Chapter 11
A Bayesian Nonparametric Model for Test Equating

George Karabatsos and Stephen G. Walker

11.1 Introduction

In observed score equating, the aim is to infer the *equating* function, $e_Y(X)$, which gives the score on test Y that is equivalent to any chosen score x on test X. Equipercentile equating is based on the premise that test scores x and y are equivalent if and only if $F_X(x) = F_Y(y)$, and therefore assumes that the equating function is defined by:

$$e_Y(x) = F_Y^{-1}(F_X(x)) = y,$$

where (F_X, F_Y) denote the cumulative distribution functions (CDFs) of the scores of test X and test Y. Of course, in order for such equating to be sensible, certain assumptions are required about the tests and the examinee populations. Also, in the practice of test equating, examinee scores on the two tests are collected according to one of the three major types of equating designs, namely, (a) the single-group design, (b) the equivalent-groups design, and (c) the nonequivalent-groups design. The single-group design may be counterbalanced, and either the equivalent-groups or the nonequivalent-groups design may make use of an internal- or external-anchor test. For more details about the aforementioned assumptions and concepts of test equating, see the textbooks by von Davier, Holland, and Thayer (2004b) and Kolen and Brennan (2004).

G. Karabatsos (✉)
College of Education, 1040 W. Harrison St. (MC 147), Chicago, IL 60607-7133, USA
e-mail: georgek@uic.edu

S.G. Walker
Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NZ, United Kingdom
e-mail: S.G.Walker@kent.ac.uk

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_11,

If the CDFs F_X and F_Y are discrete distributions, then for virtually any score x on test X, the CDF probability $F_X(x)$ does not coincide with the CDF probability $F_X(y)$ of any possible score y of test Y. Then, the equipercentile equating function is ill-defined. This poses a challenge in equipercentile equating, because in psychometric practice observed test scores are discrete. A solution to this problem is to model (F_X, F_Y) as continuous distributions and treating them as smoothed versions of discrete test score distributions (G_X, G_Y), respectively. This approach is taken by current methods of observed-score equating. In the kernel method of equating (von Davier et al., 2004b), (F_X, F_Y) are each modeled by a mixture of normal densities (the kernels) with mixing weights defined by estimates of (G_X, G_Y), and the number of mixing components (for test X and for test Y) is decided by some model-selection procedure. If the kernels are uniform distributions, then classical equating with the percentile rank method is obtained (Holland & Thayer, 1989; also see Chapter 5 of this book by Lee & von Davier). Also, the classical methods of linear equating and mean equating each provide an approach to equipercentile equating under the assumption that (F_X, F_Y) are distributions with the same shape (Karabatsos & Walker, 2009a). However, in practice, it does not seem reasonable to assume that the (continuized) population distributions of test scores are normal or come from a mixture of specific uniform distributions.

In this chapter we present a Bayesian nonparametric model for test equating, which can be applied to all the major equating designs (see Section 2.4 for details), and we illustrate the model in the analysis of two data sets. This equating model, first introduced by Karabatsos and Walker (2009a), involves the use of a bivariate Bernstein polynomial prior distribution for (F_X, F_Y) that supports the entire space of (random) continuous distributions. In particular, the model specifies (F_X, F_Y) by a mixture of beta densities via a Bernstein polynomial, where (G_X, G_Y) provide mixing weights and are modeled by a bivariate Dirichlet process prior distribution (Walker & Muliere, 2003). Also, the number of mixing components (for test X and for test Y) are modeled as random and assigned a prior distribution. Under Bayes theorem, these priors combine with the data to yield the posterior distribution of (F_X, F_Y) and of the equating function $e_Y(x) = F_Y^{-1}(F_X(x))$. As proven by Diaconis and Ylvisaker (1985), for a sufficiently large number of mixture components, a mixture of beta distributions can approximate arbitrarily well any distribution on a closed interval. For reviews of the many theoretical studies and practical applications of Bayesian nonparametrics, see, for example, Walker, Damien, Laud, and Smith (1999) and Müller and Quintana (2004). Also, see Karabatsos and Walker (2009b) for a review from the psychometric perspective.

The Bayesian nonparametric model for equating provides important advantages over the existing approaches to observed-score equating. In the approach, the Dirichlet process prior distribution can be specified to account for any dependence between (G_X, G_Y), and thus it accounts for any dependence between the continuized test score distributions (F_X, F_Y). This dependence can even be specified for the equivalent-groups design, where the other equating methods assume independence. However, independence is a questionable assumption, especially considering that the two tests to be equated are designed to measure the same construct (e.g., math

ability). Moreover, unlike the existing approaches to observed score equating, the Bayesian nonparametric model provides an approach to symmetric equating that always equates scores that fall within the correct range of test scores. Also, using the Bayesian nonparametric model, the posterior distribution of the equating function $e_Y(x) = F_Y^{-1}(F_X(x))$ provides inference of the 95% credible interval of the equated score. Thus, the Bayesian model provides a way to fully account for the uncertainty in the equated scores, for any sample size. In contrast, all the previous approaches to observed-score equating only rely on large-sample approximations to estimate the confidence interval of the equated score.

We present the Bayesian nonparametric equating model in the next section and describe the key concepts of this model, including the Dirichlet process, the bivariate Dirichlet process, the random Bernstein polynomial prior distribution, and the bivariate Bernstein prior distribution. In Section 11.3 we illustrate the Bayesian nonparametric equating model in the analysis of two data sets generated from the equivalent-groups design and the nonequivalent-groups design with internal anchor, respectively. In the first application, we compare the equating results of the Bayesian model against the results obtained by the four other approaches to observed-score equating. We conclude in Section 11.4.

11.2 Bayesian Nonparametric Equating Model

11.2.1 Dirichlet Process Prior

The Dirichlet process prior (Ferguson, 1973) is conveniently described through Sethuraman's (1994) representation, which is based on a countably infinite sampling strategy. So let θ_j, for $j = 1, 2, \ldots$, be independent and identically distributed from a fixed distribution function G_0, and let v_j, for $j = 1, 2, \ldots$, be independent and identically distributed from the Beta $(1, m)$ distribution. Then a random distribution function chosen from a Dirichlet process prior with parameters (m, G_0) can be constructed via

$$G(x) = \sum_{j=1}^{\infty} \omega_j \, 1(\theta_j \leq x),$$

where $\omega_1 = v_1$ and for $j > 1$, $\omega_j = v_j \prod_{l<j} (1 - v_l)$, and $1(\cdot)$ is the indicator function. In other words, realizations of the Dirichlet process can be represented as infinite mixtures of point masses. The locations θ_i of the point masses are a sample from G_0. It is obvious from the above construction that any random distribution G generated from a Dirichlet process prior is discrete with probability 1.

Also, for any value x from a sample space X, the random distribution (CDF) $G(x)$, modeled under the Dirichlet Process prior, has a beta distribution,

$$G(x) \sim \text{Beta}(mG_0(x), m\{1 - G_0(x)\}),$$

with prior mean $\mathrm{E}[G(x)] = G_0(x)$, and prior variance

$$\mathrm{Var}[G(x)] = \frac{G_0(x)[1 - G_0(x)]}{m+1}.$$

Hence m (the precision parameter) acts as an uncertainty parameter, increasing the variance of G as m becomes small. Given a set of data $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ having empirical distribution $\hat{G}(\cdot)$, the posterior distribution of G is also a Dirichlet process, with updated parameters given by $m \to m + n$. The posterior distribution is given by

$$G(x)|\boldsymbol{x}_n \sim \mathrm{Beta}(mG_0(x) + n\hat{G}(x), m[1 - G_0(x)] + n[1 - \hat{G}(x)]),$$

and it has mean $\mathrm{E}[G(x)|\boldsymbol{x}_n] = \frac{mG_0(x)+n\hat{G}(x)}{m+n}$. Thus, the posterior mean is a mixture of the prior guess (G_0) and the data ($\hat{G}$), with the weights of the mixture given by the precision parameter (m) and the sample size (n), respectively.

11.2.2 Random Bernstein Polynomial Prior

As mentioned earlier, the Dirichlet process prior fully supports discrete distributions. Here, a nonparametric prior is described, called the random Bernstein polynomial prior, which gives support to the entire space of continuous distributions and will provide a smooth method for equating test scores. As the name suggests, the random Bernstein polynomial prior distribution depends on the Bernstein polynomial (Lorentz, 1953). For any function G defined on [0,1] (not necessarily a distribution function) such that $G(0) = 0$, the Bernstein polynomial of order p of G is defined by

$$B(x; G, p) = \sum_{k=0}^{p} G\left(\frac{k}{p}\right) \binom{p}{k} x^k (1-x)^{p-k} \tag{11.1}$$

$$= \sum_{k=1}^{p} \left[G\left(\frac{k}{p}\right) - G\left(\frac{k-1}{p}\right) \right] \mathrm{Beta}(x|k, p-k+1) \tag{11.2}$$

$$= \sum_{k=1}^{p} w_{k,p} \mathrm{Beta}(x|k, p-k+1), \tag{11.3}$$

and it has derivative

$$f(x; G, p) = \sum_{k=1}^{p} w_{k,p} \beta(x|k, p-k+1),$$

where $\beta(\cdot|a,b)$ denotes the density corresponding to the CDF of the beta distribution, Beta(a,b). Also, $w_{k,p} = G(k/p) - G((k-1)/p),\ k = 1,\ldots,p$.

Note that if G is a CDF on [0,1], $B(x;G,p)$ is also a CDF on [0,1], corresponding to probability density function $f(x; G, p)$, defined by a mixture of p beta CDFs with mixing weights $(w_{1,p},\ldots,w_{p,p})$. Therefore, if G and p are random, then $B(x;G,p)$ is a random continuous CDF, with corresponding random probability density function $f(x;G,p)$. The random Bernstein-Dirichlet polynomial prior distribution of Petrone (1999) has G as a Dirichlet process with parameters (m,G_0), with p assigned an independent discrete prior distribution $\pi(p)$ defined on {1,2,...}. Her work extended from the results of Dalal and Hall (1983) and Diaconis and Ylvisaker (1985), who proved that, for sufficiently large p, mixtures of the form given in Equations 11.1–11.3 can approximate any CDF on [0,1], to any arbitrary degree of accuracy. Moreover, as Petrone (1999) has shown, the Bernstein polynomial prior distribution must treat p as random to guarantee that the prior supports the entire space of (Lebesgue-measurable) continuous densities on [0,1]. Suppose that a set of data $x_1,\ldots,x_n \in [0,1]$ are independent and identically distributed samples from a true density, denoted by f_0. Standard arguments of probability theory involving Bayes theorem can be used to show that the data update the Bernstein prior to yield a posterior distribution of the random density f (via the posterior distribution of (G,p)). Walker (2004, Section 6.3) proved the posterior consistency of the random Bernstein model (prior), in the sense that as $n \to \infty$, the posterior distribution of the model converges to a point mass at the true f_0. In fact (Walker, Lijoi, & Prünster, 2007), if the choice of prior distribution $\pi(p)$ satisfies $\pi(p)<\exp(-4p\log p)$, the convergence rate of the posterior matches the convergence rate of the sieve maximum likelihood estimate of f_0.

11.2.3 Dependent Bivariate Model

A model for constructing a bivariate Dirichlet process has been given in Walker and Muliere (2003). The idea is as follows: Take $G_X \sim \Pi(m, G_0)$ and then, for some fixed $r \in \{0,1,2,\ldots\}$, take $z_1,\ldots,z_r$ to be independent and identically distributed from G_X. Then take

$$G_Y \sim \Pi(m+r, (mG_0 + r\widehat{F}_r)/(m+r)),$$

where $\hat{F}_r$ is the empirical distribution of $\{z_1,\ldots, z_r\}$. Walker and Muliere (2003) showed that the marginal distribution of G_Y is $\Pi(m,G_0)$. It is possible to have the marginals from different Dirichlet processes. However, it will be assumed that the priors for the two random distributions are the same. It is also easy to show that for any measurable set A, the correlation between $G_X(A)$ and $G_Y(A)$ is given by

$$\text{Corr}(G_X(A), G_Y(A)) = r/(m+r),$$

and hence this provides an interpretation for the prior parameter r.

For modeling continuous test score distributions (F_X, F_Y), it is possible to construct a bivariate random Bernstein polynomial prior distribution on (F_X, F_Y) via the random distributions:

$$F_X(\cdot; G_X, p_X) = \sum_{k=1}^{p_X} \left[G_X\left(\frac{k}{p_X}\right) - G_X\left(\frac{k-1}{p_X}\right) \right] \text{Beta}(\cdot|k, p_X - k + 1),$$

$$F_Y(\cdot; G_Y, p_Y) = \sum_{k=1}^{p_Y} \left[G_Y\left(\frac{k}{p_Y}\right) - G_Y\left(\frac{k-1}{p_Y}\right) \right] \text{Beta}(\cdot|k, p_Y - k + 1).$$

with (G_X, G_Y) coming from the bivariate Dirichlet Process model, and with independent prior distributions $\pi(p_X)$ and $\pi(p_Y)$. Each of these random distributions is defined on (0,1]. However, without loss of generality, it is possible to model observed test scores after transforming each of them into (0,1). For example, if $x_{\min}$ and $x_{\max}$ denote the minimum and maximum possible scores on a test X, each observed test score x can be mapped into (0,1) by the equation $x' = (x - x_{\min} + \varepsilon)/(x_{\max} - x_{\min} + 2\varepsilon)$, where $\varepsilon > 0$ is a very small constant. The scores can be transformed back to their original scale by taking $X = X'(x_{\max} - x_{\min} + 2\varepsilon) + x_{\min} - \varepsilon$.

Given samples of observed scores $\boldsymbol{x}_{n(X)} = \{x_1, \ldots, x_{n(X)}\}$ and $\boldsymbol{y}_{n(Y)} = \{y_1, \ldots, y_{n(Y)}\}$ on the two tests (assumed to be mapped onto a sample space (0,1)), the random bivariate Bernstein polynomial prior combines with these data to define a joint posterior distribution, which we denote by $F_X, F_Y|\, \boldsymbol{x}_{n(X)}, y_{n(Y)}$. As proven by Walker et al. (2007), posterior consistency of the bivariate model is obtainable when the independent prior distributions $(\pi_X(p_X), \pi_Y(p_Y))$ satisfy $\pi(p_X, p_Y) \propto \exp(-4p_X \log p_X)\exp(-4p_Y \log p_Y)$. Also, this posterior consistency implies consistent estimation of the posterior distribution of the equating function $e_Y(\cdot) = F_{0Y}^{-1}(F_{0X}(\cdot))$, as desired. Karabatsos and Walker (2009b) described a Gibbs sampling algorithm that can be used to infer the posterior distribution F_X, $F_Y|\, \boldsymbol{x}_{n(X)}, y_{n(Y)}$, which is an extension of Petrone's (1999) Gibbs algorithm. We wrote a MATLAB program to implement the algorithm, and it can be obtained through correspondence with the first author.

At each iteration of this Gibbs algorithm, a current set of $\{p_X, G_X\}$ for test X and $\{p_Y, G_Y\}$ for test Y is available, from which it is possible to construct random distribution functions (F_X, F_Y) and the random equating function

$$e_Y(x) = F_Y^{-1}(F_X(x)) = y.$$

Hence, for each score x on test X, a posterior distribution for the equated score on test Y is available. A (finite-sample) 95% credible ("confidence") interval of an equated score $e_Y(x) = F_Y^{-1}(F_X(x))$ is easily obtained from the samples of posterior distribution $F_X, F_Y|\, \boldsymbol{x}_{n(X)}, y_{n(Y)}$. A point estimate of an equated score $e_Y(X)$ also can be obtained from this posterior distribution. While one conventional choice of point

estimate is given by the posterior mean of $e_Y(X)$, the posterior median point estimate of $e_Y(\cdot)$ has the advantage that it is invariant over monotone transformations. This invariance is important considering that the test scores are transformed into the (0,1) domain and back onto the original scale of the test scores.

11.2.4 Applying the Model to Different Equating Designs

The Bayesian nonparametric equating method presented in Section 11.2.3 readily applies to the equivalent-groups design with no anchor test and applies to the single-group design. However, with minor modifications, this method can be easily extended to an equivalent-groups or nonequivalent-groups design with an anchor test, or to a counterbalanced design.

For an equating design having an anchor test, it is possible to adopt the idea of chained equipercentile equating (Angoff, 1971). In particular, let $\boldsymbol{x}_{n(X)}$ and $\boldsymbol{v}_{n(V_1)}$ denote the set of scores observed from examinee Group 1 who completed test X and an anchor test V, and let $Y_{n(Y)}$ and $\boldsymbol{v}_{n(V_2)}$ denote sets of scores observed from examinee Group 2 who completed test Y and the same anchor test V. Then, under an equating design with anchor test, it is possible to infer the posterior distribution of the random equating functions $e_Y(x) = F_Y^{-1}(F_{V_2}(e_{V_1}(x)))$ and $e_{V_1}(x) = F_{V_1}^{-1}(F_{X_1}(x))$, based on samples from the posterior distributions $F_X, F_{V_1}|x_{n(X)}, v_{n(V_1)}$ and $F_Y, F_{V_2}|y_{n(Y)}, v_{n(V_2)}$, each modeled under a bivariate Bernstein prior.

For a counterbalanced design, it is possible to adopt the ideas from von Davier et al. (2004b, Section 2.3), to combine the information of the two examinee Groups 1 and 2. Specifically, the inference of the posterior distribution of the random equating function $e_Y(x) = F_Y^{-1}(F_X(x))$ is obtained by taking $F_X(\cdot) = \varpi_X F_{X_1}(\cdot) + (1 - \varpi_X)F_{X_2}(\cdot)$ and $F_Y(\cdot) = \varpi_Y F_{Y_1}(\cdot) + (1 - \varpi_Y)F_{Y_2}(\cdot)$, where $(F_{X_1}, F_{X_2}, F_{Y_1}, F_{Y_2})$ are from the posterior distributions $F_{X_1}, F_{Y_2}|x_{n(X_1)}, y_{n(Y_2)}$ and $F_{X_2}, F_{Y_1}|x_{n(X_2)}, y_{n(Y_1)}$ under two bivariate Bernstein models. Also, $0 \leq \varpi_X, \varpi_Y \leq 1$ are chosen weights, and they can be varied to determine how much they change the posterior distribution of $e_Y(\cdot)$.

11.3 Illustrations

The following two subsections illustrate the Bayesian model for the equating of test scores arising from the equivalent-groups design, the counterbalanced design, and the nonequivalent-groups design, respectively. In applying our Bayesian model to each of the three data sets, we assumed the following specification of the prior distributions. In particular, we assumed the bivariate Dirichlet process to have baseline distribution G_0 that equals the Beta(1,1) distribution, and we assumed a relatively noninformative prior by taking $m = 1$ and $r = 4$, reflecting the (rather uncertain) prior belief that the correlation of the scores between two tests is

$0.8 = r/(m + r)$. In particular, the choice of "prior sample size" of $m = 1$ leads to a posterior distribution of F_X, F_Y that is primarily determined by the observed data. Furthermore, up to a constant of proportionality, we specify an independent prior distribution of $\pi(p) \propto \exp(-4p\log p)$ for p_x and for p_y. As discussed in Section 11.2.3, this choice of prior ensures the consistency of the posterior distribution of (F_X, F_Y). Also, for each data set analyzed with the Bayesian model, we implemented the Gibbs sampling algorithm to generate 10,000 samples from the posterior distribution of (F_X, F_Y), including (P_X, P_Y), after discarding the first 2,000 Gibbs samples as burn-in. We found through separate analyses that 10,000 samples had converged to samples from the posterior distribution.

11.3.1 Equivalent-Groups Design

The Bayesian nonparametric equating model is demonstrated in the analysis of a large data set generated from an equivalent-groups design. This data set, obtained from von Davier et al. (2004b, p. 100), consists of 1,453 examinees who completed test X and 1,455 examinees completing test Y of a national mathematics exam. Each test has 20 items and is scored by number correct. The average score on test X is 10.82 (SD = 3.81), and the average score on test Y is 11.59 (SD = 3.93), and so the second test is easier than the first. For the Bayesian nonparametric model for equating, the marginal posterior distributions of P_X and of P_Y concentrated on 1 and 2, respectively. Figure 11.1 presents the posterior median estimate of the equating function under the Bayesian equating model.

Figure 11.1 also presents four more estimates of the equating functions, obtained by the kernel, percentile-rank, linear, and mean methods of equating, respectively. We use the kernel estimate that is reported in Chapter 7 of von Davier et al. (2004b). According to the figure, the Bayesian estimate differs substantially from the estimate obtained by the other four methods. This difference suggests that in the

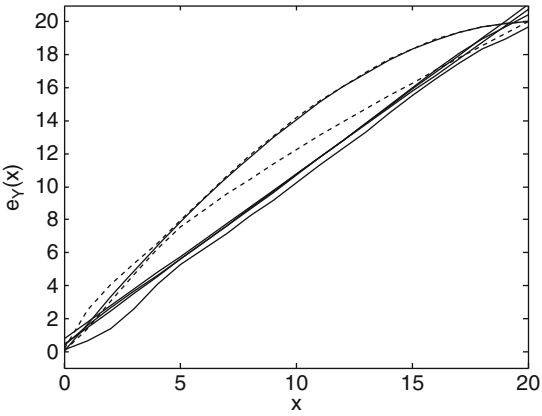

Fig. 11.1 The posterior median estimate of $e_Y(\cdot)$ given by the top solid line, enveloped by the 95% posterior credible interval (dotted lines). The other solid lines give the point-estimates of $e_Y(\cdot)$ obtained via the kernel, percentile-rank (bottom solid line), linear, and mean methods of equating

data, the test score distributions (F_X, F_Y) are correlated. In particular, the Bayesian model accounts for the dependence (correlation) between F_X and F_Y, whereas in the other four methods, they are assumed to be independent (zero correlation) under the equivalent-groups design. However, there must be correlation between the two test forms, given that the two forms were designed to measure the same construct of math ability. Moreover, Figure 11.1 shows that the kernel, linear, and mean equating methods equate some scores on test X that fall above the 0–20 range of possible scores on test Y. In contrast, the Bayesian nonparametric model equated scores on test X with scores that fall inside the range of test Y, as it will always do.

11.3.2 Nonequivalent-Groups Design and Chained Equating

In this section we apply the Bayesian nonparametric equating model to analyze a classic data set arising from a nonequivalent-groups design with internal anchor, obtained from Kolen and Brennan (2004). The first group of examinees completed test X, and the second group of examinees completed test Y, both groups being random samples from different populations. Here, test X and test Y each have 36 items and is scored by number correct, and both tests have 12 items in common. These 12 common items form an internal anchor test because they contribute to the scoring of test X and of test Y. While the two examinee groups come from different populations, the anchor test provides a way to link the two groups and the two tests. The anchor test completed by the first examinee group (population) is labeled as V_1 and the anchor test completed by the second examinee group is labeled as V_2, even though both groups completed the same anchor test. The first group of 1,655 examinees had a mean score of 15.82 (SD = 6.53) on test X, and a mean score of 5.11 (SD = 2.38) for the anchor test. The second group of examinees had a mean score of 18.67 (SD = 6.88) on test Y and a mean score of 5.86 (SD = 2.45) on the anchor test.

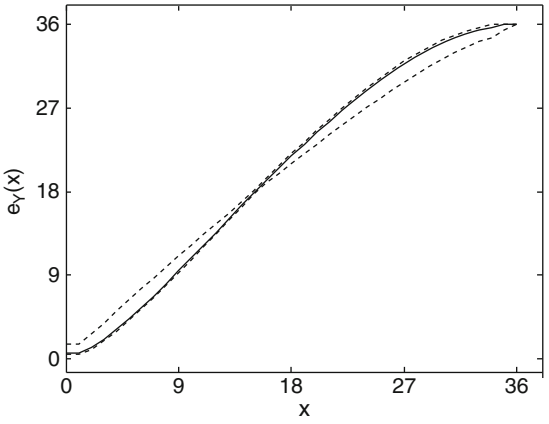

Fig. 11.2 The posterior median estimate of $e_Y(\cdot)$ (solid line), enveloped by the 95% posterior credible interval (dotted lines)

In the analysis of these data from the nonequivalent-groups design, chained equipercentile equating was used with the Bayesian nonparametric model, as described in Section 11.2.4. The marginal posterior distribution of $p_{X_1}, p_{V_1}, p_{V_2}$, and p_{Y_2} concentrated on values of 6, 1, 3, and 5 respectively. Figure 11.2 presents the posterior median estimate of the equating function estimate, along with the corresponding 95% confidence interval from the posterior distribution.

11.4 Conclusions

This study introduced a Bayesian nonparametric model for test equating. It is defined by a bivariate Bernstein polynomial prior distribution for (F_X, F_Y) that supports the entire space of (random) continuous distributions, with this prior depending on the bivariate Dirichlet process. The Bayesian equating model has important theoretical and practical advantages over all the previous approaches to observed score equating. A key advantage of the Bayesian equating model is that in equivalent-groups designs, it accounts for the realistic situation that the two distributions of test scores (F_X, F_Y) are correlated, instead of independent, as is often assumed in the previous methods of observed score equating. This dependence seems reasonable, considering that in practice, the two tests that are to be equated are designed to measure the same psychological construct (e.g., ability in some math domain). We also note that the Bayesian model provides a method of symmetric equating which yields equated scores within the range of test scores, something which could not be said about the other methods of observed score equating. Finally, through the posterior distribution, the Bayesian model provides a 95% credible interval for the equated score. Thus, unlike previous approaches to observed score equating, the Bayesian model fully accounts for uncertainly in the equating score, for any given sample size.