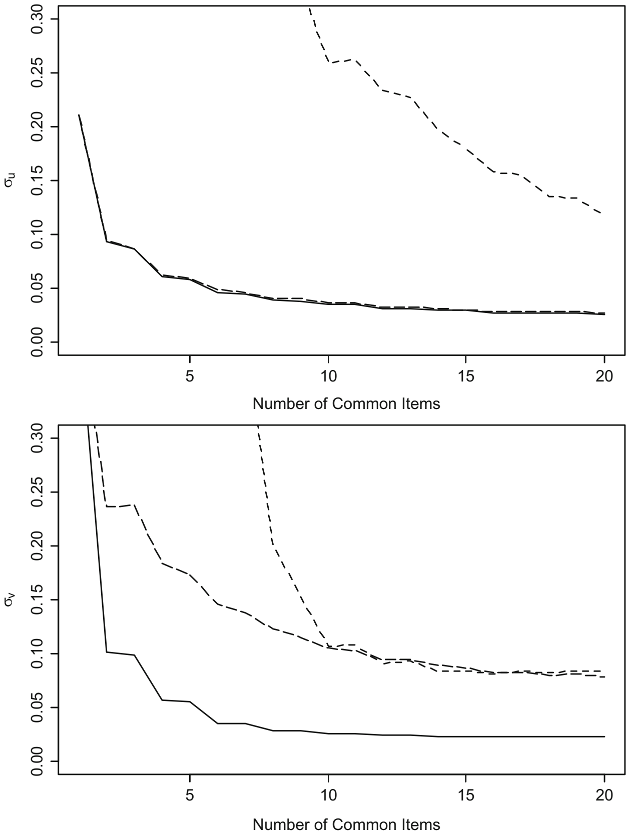

FIGURE 3.
Estimated standard errors for linking parameters u and v for the precision-weighted (*solid*), mean/mean (*longdash*), and mean/sigma methods (*shortdash*) as a function of the number of common items in the linking design for a different order of the items than in Table 2.

method, including comparative studies with other linking methods using empirical data, attempts to explain the aberrances produced by the traditional linking methods, and an analysis of the consequences of the confounding of linking error with estimation error in the c_i parameters by the response-function methods is currently conducted. Also, as just noted, further research is needed on how to find the optimal linking design given a set of candidate common items.

7. Concluding Remarks

We began this article with a review of the traditional conception of parameter linking in IRT, which appears to have been motivated largely by the notion of equating θ scores on scales with an indeterminate zero and unit, that is, interval scales in the tradition of Stevens' (1946) classification. An obvious way to remove the indeterminacy of a θ scale, according to the tradition, would be to set its zero and unit equal to the mean and standard deviation of the abilities for a population of test takers, and the only reason why we need to link parameters from different calibrations would be to correct for differences between population distributions. Besides, because they are not assumed to be affected by this choice of zero and unit, the c_i parameters could be treated as invariant. Lord (1980, sect. 3.5) is quite direct in his claim as to this last point.

The fundamental notion underlying the necessity of parameter linking in IRT, however, is not that of an "interval scale" for the θ parameters in the response model, but possible lack of identifiability of any of its parameters. We have been able to present several results due to this reconceptualization. First, although never formally derived before, the linking functions for the standard reparameterization of the 3PL model (Theorem 4) appear to have the general linear form for the a_i, b_i, and θ_p parameters assumed in the current literature, but with the different slope and intercept parameters u and v in (37) and (38), respectively, and the component function $\varphi_c(c) = c$ added for the c_i parameters. The definitions of u and v allowed us to derive the different solutions for a few minimal linking designs in (44)–(48). Second, the alternative slope-intercept reparameterization of the 3PL model appears to have a serious impact on the linking problem. Not only have the linking functions for its α_i, β_i, and ϑ_p parameters a different form with one more unknown linking parameter w (Theorem 6), its parameters are identifiable only for the unpractical type of design that has both common items and common persons (Theorem 7). Third, as a more general result, we now know from Theorem 3 that for any other monotone, continuous response model, the linking functions take the general form of a componentwise monotone vector function—a fact that will simplify our explorations of the linking functions required for nearly every other IRT model currently used in educational and psychological testing. Fourth, although all linking functions were derived for true model parameters and their subsequent estimation was not the focus of this paper, it is already clear that we will have to deviate from the estimation methods currently practised. For instance, as illustrated by our example, rather than estimating linking parameter u as the ratio of the mean of estimates of the a_i parameters in (6) for the common items, it is much more efficient to use an estimate based on the (precision-weighted) mean of their ratios. Fifth, the derivation of the linking function for the c_i parameters in (35) helps us to evaluate their role in the currently used estimation methods discussed in the introductory section. The Stocking-Lord and Haebara methods admit estimation error in the c_i parameters into their estimates of the linking parameters, but the mean/mean and mean/sigma methods ignore these parameters entirely. At first sight, the lack of identifiability of the c_i parameters seems to suggest the choice of a method from the former rather than the latter category. But the fact that their linking function is the identity function $\varphi_c(c) = c$ implies that, once they have been made identifiable, no further linking is necessary. Consequently, unlike the mean/mean and mean/sigma methods, the Haebara and Stocking-Lord methods confound linking error in the a_i, b_i, and θ_p parameters with estimation error in the c_i parameters.

Any choice of identifiability restrictions has an element of arbitrariness to it, and the practice of making the 3PL model identifiable using restrictions that include the mean ($\mu_\theta = 0$) or standard deviation ($\sigma_\theta = 1$) of the ability parameters for the test takers in the calibration study therefore cannot be wrong. Nevertheless, the reliance on the notion of randomly sampling from some population of test takers sometimes automatically associated with it is potentially dangerous. For instance, it easily leads to the idea that we now estimate a population mean and standard deviation and therefore have to account for their sampling error. Indeed, a recent study advocated this idea, along with the claim that large-scale educational assessments tend to overlook the design effects on linking error due to the typical clustering of test takers during sampling (Doorey, 2011, p. 6). However, as demonstrated by Theorem 4, the shape of the true linking functions for the 3PL model does not depend on the actual identifiability restrictions imposed on the calibration studies, let alone on any population parameters adopted for them, or even a specific choice of sampling design used to estimate such parameters.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*, 261–269.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*, 147–172.

Bartels, R. (1985). Identification in econometrics. *The American Statistician*, *39*, 102–104.

Bechger, T. M., Verhelst, N. D., & Verstralen, H. H. F. M. (2001). Identifiability of nonlinear logistic models. *Psychometrika*, *66*, 357–372.

Bechger, T. M., Verstralen, H. H. F. M., & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika*, *67*, 123–136.

Bekker, P. A., Merckens, A., & Wansbeek, T. J. (1994). *Identification, equivalent models, and computer algebra*. Boston: Academic Press.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Doorey, N. A. (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments*. Washington, DC: Council of Chief State School Officers.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Fischer, G. H. (2004). Remarks on "Equivalent linear logistic test models" by Bechger, Verstralen, and Verhelst (2002). *Psychometrika*, *69*, 305–315.

Fisher, F. M. (1961). Identifiability criteria in nonlinear systems. *Econometrica*, *29*, 574–590.

Fisher, F. M. (1965). Identifiability criteria in nonlinear systems: A further note. *Econometrica*, *33*, 197–205.

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics*, *8*, 261–263.

Glas, C. A. W. (2010). *MIRT: Multidimensional item response theory, version 1.01* [*Computer software and manual*]. Enschede, The Netherlands: University of Twente. Retrieved from http://www.utwente.nl/gw/omd/en/employees/employees/glas.doc/.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Haberman, S. (2009). *Linking parameter estimates derived from item response model through separate calibration (Research Report 09–40)*. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Rubin, D. B. (Eds.). (1982). *Test equating*. New York: Academic Press.

Kim, S., Harris, D. J., & Kolen, M. J. (2010). Equating with polytomous item response models. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous items response theory models* (pp. 257–291). New York: Taylor & Francis.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica*, *17*, 125–144.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193.

Luijben, T. C. W. (1991). Equivalent models in covariance structure analysis. *Psychometrika*, *56*, 653–665.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160.

Maris, G. (2002). *Concerning the identification of the 3PL model (Measurement and Research Department Reports 2002–3)*. Arnhem: Cito.

Maris, G., & Bechger, T. (2004). Equivalent MIRD models. *Psychometrika*, *69*, 627–639.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three-parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives*, *7*, 75–88.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reiersøl, O. (1950). On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika*, *15*, 121–149.

Revuelta, J. (2009). Identifiability and equivalence of GLLIRM models. *Psychometrika*, *74*, 257–272.

Richmond, J. (1974). Identifiability in linear models. *Econometrica*, *42*, 731–736.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, *39*, 577–591.

Sahoo, P. K., & Kannappan, P. (2011). *Introduction to functional equations*. Boca Raton, FL: Chapman & Hall/CRC.

San Martín, E., Gonzáles, J., & Tuerlinkckx, F. (2009). Identified parameters, parameters of interest and their relationships. *Measurement: Interdisciplinary Research and Perspective*, *7*, 97–105.

San Martín, E. Gonz áles, J., & Tuerlinkckx, F. (2015). On the identifiability of the fixed-effects 3PL model. *Psychometrika*, *80*, 450–467.

San Martín, E., Jara, A., Rolin, J.-M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, *76*, 385–409.

San Martín, E., Rolin, J.-M., & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika*, *78*, 341–379.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Small, C. G. (2007). *Functional equations and how to solve them*. New York: Springer.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

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

Tsai, R.-C. (2000). Remarks on the identifiability of Thurstonian ranking models: Case V, Case III, or neither? *Psychometrika*, *65*, 233–240.

van der Linden, W. J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement*, *47*, 92–114.

Volodin, N., & Adams, R. J. (2002). *The estimation of polytomous item response models with many dimensions (Internal Report)*. Parkville, VIC: Faculty of Education, University of Melbourne.

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating scaling, and linking*. New York: Springer.

von Davier, M., & von Davier, A. A. (2011). A general model for IRT scale linking and scale transformations. In A. A. von Davier (Ed.), *Statistical models for test equating scaling, and linking* (pp. 225–242). New York: Springer.

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

Manuscript Received: 18 SEP 2014
Published Online Date: 9 JUL 2015