Chapter 15
Linking With Nonparametric IRT Models

Xueli Xu, Jeff A. Douglas, and Young-Sun Lee

15.1 Introduction

In educational testing, it is a common practice to produce test forms under a nonequivalent groups with anchor test (NEAT) design. In this design, the two test forms share a certain number of common items, while the populations who take the test forms might not be equivalent to each other. Test linking is conducted to establish the equivalency of ability scales from separate item response theory (IRT) calibrations of two test forms. Among existing linking methods, two approaches are related to the study in this paper: IRT model-based linking (Loyd & Hoover, 1980; Marco, 1977; Stocking & Lord, 1983) and equipercentile linking (Kolen & Brennan, 1995; von Davier, Holland, & Thayer, 2004b).

From the viewpoint of IRT models, the necessity of linking is embedded in the models themselves. That is, the linear indeterminancy of ability requires a linear transformation to ensure the equivalency between the two test forms when separately calibrated. If a three-parameter logistic (3PL) model

$$P(\theta; a, b, c) = c + \frac{1 - c}{1 + exp(-a(\theta - b))} \tag{15.1}$$

is used in item calibration, then $P(A\theta + B; a/A, Ab + B, c)$ remains the same value as $P(\theta; a, b, c)$ by an appropriate linear transformation. Thus, the linear transformation is

X. Xu (✉)
Educational Testing Service, Princeton, Rosedale Rd, Princeton, NJ 08541, USA
e-mail: xxu@ets.org

J.A. Douglas
101 Illini Hall, 725 S. Wright St, Champaign, IL 61820, USA
e-mail: jeffdoug@uiuc.edu

Y.-S. Lee
Teachers College, Columbia University, 525 W.120th St, New York, NY 10027, USA
e-mail: yslee@tc.columbia.edu

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_15,

used to maintain the item characteristic curves (ICCs) of common items. However, the linear transformation might not be appropriate if the parametric form is incorrect, or if the difference between ability distributions involves something more than a location and scale change. When assuming that the target score distributions from the two test forms are equivalent when the common-item score is held constant, this issue has been addressed by the equipercentile linking. This method is implemented by transforming raw scores from one test to the scale of raw scores of another test. Obviously, this method is model independent. However, the assumption of equivalence is likely to fail when the two groups differ substantially in ability, age, or other demographic information (Liou, 1998).

An alternative to these two approaches (IRT model-based linking and equipercentile linking) is fitting a more flexible model to the data and conducting linking by using nonparametrically estimated items. Currently, no methods for linking are available when using nonparametrically estimated IRT models. This is a serious practical limitation. These flexible nonparametric models will prove most useful when no single parametric family fits the entire set of ICCs well, and in situations like this, nonparametric methods of linking will be required. Our aim is to make such methods available so that nonparametrically estimated models can be considered to be practical alternatives for operational use.

15.2 Estimating the ICCs Nonparametrically

Methods of nonparametric ICC estimation include kernel smoothing with a selected scale for the latent trait (Douglas, 1997; Ramsay, 1991), isotonic regression (Lee, 2002), monotone splines (Ramsay & Abrahamowicz, 1989), penalized maximum likelihood estimation (Rossi, Wang, & Ramsay, 2002), as well as several others. For the linking methods to be presented later, most of the nonparametric estimation methods will apply. However, due to the need to compute the inverse function of an ICC, monotone methods are preferred. In this paper, we use kernel smoothing to obtain the initial estimates of ICCs and then smooth the estimates once again using a B-spline smoother constrained to be monotone. First we review kernel smoothing and constrained B-spline smoothing.

15.2.1 Kernel Smoothing

Suppose N examinees (i = 1,2,. . ., N) are randomly sampled and take a test of length n (j = 1,2,. . ., n). The kernel smoothed estimate of the ICC of item j, $P_j(\theta)$, is the weighted average of the response vector $\{Y_j\} = \{Y_{1j}, \ldots, Y_{Nj}\}$,

$$P_j(\theta) = \sum_{i=1}^{N} w_i Y_{ij}, \tag{15.2}$$

where the weights w_i of examinee i are defined in a certain way so that they are nonnegative and reach a maximum when $\theta = \theta_i$ and will approach or equal zero as $|\theta - \theta_i|$ increases. In order to keep $P_j(\theta)$ within [0,1], the weights should, at the same time, satisfy two conditions: $w_i \geqslant 0$ and $\sum_i w_i = 1$. Thus, it is preferable to use nonnegative kernel functions and Nadaraya-Watson weights (Nadaraya, 1964; Watson, 1964):

$$w_i = \frac{K(\frac{\theta-\theta_i}{h})}{\sum_i K(\frac{\theta-\theta_i}{h})}, \tag{15.3}$$

where h is a smoothing parameter, θ is a grid point along a desired latent scale, θ_i is the ability of examinee i, and $K(\cdot)$ is a kernel function.

The kernel smoothing estimator of $P_j(\theta)$ is consistent when θ_i can be estimated without error. However, the latent trait values of θ_i are not observable. The Nadaraya-Watson weights still can be used after substituting θ_i for true θ_i. A common and appropriate way to estimate θ_i is to transform the ranked raw scores to the corresponding quantiles of the chosen latent ability distribution, which is usually on a standard uniform U(0,1) scale. This leads to the kernel smoothed estimate

$$P_j(\theta) = \frac{\sum_{i=1}^{N} K(\frac{\theta-\hat{\theta}_i}{h}) Y_{ij}}{\sum_{i=1}^{N} K(\frac{\theta-\hat{\theta}_i}{h})}, \tag{15.4}$$

proposed by Ramsay (1991) and implemented in *TestGraf* (Ramsay, 2001). The consistency of this estimate was proved by Douglas (1997).

In kernel smoothing, the bandwidth h is used to control the balance between the bias and variance of estimation. At this point, there is no theorem on an optimal bandwidth for ICC estimation. However, we can use results from simpler models where the covariate is measured without error as a guideline. For example, Ramsay (1991) suggested that $h = N^{-1/5}$ works well when using a Gaussian kernel.

15.2.2 Constrained B-spline Smoothing

A simple but effective monotone smoothing method using splines was proposed to solve the nonparametric regression problem (He & Shi, 1998). They proposed a method based on the constrained least absolute deviation in the space of B-spline functions. The idea of this method is to characterize the monotonicity by linear constraints as well as to solve the least absolute deviation efficiently via linear programming (He & Shi, 1998). Suppose n pairs of observations (x_i, y_i) are used to estimate the nondecreasing regression curve $g(x)$. The model to be estimated is

$$y_i = g(x_i) + u_i, \quad i = 1, \ldots, n \tag{15.5}$$

where u_i is random error. Assume that $g(x)$ is uniformly continuous and has a second-order derivative. Then the function g and its first-order derivative function g' can be approximated adequately by B-splines and their derivatives. Assuming $x \in [a, b]$, letting the knots t_s be selected as $a = t_1 < t_2, \ldots, < t_{k_n} = b$; He and Shi (1998) chose to use quadratic B-splines with degree of 2. Let $B(x) = (B_1(x), B_2(x), \ldots, B_Q(x))^T$ be the normalized B-splines based on the knots t_s, where $Q = k_n + p$, with $p + 1$ being the order of the B-spline. The estimate of $g(x)$, denoted by $g_n(x) = B(x)^T\hat{\alpha}$ for $\hat{\alpha} \in \mathbf{R}^Q$, is obtained by minimizing

$$\sum_{i=1}^{n} |y_i - B(x_i)^T\alpha|, \tag{15.6}$$

subject to the linear constraint to ensure monotonicity,

$$B'(t_s)^T\alpha > 0 \tag{15.7}$$

where $s = 1,..., k_n$ and is subject to other linear constraints, such as those for the boundary points. Here, $B'(\cdot)$ is a vector of the first derivative functions of $B(\cdot)$. This technique is implemented by an R function *SCOBS* (He & Ng, 1998). The consistency of function estimation and effectiveness of this method have been explored in several papers (He & Shi, 1998; Koenker, Ng, & Portnoy, 1994).

Though this constrained B-spline method cannot be used directly to estimate nonparametric IRT models for binary responses, it can be used as a postsmoother after a nonmonotone method such as kernel smoothing is used to estimate ICCs. In particular, we treat the kernel smoothed estimate $P_j(\theta_m)$ and θ_m as a pair of observations without error, in which θ_m is a grid point on a desired scale.

15.3 Nonparametric IRT Linking

Nonparametrically estimated ICCs provide us not only with more flexible forms to fit the data but also a platform to conduct linking. Two approaches are proposed to conduct linking under nonparametric IRT models. One of them conducts linking on a uniform U(0,1) scale, the other on a normal N(0,1) scale. These two approaches are introduced in the following sections.

15.3.1 Constrained Spline Linking on a U(0,1) Scale

Let η be a point in the sample space of a latent variable and $P(\eta)$ be the probability of giving a correct answer conditional on η. Let F_1 and F_2 be the cumulative

distribution functions of two nonequivalent testing populations, with $F_1(\eta) = \theta_1$ and $F_2(\eta) = \theta_2$ being the correspondence in different populations. As mentioned in the introduction to the kernel smoothing method, the nonparametric calibration process implicitly transforms η from its original scale to a U(0,1) scale through the cumulative distribution function. For illustration, Equations 15.8 and 15.9 describe the process of calibration with respect to F_1 and F_2, respectively.

$$P(\eta) = P(F_1^{-1}(\theta_1)) = P_1(\theta_1) \tag{15.8}$$

$$P(\eta) = P(F_2^{-1}(\theta_2)) = P_2(\theta_2). \tag{15.9}$$

Thus, we in fact put the latent variable η on a U(0,1) scale relative to the different groups. This will lead to the "pseudo difference" expressed in ICCs P_1 and P_2. Linking means finding the transformation

$$\theta_2 = F_2 \circ F_1^{-1}(\theta_1) = P_2^{-1} \circ P_1(\theta_1) = g(\theta_1) \tag{15.10}$$

on the U(0,1) scale. Based on the fact that $\hat{P}_1(\theta)$and $\hat{P}_2(\theta)$are consistent estimates of $P_1(\theta)$and $P_2(\theta)$(Douglas, 1997), we can obtain an estimate of the linking function $g(\theta)$ by minimizing the loss function:

$$\int \sum_{j \in common} |\hat{g}(\theta) - \hat{P}_{2j}^{-1} \circ \hat{P}_{1j}(\theta)| d\theta, \tag{15.11}$$

where $P_{1j}(\theta)$ is the estimate of ICC for item j and Group 1, $P_{2j}(\theta)$ for item j and Group 2, and the summation is taken over all the common items in the two test forms. In our application, $\hat{P}_{2j}^{-1} \circ \hat{P}_{1}j(\theta)$ is taken as an observation without error. The estimate $\hat{g}(\theta)$ is obtained as the best solution in the span of a family of constrained B-splines. One approximates $\hat{g}(\theta)$ by $\sum_{m=1}^{M} \beta_m B_m(\theta)$, subject to

$$\beta_m B'_m(\theta) > 0,$$

$$g'(\theta) > 0,$$

$$g(0) = 0,$$

$$g(1) = 1,$$

where $B'_m(\cdot)$ and $g'(\cdot)$ are first derivatives of $B_m(\cdot)$ and $g(\cdot)$ defined earlier.

The additional constraints and possible penalty functions make it difficult to derive an explicit standard error of the linking function. However, the bootstrap method can be used to obtain pointwise estimates of the standard error.

15.3.2 Linear Linking on a N(0,1) Scale

It is well known that the linear linking is appropriate for the logistic and the probit IRT models with normally distributed latent traits. Even though the linking function in the nonparametric setting need not be constrained to be linear, linear linking on a N(0,1) scale is still possible when using nonparametrically estimated ICCs. Under a nonparametric IRT framework, we can change the scale of the latent variable if desired. To illustrate this issue, we revisit Equations 15.8 and 15.9 and transform the U(0,1) scale to a N(0,1) scale by inserting the cumulative distribution function of a standard normal random variable, denoted by Φ:

$$P(\eta) = P_1(\theta_1) = P_1(\Phi(\theta_{11})) \tag{15.12}$$

$$P(\eta) = P_2(\theta_2) = P_2(\Phi(\theta_{21})). \tag{15.13}$$

Thus, θ_{11} and θ_{21} are now on a N(0,1) scale. The linear linking is done by finding A and B to minimize the loss function:

$$\int \sum_{j\in(common)} [\theta_i - A\Phi^{-1} \circ {P_{2j}}^{-1} \circ P_{1j} \circ \Phi(\theta_i) - B]^2 d\theta \tag{15.14}$$

where $P_{1j}(\cdot)$ and $P_{2j}(\cdot)$ are the consistent estimates of $P_{1j}(\cdot)$ and $P_{2j}(\cdot)$ for item j in common item set, respectively.

15.4 Simulation Study

15.4.1 Design

In our simulation study, two parallel test forms and four pairs of populations were selected to study the behaviors of the proposed approaches. Each test form contained 30 common items and 10 unique items. The items were characterized by a 2PL model. The parameter a is generated from U(0.75,2.5) and the parameter b is generated from a N(0,1). These two test forms are denoted as form 1 and form 2. The four pairs of populations in this study are:

- N(0,1) vs. N(0.25,1)
- truncated normal N(0,1)I[−2,2] vs. U(−2,2)
- Beta(1,1) vs. Beta(1.5,0.5)
- Beta(1,2) vs. Beta(2,1)

Their corresponding density functions are shown in Figure 15.1.

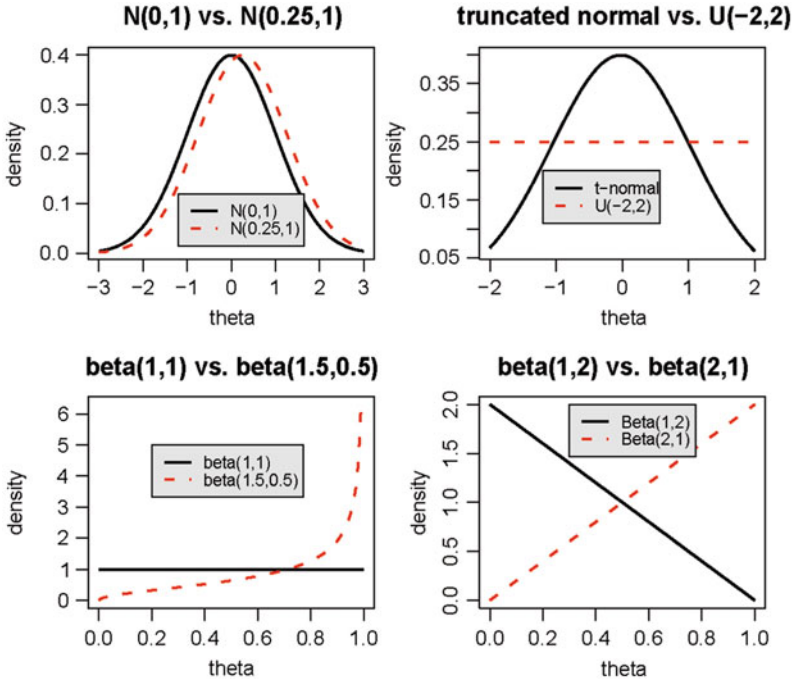

Fig. 15.1 Four pairs of comparison

Figure 15.1 gives us an understanding of how these four pairs of populations compare. In the first pair, both populations have the same shape except for the location. In the second pair, both populations are symmetric about point 0 but have different shapes. For the third pair, one population is symmetric, whereas the other has extreme mass on the high end of the ability scale. In the final pair, both populations have larger mass on the end points of the ability scale, but in different directions. The first pair is an ideal situation for real practice and is more realistic than the other three pairs. These more extreme cases are included to examine how the proposed approaches behave in such situations. If the proposed approaches can work in these extreme situations, then they can be expected to work in less extreme conditions.

Within each pair, the first population is considered the target population; the second one will be linked to the target population. The true linking function in each pair can be derived from the true distribution forms. Suppose F_1 and F_2 are the cumulative distribution functions in each pair. They are on a U(0,1) scale. The true linking function is $F_2^{-1} \circ F_1$, if the first distribution is taken as the target. The true linking function also can be obtained by smoothing techniques such as the constrained splines if its closed form is hard to get. Within each pair, 3,000 examinees are generated from the specified populations respectively and the data are generated

according to the 2PL model. Once the responses are generated, these two test forms are calibrated on a U(0,1) scale and a N(0,1) scale for their corresponding populations. The kernel smoothing method is then used to get the initial estimates of the ICCs, with the bandwidth set as 0.20, which is roughly $3000^{-1/5}$. Then *SCOBS* is utilized to make the estimated ICCs monotone and to estimate the linking function $\hat{g}$ when using the constrained spline approach. The maximum number of knots is set at 6, which was recommended by He and Shi (1998). In the linear approach, the linking parameters A and B are estimated from a least squares criterion displayed in Equation 15.14. The entire procedure is repeated 100 times for each pair in order to calculate the linking errors.

Three statistics are used to measure the efficacy of the proposed linking methods. One is the root mean-square error (RMSE) of the linking function along the U(0,1) scale,

$$RMSE = \sqrt{\sum_m \sum_r [g(\theta_m) - \hat{g}_r(\theta_m)]^2 / R / M}, \tag{15.15}$$

where R is the number of replications, M is the number of grid points on θ scale, and m and r are indices for M and R. Here $\hat{g}_r$ is the estimated linking function from the rth repetition, and g is the true linking function. This statistic is used to examine how well the true linking function is recovered by the estimated one. The other is the root mean-square difference (RMSD) of test functions for the anchor items,

$$RMSD = \sqrt{\sum_m \left(\sum_j P_{2j}(\hat{g}(\theta_m)) - \sum_j P_{1j}(\theta_m)\right)^2 / M}, \tag{15.16}$$

where j is the index for common items and M has the same meaning as above. This statistic is employed to examine the recovery of the true test characteristic function. The third statistic is called an improvement ratio (IR), which is similar to the statistic used in Kaskowitz and De Ayala (2001).

$$IR = 1 - \frac{F_{equate}}{F_{original}}, \tag{15.17}$$

where

$$F_{equate} = \sum_m \sum_j [P_{2j}(\hat{g}(\theta_m)) - P_{1j}(\theta_m)]^2 / (J * M) \tag{15.18}$$

and

$$F_{original} = \sum_m \sum_j [P_{2j}(\theta_m) - P_{1j}(\theta_m)]^2 / (J * M). \tag{15.19}$$

Here j is the total number of anchor items and M has the same meaning as above. $F_{original}$ represents the largest discrepancy between item response functions on the common set, whereas F_{equate} stands for the discrepancy between item response functions due to linking. This ratio reflects the improvement due to linking.

15.4.2 Results and Discussion

The summary of the results is shown in Tables 15.1–15.3 and Figures 15.2 and 15.3. Table 15.1 presents the RMSE of the estimated linking function. Since the scales of the constrained spline linking and the linear linking are different, the scale of linear linking is transformed to the U(0,1) scale before calculating the RMSE. The label $RMSE_{CS}$ represents the RMSE using the constrained spline linking, whereas $RMSE_L$ represents the RMSE using linear linking. Table 15.1 reveals that linear linking has similar or smaller RMSE in the estimate of the linking function, but all the RMSE

Table 15.1 Root Mean-Square Error (RMSE) Under Two Proposed Approaches Using 3,000 Examinees

	$RMSE_{CS}$	$RMSE_L$
N(0,1) vs. N(0.25,1)	0.0167	0.0134
N(0,1)I[−2,2] vs. U(−2,2)	0.0161	0.0218
Beta(1,1) vs. Beta(1.5,0.5)	0.0267	0.0263
Beta(1,2) vs. Beta(2,1)	0.0229	0.0131
Average	0.0206	0.0187

Note. CS = constrained spline linking; L = linear linking.

Table 15.2 Root Mean-Squared Difference (RMSD) Under Two Proposed Approaches Using 3,000 Examinees

	$RMSD_{CS}$	$RMSD_L$
N(0,1) vs. N(0.25,1)	0.2130	0.2214
N(0,1)I[−2,2] vs. U(−2,2)	0.3363	0.3416
Beta(1,1) vs. Beta(1.5,0.5)	0.3780	0.3760
Beta(1,2) vs. Beta(2,1)	0.3233	0.3186
Average	0.3127	0.3144

Note. CS = constrained spline linking; L = linear linking.

Table 15.3 Improvement Ratio (IR) Under Two Proposed Approaches Using 3,000 Examinees

	IR_{CS}	IR_L
N(0,1) vs. N(0.25,1)	0.868	0.816
N(0,1)I[−2,2] vs. U(−2,2)	0.963	0.903
Beta(1,1) vs. Beta(1.5,0.5)	0.956	0.946
Beta(1,2) vs. Beta(2,1)	0.821	0.808
average	0.902	0.880

Note. CS = constrained spline linking; L = linear linking.

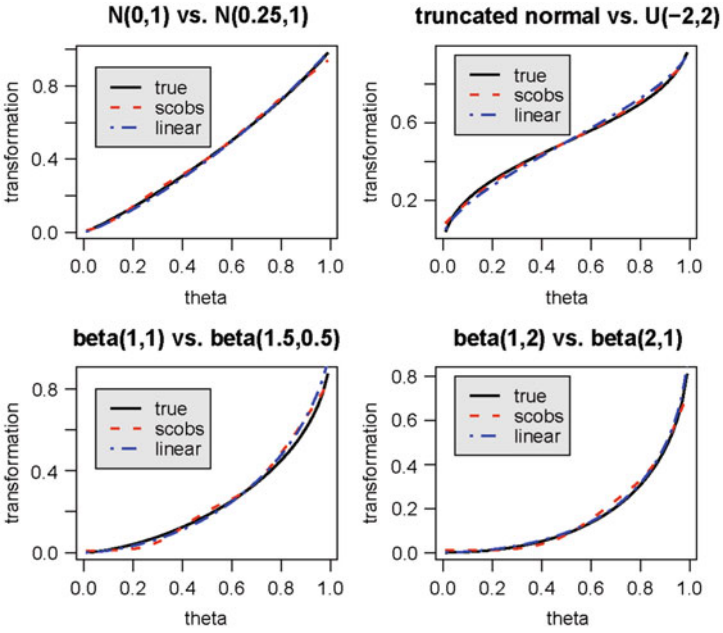

Fig. 15.2 Recovery of true linking functions

are below the nominal level 0.05. Table 15.2 presents the RMSD for the common-item test. The scale of the linear linking is transformed to U(0,1) scale before calculating the RMSD. The label $RMSE_{CS}$ is the RMSD of the test function by constrained spline linking, whereas $RMSE_L$ is the RMSD by linear linking. Notice that the RMSD of the test functions resulted from these two methods are similar to each other. Table 15.3 gives us the IR results for these two proposed linking approaches. Both methods have similar IRs in all cases. Figures 15.2 and 15.3 illustrate the results. Figure 15.2 presents the recovery of the true linking function, and Figure 15.3 presents the recovery of the true test function of the common items. The solid line represents the true function (either a linking function or a test function), and the dashed and the dashed-dotted lines are for the corresponding estimated functions by using the constrained B-spline and the linear methods, respectively. Given the figures are printed in black and white, it is hard to distinguish the lines in some graphs. Specifically, there are three lines in each graph. The solid line represents the true function (either a linking function or a test function), and the dashed and the dashed-dotted lines are for the corresponding estimated functions by using the constrained B-spline and the linear methods, respectively.

The results enable us to say that both proposed methods work well in all four situations in terms of recovering the true linking function. Both methods show similar performances in terms of the RMSD of the estimated test functions and IR. Furthermore, results show that the large population differences in the last two pairs have little impact in recovering the true linking functions but have impact in recovering the true test functions of the common items. When two extremely different populations are linked, it is expected that the test functions of these two

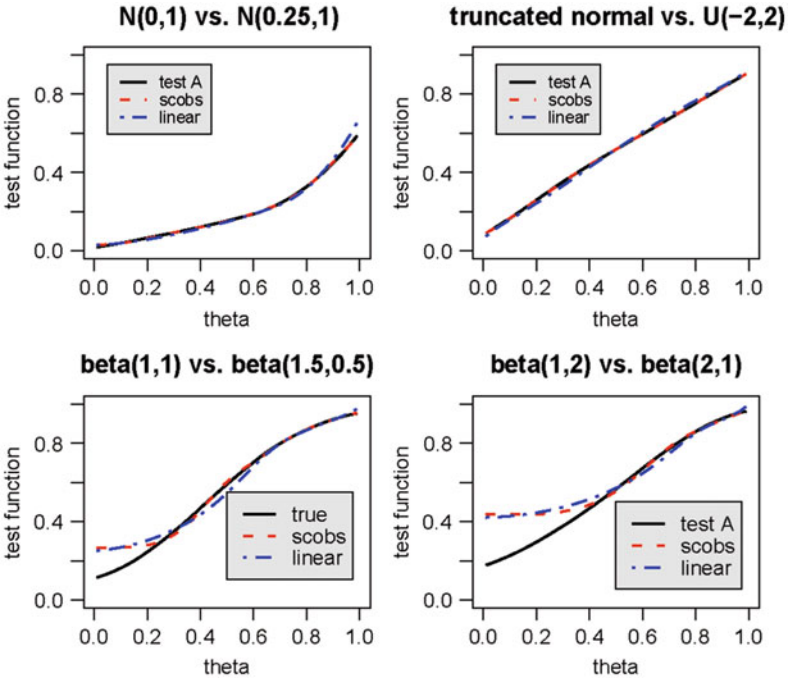

Fig. 15.3 Recovery of true test functions of common items

populations differ on the end points of the ability scale. It turns out that for the last two pairs, the true test functions have a large difference on the lower end of the ability scale. This fact leads to poor recovery of test functions on the lower end of ability, which is displayed in the last two pairs. From the results of the simulation study, we expect that the proposed methods are able to recover the true linking function even when the parametric models do not fit the data or when the testing populations of interest have large discrepancies.

15.5 Real Data Example

A real data example is also used to compare the two proposed methods and the test characteristic curve (TCC; Stocking & Lord, 1983) method. The TCC method is an IRT model-based approach. As described in the introduction, an IRT model-based method is initiated from the linear indeterminancy of ability in IRT functional form:

$$P(\theta; a, b, c) = P(A\theta + B; a/A, Ab + B, c). \tag{15.20}$$

Under the NEAT design, the common items are calibrated separately for each population. According to the TCC method, the slope A and the intercept B are obtained by minimizing the overall squared differences between ICCs for the

common items from separate calibrations. The objective function is shown in the equation below, denoted as *SL(θ)*.

$$SL(\theta) = \int [\sum_{j \in Common} P_j(\theta; a_{1j}, b_{1j}, c_{1j}) - \sum_{j \in Common} P_j(\theta; a_{2j}/A, Ab_{2j} + B, c_{1j})]^2 f_1(\theta) d\theta, \quad (15.21)$$

where a_j is the parameter a of item j for Group 1 or 2. The same indices are for parameter b and c. In this study, the programs BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and EQUATE (Baker, 1990) are used to calibrate item parameters and to implement the TCC method, respectively.

The data were taken from the responses to a mathematics placement test administered in 1993 in the University of Wisconsin system. Two forms of 46 multiple-choice-item tests with five alternatives, Form 1 and Form 2, are used in this study. Every 10th item is a pilot item, and all the other items on the test are common items. Omitted responses or not reached items were scored as incorrect. Form 1 was administered to 1,938 male students and Form 2 to 1,716 female students. After some preliminary analysis, one item was deleted from the analysis because its ICC was a decreasing function of the latent ability. It is assumed that the common items should behave similarly across testing populations. Thus SIBTEST (Shealy & Stout, 1993) was used to detect any possible differentially functioning items. In this study, the males were considered as the focal group and the females as the reference group. The critical value was set at α=0.05. After this examination, 21 items were left to construct the common-item group. To summarize, each form contained 45 items, with 21 common to both forms.

The empirical raw-score distributions of these two testing samples are shown in Figure 15.4. An empirical transformation function was derived by two steps. Specifically, the first step was to find the empirical raw-score distributions and convert them to a U(0,1) scale and then to find the transformation between these two transformed empirical raw-score distributions. This empirical linking function is shown as the solid line in Figure 15.5. The other estimated linking functions obtained from other methods (represented by the dashed or dashed-dotted lines) are compared to this empirical linking function.

Form 1 and Form 2 were calibrated with male and female groups, respectively. For each group, the items were calibrated parametrically by a 3PL model and nonparametrically on the U(0,1) scale and on the N(0,1) scale, and then transformed back to U(0,1) scale.

After item calibration, constrained spline linking, linear linking, and TCC method linking were used to link the female group to the male group. In the constrained spline approach, we specified the maximum number of knots as 6 and the bandwidth of smoothing as 0.22, which is approximately $1938^{-1/5}$. These methods were compared in terms of (a) the RMSE of the estimated linking functions, taking the empirical linking function as the truth, and (b) linking error, which

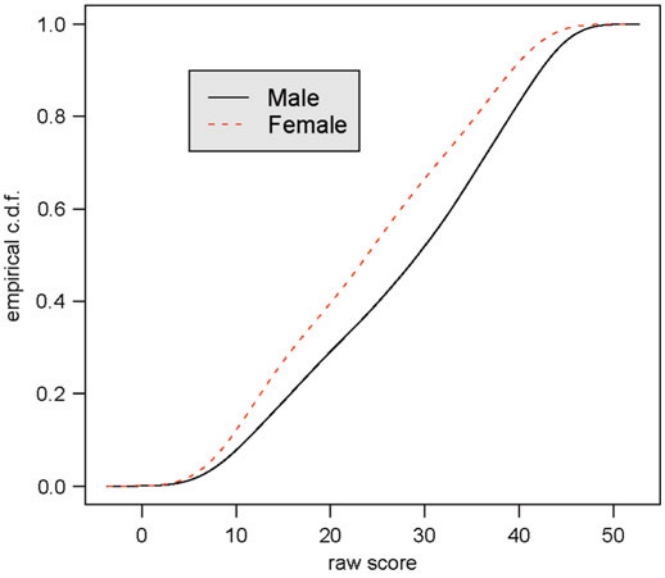

Fig. 15.4 Empirical score distributions for real data

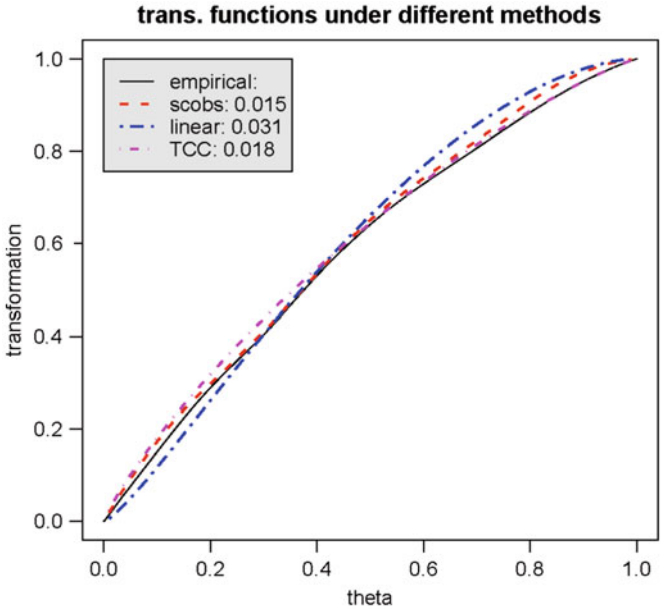

Fig. 15.5 Real data analysis: Transformation functions under different methods

for the constrained spline linking and linear linking is obtained from 100 bootstrap samples. Figure 15.5 depicts how close the estimated linking functions are to the empirical linking function.

The solid line in Figure 15.5 is the empirical relationship between abilities of the male and female groups. The other three curves are the relationship curves after conducting the linking procedures. The legend of the figure gives you the RMSE of these three linking approaches. For example, the RMSE by using constrained spline linking was 0.015, while linear aproach resulted in RMSE 0.031, and the RMSE of the TCC method was 0.018. We noticed that the constrained spline approach had a similar RMSE to the TCC method, whereas the linear approach on N(0,1) scale had a relatively higher RMSE.

Figure 15.6 shows the linking error for the constrained spline linking and linear linking. These linking errors were obtained from the bootstrapping method. On average, the linear approach on the N(0,1) scale had a smaller linking error than the constrained spline approach, though both were within the nominal level.

As for the linking under the TCC method, Table 15.4 summarizes the results. The linking slope is 1.004, and the intercept is -0.3708. For the purpose of comparison, the linking slope obtained from linear linking on the N(0,1) scale was 0.867, and the intercept from this approach was -0.362.

In this real data analysis, we actually employed four different linking methods. The empirical linking function was in fact derived by the equipercentile technique,

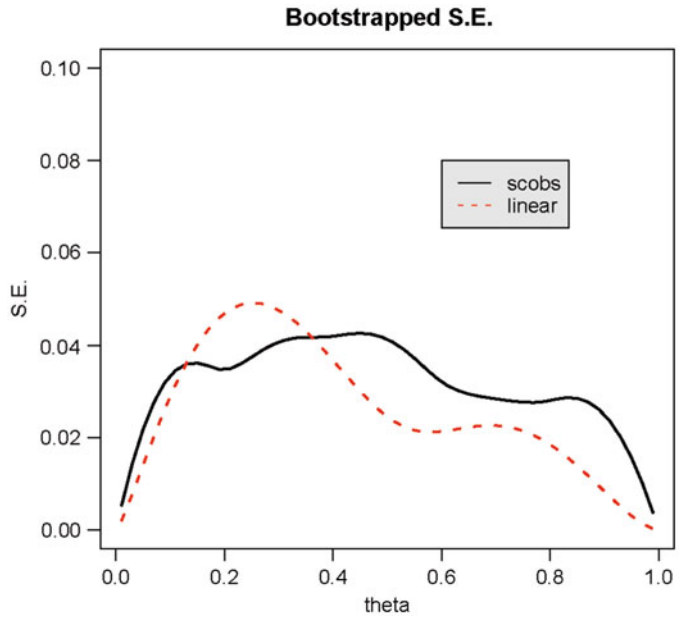

Fig. 15.6 Real data analysis: Bootstrapped *SE*

Table 15.4 Means and Standard Deviations of Item Parameter Estimates and Linking Constants for the Common Items

	Form A			Form B		
	a	b	c	a	b	c
Item *M*	2.284	−0.1316	0.1943	2.2246	0.2218	0.1872
SD	0.6073	0.3246	0.0524	0.6165	0.3205	0.0555
Linking constants	A = 1.0004			B = −0.3696		
Transformed items *M*				2.2236	−0.1477	0.1872
Transformed items *SD*				0.6163	0.3206	0.0555

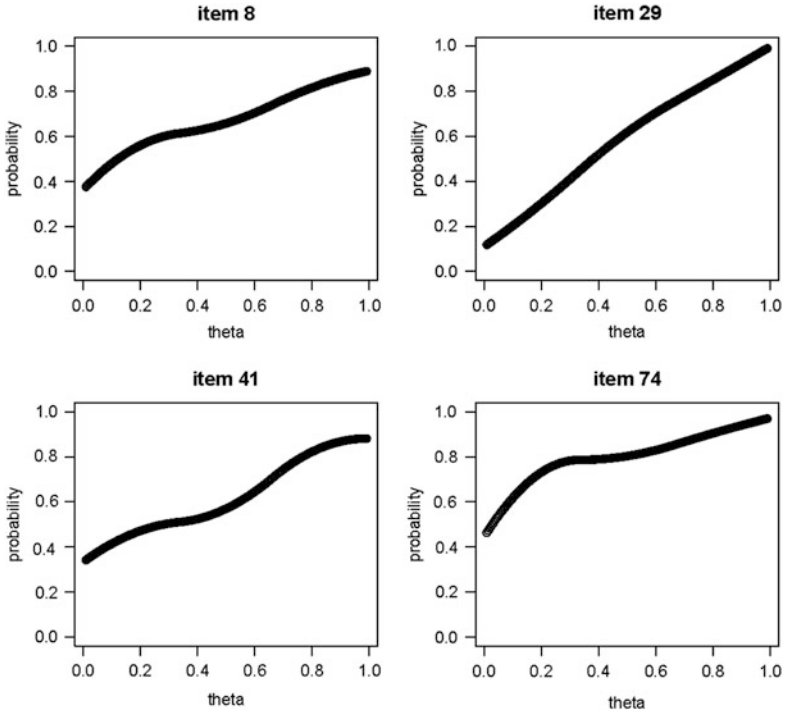

Fig. 15.7 Four nonparametric item characteristic curves

which is a core method in observed score linking. The two proposed methods and the TCC method were compared with the empirical linking function. The results in this study showed that there is not much difference among these four methods, except that the linear linking on the N(0,1) scale is slightly off. This is probably due to the good fit of the parametric model. When the parametric model is appropriate for the data, all methods will converge and show similar results.

15.6 Discussion

It is well admitted that no parametric model for item responses is perfect, even in a large-scale assessment. Figure 15.7 gives us four example items, which are taken from the data of a Psychology 101 exam at McGill University that are featured in the manual of *TestGraf* (Ramsay, 2001). Although these items are often treated as "bad" items relative to the parametric models, we still want to (and have to) include them as a part of data analysis.

Nonparametric IRT models have been developed to fit the data with more flexible functional forms. However, the nonparametric IRT models are not widely used in testing practice. One reason is little knowledge about the applications of this model, such as in linking applications, among other complications. Through the simulation study and the real data analysis, we have shown that both proposed methods are able to recover the true linking functions, even when the nonequivalent populations differ substantially. When a parametric model fits the data well, both of the proposed methods will behave similarly with traditional methods. The results of this study give us hope that we can do real applications with nonparametric IRT models.

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