Chapter 14
A General Model for IRT Scale Linking and Scale Transformations

Matthias von Davier and Alina A. von Davier

14.1 Introduction

The need for equating arises when two or more tests on the same construct or subject area can yield different scores for the same examinee. The goal of test equating is to allow the scores on different forms of the same tests to be used and interpreted interchangeably. Item response theory (IRT; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Thissen & Wainer, 2001) has provided new ways to approach test equating. Using IRT in the equating process usually also requires some sort of linking procedure to place the IRT parameter estimates on a common scale.

In this chapter we focus on the IRT linking procedures used for data collection designs that involve common items. The data collection designs that use this method are nonequivalent groups with anchor test (NEAT) designs and can have both internal and external anchor tests (see, e.g., Kolen & Brennan, 1995; von Davier, Holland, & Thayer, 2004b).

The NEAT design has two populations of test takers, populations *P* and *Q* (*P* and *Q*) of test takers and a sample of examinees from each. The sample from *P* takes test form X, or *X*.The sample from *Q* takes test form Y, or *Y* and both samples take a set of common items, the anchor test *V*. This design is often used when only one test form can be administered at one test administration because of test security or other practical concerns. The two populations may not be equivalent in that the two samples are not from a common population.

The two test forms X and Y and the anchor *V* are, in general, not *parallel* test forms, that is, their conditional expectations and error variances for a given examinee will not be identical. More specifically, the anchor test *V* is usually shorter and less reliable than either *X* or *Y*. Angoff (1971/1984) gave advice on

M. von Davier (✉) and A.A. von Davier
Educational Testing Service, Rosedale Rd, Princeton, New Jersey 08541, USA
e-mail: mvondavier@ets.org

A.A. von Davier (ed.), *Statistical Models for Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-0-387-98138-3_14, © Springer Science+Business Media, LLC 2011

designing anchor tests. For a comparison of a variety of methods for treating the NEAT design, see Kolen and Brennan (1995); Marco, Petersen, and Stewart (1983b); and Petersen, Marco, and Stewart (1982).

In this chapter we examine the IRT scale transformation and IRT linking methods used in the NEAT design to link *X* and *Y* More exactly, we propose a unified approach to the IRT linking methods: mean-sigma and mean-mean, concurrent calibration, fixed-parameters calibration, the Stocking and Lord characteristic-curves approach, and the Haebara characteristic-curves approach (see Kolen & Brennan, 1995, Ch. 6, for a detailed description of these methods). Moreover, we believe that our view of IRT linking can be extended to cover other flavors of IRT scaling and linking procedures.

In our approach, the parameter space is described by all the parameters of the IRT model fitted to the data from both populations in a marginal maximum likelihood framework. Under the usual assumptions for the NEAT design, which are described later, the joint log-likelihood function for this model on the data from both populations can be expressed as the sum of two log-likelihood functions corresponding to each of the two groups of data and parameters.

The main idea in our approach is to view any linking method as a restriction function on the joint parameter space of the instruments to be equated. Once this is understood, rewriting the joint log-likelihood function by including a term for each restriction and an appropriately implemented maximization procedure will accomplish the linking. The maximization is carried out using a vector of Lagrange multipliers (see, e.g., Aitchison & Silvey, 1958; Glas, 1999; von Davier, 2003a).

We will show that the new approach is general enough to cover the usual item response models—the one-parameter logistic (1PL), 2PL, and 3PL models—as well as polytomous, unidimensional IRT models like the generalized partial-credit model.

This new perspective on IRT linking has advantages. First, providing a common framework for all IRT scale linking methods yields a better understanding of the differences between the approaches, which opens paths to more flexible methods of IRT linking. Also, viewing the IRT linking as a restriction function allows us to control the strength of the restriction. For example, the concurrent calibration with fixed item parameters is the most restrictive IRT linking method, as it assumes the equality of all parameters in the anchor test. When such a strong restriction is not appropriate, the proposed method provides alternatives. Moreover, the method provides a family of linking functions that ranges from the most restrictive one, the concurrent calibration with fixed item parameters, to separate calibration (without additional restrictions, i.e., to no linking at all). Finally, the new perspective allows the development of methods to check the IRT linking (such as Lagrange multiplier tests) for appropriateness of different methods. For this, similar principles as developed in Glas (1999, 2006) could be applied to check the invariance of certain parameters or types of parameters.

This chapter, a summarized version of von Davier and von Davier (2007), describes the theoretical framework and derivations of a general approach to IRT linking. It generalizes a linking method implemented and utilized by von Davier

& Yamamoto (2004) to link IRT scales across three student populations. The rest of the chapter is structured as follows. First we introduce our notation and briefly describe the well-known IRT linking methods. Then, we investigate the joint log-likelihood function and the restriction function more formally and for several IRT linking methods. Finally, we discuss the advantages of this perspective on the IRT linking.

14.2 The NEAT Design and IRT Linking

14.2.1 The NEAT Design

The data structure and assumptions needed for the NEAT design are described in von Davier et al. (2004b). Briefly, population *P* yields Sample 1, taking test form X; population *Q* yields Sample 2, taking test form Y. Both samples take anchor test *V*. We denote the matrices of observed item responses to the tests X, V, and Y by *X*, *V* and *Y*. The subscripts *P* and *Q* denote the populations.

The analysis of the NEAT design usually makes the following assumptions:

1. There are two populations of examinees *P* and *Q*, each of which can take one of the tests and the anchor.
2. The two samples are independently and randomly drawn from *P* and *Q*, respectively. In the NEAT design *X* is not observed in population *Q*, and *Y* is not observed in population *P*. To overcome this feature, all linking methods developed for the NEAT design (both observed-score and IRT methods) must make additional assumptions of a type that does not arise in the other linking designs.
3. The tests to be equated, *X* and *Y*, and the anchor *V*, are all unidimensional (i.e., all items measure the same unidimensional construct), carefully constructed tests, in which the local independence assumption holds (Hambleton et al., 1991).

These three assumptions are sufficient for our exposition. We will not impose any constraints on the distributions of *X*, *Y* or *V*, that is, the score distribution will be assumed to be multinomial. Alternatives, such as log-linear models for observed score distributions (Rost & von Davier, 1992, 1995; von Davier, 1994, 2000) and latent skill distributions (Xu & von Davier, 2008) have been discussed and are also used in observed-score equating (von Davier et al., 2004b).

14.2.2 Unidimensional IRT Models

IRT models rely on the assumptions of monotonicity, undimensionality, and local independence (Hambleton et al., 1991). These models express the probability of a response x_{ni} of a given person, n ($n = 1, ..., N$), to a given item, i ($i = 1, ..., I$), as a function of the person's ability θ_n and a possibly vector valued item parameter, β_i,

$$P_{ni} = P(X = x_{ni}) = f(x_{ni}, \theta_n, \beta_i) \qquad (14.1)$$

In the case of the well-known 3PL model (Lord & Novick, 1968), the item parameter is three-dimensional and consists of the slope, the difficulty, and the guessing parameter, $\beta_i^t = (a_i, b_i, c_i)$.The 3PL model, which serves as the standard example of an item response model in this paper, is given by

$$P(x_i = 1|\theta, a_i, b_i, c_i) = c_i + (1 - c_i)\text{logit}^{-1}[a_i(\theta - b_i)], \qquad (14.2)$$

with $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$. However, most results presented here do not depend on the specific choice of the item response model and apply to models for both dichotomous and polytomous data.

14.2.3 IRT Linking

When conducting IRT scale linking in the NEAT design, the parameters of the item response model from different test forms need to be, or need to be brought, on the same IRT scale. IRT models without scale constraints are undetermined: linear transformations of the item, and person parameters do not change the likelihood of observed responses. Therefore, the potentially different scales from separate calibrations can be brought on the same scale in one of the following ways: Assuming that the calibration was carried out separately on the two samples from the two different populations P and Q, two sets of parameter estimates for the anchor test V will be available for examinees in the two groups. These separate parameter estimates of the anchor in the two groups serve as the basis for the scale transformation (mean-mean, mean-sigma methods, or characteristic-curves methods, such as Stocking and Lord or Haebara methods).

As an alternative, the item parameters from X, V (in both populations), and Y can be estimated jointly, coding the items that an examinee did not receive as "not administered," since these outcomes were unobserved and are missing by design. The item parameters are estimated simultaneously, and separate ability distributions are assumed in the two populations, while the parameters of the anchor are assumed to be identical in both populations; this IRT scaling is usually referred to as *concurrent calibration*. Another calibration method is the *fixed-parameters method*. This approach differs from concurrent calibration in that common items whose parameters are known (either from a previous-year calibration or a separate calibration) are anchored or fixed to their known values, often estimates from a previous calibration, during calibration of other forms, or the same common items within a form administered in a different year or in a different population.

By treating the common item parameters as known, they are not estimated, and the item parameters from the unique items are forced onto the same scale as the fixed, common items. This procedure is implemented in IRT calibration using models referred to as multiple-group models with constraints of common items across groups (populations, forms). For more details, see Kolen and Brennan (1995), Stocking and Lord (1983), Haebara (1980), or Hambleton et al. (1991).

14.3 A Lagrangean Approach to IRT Linking

Let the sample size of the group from P that takes (X, V) be denoted by N_P, let the sample size of the group from Q that takes (Y, V) be N_Q and denote $N = N_P + N_Q$. We will use the following notation for the item parameters in the different test forms and populations:

$$\beta_i = \begin{cases} \beta_{X_{Pj}}, & 1 \leq i \leq J, (1) \\ \beta_{V_{Pl}}, & J + 1 \leq i \leq (J + L), (2) \\ \beta_{V_{Ql}}, & (J + L) + 1 \leq i \leq (J + 2L), (3) \\ \beta_{Y_{Qk}}, & (J + 2L) + 1 \leq i \leq \text{IP}, \end{cases} \tag{14.3}$$

where $\text{IP} = J + 2L + K$ denotes the total number of items and J, L, and K are the number of the items in the tests X, V, and Y respectively. For example, $\beta_{X_{Pj}}$ denotes the (possible vector-valued) item parameter for item j from the set of items from the test X that was taken by the examinees from P. Similarly, $\beta_{V_{Pl}}$ denotes the (possibly vector-valued) item parameter for item l from the set of items from the anchor test V that was taken by the examinees from P, and so forth.

The total number of the item parameters (TNIP) is the dimension of the vector of the item parameters times the number of parameters per item. For example, if all items are modeled via the Rasch model, $\text{TNIP} = 1 \times \text{IP}$; for a 2PL model, $\text{TNIP} = 2 \times \text{IP}$; and for the 3PL model, $\text{TNIP} = 3 \times \text{IP}$; For mixtures of item model types in one test, TNIP is the sum of individual item parameter dimensions (1, 2, or 3 for dichotomous items and 2 or more for partial-credit items) over all items.

14.3.1 Separate Calibration

When estimating separately, the item and ability distribution parameters for population P are obtained given data (X, V_P), separately from the item and ability distribution parameters for population Q given data (Y, V_Q). Technically, this can be accomplished by fitting one IRT model to the combined data without assuming that the common items have the same item parameters in both populations.

As mentioned before, the parameter space is described by all the parameters in the IRT model fitted to the data from both populations, in a marginal maximum likelihood framework. Software such as MULTILOG (Thissen, 1991) and PARSCALE (Muraki & Bock, 1991) as well as the more recent *mdltm* (von Davier, 2005) may be used to perform such calibrations.[1]

Let π_P and π_Q denote the parameters used to model the ability distribution. We may think of them as $\pi_{P,Q} = (\mu_{P,Q}, \sigma_{P,Q})$ in the case where we assume normal ability distributions. In somewhat more flexible models, we may assume that the $\pi_{P,Q}$ is a set of multinomial probabilities over quadrature points approximating arbitrary distribution shapes. Hence, the complete parameter space is contained in the (transposed) parameter vector

$$\eta^t = (\beta_{X_P}, \beta_{V_P}, \pi_P, \beta_{Y_Q}, \beta_{V_Q}, \pi_Q), \tag{14.4}$$

Given Assumptions 1 and 2 and the properties of the logit and logarithm functions, we can rewrite the joint log-likelihood function for the IRT model applied to the data from both populations as the sum of the two log-likelihood functions,

$$L\big(\eta; X, V_{P,} Y, V_Q\big) = L\big(\beta_{X_P}, \beta_{V_P}, \pi_p; X, V_P\big) + L(\beta_{Y_Q}, \beta_{V_Q}, \pi_Q; Y, V_Q) \tag{14.5}$$

In other words, the two separate models are estimated and the two log-likelihood functions are maximized jointly using marginal maximum likelihood. Now, it is easy to conceive *any* linking function as a restriction function on the parameter space and any linking process as a maximization of Equation 14.5 under the linking restriction. Later we will illustrate in detail how this approach works for each linking method. Next, we illustrate the concurrent calibration method in some detail and then outline how this approach translates to each of the other IRT linking methods: mean-mean, mean-sigma, Stocking and Lord, Haebara, and fixed parameters.

14.3.2 Lagrangean Concurrent Calibration

When estimating concurrently, the item and ability distribution parameters for population P are obtained given data (*X*,V_P) simultaneously with the item and ability distribution parameters for population Q given data *X*, V_Q.Technically, two separate ability distributions are estimated and the two log-likelihood

[1] The *mdltm* software is a command line controlled program that runs on various operating systems. Executables can be made available for noncommercial purposes upon request; please contact the first author of this chapter for details.

functions are maximized jointly with certain restrictions on the item parameters, namely

$$\beta_{V_{Pl}} = \beta_{V_{Ql}} \tag{14.6}$$

for $l = 1, \ldots, L$.

Let R denote the L-dimensional restriction function with the components given by

$$R_l(\eta) = k_l(\beta_{V_{Pl}} - \beta_{V_{Ql}}), \tag{14.7}$$

with $k_l = 1$ for active restrictions on item l.

For the 2PL and 3PL, the restrictions may be imposed only on the b parameters and not on the slope and guessing parameters. This can be achieved by first using projections, h, and then imposing the same constraints as before. The projection, say a mapping h would isolate the item difficulty b out of the three-dimensional item parameter in the case of the 3PL, and then apply the restriction function on this parameter only. That is, $b_{Vl} = h(\beta_{Vl})$ to obtain the difficulty, and then use $R_l(\eta) = k_l(h(\beta_{V_{Pl}}) - h(\beta_{V_{Ql}}))$ to impose constraints on the difficulties only.

Hence, the concurrent calibration refers to maximizing Equation 14.5 under the restriction

$$R_l(\eta) = 0. \tag{14.8}$$

This setup, maximizing Equation 14.5 under the restriction Equation 14.8, is used whenever certain item parameters are assumed to be equal across populations, in our case across P and Q.

Given a vector λ of Lagrangean multipliers, the linking process can be viewed as maximizing the modified log-likelihood function

$$\Lambda\left(\eta,\lambda;X,V_p,Y,V_Q\right)=L\left(\beta_{Xp},\beta_{Vp},\pi_P;X,V_P\right)+L(\beta_{YQ},\beta_{VQ},\pi_Q;Y,V_Q)-\lambda^t R(\eta). \tag{14.9}$$

Note that if we choose $k_l = 0$ for all l the restriction functions, all $R_l(\eta) = k_l(\beta_{V_{Pl}} - \beta_{V_{Ql}})$ vanish. In that case, the β_{V_P} and β_{V_Q} are no longer constrained to be equal; instead of concurrent calibration with equality constraints, maximizing the likelihood simultaneously now yields separate calibrations and allows the item parameters in the anchor test V to differ between P and Q.

The function in Equation 14.9 is then maximized with respect to parameters η and λ.

In concurrent calibration, Equation 14.9 includes a term R_l for each item $l = 1, \ldots, L$ in the anchor test V. This term enables the imposition of equality constraints on the parameters β_{V_P} and β_{V_Q}.

14.3.3 Lagrangean Fixed-Parameters Scale Linkage

In this method, common items whose parameters are known (for example, from a previous administration calibration or a separate calibration) are anchored or fixed to their known estimates, $w_l,\ l = 1, \ldots L,$ during calibration of other forms, or forms with common items in other assessment years or populations. These common item parameters are treated as known and therefore are not estimated; the item parameters from the items that are not common to the forms are forced onto the same scale as the fixed items. This calibration procedure is more restrictive than concurrent calibration.

As before, let now R denote the $2L$-dimensional restriction function with the components given by

$$R(\eta)^t = (R_{Pl}(\eta), R_{Ql}(\eta)). \tag{14.10}$$

where

$$\begin{aligned} R_{Pl}(\eta) &= k_l(\beta_{V_{Pl}} - w_l), \\ R_{Ql}(\eta) &= k_l(\beta_{V_{Ql}} - w_l), \end{aligned} \tag{14.11}$$

and hence the concurrent calibration refers to maximizing Equation 14.9 under the restriction

$$R(\eta) = 0. \tag{14.12}$$

14.3.4 Lagrangean Mean-Mean IRT Scale Linking

If an IRT model fits the data, any linear transformation[2] (with slope A and intercept B) of the θ-scale also fits these data, provided that the item parameters are also transformed (see, e.g., Kolen & Brennan, 1995, pp. 162–167). In the NEAT design, the most straightforward way to transform scales when the parameters were estimated separately is to use the means and standard deviations of the item parameter estimates of the common items for computing the slope and the intercept of the linear transformation. Loyd and Hoover (1980) described the mean-mean method, where the mean of the a-parameter estimates for the common items is used to estimate the slope of the linear transformation. The mean of the b-parameter

[2] A more general result holds: All strictly monotone transformations of θ are also permissible. This feature, however, will not be pursued further in this chapter.

estimates of the common items is then used to estimate the intercept of the linear transformation (see Kolen & Brennan, 1995, p. 168).

Lagrange multipliers also may be used to achieve IRT scale linking according to the mean-mean approach. Again, maximizing the modified log-likelihood function Λ given in Equation 14.9 with a different set of restrictions does the trick. For the mean-mean IRT linking, the restriction function is two-dimensional with the components R_a and R_b, $R^t = (R_\mathrm{a}, R_\mathrm{b})$ To match the mean of anchor parameters of population P, define

$$R_a(\eta) = \left(\sum_{l=1}^{L} h_a(\beta_{V_{Ql}}) - A_P \right) \tag{14.13}$$

with a constant term $A_P = \sum h_a(\beta_{V_{Pl}})$, which is not viewed as a function of the β_{VP} (but recomputed at each iteration during maximization) in order to allow unconstrained maximization for P and enforce the mean of β_{VQ} to match this mean in P. As has been explained, h is a projection.

The same is done with the difficulty parameters $b_l = h_b(\beta_l)$:

$$R_b(\eta) = \left(\sum_{l=1}^{L} h_b(\beta_{V_{Ql}}) - B_P \right). \tag{14.14}$$

This new approach to IRT linking includes also a more general approach that handles populations P and Q symmetrically using

$$R_a(\eta) = \left(\sum_{l=1}^{L} h_a(\beta_{V_{Ql}}) - h_a(\beta_{V_{Pl}}) \right) \tag{14.15}$$

with $h_a(\beta_i) = a_i$ and

$$R_b(\eta) = \left(\sum_{l=1}^{L} h_b(\beta_{V_{Ql}}) - h_b(\beta_{V_{Pl}}) \right) \tag{14.16}$$

with $h_b(\beta_i) = b_i$. This avoids the arbitrary choice whether to match P or Q's slope and difficulty means on the anchor test V.

14.3.5 Lagrangean Mean-Var IRT Scale Linking

The mean-var IRT scale linkage (Marco, 1977) obviously can be implemented in the same way, with only a slight difference in the restrictions used. The means and the standard deviations of the b-parameters are used to estimate the slope and the intercept of the linear transformation.

Again, a two-dimensional restriction function with components R_a and R_b is needed. In order to match the mean and variance of the anchor test's difficulty parameter in population P, we define

$$R_a(\eta) = \left(\sum_{l=1}^{L} h_a(\beta_{V_{Ql}}) - B_P \right) \tag{14.17}$$

with a constant term $B_P = \sum h_a(\beta_{V_{Pl}})$, which again is not viewed as a function of the β_{V_P}. The same is done with the squared difficulties, $b_l^2 = h_b^2(\beta_l)$,

$$R_b(\eta) = \left(\sum_{l=1}^{L} h_b^2(\beta_{V_{Ql}}) - B_P^2 \right) \tag{14.18}$$

where $B_P^2 = \sum h_b^2(\beta_{V_{Pl}})$.

As before, a more general approach handles populations P and Q symmetrically using

$$R_a(\eta) = \left(\sum_{l=1}^{L} h_b(\beta_{V_{Ql}}) - \sum_{l=1}^{L} h_b(\beta_{V_{Pl}}) \right) \tag{14.19}$$

with $h_b(\beta_i) = b_i$ and

$$R_b(\eta) = \left(\sum_{l=1}^{L} h_b^2(\beta_{V_{Ql}}) - \sum_{l=1}^{L} h_b^2(\beta_{V_{Pl}}) \right) \tag{14.20}$$

with $h_b^2(\beta_i) = b_i^2$.

14.3.6 Lagrangean Stocking and Lord Scale Linkage

Characteristic-curves transformation methods were proposed (Haebara, 1980; Stocking & Lord, 1983) in order to avoid some issues related to the mean-mean and mean-var approaches. For the mean-mean and mean-var approaches, various combinations of the item parameter estimates produce almost identical item-characteristic curves over the range of ability at which most examinees score.

The Stocking and Lord (1983) IRT scale linkage finds parameters for the linear transformation of item parameters in one population (say, Q) that matches the test characteristic function of the anchor in the reference population (say, P). The Stocking and Lord transformation finds a linear transformation (a slope A and an intercept B) of the item parameters—difficulties and slopes—in one population based on a matching of test characteristic curves. Expressing this in the marginal maximum likelihood framework yields

$$(A,B) = \min\left[\sum_{g=1}^{G} \pi_g^*\left(\sum_{l=1}^{L}\left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, B + A\beta_{VQl}\right)\right)\right)^2\right], \tag{14.21}$$

where the weights π_g^* of the quadrature points θ_g for $g = 1, \ldots,$ G. are given by

$$\pi_g^* = \frac{n_{Pg}\pi_{Pg} + n_{Qg}\pi_{Qg}}{n_{Pg} + n_{Qg}}. \tag{14.22}$$

We propose using a method employing the same rationale as the Stocking and Lord (1983) approach, namely optimizing the match of the test characteristic curves between the anchors V_P and V_Q. In the proposed framework, the primitive of these functions, which is the criterion to be minimized to match the two test characteristic functions as closely as possible, is defined as

$$R^{SL}(\eta) = \left[\sum_{g=1}^{G} \pi_g^*\left(\sum_{l=1}^{L}\left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right)\right)\right)^2\right]. \tag{14.23}$$

In order to minimize Equation 14.23, we implement the Lagrangean in such a way that

$$\Lambda(\eta, \lambda) = L(X, V_P) + L(Y, V_Q) - \lambda \mathrm{J}_{R^{SL}}(\eta), \tag{14.24}$$

or, more explicitly,

$$\Lambda(\eta, \lambda) = L(X, V_P) + L(Y, V_Q) - \lambda_P^T \mathrm{J}_{P,R^{SL}}(\eta) - \lambda_Q^T \mathrm{J}_{Q,R^{SL}}(\eta). \tag{14.25}$$

The components of interest of the Jacobian $\mathrm{J}_{R^{SL}}(\eta)$ are defined by components for the anchor item parameters in P,

$$\frac{\partial R^{SL}(\eta)}{\partial \beta_{i,P}} = \sum_{g=1}^{G} \pi_g^* \frac{\partial p\left(\theta_g, \beta_{VPi}\right)}{\partial \beta_{iVP}} 2\left[\sum_{l=1}^{L}\left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right)\right)\right], \tag{14.26}$$

and for the components representing the item parameters in Q we find

$$\frac{\partial R^{SL}(\eta)}{\partial \beta_{i,Q}} = -\sum_{g=1}^{G} \pi_g^* \frac{\partial p\left(\theta_g, \beta_{VQi}\right)}{\partial \beta_{iVP}} 2\left[\sum_{l=1}^{L}\left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right)\right)\right] \tag{14.27}$$

because of the negative sign of all $\beta_{\bullet,Q}$ terms. Let $W \in \{P, Q\}$ denote that the following exposition applies to both P and Q. The derivatives in the equations above actually represent vector-valued derivatives if $\beta_{i,W}$ is vector valued. For example, we have

$$\frac{\partial R^{SL}(\eta)}{\partial \beta_{i,W}} = \left(\frac{\partial R^{SL}(\eta)}{\partial a_{i,W}}, \frac{\partial R^{SL}(\eta)}{\partial b_{i,W}}, \frac{\partial R^{SL}(\eta)}{\partial c_{i,W}} \right)$$

in the case of the 3PL model, where W stands for either P and Q.

By maximizing Equation 14.19 we will find the transformation of the difficulties and the slopes in one population based on matching test characteristics. Note that in our approach this transformation does not need to be linear (although it will be linear if the model fits the data).

14.3.7 Lagrangean Haebera Scale Linkage

Haebara (1980) expressed the differences between the characteristic curves as the sum of the squared differences between the item characteristic functions for each item over the common items for examinees of a particular ability θ_n. The Haebara method is more restrictive than the Stocking and Lord (1983) method because the restrictions take place at the item level (i.e., for each item), whereas the Stocking and Lord approach poses a global restriction at the test level. The slope and the intercept of the linear transformation can be found by minimizing the expression on the right-hand side of Equation 14.28:

$$(A, B) = \min \left[\sum_{g=1}^{G} \pi_g^* \sum_{l=1}^{L} \left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, B + A\beta_{VQl}\right) \right)^2 \right], \tag{14.28}$$

(see, e.g., Kolen & Brennan, 1995, p. 170).

The algorithm we are proposing is similar to the one described previously for the Stocking and Lord scale linking; the only difference (from the computational point of view) is in the form of the restriction function:

$$R^H(\eta) = \left[\sum_{g=1}^{G} \pi_g^* \sum_{l=1}^{L} \left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right) \right)^2 \right] \tag{14.29}$$

As before, in order to minimize Equation 14.29, we implement the Lagrangeans in such a way that

$$\Lambda(\eta, \lambda) = L(X, V_P) + L(Y, V_Q) - \lambda \mathrm{J}_{R^H}(\eta), \tag{14.30}$$

or, more explicitly,

$$\Lambda(\eta, \lambda) = L(X, V_P) + L(Y, V_Q) - \lambda_P^T \mathrm{J}_{PR^H}(\eta) - \lambda_Q^T \mathrm{J}_{QR^H}(\eta). \tag{14.31}$$

The components of interest of the Jacobian $\mathrm{J}_{R^H}(\eta)$ are

$$\frac{\partial R^H(\eta)}{\partial \beta_{lP}}, \frac{\partial R^H(\eta)}{\partial \beta_{lQ}}. \tag{14.32}$$

Again, the partial derivatives may be vector valued for each $\beta_{i,P|Q}$, so that the dimension of the restriction is approximately $2L$ times average number of item parameter dimensions, 3 if only the 3PL model is used but maybe higher when generalized partial-credit items or other polytomous item response models are present.

14.4 An Illustration of Concurrent Calibration

The illustration reuses data from a pilot language test that was described in detail by von Davier (2005), where the data were analyzed using diagnostic models as well as IRT models. For the selection of models tested, von Davier (2005) found that the mixed 2PL, generalized partial-credit model fits the data best among the models tested in this study. For illustration of some of the concepts outlined in this chapter, we reproduce only the slopes and difficulties of the dichotomous items, so that the tables of this example will consist of subset items, all with two response categories. Items 11, 25, and 38 were items with a polytomous response format. They were part of the calibration but are not reproduced here in order to keep the tables easy to read and to allow comparisons of constraints based on one slope and one difficulty parameter only.

The data came from different subsamples, each student receiving an identification code number that included information about the testing site. For the example presented here, students were split into two subsamples comprising of identification codes lower than 3,000 and equal to or higher than 3,000. This produced two subsamples that represented similar populations such as the ones observed when data are collected over 2 days of test administration, a common feature of large-scale testing programs. In our example, the data split in such a way produced two subsamples that may differ slightly in average ability, although not expected to demonstrate systematic differences in response rates. The first subsample, from population P, consists of $N_P = 1{,}463$, and the subsample that constitutes population Q consists of $N_Q = 1{,}257$ students. All item calibrations were performed using the *mdltm* software (von Davier, 2005), which incorporates standard ways of constraining the average of item parameters, or the mean and variance of the ability distribution using parameter adjustments in the maximization methods employed, in order to remove the indeterminacy of the IRT scale. More specifically, the estimates of item difficulties and slopes from the current iteration of the maximization procedure are adjusted either to match certain constants or to match the current estimate of the mean and variance of the ability distribution. Then, maximization continues in the following iteration. These steps, or steps equivalent to these, correspond to Equations 14.13 and 14.14 and are necessary to remove the indeterminacy (invariance under linear transformations) of the IRT scale.[3]

For the second calibration in this example, equality constraints across populations were used to link the scales across multiple groups. The equality constraints of item difficulties and slope parameters across two or more populations in a multiple-group IRT model, as shown in the example below, are implemented in *mdltm* in a way that is equivalent to the Lagrangean multiplier approach presented in Section 14.3.3. and 14.3.4.

Table 14.1 shows the item parameter estimates in these two subsamples for a multiple-group, 2PL IRT model with the same set of constraints on the mean of item difficulties and slopes in both populations.

Table 14.2 shows item parameter estimates of a concurrent calibration that sets equality constraints only on a subset of the items, in this example imposing equality of item parameters on the first half, or 19 items. This is, as outlined above, a form of IRT scale linkage that involves setting equality constraints on subsets of items and allowing unique parameters for other items across the groups. If all items were assumed to have the same parameters, this would represent a concurrent calibration with complete equality of all item parameters (see the relevant subsections in Section 14.3).

The ability distributions in the two different calibrations are given in Table 14.3. The constraints on the average difficulty and slope parameters were chosen to match 0.0 for the average difficulty and 1.0 for the average slope. This did not change overall results, because IRT scales are invariant under linear transformations.

Table 14.3 shows that the means and variances of the ability distributions in the two different conditions are comparable. The constraints used in Condition A are weaker than in Condition B. The number of constraints set in Condition B is 41 (19 slopes + 18 dichotomous difficulties + 4 thresholds for the five-category Item 11[4]) and amounts effectively to a reduction of free parameters for the two-group model to be estimated from 170 in Condition A to 129 in Condition B.

The fit of the models estimated in the two conditions is comparable, the log likelihood is $L_a = -56994.954$ ($\text{BIC}_a = 115334.33$) for Condition A and $L_a = -57018.37$ (BIC_b=115056.93) for Condition B. Taking the reduction of estimated parameters into account, the smaller Bayesian information criterion

[3] An alternative to procedures relying on these averages is to remove the indeterminacy by setting one item difficulty and the slope of that item to prespecified constants, and fix these values for that item without updating in the maximization.

[4] The polytomous items are not shown in the tables but were part of the data, with mixed item format in both calibrations.

Table 14.1 Multiple Group Calibration in Two Populations, Separately Calibrated

Item	Mean slopes		Mean difficulties	
	Pop. *P*	Pop. *Q*	Pop. *P*	Pop. *Q*
1	0.789	0.867	−0.345	−0.500
2	0.976	0.893	−0.042	0.190
3	1.222	1.503	0.603	0.593
4	1.304	1.405	0.670	0.759
5	0.660	0.770	−0.512	−0.461
6	1.042	1.059	0.710	0.584
7	0.701	0.678	−0.039	−0.113
8	1.672	1.471	0.152	0.351
9	0.373	0.418	0.820	0.749
10	0.830	0.838	−0.804	−0.907
12	1.132	0.872	0.635	0.845
13	2.270	1.787	2.740	2.604
14	0.581	0.530	0.329	0.383
15	1.091	1.243	−0.245	−0.467
16	1.057	0.900	−1.262	−1.052
17	0.751	0.912	−0.919	−1.042
18	1.449	1.670	−1.158	−1.365
19	1.365	1.305	2.021	2.034
20	2.011	1.592	1.937	2.212
21	0.815	0.779	0.328	0.342
22	1.234	1.598	0.850	0.872
23	1.248	1.231	0.329	0.063
24	1.019	1.042	0.301	0.124
26	0.886	0.930	−0.576	−0.550
27	0.979	0.842	−0.494	−0.394
28	0.853	0.930	0.627	0.515
29	0.563	0.609	0.433	0.269
30	1.845	1.768	0.331	0.614
31	0.478	0.530	−0.839	−0.961
32	1.036	1.199	−0.367	−0.468
33	0.628	0.588	0.013	0.307
34	0.645	0.600	0.381	0.606
35	0.914	0.886	−0.590	−0.523
36	0.686	0.904	1.291	0.967
37	0.661	0.504	−0.941	−0.862

(Schwarz, 1978) for Condition B indicates that the model with 19 items constrained to have equal item parameters across groups fits the data relatively better in terms of balancing parsimony and model-data fit.

14.5 Discussion

This chapter presents a new perspective on IRT linking. We introduce a unified approach to IRT linking, emphasizing the similarities between different methods. We show that IRT linking might consist of a family of IRT linking functions, where

Table 14.2 Concurrent Calibration With First 19 Items Constrained to Same Slopes and Difficulties in Both Populations

Item	Slopes		Difficulties	
	Pop. *P*	Pop. *Q*	Pop. *P*	Pop. *Q*
1	0.833	0.833	−0.423	−0.423
2	0.929	0.929	0.075	0.075
3	1.349	1.349	0.588	0.588
4	1.347	1.347	0.711	0.711
5	0.709	0.709	−0.491	−0.491
6	1.059	1.059	0.648	0.648
7	0.694	0.694	−0.073	−0.073
8	1.564	1.564	0.259	0.259
9	0.397	0.397	0.784	0.784
10	0.838	0.838	−0.853	−0.853
12	0.995	0.995	0.750	0.750
13	1.997	1.997	2.669	2.669
14	0.555	0.555	0.359	0.359
15	1.170	1.170	−0.354	−0.354
16	0.979	0.979	−1.153	−1.153
17	0.828	0.828	−0.981	−0.981
18	1.554	1.554	−1.256	−1.256
19	1.338	1.338	2.031	2.031
20	1.993	1.602	1.946	2.206
21	0.821	0.779	0.332	0.338
22	1.241	1.603	0.858	0.865
23	1.256	1.227	0.336	0.058
24	1.023	1.045	0.308	0.118
26	0.894	0.927	−0.574	−0.552
27	0.985	0.843	−0.489	−0.398
28	0.861	0.928	0.631	0.512
29	0.566	0.609	0.436	0.267
30	1.848	1.768	0.343	0.606
31	0.485	0.527	−0.842	−0.961
32	1.046	1.196	−0.365	−0.472
33	0.633	0.588	0.016	0.304
34	0.651	0.601	0.383	0.603
35	0.923	0.884	−0.587	−0.526
36	0.691	0.900	1.295	0.963
37	0.667	0.505	−0.940	−0.865

Table 14.3 Comparison of Mean and Variance Estimates for the Ability Parameter Across the Concurrent Calibration With 19 Items Constrained and the Mean-Mean Linkage Constraining Item Parameters

Population	Mean	Standard deviation
A. Constraints on mean of item difficulties (0.0) and slopes (1.0)		
P	1.041	1.269
Q	0.898	1.369
B. Equality constraints on first 19 items		
P	1.028	1.260
Q	0.904	1.371

restrictions can be turned on or off, according to what the data suggest. Moreover, this new approach allows both generalizations and exactly matching implementations of the existing methods, as the existing IRT linking methods are included as special cases in this new family of IRT linking functions.

We believe that this approach will allow the development of statistical tests (such as Lagrange multiplier tests) for checking the appropriateness of different IRT linking methods (see Glas, 1999, for a similar approach used for investigating nested IRT models). Such a test would allow checking whether lifting certain restrictions will yield a significant improvement in model-data fit, for example in a case where Lagrangean concurrent calibration is used for all anchor items in a vertical linkage and a certain set of items exposes parameter drift over time.

This approach to IRT linking can be easily viewed in a Markov chain Monte Carlo (MCMC) framework, where, by specifying appropriate prior distributions, the estimation of the modified likelihood functions is straightforward. At the same time, the view of any linking function as a restriction function implies a larger flexibility in the linking process: When dealing with vertical linking, this method can incorporate the modeling of growth, possibly expressed as a hierarchical structure of the item parameters in the anchor.

Such a hierarchical structure was proposed by Patz, Yao, Chia, Lewis, and Hoskens (2003), who used the MCMC estimation method. The hierarchical approach Patz et al. proposed is “a more general version of concurrent estimation of the unidimensional IRT model" (p. 40), and their motivation has similarities with ours: to unify the two most commonly used linking methods for vertical equating, the very restrictive concurrent calibration method and separate calibration followed by a test characteristic-curves linking.

If we recast this hierarchical approach of the proficiencies across grades into a hierarchical structure of the (common) item difficulties, a short summary of the Patz et al. (2003) approach (slightly generalized) in our notation is

$$R_l(\eta) = k_l(h(\beta_{V_{Pl}}) - f(h(\beta_{Q_{Pl}}))), \tag{14.33}$$

where $k_l = 1$ for active restrictions on item l, R_l denotes the component l of the restriction function, h is the projection described before, and f is a function of the common item parameters of the old administration (or previous grade). In order to obtain the hierarchical structure at the level of the difficulties of the common items, we consider

$$h(\beta_{V_{Pl}}) = b_{V_{Pl}}.$$

Following the approach of Patz et al. (2003), the relationship between the difficulty of the item parameters across grades can be expressed as a quadratic function,

$$f(h(\beta_{Q_{Pl}})) = f(b_{Q_{Pl}}) = \alpha b_{Q_{Pl}}^2 + \gamma b_{Q_{Pl}} + \delta, \tag{14.34}$$

where α, γ, δ are additional parameters of the model that need to be estimated. Furthermore, the modified likelihood function, with a restriction function described in Equation 14.33, can be maximized using the Lagrange multipliers in the same way as explained for the other linking methods. Note that from a computational point of view, this is only a slight generalization of the restriction functions described for the mean-mean and mean-var linking methods.

Obviously, additional investigations are necessary in order to insure that the model is identified and to insure the convergence of the maximization algorithm. Although here we propose an analytical approach and will try to use an expectation-maximization algorithm, a MCMC estimation method would be straightforward to implement.

Moreover, the approach presented in this paper may easily be extended to multidimensional IRT models, at least for simple-structure, multiscale IRT models (like the one used in the National Assessment of Educational Progress and other large-scale assessments). There is no additional formal work necessary, and the method proposed in this report can be readily applied. Patz et al. (2003) also investigated multidimensional IRT models for vertical linking and used the MCMC estimation method. However, implementing and maximizing modified likelihood functions under such restrictions using analytical methods are of interest for future research.

Longitudinal studies also may benefit from these two approaches: one that assumes a hierarchical structure in the item parameters of the anchor and one that assumes a multidimensionality of the proficiencies (or of the common item difficulties) across school grades. This flexibility also may be a desirable feature in educational large-scale assessments, where in some instances it is necessary to relax the restriction of equality of all item parameters. In conclusion, this new approach is very promising for assessment programs that use IRT linking.

Author Note: Any opinions expressed in this chapter are those of the author and not necessarily of Educational Testing Service.