

## A BAYESIAN NONPARAMETRIC APPROACH TO TEST EQUATING

GEORGE KARABATSOS

UNIVERSITY OF ILLINOIS-CHICAGO

STEPHEN G. WALKER

UNIVERSITY OF KENT

A Bayesian nonparametric model is introduced for score equating. It is applicable to all major equating designs, and has advantages over previous equating models. Unlike the previous models, the Bayesian model accounts for positive dependence between distributions of scores from two tests. The Bayesian model and the previous equating models are compared through the analysis of data sets famous in the equating literature. Also, the classical percentile-rank, linear, and mean equating models are each proven to be a special case of a Bayesian model under a highly-informative choice of prior distribution.

Key words: Bayesian nonparametrics, bivariate Bernstein polynomial prior, Dirichlet process prior, test equating, equipercentile equating, linear equating.

### 1. Introduction

Often, in psychometric applications, two (or more) different tests of the same trait are administered to examinees. The trait may refer to, for example, ability in an area of math, verbal ability, quality of health, quality of health care received, and so forth. Different tests are administered to examinees for a number of reasons. For example, to address any concerns about the security of the content of the test items, to address time efficiency in the examination process, or the two different tests may have identical items administered at different time points. Under such scenarios, the two tests, label them Test  $X$  and Test  $Y$ , may have different numbers of test items, may not have any test items in common, and possibly, each examinee completes only one of the tests. Test equating makes it possible to compare examinees' scores on a common frame of reference, when they take different tests. The goal of test equating is to infer the "equating" function,  $e_Y(x)$ , which states the score on Test  $Y$  that is equivalent to a chosen score  $x$  on Test  $X$ , for all possible scores  $x$  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 thus defines the equating function:

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

where  $(F_X(\cdot), F_Y(\cdot))$  are the cumulative distribution functions (c.d.f.s) of the scores of Test  $X$  and Test  $Y$ . There are basic requirements of test equating that are commonly accepted (Kolen & Brennan, 2004, Section 1.3; Von Davier, Holland, & Thayer 2004, Section 1.1). They include the equal construct requirement (Test  $X$  and Test  $Y$  measure the same trait), the equal reliability requirement (Test  $X$  and Test  $Y$  have the same reliability), the symmetry requirement ( $e_X(e_Y(x)) = x$  for all possible scores  $x$  of Test  $X$ ), the equity requirement (it should be a matter of indifference for an examinee to be tested by either Test  $X$  or Test  $Y$ ), and the population invariance requirement ( $e_Y(x)$  is invariant with respect to any chosen subgroup in the population

Requests for reprints should be sent to George Karabatsos, College of Education, University of Illinois-Chicago, 1040 W. Harrison St. (MC 147), Chicago, IL 60607, USA. E-mail: [georgek@uic.edu](mailto:georgek@uic.edu)

PSYCHOMETRIKA

51 of examinees). Also, we believe that it is reasonable for test equating to satisfy the range require-  
 52 ment, that is, for all possible scores  $x$  of Test  $X$ ,  $e_Y(x)$  should fall in the range of possible scores  
 53 in Test  $Y$ .

54 Also, in the practice of test equating, examinee scores on the two tests are collected accord-  
 55 ing to one of the three major types of equating designs, with each type having different versions  
 56 (von Davier et al., 2004, Chap. 2; Kolen & Brennan, 2004, Section 1.4). The first is the Sin-  
 57 gle Group (SG) design where a random sample of common population of examinees complete  
 58 both Test  $X$  and Test  $Y$ , the second is an Equivalent Groups (EG) design where two independent  
 59 random samples of the same examinee population complete Test  $X$  and Test  $Y$ , respectively,  
 60 and the third is the nonequivalent groups (NG) design where random samples from two differ-  
 61 ent examinee populations complete Test  $X$  and Test  $Y$ , respectively. A SG design is said to be  
 62 counterbalanced (a CB design) when one examinee subgroup completes Test  $X$  first and the re-  
 63 maining examinees Test  $Y$  first. NG designs, and some EG designs, make use of an anchor test  
 64 consisting of items appearing in both Test  $X$  and Test  $Y$ . The anchor test, call it Test  $V$  (with  
 65 possible scores  $v_1, v_2, \dots$ ), is said to be internal when the anchor items contribute to the scores  
 66 in Test  $X$  and in Test  $Y$ , and is said to be external when they do not contribute.

67 If the c.d.f.s  $F_X(\cdot)$  and  $F_Y(\cdot)$  are discrete, then often  $F_X(x)$  will not coincide with  $F_Y(y)$   
 68 for any possible score on Test  $Y$ , in which case the equipercentile equating function in (1)  
 69 is ill-defined. This poses a challenge in equipercentile equating, because often in psychome-  
 70 tric practice, test scores are discrete. Even when test scores are on a continuous scale, the  
 71 empirical distribution estimates of  $(F_X, F_Y)$  are still discrete, with these estimates given by  
 72  $\hat{F}_X(x) = \frac{1}{n(X)} \sum \mathbf{1}(x_i \leq x)$  and  $\hat{F}_Y(y) = \frac{1}{n(Y)} \sum \mathbf{1}(y_i \leq y)$ , with  $\mathbf{1}(\cdot)$  the indicator function. A  
 73 solution to this problem is to model  $(F_X, F_Y)$  as continuous distributions, being smoothed ver-  
 74 sions of discrete test score c.d.f.s  $(G_X, G_Y)$ , respectively (with corresponding probability mass  
 75 functions  $(g_X, g_Y)$ ). This approach is taken by the four well-known methods of observed-score  
 76 equating. They include the traditional methods of percentile-rank equating, linear equating, mean  
 77 equating (e.g., Kolen & Brennan, 2004), and the kernel method of equating (von Davier et al.,  
 78 2004). While in the equating literature they have been presented as rather distinct methods, each  
 79 of these methods corresponds to a particular mixture model for  $(F_X, F_Y)$  having the common  
 80 form:

$$81 \quad F_X(\cdot) = \sum_{k=1}^{p(X)} w_{k,p(X)} F_{Xk}(\cdot); \quad F_Y(\cdot) = \sum_{k=1}^{p(Y)} w_{k,p(Y)} F_{Yk}(\cdot),$$

85 where for each Test  $X$  and Test  $Y$ , the  $F_k(\cdot)$  are continuous c.d.f.s and  $\mathbf{w}_p = (w_{1,p}, \dots, w_{p,p})$ ,  
 86  $\sum_{k=1}^p w_{k,p} = 1$ , are mixture weights defined by a discrete test score distribution  $G$ . For ex-  
 87 ample, the linear equating method corresponds to the simple mixture model defined by  $p(X) =$   
 88  $p(Y) = 1$ , with  $F_X(\cdot) = \text{Normal}(\cdot | \mu_X, \sigma_X^2)$  and  $F_Y(\cdot) = \text{Normal}(y | \mu_Y, \sigma_Y^2)$ , and the mean equat-  
 89 ing method corresponds to the same model with the further assumption that  $\sigma_X^2 = \sigma_Y^2$ . While the  
 90 equating function for linear or mean equating is usually presented in the form  $e_Y(x) = \frac{\sigma_Y}{\sigma_X}(x -$   
 91  $\mu_X) + \mu_Y$ , this equating function coincides with  $e_Y(x) = F_Y^{-1}(F_X(x))$ , when  $F_X(\cdot)$  and  $F_Y(\cdot)$   
 92 are normal c.d.f.s with parameters  $(\mu_X, \sigma_X^2)$  and  $(\mu_Y, \sigma_Y^2)$ , respectively. Now, let  $\{x_{k-1}^* < x_k^*, k =$   
 93  $2, \dots, p(X)\}$  be the possible scores in Test  $X$  assumed to be consecutive integers, with  $x_0^* =$   
 94  $x_1^* - \frac{1}{2}$ ,  $x_{p(X)+1}^* = x_{p(X)}^* + \frac{1}{2}$ , and  $\{y_{k-1}^* < y_k^*, k = 2, \dots, p(Y)\}$ ,  $y_0^*$ , and  $y_{p(Y)+1}^*$  are similarly de-  
 95 fined for Test  $Y$ . The percentile-rank method corresponds to a model defined by a mixture of uni-  
 96 form c.d.f.s, with  $F_{Xk}(\cdot) = \text{Uniform}(\cdot | x_{k-1}^* - \frac{1}{2}, x_k^* + \frac{1}{2})$  and  $w_{k,p(X)} = g_X(x_k^*)$  ( $k = 1, \dots, p(X)$ ),  
 97 along with  $F_{Yk}(\cdot) = \text{Uniform}(\cdot | y_{k-1}^* - \frac{1}{2}, y_k^* + \frac{1}{2})$  and  $w_{k,p(Y)} = g_Y(y_k^*)$  ( $k = 1, \dots, p(Y)$ ) (for  
 98 a slightly different view, see Holland and Thayer 2000). Finally, the kernel method corre-  
 99 sponds to a model defined by a mixture of normal c.d.f.s, with  $(p_X, p_Y)$  defined similarly,

GEORGE KARABATSOS AND STEPHEN G. WALKER

101  $F_{Xk}(\cdot; h_X) = \text{Normal}(R_{Xk}(\cdot; h_X)|0, 1)$  and  $F_{Yk}(\cdot; h_Y) = \text{Normal}(R_{Yk}(\cdot; h_Y)|0, 1)$ , where  $h_X$   
 102 and  $h_Y$  are fixed bandwidths. (In particular,  $R_{Xk}(x; h_X) = [x - x_k a_X - \mu_X(1 - a_X)](h_X a_X)^{-1}$ ,  
 103 with  $a_X^2 = [\sigma_X^2 / (\sigma_X^2 + h_X^2)]^{1/2}$ . The function  $R_{Yk}(y; h_Y)$  is similarly defined). In the kernel  
 104 model, the mixing weights are defined by  $(G_X, G_Y)$  with  $\{w_{k,p(X)} = g_X(x_k^*) : k = 1, \dots, p(X)\}$   
 105 and  $\{w_{k,p(Y)} = g_Y(y_k^*) : k = 1, \dots, p(Y)\}$ , where  $(G_X, G_Y)$  are marginals of the bivariate dis-  
 106 tribution  $G_{XY}$  assumed to follow a log-linear model. In the kernel equating model, the possi-  
 107 ble scores of each test (i.e., all the  $x_k^*$  and all the  $y_k^*$ ) need not be represented as consecu-  
 108 tive integers. In general, for all these models and their variants, it is assumed that the contin-  
 109 uous distributions  $(F_X, F_Y)$  are governed by some chosen function  $\varphi$  of a finite-dimensional  
 110 parameter vector  $\theta$  (Kolen & Brennan, 2004; von Davier et al., 2004). Also, in practice, a  
 111 point-estimate of the equating function at  $x$  is obtained by  $e_Y(x; \hat{\theta}) = F_Y^{-1}(F_X(x; \hat{\theta})|\hat{\theta})$  via  
 112 the point-estimate  $\hat{\theta}$ , usually a maximum-likelihood estimate (MLE) (Kolen & Brennan, 2004;  
 113 von Davier et al., 2004). Asymptotic (large-sample) standard errors and confidence intervals for  
 114 any given equated score  $e_Y(x; \theta)$  (for all values of  $x$ ) can be obtained with either analytic or  
 115 bootstrap methods (Kolen & Brennan, 2004; von Davier et al., 2004). In the kernel equating  
 116 model, given the MLE  $\hat{G}_{XY}$  obtained from a selected log-linear model, the marginals  $(\hat{G}_X, \hat{G}_Y)$   
 117 are derived, and then the bandwidth parameter estimates  $(\hat{h}_X, \hat{h}_Y)$  are obtained by minimizing a  
 118 squared-error loss criterion  $(\hat{h} = \arg \min_{h>0} \sum_{k=1}^p (\hat{w}_{k,p} - \hat{f}(x_k^*, h))^2)$ .

119 The percentile-rank, linear, mean, and kernel equating models each have seen many suc-  
 120 cessful applications in equating. However, for five reasons, they are not fully satisfactory, and  
 121 these reasons invite the development of a new model of test equating. First, under any of these  
 122 four equating models, the mixture distributions  $(F_X, F_Y)$  support values outside the range of  
 123 scores of Test  $X$  and/or of Test  $Y$ , the ranges given by  $[x_1^*, x_{p(X)}^*]$  and  $[y_1^*, y_{p(Y)}^*]$ , respec-  
 124 tively. The kernel, linear, and mean equating models each treat  $(F_X, F_Y)$  as normal distribu-  
 125 tions on  $\mathbb{R}^2$ , and the percentile-rank model treats  $(F_X, F_Y)$  as continuous distributions on  
 126  $[x_0^*, x_{p(X)+1}^*] \times [y_0^*, y_{p(Y)+1}^*]$ . As a consequence, under either of these four models, the estimate  
 127 of the equating function  $e_Y(\cdot; \hat{\theta}) = F_Y^{-1}(F_X(\cdot; \hat{\theta})|\hat{\theta})$  can equate scores on Test  $X$  with scores on  
 128 Test  $Y$  that fall outside the correct range  $[y_1^*, y_{p(Y)}^*]$ . While it is tempting to view this as a minor  
 129 problem which only affects the equating of Test  $Y$  scores around the boundaries of  $[y_1^*, y_{p(Y)}^*]$ ,  
 130 the extra probability mass that each of these models assign to values outside of the sample space  
 131  $[x_1^*, x_{p(X)}^*] \times [y_1^*, y_{p(Y)}^*]$  can also negatively impact the equating of scores that lie in the middle  
 132 of the  $[y_1, y_{p(Y)}]$  range. Second, the standard errors and confidence intervals of equated scores  
 133 are asymptotic, and thus are valid for only very large samples (e.g., von Davier et al., 2004,  
 134 pp. 68–69). Third, the kernel and percentile-rank models do not guarantee symmetric equating.  
 135 Fourth, the percentile-rank, linear, and mean equating models each make overly-restrictive as-  
 136 sumptions about the shape of  $(F_X, F_Y)$ . There is no compelling reason to believe that in practice,  
 137 (continuized) test scores are truly normally distributed or truly a mixture of specific uniform  
 138 distributions. Fifth, each of the four equating models carry the assumption that the test score dis-  
 139 tributions  $(F_X, F_Y)$  are independent (e.g., von Davier et al., 2004, Assumption 3.1, p. 49). This  
 140 is not a realistic assumption in practice, since the two tests to be equated are designed to measure  
 141 the same trait, under the “equal construct” requirement of equating mentioned earlier. The equal  
 142 construct requirement implies the prior belief that  $(F_X, F_Y)$  are highly correlated (dependent),  
 143 i.e., that  $(F_X, F_Y)$  have a similar shapes, in the sense that the shape of  $F_X$  provides information  
 144 about the shape of  $F_Y$ , and vice versa. A correlation of 1 represents the extreme case of depen-  
 145 dence, where  $F_X = F_Y$ . A correlation of 0 represents the other extreme case of independence,  
 146 where the shape of  $F_X$  provides absolutely no information about the shape of  $F_Y$ , and vice versa.  
 147 However, the assumption is not strictly true in practice and is not warranted in general.

148 In this paper, a novel Bayesian nonparametric model for test equating is introduced (Kara-  
 149 batsos & Walker, 2007), which address all five issues of the previous equating models, and can  
 150 be applied to all the equating designs. Suppose with no loss of generality that the test scores are

151 mapped into the interval  $[0, 1]$ , so that each test score can be interpreted as a “proportion-correct”  
 152 score (a simple back-transformation gives scores on the original scale). In the Bayesian equating  
 153 model, continuous test score distributions  $(F_X, F_Y)$  are modeled nonparametrically and as de-  
 154 pendent, through the specification of a novel, bivariate Bernstein polynomial prior distribution.  
 155 This prior supports the entire space  $\{F_{XY}\}$  of continuous (measurable) distributions on  $[0, 1]^2$   
 156 (with respect to the Lebesgue measure), where each  $F_{XY}$  corresponds to univariate marginals  
 157  $(F_X, F_Y)$ . The bivariate Bernstein prior distribution is a very flexible nonparametric model,  
 158 which defines a (random) mixture of Beta distributions (c.d.f.s) for each marginal  $(F_X, F_Y)$ .  
 159 In particular, the  $(p(X), p(Y))$  are random and assigned an independent prior distribution. Also,  
 160 the vectors of mixing weights  $\{\mathbf{w}_{p(X)}, \mathbf{w}_{p(Y)}\}$  are random and defined by discrete score distribu-  
 161 tions  $(G_X, G_Y)$  which themselves are modeled nonparametrically and as dependent by a bivari-  
 162 ate Dirichlet Process (Walker & Muliere, 2003). The Dirichlet process modeling of dependence  
 163 between  $(G_X, G_Y)$  induces the modeling of dependence between  $(F_X, F_Y)$ . Under Bayes’ the-  
 164 ore, the bivariate Bernstein prior distribution combines with the data (the observed scores on  
 165 Tests  $X$  and  $Y$ ) to yield a posterior distribution of the random continuous distributions  $(F_X, F_Y)$ .  
 166 For every sample of  $(F_X, F_Y)$  from the posterior distribution, the equating function is simply  
 167 obtained by  $e_Y(\cdot) = F_Y^{-1}(F_X(\cdot))$ , yielding a posterior distribution of  $e_Y(\cdot)$ . It is obvious that  
 168 every posterior sample of  $(F_X, F_Y)$ , the equating function  $e_Y(\cdot) = F_Y^{-1}(F_X(\cdot))$  is symmetric and  
 169 equates scores on Test  $X$  with a score that always falls in the  $[0, 1]$  range of scores on Test  $Y$ .  
 170 Also, the posterior distribution of the equating function easily provides finite-sample confidence  
 171 interval estimates of the equated scores, and fully accounts for the uncertainty in all the paramet-  
 172 ers of the bivariate Bernstein model. Furthermore, the Bayesian nonparametric method of test  
 173 equating can be applied to all major types of data collection designs for equating, with no special  
 174 extra effort (see Section 2.3). The bivariate Bernstein polynomial prior distribution is just one  
 175 example of a nonparametric prior distribution arising from the field of Bayesian nonparametrics.  
 176 For reviews of the many theoretical studies and practical applications of Bayesian nonparamet-  
 177 rics, see, for example, Walker, Damien, Laud, and Smith (1999) and Müller and Quintana (2004).  
 178 Also, see Karabatsos and Walker (2009) for a review from the psychometric perspective.

179 The Bayesian nonparametric equating model is presented in the next section, including the  
 180 Dirichlet process, the bivariate Dirichlet process, the random Bernstein polynomial prior distri-  
 181 bution, and the bivariate Bernstein prior distribution. It is proven that the percentile-rank, lin-  
 182 ear, and mean equating models are special cases of the Bayesian nonparametric model, under  
 183 highly informative choices of prior distribution for  $(F_X, F_Y)$ . Also, it is shown that under rea-  
 184 sonable conditions the Bayesian model guarantees consistent estimation of the true marginal  
 185 distributions  $(F_X, F_Y)$ , and as a consequence, guarantees consistent estimation of the true equat-  
 186 ing function  $e_Y(\cdot)$ . Furthermore, a Gibbs sampling algorithm is described, which provides a  
 187 means to infer the posterior distribution of the bivariate Bernstein polynomial model. Section 3  
 188 illustrates the Bayesian nonparametric equating model in the analysis of three data sets gen-  
 189 erated from the equivalent groups design, the counterbalanced design, and the nonequivalent  
 190 groups design with internal anchor, respectively. These data sets are classic examples of these  
 191 equating designs, and are obtained from modern textbooks on equating (von Davier et al., 2004;  
 192 Kolen & Brennan, 2004). The equating results of the Bayesian nonparametric model are com-  
 193 pared against the equating results of the kernel, percentile-rank, linear, and mean equating mod-  
 194 els. Finally, Section 4 ends with some conclusions about the Bayesian equating method.

## 196 2. Bayesian Nonparametric Test Equating

### 197 2.1. Dirichlet Process Prior

199 Any prior distribution generates random distribution functions. A parametric model gener-  
 200 ates a random parameter which then fits into a family of distributions, while a nonparametric prior

201 generates random distribution functions which cannot be represented by a finite-dimensional pa-  
 202 rameter. The Dirichlet process prior, which was first introduced by Ferguson (1973), is conveni-  
 203 nently described through Sethuraman’s (1994) representation, which is based on a countably-  
 204 infinite sampling strategy. So, let  $\theta_j$ , for  $j = 1, 2, \dots$ , be independent and identically distrib-  
 205 uted (i.i.d.) from a fixed distribution function  $G_0$ , and let  $v_j$ , for  $j = 1, 2, \dots$ , be independent  
 206 and identically distributed from the Beta(1,  $m$ ) distribution. Then a random distribution function  
 207 chosen from a Dirichlet process prior with parameters  $(m, G_0)$  can be constructed via

208  
209  
210  
211

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

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

217 Also, for any measurable subset  $A$  of a sample space  $\mathcal{X}$ ,

218  
219  
220

$$F(A) \sim \text{Beta}(mG_0(A), m\{1 - G_0(A)\})$$

221 with prior mean  $E[F(A)] = G_0(A)$ , and prior variance

222  
223  
224

$$\text{Var}[F(A)] = \frac{G_0(A)[1 - G_0(A)]}{m + 1}.$$

225 Hence,  $m$  acts as an uncertainty parameter, increasing the variance as  $m$  becomes small. The  
 226 parameter  $m$  is known as a precision parameter, and is often referred to as the “prior sample  
 227 size.” It reflects the prior degree of belief that the chosen baseline distribution  $G_0$  represents  
 228 the true distribution. An alternate representation of the Dirichlet process involves the Dirichlet  
 229 distribution. That is,  $F$  is said to arise from a Dirichlet process with parameters  $m$  and  $G_0$  if  
 230 for every possible partition  $A_1, \dots, A_p$  of the sample space,  $F(A_1), \dots, F(A_p)$  is distributed as  
 231 Dirichlet  $(mG_0(A_1), \dots, mG_0(A_p))$ .

232 With the Dirichlet process being a conjugate prior, given a set of data  $\mathbf{x}_n = \{x_1, \dots, x_n\}$   
 233 with empirical distribution  $\widehat{F}(x)$ , the posterior distribution of  $F$  is also a Dirichlet process, with  
 234 updated parameters given by  $m \rightarrow m + n$ , and

235  
236  
237

$$F(A)|\mathbf{x}_n \sim \text{Beta}(mG_0(A) + n\widehat{F}(A), m[1 - G_0(A)] + n[1 - \widehat{F}(A)])$$

238 for any measurable subset  $A$  of a sample space  $\mathcal{X}$ . It follows that the posterior distribution of  $F$   
 239 under a Dirichlet process can be represented by:

240  
241

$$F(A_1), \dots, F(A_p)|\mathbf{x}_n \sim \text{Dirichlet}(mG_0(A_1) + n\widehat{F}(A_1), \dots, mG_0(A_p) + n\widehat{F}(A_p)),$$

242 for every measurable partition  $A_1, \dots, A_p$  of a sample space  $\mathcal{X}$ . The posterior mean under the  
 243 Dirichlet process posterior is given by

244  
245  
246  
247

$$F_n(x) = E[F(x)|\mathbf{x}_n] = \int F(x)\Pi_n(dF) = \frac{mG_0(x) + n\widehat{F}(x)}{m + n}.$$

248 In the equation above,  $\Pi_n$  denotes the Dirichlet process posterior distribution over the space of  
 249 sampling distributions  $\{F\}$  defined on a sample space  $\mathcal{X}$ . Hence, the Bayes estimate, the posterior  
 250 mean, is a simple mixture of the data, via the empirical distribution function and the prior mean,

251  $G_0$ . As seen in the above equation,  $F_n$  is the posterior expectation and is thus the optimal point-  
 252 estimate of the true sampling distribution function (of the data), under squared-error loss.

253 In general, if  $F_X$  is modeled with a Dirichlet process with parameters  $(m(X), G_{0X})$ , and  $F_Y$   
 254 is modeled with parameters  $(m(Y), G_{0Y})$ , then given data  $\mathbf{x}_{n(X)} = \{x_1, \dots, x_{n(X)}\}$  and  $\mathbf{y}_{n(Y)} =$   
 255  $\{y_1, \dots, y_{n(Y)}\}$ ,

$$256 \begin{aligned} 257 P(F_Y^{-1}(F_X(x)) > y | F_X, \mathbf{x}_{n(X)}, \mathbf{y}_{n(Y)}) \\ 258 = \text{Beta}(F_X(x); (m(Y)G_{0Y} + n(Y)\widehat{F}_Y)(y), (m(Y)[1 - G_{0Y}] + n(Y)[1 - \widehat{F}_Y])(y)), \end{aligned}$$

259 where  $n(Y)$  is the number of observations on test  $Y$ , and  $\text{Beta}(t; \cdot, \cdot)$  denotes the c.d.f. of a beta  
 260 distribution. See Hjort and Petrone (2007) for this result. From this, the density function for  
 261  $F_Y^{-1}(F_X(x))$  is available (see (4) in Hjort and Petrone 2007), and so an alternative score for Test  
 262  $Y$  which corresponds to the score of  $x$  for Test  $X$  can be the mean of this density. A sampling  
 263 approach to evaluating this is as follows: take  $F_X(x)$  from the beta distribution with parameters

$$264 (m(X)G_{0X}(x) + n(X)\widehat{F}_X(x), m(X)[1 - G_{0X}(x)] + n(X)[1 - \widehat{F}_X(x)])$$

265 and then take  $y = y_{(i)}$  with probability

$$266 (n(Y) + m(Y))^{-1} \beta(F_X(x); m(Y)G_{0Y}(y_{(i)}) + i, m(Y)[1 - G_{0Y}(y_{(i)})] + n(Y) - i + 1),$$

267 where  $\beta(\cdot, \cdot)$  denotes the density function of the beta distribution (see Hjort & Petrone, 2007).  
 268 Here,  $y_{(1)} < \dots < y_{(n)}$  are the ordered observations and assumed to be distinct.

269 This Dirichlet process model assumes  $(F_X, F_Y)$  are independent, and for convenience, this  
 270 model is referred to as the Independent Dirichlet Process (IDP) model. As proven in Appendix I,  
 271 the linear equating, mean equating, and percentile-rank models are each a special case of the  
 272 IDP model for a very highly-informative choice of prior distribution. This highly-informative  
 273 choice of prior distribution is defined by precision parameters  $m(X), m(Y) \rightarrow \infty$  which lead to  
 274 an IDP model that gives full support to baseline distributions  $(G_{0X}, G_{0Y})$  that define the given  
 275 (linear, mean, or percentile-rank) equating model (see Section 1). Moreover, while the kernel  
 276 equating model cannot apparently be characterized as a special case of the IDP, like the mean,  
 277 linear, and percentile-rank equating models, it carries the assumption that the true  $(F_X, F_Y)$  are  
 278 independent. However, as we explained in Section 1, there is no compelling reason to believe  
 279 that real test data are consistent with the assumption of independence or, for example, that the  
 280 test score distributions are symmetric.

281 The next subsection describes how to model  $(F_X, F_Y)$  as continuous distributions using  
 282 Bernstein polynomials. The subsection that follows describes the bivariate Bernstein polyno-  
 283 mial prior that allows the modeling of dependence between  $(F_X, F_Y)$  via the bivariate Dirichlet  
 284 process.

## 2.2. Random Bernstein Polynomial Prior

285 As mentioned before, the Dirichlet process prior fully supports discrete distributions. Here,  
 286 a nonparametric prior is described, which gives support to the space of continuous distributions,  
 287 and which leads to a smooth method for equating test scores. As the name suggests, the random  
 288 Bernstein polynomial prior distribution depends on the Bernstein polynomial (Lorentz, 1953).  
 289 For any function  $G$  defined on  $[0, 1]$  (not necessarily a distribution function), such that  $G(0) = 0,$

301 the Bernstein polynomial of order  $p$  of  $G$  is defined by

302  
303  
304 
$$B(x; G, p) = \sum_{k=0}^p G\left(\frac{k}{p}\right) \binom{p}{k} x^k (1-x)^{p-k} \tag{2}$$

305  
306  
307 
$$= \sum_{k=1}^p \left[ G\left(\frac{k}{p}\right) - G\left(\frac{k-1}{p}\right) \right] \text{Beta}(x|k, p-k+1) \tag{3}$$

308  
309  
310 
$$= \sum_{k=1}^p w_{k,p} \text{Beta}(x|k, p-k+1), \tag{4}$$

311  
312 and it has derivative:

313  
314 
$$f(x; G, p) = \sum_{k=1}^p w_{k,p} \beta(x|k, p-k+1).$$

315  
316 Here,  $w_{k,p} = G(k/p) - G((k-1)/p)$  ( $k = 1, \dots, p$ ), and  $\beta(\cdot|a, b)$  denotes the density of the Beta( $a, b$ ) distribution, with c.d.f. denoted by  $\text{Beta}(\cdot|a, b)$ .

317 Note that if  $G$  is a c.d.f. on  $[0, 1]$ ,  $B(x; G, p)$  is also a c.d.f. on  $[0, 1]$  with probability density  
318 function (p.d.f.)  $f(x; k, G)$ , defined by a mixture of  $p$  beta c.d.f.s with mixing weights  $\mathbf{w}_p =$   
319  $(w_{1,p}, \dots, w_{p,p})$ , respectively. Therefore, if  $G$  and  $p$  are random, then  $B(x; G, p)$  is a random  
320 continuous c.d.f., with corresponding random p.d.f.  $f(x; G, p)$ . The random Bernstein–Dirichlet  
321 polynomial prior distribution of Petrone (1999) has  $G$  as a Dirichlet process with parameters  
322  $(m, G_0)$ , with  $p$  assigned an independent discrete prior distribution  $\pi(p)$  defined on  $\{1, 2, \dots\}$ .  
323 Her work extended from the results of Dalal and Hall (1983) and Diaconis and Ylvisaker (1985)  
324 who proved that for sufficiently large  $p$ , mixtures of the form given in (2) can approximate  
325 any c.d.f. on  $[0, 1]$  to any arbitrary degree of accuracy. Moreover, as Petrone (1999) has shown,  
326 the Bernstein polynomial prior distribution must treat  $p$  as random to guarantee that the prior  
327 supports the entire space of continuous densities with domain  $[0, 1]$ . This space is denoted by  
328  $\Omega = \{f\}$ , and all densities in  $\Omega$  are defined with respect to the Lebesgue measure.

329 We can elaborate further: A set of data  $x_1, \dots, x_n \in [0, 1]$  are i.i.d. samples from a true  
330 density, denoted by  $f_0$ , where  $f_0$  can be any member of  $\Omega$ . With the true density unknown in  
331 practice, the Bayesian assigns a prior distribution  $\Pi$  on  $\Omega$  and, for example,  $\Pi$  could be chosen  
332 as the random Bernstein polynomial prior distribution defined on  $\Omega$ . Under Bayes’ theorem, this  
333 prior combines with a set of data  $x_1, \dots, x_n \in [0, 1]$  to define a posterior distribution  $\Pi_n$ , which  
334 assigns mass:

335  
336 
$$\Pi_n(A) = \frac{\int_A \prod_{i=1}^n f(x_i) \Pi(df)}{\int_{\Omega} \prod_{i=1}^n f(x_i) \Pi(df)}$$

337 to any given subset of densities  $A \subseteq \Omega$ . Let  $f_0$  denote the true density of the data, which can be  
338 any member of  $\Omega$ . As proved by Walker (2004, Section 6.3), the random Bernstein prior satisfies  
339 strong (Hellinger) posterior consistency, in the sense that  $\Pi_n(A_\epsilon)$  converges to zero as the sample  
340 size  $n$  increases, for all  $\epsilon > 0$ , where:

341  
342 
$$A_\epsilon = \{f \in \Omega : H(f, f_0) > \epsilon\}$$

343 is a Hellinger neighborhood around the true  $f_0$ , and  $H(f, f_0) = \{\int (\sqrt{f(x)} - \sqrt{f_0(x)})^2 dx\}^{1/2}$   
344 is the Hellinger distance. This is because the random Bernstein prior distribution,  $\Pi$ , satisfies  
345 two conditions which are jointly sufficient for this posterior consistency (Walker, 2004,  
346 Theorem 4). First, the Bernstein prior satisfies the Kullback–Leibler property, that is,  $\Pi(\{f :$

PSYCHOMETRIKA

351  $D(f, f_0) < \epsilon\}) > 0$  for all  $f_0 \in \Omega$  and all  $\epsilon > 0$ , where  $D(f, f_0) = \int \log[f_0(x)/f(x)]f_0(x) dx$   
 352 is the Kullback–Leibler divergence. Second, the prior satisfies  $\sum_k \Pi(A_{k,\delta})^{1/2} < \infty$  for all  $\delta > 0$ ,  
 353 where the sets  $A_{k,\delta} = \{f : H(f, f_k) < \delta\}$  ( $k = 1, 2, \dots$ ) are a countable number of disjoint  
 354 Hellinger balls having radius  $\delta \in (0, \epsilon)$  and which cover  $A_\epsilon$ , where  $\{f_k\} = A_\epsilon$ . Moreover, Walker,  
 355 Lijoi, and Prünster (2007) proved that consistency is obtained by requiring the prior distribution  
 356  $\pi(p)$  on  $p$  to satisfy  $\pi(p) < \exp(-4p \log p)$  for all  $p = 1, 2, \dots$ . Also, assuming such a prior  
 357 for  $p$  under the random Bernstein polynomial prior, the rate of convergence of the posterior distri-  
 358 bution is  $(\log n)^{1/3}/n^{1/3}$ , which is the same convergence rate as the sieve maximum likelihood  
 359 estimate (Walker et al., 2007, Section 3.2).

360 In practice, Petrone’s (1999) Gibbs sampling algorithm may be used to infer the posterior  
 361 distribution of the random Bernstein polynomial model. This algorithm relies on the introduction  
 362 of an auxiliary variable  $u_i$  for each data point  $x_i$  ( $i = 1, \dots, n$ ), such that  $u_1, \dots, u_n | p, G$  are  
 363 i.i.d. according to  $G$ , and that  $x_1, \dots, x_n | p, G, u_1, \dots, u_n$  are independent, with joint (likelihood)  
 364 density:

$$365 \prod_{i=1}^n \beta(x_i | \theta(u_i, p), p - \theta(u_i, p) + 1),$$

366 where for  $i = 1, \dots, n$ ,  $\theta(u_i, p) = \sum_{k=1}^p k \mathbf{1}(u_i \in A_{k,p})$  indicates the bin number of  $u_i$ , where  
 367  $A_{k,p} = ((k-1)/p, k/p]$ ,  $k = 1, \dots, p$ . Then for the inference of the posterior distribution, Gibbs  
 368 sampling proceeds by drawing from the full-conditional posterior distributions of  $G$ ,  $p$ , and  
 369  $u_i$  ( $i = 1, \dots, n$ ), for a very large number of iterations. For given  $p$  and  $u_i$  ( $i = 1, \dots, n$ ), as  
 370 suggested by the Dirichlet distribution representation of the Dirichlet process in Section 2.1, the  
 371 full conditional posterior of  $\mathbf{w}_p = (w_{1,p}, \dots, w_{p,p})$  is  $\text{Dirichlet}(\mathbf{w}_p | \alpha_{1,p}, \dots, \alpha_{p,p})$ , with  
 372  $\alpha_{k,p} = mG_0(A_{k,p}) + n\widehat{F}_u(A_{k,p})$ ,  $k = 1, \dots, p$ ,  
 373 where  $G_0$  denotes the baseline distribution of the Dirichlet process for  $G$ , and  $\widehat{F}_u$  denotes the  
 374 empirical distribution of the latent variables. For given  $u_i$  ( $i = 1, \dots, n$ ), the full conditional  
 375 posterior distribution of  $p$  is proportional to

$$376 \pi(p) \prod_{i=1}^n \beta(x_i | \theta(u_i, p), p - \theta(u_i, p) + 1).$$

377 Also, it is straightforward to sample from the full conditional posterior distribution of  $u_i$ , for  
 378  $i = 1, \dots, n$  (for details, see Petrone, 1999, p. 385).

379 **2.3. Dependent Bivariate Model**

380 A model for constructing a bivariate Dirichlet process has been given in Walker and Muliere  
 381 (2003). The idea is as follows: Take  $G_X \sim \Pi(m, G_0)$  and then for some fixed  $r \in \{0, 1, 2, \dots\}$ ,  
 382 and take  $z_1, \dots, z_r$  to be independent and identically distributed from  $G_X$ . Then take

$$383 G_Y \sim \Pi(m + r, (mG_0 + r\widehat{F}_r)/(m + r)),$$

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

$$389 \text{Corr}(G_X(A), G_Y(A)) = r/(m + r)$$

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  $\pi(p(Y))$  for  $p(X)$  and  $p(Y)$ . Each of these random distributions are defined on  $(0, 1]$ . However, without loss of generality, it is possible to model observed test scores after transforming each of them into  $(0, 1)$ . 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  $x' = (x - x_{\min} + \epsilon) / (x_{\max} - x_{\min} + 2\epsilon)$ , where  $\epsilon > 0$  is a very small constant, with  $x_{\min}$  and  $x_{\max}$  denoting the minimum and maximum possible scores on the test. The scores can be transformed back to their original scale by taking  $x = x'(x_{\max} - x_{\min} + 2\epsilon) + x_{\min} - \epsilon$ ; similarly, for  $y$  and  $y'$  for Test  $Y$ .

Under Bayes' theorem, given observed scores  $\mathbf{x}_{n(X)} = \{x_1, \dots, x_{n(X)}\}$  and  $\mathbf{y}_{n(Y)} = \{y_1, \dots, y_{n(Y)}\}$  on the two tests (assumed to be mapped onto a sample space  $[0, 1]$ ), the random bivariate Bernstein polynomial prior distribution combines with these data to define a posterior distribution. This posterior distribution,  $\Pi_n$ , assigns mass:

$$\Pi_n(A) = \frac{\int_A \{ \prod_{i=1}^{n(X)} f_X(x_i) \prod_{i=1}^{n(Y)} f_Y(y_i) \} \Pi(d f_{XY})}{\int_{\Omega} \{ \prod_{i=1}^{n(X)} f_X(x_i) \prod_{i=1}^{n(Y)} f_Y(y_i) \} \Pi(d f_{XY})}$$

to any given subset of bivariate densities  $A \subseteq \Omega = \{f_{XY}\}$  defined on  $[0, 1]^2$  (with respect to Lebesgue measure). For notational convenience, the posterior distribution of the bivariate Bernstein model is represented by  $F_X, F_Y | \mathbf{x}_{n(X)}, \mathbf{y}_{n(Y)}$ , where  $(F_X, F_Y)$  (with densities  $f_X$  and  $f_Y$ ) are the marginal distributions of  $F_{XY}$ . Recall from Section 2.2 that this posterior distribution can also be represented by  $\mathbf{w}_X, \mathbf{w}_Y, p(X), p(Y) | \mathbf{x}_{n(X)}, \mathbf{y}_{n(Y)}$ , where the mixing weights  $\mathbf{w}_X$  and  $\mathbf{w}_Y$  each follow a Dirichlet distribution.

It is natural to ask whether the random bivariate Bernstein prior satisfies strong (Hellinger) posterior consistency, in the sense that  $\Pi_n(A_\epsilon)$  converges to zero as the sample size  $n$  increases, for all  $\epsilon > 0$ . Here,  $A_\epsilon$  is a Hellinger neighborhood around  $f_{0XY}$ , denoting a true value of  $f_{XY}$ . It so happens that this consistency follows from the consistency of each univariate marginal density  $f_X$  and  $f_Y$ , since as the sample sizes go to infinity, the dependence between the two models disappears to zero (since  $r$  is fixed). As mentioned in Section 2.2, posterior consistency of the random Bernstein prior in the univariate case was established by Walker (2004) and Walker et al. (2007). Moreover, as proven by Walker et al. (2007), consistency of the bivariate model is obtained by allowing the prior distribution  $\pi(p_X)$  to satisfy  $\pi(p_X) < \exp(-4p_X \log p_X)$  for all large  $p_X$ , and similarly for  $p_Y$ . A desirable consequence of this posterior consistency of the true marginal densities,  $(f_{0X}, f_{0Y})$ , corresponding to marginal distribution functions  $(F_{0X}, F_{0Y})$ , posterior consistency is achieved in the estimation of the true equating function, given by  $e_Y(\cdot) = F_{0Y}^{-1}(F_{0X}(\cdot))$ .

Inference of the posterior distribution of the bivariate Bernstein model requires the use of an extension of Petrone's (1999) Gibbs sampling algorithm, which is described in Appendix II. A MATLAB program was written to implement the Gibbs sampling algorithm, to infer the posterior distribution of the bivariate Bernstein polynomial model. This program can be obtained

451 through correspondence with the first author. At each iteration of this Gibbs algorithm, a current  
 452 set of  $\{p(X), \mathbf{w}_X\}$  for Test  $X$  and  $\{p(Y), \mathbf{w}_Y\}$  for Test  $Y$  is available, from which it is possible to  
 453 construct the random equating function

$$e_Y(x) = F_Y^{-1}(F_X(x)) = y. \tag{5}$$

456 Hence, for each score  $x$  on Test  $X$ , a posterior distribution for the equated score on Test  $Y$  is  
 457 available. A (finite-sample) 95% confidence interval of an equated score  $e_Y(x) = F_Y^{-1}(F_X(x))$   
 458 is easily attained from the samples of posterior distribution  $F_X, F_Y | \mathbf{x}_n, \mathbf{y}_n$ . A point estimate of  
 459 an equated score  $e_Y(x)$  can also be obtained from this posterior distribution. While one conven-  
 460 tional choice of point-estimate is given by the posterior mean of  $e_Y(x)$ , the posterior median  
 461 point-estimate of  $e_Y(\cdot)$  has the advantage that it is invariant over monotone transformations. This  
 462 invariance is important considering that the test scores are transformed into the (0, 1) domain,  
 463 and back onto the original scale of the test scores.

464 As presented above, the Bayesian nonparametric equating method readily applies to the EG  
 465 design with no anchor test, and the SG design. However, with little extra effort, this Bayesian  
 466 method can be easily extended to a EG or NG design with an anchor test, or to a counterbal-  
 467 anced design. For an equating design having an anchor test, it is possible to implement the  
 468 idea of chained equipercentile equating (Angoff, 1971) to perform posterior inference of the  
 469 random equating function. In particular, if  $\mathbf{x}_{n(X)}$  and  $\mathbf{v}_{n(V_1)}$  denote the set of scores observed  
 470 from examinee group 1 who completed Test  $X$  and an anchor Test  $V$ , and  $\mathbf{y}_{n(Y)}$  and  $\mathbf{v}_{n(V_2)}$   
 471 sets of scores observed from examinee group 2 who completed Test  $Y$  and the same anchor  
 472 Test  $V$ , then it is possible to perform posterior inference of the random equating functions  
 473  $e_Y(x) = F_Y^{-1}(F_{V_2}(e_{V_1}(x)))$  and  $e_{V_1}(x) = F_{V_1}^{-1}(F_{X_1}(x))$ , based on sample from the posterior dis-  
 474 tributions  $F_X, F_{V_1} | \mathbf{x}_{n(X)}, \mathbf{v}_{n(V_1)}$  and  $F_Y, F_{V_2} | \mathbf{y}_{n(Y)}, \mathbf{v}_{n(V_2)}$  each under a bivariate Bernstein prior.  
 475 For a counterbalanced design, to combine the information of the two examine subgroups 1 and 2  
 476 in the spirit of von Davier et al. (2004, Section 2.3), posterior inference of the random equating  
 477 function  $e_Y(x) = F_Y^{-1}(F_X(x))$  is attained by taking  $F_X(\cdot) = \varpi_X F_{X_1}(\cdot) + (1 - \varpi_X) F_{X_2}(\cdot)$  and  
 478  $F_Y(\cdot) = \varpi_Y F_{Y_1}(\cdot) + (1 - \varpi_Y) F_{Y_2}(\cdot)$ , where  $(F_{X_1}, F_{X_2}, F_{Y_1}, F_{Y_2})$  are from the posterior distribu-  
 479 tions  $F_{X_1}, F_{Y_2} | \mathbf{x}_{n(X_1)}, \mathbf{y}_{n(Y_2)}$  and  $F_{X_2}, F_{Y_1} | \mathbf{x}_{n(X_2)}, \mathbf{y}_{n(Y_1)}$  under two bivariate Bernstein models,  
 480 respectively. Also,  $0 \leq \varpi_X, \varpi_Y \leq 1$  are chosen weights, which can be varied to determine how  
 481 much they change the posterior inference of the equating function  $e_Y(\cdot)$ .

### 3. Applications

482 The following three subsections illustrate the Bayesian nonparametric model in the equating  
 483 of test scores arising from the equivalent groups design, the counterbalanced design, and the non-  
 484 equivalent groups design for chain equating, respectively. The equating results of the Bayesian  
 485 model will also be compared against the results of the kernel, percentile-rank, linear, and mean  
 486 models of equating. In so doing, the assumption of independence will be evaluated for each of the  
 487 three data sets. Before proceeding, it is necessary to review some themes that repeat themselves  
 488 in the three applications.

- 494 1. In applying our Bayesian model to analyze each of the three data sets, we assumed the bi-  
 495 variate Dirichlet process to have baseline distribution  $G_0$  that equals the Beta(1, 1) distribu-  
 496 tion. Also, we assumed a relatively noninformative prior by taking  $m = 1$  and  $r = 4$ , re-  
 497 flecting the (rather uncertain) prior belief that the correlation of the scores between two tests  
 498 is  $0.8 = r/(m + r)$ . In particular, the Beta(1, 1) distribution reflects the prior belief that the  
 499 different possible test scores are equally likely, and the choice of “prior sample size” of  $m = 1$   
 500 will lead to a data-driven posterior distribution of  $(F_X, F_Y)$ . This is true especially considering

501 that among the three data sets analyzed, the smallest sample size for a group of examinees was  
 502 about 140 (i.e., 140 times the prior sample size), and the other two data sets had a sample size  
 503 of around 1,500 for each group of examinees. Furthermore, up to a constant of proportionality,  
 504 we specify an independent prior distribution of  $\pi(p) \propto \exp(-4p \log p)$  for each  $p(X)$  and on  
 505  $p(Y)$ . As discussed in Sections 2.2 and 2.3, this choice of prior ensures the consistency of the  
 506 posterior distribution of  $(F_X, F_Y)$ .

507 2. For each data set analyzed with the Bayesian model, we implemented the Gibbs sampling al-  
 508 gorithm (Appendix II) to generate 10,000 samples from the posterior distribution of  $(F_X, F_Y)$ ,  
 509 including  $(p(X), p(Y))$ , after discarding the first 2,000 Gibbs samples as burn-in. We found  
 510 that Gibbs sampler displayed excellent mixing in the posterior sampling. Also, while we chose  
 511 the first 2,000 Gibbs samples as burn in, this number was a conservative choice because trace  
 512 plots suggested that convergence was achieved after the first 500 samples.

513 3. For the percentile rank model, the estimate  $(\widehat{G}_X, \widehat{G}_Y)$  is given by the empirical distribution  
 514 estimates of the scores in Test  $X$  and Test  $Y$ , respectively. For either the linear or the mean  
 515 equating model, the estimate  $(\widehat{\mu}_X, \widehat{\sigma}_X^2, \widehat{\mu}_Y, \widehat{\sigma}_Y^2)$  is given by sample means and variances of the  
 516 test scores. For the kernel model, the estimate  $(\widehat{G}_X, \widehat{G}_Y)$  of the discrete test score distributions  
 517 are obtained as marginals of the estimate  $\widehat{G}_{XY}$  obtained via maximum likelihood estimation of  
 518 a chosen log-linear model. The log-linear model needs to be specified differently for each of  
 519 the three applications, since they involve different equating designs (von Davier et al., 2004).  
 520 Also, the percentile-rank, linear, and mean equating models, 95% confidence intervals of the  
 521 equated scores were estimated from 10,000 bootstrap samples, each bootstrap sample taking  
 522  $n(X)$  and  $n(Y)$  samples with replacement from the empirical distributions of test scores,  $\widehat{G}_X$   
 523 and  $\widehat{G}_Y$ , and then performing equipercetile equating using these samples. For single-group  
 524 designs,  $n = n(X) = n(Y)$  samples are taken with replacement from the bivariate empirical  
 525 distribution  $\widehat{G}_{XY}$  having univariate marginals  $(\widehat{G}_X, \widehat{G}_Y)$ . Also, unless otherwise noted, for the  
 526 kernel equating model, 95% confidence intervals of the equated scores were estimated from  
 527 10,000 bootstrap samples, each bootstrap sample involving taking  $n(X)$  and  $n(Y)$  samples  
 528 with replacement from the continuous distribution estimates of the test scores,  $\widehat{F}_X(\cdot|\widehat{\theta})$  and  
 529  $\widehat{F}_Y(\cdot|\widehat{\theta})$ , and then performing equipercetile equating using these samples. Efron and Tibshirani  
 530 (1993) suggest that at least 500 bootstrap samples are sufficient for estimating confidence  
 531 intervals.

532  
 533 *3.1. Equivalent-Groups Design*  
 534

535 The Bayesian nonparametric equating model is demonstrated in the analysis of a large data  
 536 set generated from an equivalent groups (EG) design. This data set, obtained from von Davier  
 537 et al. (2004, Chap. 7, p. 100), consists of 1,453 examinees who completed Test  $X$ , and 1,455  
 538 examinees completing Test  $Y$  of a national mathematics exam. Each test has 20 items, and is  
 539 scored by number correct. The average score on Test  $X$  is 10.82 (s.d. = 3.81), and the average  
 540 score on Test  $Y$  is 11.59 (s.d. = 3.93), and so the second test is easier than the first.

541 Figure 1 plots the 95% confidence interval of  $G_X(x) - G_Y(x)$  ( $x = 0, 1, \dots, 20$ ), estimated  
 542 from 10,000 bootstrap samples from the empirical distributions  $(\widehat{G}_X, \widehat{G}_Y)$ . The plot suggests that  
 543  $G_X$  and  $G_Y$  have quite similar shapes, and thus are not independent (as assumed by the kernel,  
 544 percentile-rank, linear, and mean models of equating). In fact, according to the plot, the 95%  
 545 confidence interval of  $G_X(x) - G_Y(x)$  envelopes 0 for all score points except for scores 5, 8,  
 546 18, and 19. Clearly, there is a need to model the similarity (correlation) between the test score  
 547 distributions. Furthermore, according to the confidence interval, the two c.d.f.s of the test score  
 548 distributions differ by about .045 in absolute value at most.

549 For the kernel model, von Davier et al. (2004, Chap. 7) reported estimates of the discrete  
 550 score distributions  $(\widehat{G}_X, \widehat{G}_Y)$  through maximum likelihood estimation of a joint log-linear model,

PSYCHOMETRIKA

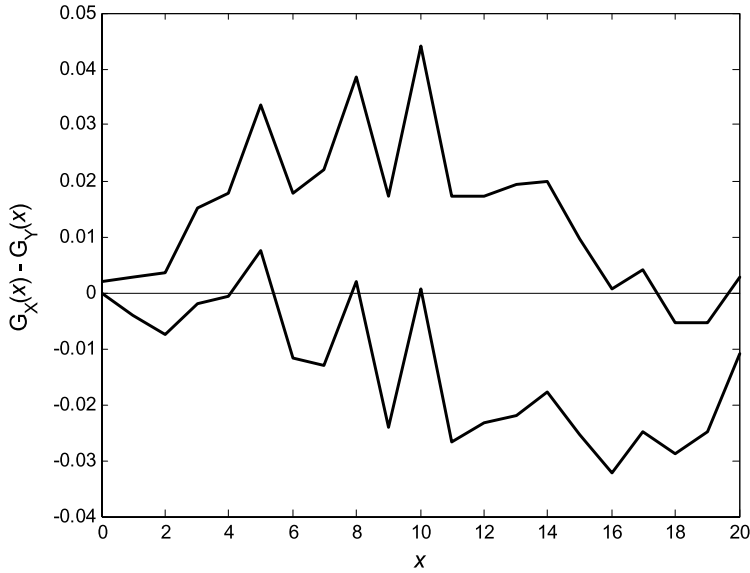


FIGURE 1.

The 95% confidence interval of the difference  $G_X(x) - G_Y(x)$  ( $x = 0, 1, \dots, 20$ ) is given by the thick jagged lines. The horizontal straight line indicates zero differences for all points of  $x$ .

TABLE 1.  
Comparison of the 95% confidence (credible) intervals of 5 equating methods: EG design.

	Bayes	Kernel	PR	Linear
Kernel	2–18			
PR	1–16, 20	19		
Linear	2–15, 20	20	19–20	
Mean	0, 2–16, 20	20	1–3, 19–20	None

and they also report the bandwidth estimates ( $\hat{h}_X = .62, \hat{h}_Y = .58$ ). For the Bayesian model, the marginal posterior distributions of  $p(X)$  and of  $p(Y)$  concentrated on 1 and 2, respectively. Figure 2 presents the posterior median estimate of the equating function for the Bayesian equating model, and the estimate of the equating functions for the other four equating models. This figure also presents the 95% confidence interval estimates of the equating functions. For the kernel model, the 95% confidence interval was obtained from the standard error of equating estimates reported in von Davier et al. (2004, Table 7.4). This figure shows that the equating function estimate of the Bayesian model differs from the estimates obtained from the other four equating models. Table 1 presents pairwise comparisons of the equating function estimates between the five equating models. This table reports the values of  $x$  which provide no overlap between the 95% confidence interval of the  $e_Y(x)$  estimate, between two models. Among other things, this table shows that the equating function estimate of the Bayesian model is very different from the equating function estimates of the other four equating models. The assumption of independence between  $(F_X, F_Y)$  may play a role, with independence not assumed in the Bayesian model, and independence assumed by the other four models. Though, as shown earlier, the data present evidence against the assumption of independence.

Also, this table shows that the equating function estimate of the kernel model do not differ much with the equating estimates of the linear and mean equating models. Furthermore, upon closer inspection, the kernel, linear, and mean equating models equated some scores of Test X

GEORGE KARABATSOS AND STEPHEN G. WALKER

601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650

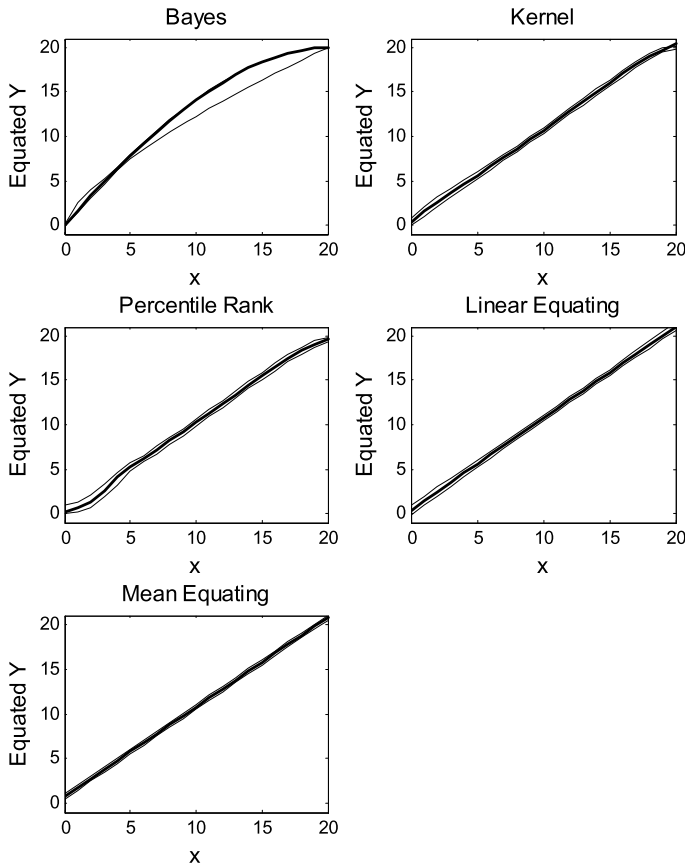


FIGURE 2.

For each of the five equating methods, the point-estimate of  $e_Y(\cdot)$  (*thick line*) and the 95% confidence interval (*thin lines*).

with scores outside the range of scores for Test  $Y$ . For example, the kernel model equated a score of 20 on Test  $X$  with a score of 20.4 on Test  $Y$ , and the Test  $X$  scores of 0 and 20 led to equated scores on Test  $Y$  with 95% confidence intervals that included values outside the 0–20 range of test scores. The linear and mean equating models had similar issues in equating.

### 3.2. Counterbalanced Design

Here, the focus of analysis is a data set generated from a counterbalanced single-group design (CB design). These data were collected from a small field study from an international testing program, and were obtained from von Davier et al. (2004, Tables 9.7–8). Test  $X$  has 75 items, Test  $Y$  has 76 items, and both tests are scored by number correct. Group 1 consists of 143 examinees completed Test  $X$  first, then Test  $Y$ . The tests of this single-group design are referred to as  $(X_1, Y_2)$ . The average score on Test  $X_1$  is 52.54 (s.d. = 12.40), and for Test  $Y_2$  it is 51.29 (s.d. = 11.0). Group 2 consists of 140 examinees completed Test  $Y$  first, then Test  $X$ , and the tests of this single-group design are referred to as  $(Y_1, X_2)$ . The average score on Test  $Y_1$  is 51.39 (s.d. = 12.18), and for Test  $X_2$  it is 50.64 (s.d. = 13.83).

Since a pair of test scores is observed from every examinee in each of the two single group designs, it is possible to evaluate the assumption of independence with the Spearman’s rho statistic. By definition, if  $G_{XY}$  is a bivariate c.d.f. with univariate margins  $(G_X, G_Y)$ , and  $(X, Y) \sim G_{XY}$ , then Spearman’s rho is the correlation between c.d.f.s  $G_X(X)$  and  $G_Y(Y)$  (see

PSYCHOMETRIKA

651 Joe, 1997, Section 2.1.9). The Spearman’s rho correlation is .87 between the scores of Test  $X_1$   
 652 and Test  $Y_2$ , and this correlation is .86 between the scores of Test  $X_2$  and Test  $Y_1$ . So clearly, the  
 653 assumption of independence does not hold in either data set. In contrast, a Spearman’s rho cor-  
 654 relation of 0 (independence) is assumed for the mean, linear, percentile-rank, and kernel models  
 655 of equating.

656 In the kernel and percentile-rank methods for the counterbalanced design, the estimate of  
 657 the equating function,

$$e_Y(x; \hat{\theta}, \varpi_X, \varpi_Y) = \hat{F}_Y^{-1}(\hat{F}_X(x; \hat{\theta}, \varpi_X) | \hat{\theta}, \varpi_Y),$$

660 is obtained through weighted estimates of the score probabilities, given by:

$$\begin{aligned} \hat{g}_X(\cdot) &= \varpi_X \hat{g}_{X_1}(\cdot) + (1 - \varpi_X) \hat{g}_{X_2}(\cdot), & \hat{g}_Y(\cdot) &= \varpi_Y \hat{g}_{Y_1}(\cdot) + (1 - \varpi_Y) \hat{g}_{Y_2}(\cdot), \\ \hat{g}_{X_1}(\cdot) &= \sum_l \hat{g}_{X_1, Y_2}(\cdot, y_l), & \hat{g}_{X_2}(\cdot) &= \sum_l \hat{g}_{X_2, Y_1}(\cdot, y_l), \\ \hat{g}_{Y_1}(\cdot) &= \sum_k \hat{g}_{X_2, Y_1}(x_k, \cdot), & \hat{g}_{Y_2}(\cdot) &= \sum_k \hat{g}_{X_1, Y_2}(x_k, \cdot), \end{aligned}$$

669 while in linear and mean equating methods, the weights are put directly on the continuous distri-  
 670 butions, with

$$\begin{aligned} \hat{F}_X(\cdot; \hat{\theta}, \varpi_X) &= \varpi_X \text{Normal}_{X_1}(\cdot | \hat{\mu}_{X_1}, \hat{\sigma}_{X_1}^2) + (1 - \varpi_X) \text{Normal}_{X_2}(\cdot | \hat{\mu}_{X_2}, \hat{\sigma}_{X_2}^2), \\ \hat{F}_Y(\cdot; \hat{\theta}, \varpi_Y) &= \varpi_Y \text{Normal}_{Y_1}(\cdot | \hat{\mu}_{Y_1}, \hat{\sigma}_{Y_1}^2) + (1 - \varpi_Y) \text{Normal}_{Y_2}(\cdot | \hat{\mu}_{Y_2}, \hat{\sigma}_{Y_2}^2). \end{aligned}$$

676 Here,  $\varpi_X, \varpi_Y \in [0, 1]$  are weights chosen to combine the information of the two groups of ex-  
 677 aminees who completed Test  $X$  and Test  $Y$  in different orders. This idea for equating in the  
 678 counterbalanced design is due to von Davier et al. (2004, Section 2.3). As they describe, the  
 679 value  $\varpi_X = \varpi_Y = 1$  represents a default choice because it represents the most conservative use  
 680 of the data in the CB design, while the choice of  $\varpi_X = \varpi_Y = 1/2$  is the most generous use of the  
 681  $(X_2, Y_2)$  data because it weighs equally the two versions of Test  $X$  and of Test  $Y$ . One approach  
 682 is to try these two different weights, and see what effect they have on the estimated equated func-  
 683 tion. von Davier et al. (2004, Chap. 9) obtain the estimate  $(\hat{G}_X, \hat{G}_Y)$  by selecting and obtaining  
 684 maximum likelihood estimates of a joint log-linear model for the single group designs  $(X_1, Y_2)$   
 685 and  $(Y_1, X_2)$ . Then conditional on that estimate, they found (von Davier’s et al. 2004, p. 143)  
 686 that the bandwidth estimates are  $(h_X = .56, h_Y = .63)$  under  $\varpi_X = \varpi_Y = 1/2$ , and the band-  
 687 width estimates are  $(h_X = .56, h_Y = .61)$  under  $\varpi_X = \varpi_Y = 1$ . Through the use of bootstrap  
 688 methods, it was found that for kernel, percentile-rank, linear, and mean equating methods, the  
 689 95% confidence intervals for bootstrap samples of  $e_Y(x; 1, 1) - e_Y(x; \frac{1}{2}, \frac{1}{2})$  enveloped 0 for all  
 690 scores  $x = 0, 1, \dots, 75$ . This result suggests that the equating function under  $\varpi_X = \varpi_Y = 1/2$   
 691 is not significantly different than the equating function under  $\varpi_X = \varpi_Y = 1$ . Also, no equating  
 692 difference was found for the Bayesian model. Using the chain equating methods described at  
 693 the end of Section 2.3, it was found that the 95% confidence interval of the posterior distribu-  
 694 tion of  $e_Y(x; 1, 1) - e_Y(x; \frac{1}{2}, \frac{1}{2})$  enveloped 0 for all scores  $x = 0, 1, \dots, 75$ . Also, the marginal  
 695 posterior mean of  $p(X_1), p(Y_2), p(X_2)$ , and  $p(Y_1)$  was 2.00 (var = .01), 2.01 (var = .19), 2.00  
 696 (var = .001), and 2.01 (var = .02), respectively.

697 Figure 3 presents the point-estimate of  $e_Y(\cdot; \frac{1}{2}, \frac{1}{2})$  of the equating function for each of the five  
 698 equating models, along with the corresponding 95% confidence (credible) intervals. As shown in  
 699 Table 2, after accounting for these confidence intervals, there were no differences in the equating  
 700 function estimates between the Bayesian equating model and the kernel equating model, and

GEORGE KARABATSOS AND STEPHEN G. WALKER

701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750

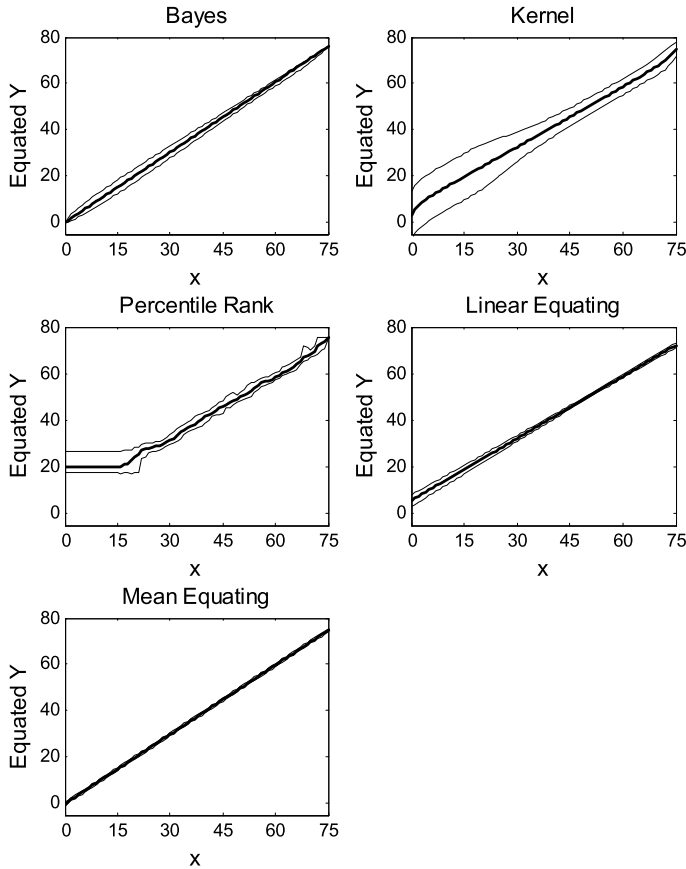


FIGURE 3.

For each of the five equating methods, the point-estimate of  $e_Y(\cdot)$  (*thick line*) and the 95% confidence interval (*thin lines*).

TABLE 2.  
Comparison of the 95% confidence (credible) intervals of 5 equating methods: CB design.

	Bayes	Kernel	PR	Linear
Kernel	None			
PR	0–14, 75	0–1		
Linear	0–6, 68–75	None	0–11, 75	
Mean	74–75	None	0–16, 22–26	0–35, 65–75

there were no differences between the kernel equating model and the linear and mean equating models. Also, for scores of  $x$  ranging from 0 to 30, the kernel and the percentile-rank models each have an equating function estimate with a relatively large 95% confidence interval, and for the percentile-rank model, the equating function estimate is flat in that score range. A closer inspection of the data reveals that this is due to the very small number of test scores in the 0 to 30 range. In fact, for each test, no more than 16 scores fall in that range. The confidence interval of the Bayesian method does not suffer from such issues. Also, among all five equating models, the kernel model had the largest 95% confidence interval. Upon closer inspection, for  $x$  scores of 0–4 and 72–75, the kernel model equated scores on Test  $Y$  having 95% confidence intervals that included values outside the  $[0, 76]$  score range for Test  $Y$ . Similar issues were observed for the

751 linear and mean equating models in equating the  $x$  score of 75, and for the mean equating model  
 752 in equating the  $x$  score of 0.

753

754 *3.3. Nonequivalent Groups Design and Chained Equating*

755

756 This section concerns the analysis of a classic data set arising from a nonequivalent groups  
 757 (NG) design with internal anchor, and it was discussed in Kolen and Brennan (2004). The first  
 758 group of examinees completed Test  $X$ , and the second group of examinees completed Test  $Y$ ,  
 759 both groups being random samples from different populations. Here, Test  $X$  and Test  $Y$  each  
 760 have 36 items and is scored by number correct, and both tests have 12 items in common. These  
 761 12 common items form an internal anchor test because they contribute to the scoring of Test  
 762  $X$  and of Test  $Y$ . While the two examinee groups come from different populations, the anchor  
 763 test provides a way to link the two groups and the two tests. The anchor test completed by the  
 764 first examinee group (population) is labeled as  $V_1$  and the anchor test completed by the second  
 765 examinee group (population) is labeled as  $V_2$ , even though both groups completed the same  
 766 anchor test. The first group of 1,655 examinees had a mean score of 15.82 (s.d. = 6.53) on Test  
 767  $X$ , and a mean score of 5.11 (s.d. = 2.38) for the anchor test. The second group of examinees  
 768 had a mean score of 18.67 (s.d. = 6.88) on Test  $Y$ , and a mean score of 5.86 (s.d. = 2.45) on  
 769 the anchor test. Also, the scores of Test  $X_1$  and Test  $V_1$  have a Spearman's rho correlation of  
 770 .84, while the Spearman's rho correlation is .87 between the scores of Test  $V_2$  and Test  $Y_2$ . In  
 771 contrast, a zero correlation (independence) is assumed for the mean, linear, percentile-rank, and  
 772 kernel models of equating.

773

774 In the analysis of these data from the nonequivalent groups design, chained equipercen-  
 775 tilate equating was used. Accordingly, under either the kernel, percentile-rank, linear, and mean equat-  
 776 ing models, the estimate of the equating function is given by  $e_Y(x; \hat{\theta}) = F_{V_2}^{-1}(F_{V_2}(e_{V_1}(x); \hat{\theta})|\hat{\theta})$   
 777 for all  $x$ , where  $e_{V_1}(\cdot; \hat{\theta}) = F_{V_1}^{-1}(F_{X_1}(\cdot; \hat{\theta})|\hat{\theta})$ , and  $\hat{\theta}$  is the parameter estimate of the correspond-  
 778 ing model. In the kernel method, to obtain the estimate of the marginals ( $\hat{G}_X, \hat{G}_Y$ ) through  
 779 log-linear model fitting, it was necessary to account for the structural zeros in the  $37 \times 13$  contin-  
 780 gency table for scores on Test  $X$  and Test  $V_1$ , and in the  $37 \times 13$  contingency table for the scores  
 781 on Test  $Y$  and Test  $V_2$ . These structural zeros arise because given every possible score  $x$  on  
 782 Test  $X$  (and every possible score  $y$  on Test  $Y$ ), the score on the internal anchor test ranges from  
 783  $\max(0, x - 24)$  to  $\min(x, 12)$  (ranges from  $\max(0, y - 24)$  to  $\min(y, 12)$ ), while for every given  
 784 score  $v$  on the anchor test, the score on Test  $X$  (Test  $Y$ ) ranges from  $v$  to  $36 - (12 - v)$ . Therefore,  
 785 for the two tables, log linear models were fit only to cells with no structural zeros (e.g., Holland  
 786 and Thayer 2000). In particular, for each table, 160 different versions of a loglinear model were  
 787 fit and compared on the basis of Akaike's Information Criterion (AIC; Akaike, 1973), and the one  
 788 with the lowest AIC was chosen as the model for subsequent analysis. Appendix III provides the  
 789 technical details about the loglinear model fitting. After finding the best-fitting loglinear model  
 790 for each of the two tables, estimates of the marginal distributions ( $\hat{G}_X, \hat{G}_Y$ ) were derived, and  
 791 then bandwidth estimates ( $\hat{h}_X = .58, \hat{h}_Y = .61, \hat{h}_{V_1} = .55, \hat{h}_{V_2} = .59$ ) were obtained using the  
 792 least-squares minimization method described in Section 1. In the analysis with the Bayesian  
 793 model, a chained equipercen-tilate method was implemented, described at the end of Section 2.3.  
 794 The marginal posterior distribution of  $p(X_1), p(V_1), p(V_2)$ , and  $p(Y_2)$  concentrated on values  
 795 of 6, 1, 3, and 5, respectively.

796

797 Figure 4 presents the equating function estimate for each of the five models, along with  
 798 their corresponding 95% confidence interval. It is shown that for the percentile-rank models, the  
 799 95% confidence interval is relatively large for  $x$  scores ranging between 20 to 36. According to  
 800 Table 3, taking into account the 95% confidence intervals, the equating function estimate of the  
 Bayesian model again differed from the estimate yielded by the other four models. The equating  
 function estimate of the kernel model did not differ much with the estimates of the linear model.

GEORGE KARABATSOS AND STEPHEN G. WALKER

801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850

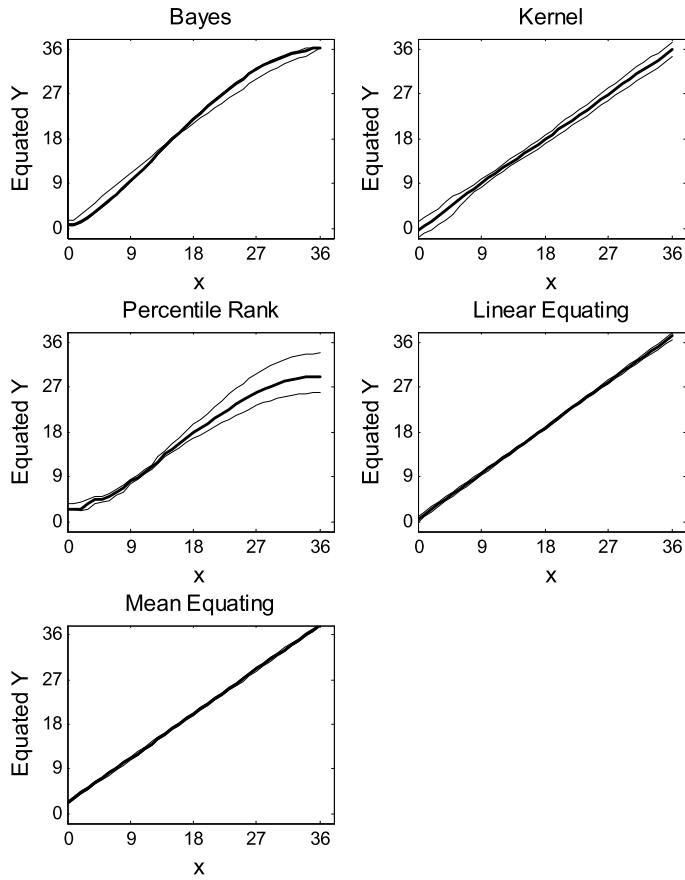


FIGURE 4.

For each of the five equating methods, the point-estimate of  $e_Y(\cdot)$  (thick line) and the 95% confidence interval (thin lines).

TABLE 3.  
Comparison of the 95% confidence (credible) intervals of 5 equating methods: NG design.

	Bayes	Kernel	PR	Linear
Kernel	11–32			
PR	0–1, 7–36	0–1, 10, 11, 36		
Linear	11–32, 36	None	0–1, 5–15, 33–36	
Mean	0–7, 15–36	0–36	0, 3–20, 31–36	0–35

Also, upon closer inspection, the kernel model equated scores of  $x = 0, 1, 2$  with scores below the range of Test  $Y$  scores. The model also equated an  $x$  score of 36 with a score on Test  $Y$  having a 95% confidence interval that includes values above the range of Test  $Y$  scores. The linear and mean equating models had similar issues.

4. Conclusions

This study introduced a Bayesian nonparametric model for test equating. It is defined by a bivariate Bernstein polynomial prior distribution that supports the entire space of (random)

PSYCHOMETRIKA

851 continuous distributions  $(F_X, F_Y)$ , with this prior depending on the bivariate Dirichlet process.  
 852 The Bayesian equating model has important theoretical and practical advantages over all previous  
 853 approaches to equating. One key advantage of the Bayesian equating model is that it accounts  
 854 for the realistic situation that the two distributions of test scores  $(F_X, F_Y)$  are correlated, instead  
 855 of independent as assumed in the previous equating models. This dependence seems reasonable,  
 856 considering that in practice, the two tests that are to be equated are designed to measure the same  
 857 psychological trait. Indeed, for each of the three data sets that were analyzed, there is strong  
 858 evidence against the assumption of independence. While perhaps the Bayesian model of equating  
 859 requires more computational effort and mathematical expertise than the kernel, linear, mean, and  
 860 percentile-rank models of equating, the extra effort is warranted considering the key advantages  
 861 of the Bayesian model. It can be argued that the four previous models make rather unrealistic  
 862 assumptions about data, and carry other technical issues such as asymmetry, out-of-range equated  
 863 scores. The Bayesian model outperformed the other models in the sense it avoids such issues.  
 864 Doing so led to remarkably different equating function estimates under the Bayesian model.  
 865 Also, the percentile-rank, linear, and mean equating models were proven to be special cases of  
 866 the Bayesian nonparametric model, corresponding to a very strong choice of prior distribution  
 867 for the continuous test score distributions  $(F_X, F_Y)$ . In future research, it may be of interest  
 868 to explore alternative Bayesian approaches to equating that are based on other nonparametric  
 869 priors that account for dependence between  $(F_X, F_Y)$ , including the priors described by De Iorio,  
 870 Müller, Rosner, and MacEachern (2004) and by Müller, Quintana, and Rosner (2004).

871  
 872

Acknowledgements

873  
 874

875 Under US Copyright, August, 2006. We thank Alberto Maydeu-Olivares (Action Editor),  
 876 Alina von Davier, and three anonymous referees for suggestions to improve the presentation of  
 877 this manuscript.

878  
 879

Appendix I. A Proof About Special Cases of the IDP Model

880  
 881

882 It is proven that the linear equating model, the mean equating model, and the percentile-rank  
 883 equating model, is each a special case the IDP model. To achieve fullest generality in the proof,  
 884 consider that a given equating model assumes that the continuous distributions  $(F_X, F_Y)$  are  
 885 governed by some function  $\varphi$  of a finite-dimensional parameter vector  $\theta$ . This way, it is possible  
 886 to cover all the variants of these three models, including the Tucker, Levine observed score,  
 887 Levine true score, and Braun–Holland methods of linear (or mean) equating, the frequency-  
 888 estimation approach to the percentile-rank equating model, all the item response theory methods  
 889 of observed score equating, and the like (for details about these methods, see Kolen & Brennan,  
 890 2004).

891  
 892

892 **Theorem 1.** *The linear equating, the mean equating model, and the percentile-rank equating*  
 893 *model is each a special case of the IDP model, where  $m(X), m(Y) \rightarrow \infty$ , and the baseline*  
 894 *distributions  $(G_{0X}, G_{0Y})$  of the IDP are defined by the corresponding equating model.*

895  
 896

896 *Proof:* Recall from Section 2.1 that in the IDP model, the posterior distribution of  $(F_X, F_Y)$  is  
 897 given by independent beta distributions, with posterior mean  $E[F_X(\cdot)] = G_{0X}(\cdot)$  and variance  
 898  $\text{Var}[F_X(\cdot)] = \{G_{0X}(A)[1 - G_{0X}(A)]\}/(m(X) + 1)$ , and similarly for  $F_Y$ . First, define the base-  
 899 line distributions  $(G_{0X}, G_{0Y})$  of the IDP model according to the distributions assumed by the lin-  
 900 ear equating model, with  $G_{0X}(\cdot) = \text{Normal}(\cdot|\mu_X, \sigma_X^2)$  and  $G_{0Y}(\cdot) = \text{Normal}(\cdot|\mu_Y, \sigma_Y^2)$ , where

901  $(\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2)$  is some function  $\varphi$  of a parameter vector  $\theta$ . Taking the limit  $m(X), m(Y) \rightarrow \infty$   
 902 leads to  $\text{Var}[F_X(\cdot)] = \text{Var}[F_Y(\cdot)] = 0$ , and then the posterior distribution of  $(F_X, F_Y)$  assigns  
 903 probability 1 to  $(G_{0X}, G_{0Y})$ , which coincide with the distributions assumed by the linear equat-  
 904 ing model.

905 The same is true for the mean equating model, assuming  $\sigma_X^2 = \sigma_Y^2$ . The same is also true  
 906 for the percentile-rank model, assuming  $G_{0X}(\cdot) = \sum_{k=1}^{p(X)} g_X(\cdot; \varphi(\theta)) \text{Uniform}(\cdot | x_k^* - \frac{1}{2}, x_k^* + \frac{1}{2})$   
 907 and  $G_{0Y}(\cdot) = \sum_{k=1}^{p(Y)} g_Y(\cdot; \varphi(\theta)) \text{Uniform}(\cdot | y_k^* - \frac{1}{2}, y_k^* + \frac{1}{2})$ , for some function  $\varphi$  depending on  
 908 a parameter vector  $\theta$ . This completes the proof.  $\square$   
 909

910  
 911 Appendix II. The Gibbs Sampling Algorithm for Bivariate Bernstein Model  
 912

913 As in Petrone's (1999) Gibbs algorithm for the one-dimensional model, latent vari-  
 914 ables are used to sample from the posterior distribution. In particular, an auxiliary vari-  
 915 able  $u_i$  is defined for each data point  $x_i$  ( $i = 1, \dots, n(X)$ ), and an auxiliary variable  $u_i(Y)$   
 916 is defined for each data point  $y_i$  ( $i = 1, \dots, n(Y)$ ), such that  $u_{1(X)}, \dots, u_{n(X)} | p(X), G_X$   
 917 are i.i.d. according to  $G_X$ , and  $u_{1(Y)}, \dots, u_{n(Y)} | p(Y), G_Y$  are i.i.d. according to  $G_Y$ . Then  
 918  $x_1, \dots, x_{n(X)} | p(X), G_X, \{u_{1(X)}, \dots, u_{n(X)}\}$  are independent, and  $y_1, \dots, y_{n(Y)} | p(Y), G_Y$ , and  
 919  $\{u_{1(Y)}, \dots, u_{n(Y)}\}$  are also independent, with joint (likelihood) density:

$$920$$

$$921 \prod_{i=1}^{n(X)} \beta(x_i | \theta(u_{i(X)}, p(Y)), p(X) - \theta(u_{i(X)}, p(Y)) + 1)$$

$$922$$

$$923$$

$$924 \times \prod_{i=1}^{n(Y)} \beta(y_i | \theta(u_{i(Y)}, p(Y)), p(Y) - \theta(u_{i(Y)}, p(Y)) + 1).$$

$$925$$

$$926$$

927 Then for the inference of the posterior distribution, Gibbs sampling proceeds by drawing from the  
 928 full-conditional posterior distributions of  $G_X, G_Y, p(X), p(Y), u_{i(X)}$  ( $i = 1, \dots, n(X)$ ),  $u_{i(Y)}$   
 929 ( $i = 1, \dots, n(Y)$ ), and  $z_j$  ( $j = 1, \dots, r$ ), for a very large number of iterations.

930 The sampling of the conditional posterior distribution of  $G_X$  is performed as follows. Note  
 931 that given  $p(X)$ , the random Bernstein polynomial density  $f_X(x; G_X, p(X))$  depends on  $G_X$   
 932 only through the values  $G_X(k/p(X))$ ,  $k = 0, 1, \dots, p(X)$ , and thus  $G_X$  is fully described by the  
 933 random vector  $\mathbf{w}_{p(X)} = (w_{1,p(X)}, \dots, w_{p(X),p(X)})'$ , with  $w_{k,p(X)} = G_X(k/p(X)) - G_X((k -$   
 934  $1)/p(X))$ ,  $k = 1, \dots, p(X)$ , that random vector having a Dirichlet distributions (e.g., see Sec-  
 935 tion 2.1). This considerably simplifies the sampling of the conditional posterior distribution of  
 936  $G_X$ . So given  $p(X)$ ,  $\{u_{1(X)}, \dots, u_{n(X)}\}$  and  $\{z_1, \dots, z_r\}$ , the full conditional posterior distribu-  
 937 tion of  $\mathbf{w}_{p(X)}$  (i.e., of  $G_X$ ) is Dirichlet( $\mathbf{w}_{p(X)} | \alpha_{1,p(X)}, \dots, \alpha_{p(X),p(X)}$ ), with

$$938$$

$$939 \alpha_{k,p(X)} = mG_0(A_{k,p}) + r\widehat{F}_r(A_{k,p(X)}) + n(X)\widehat{F}_{u(X)}(A_{k,p(X)}), \quad k = 1, \dots, p(X)$$

940 where  $\widehat{F}_{u(X)}$  is the empirical distribution of  $\{u_{1(X)}, \dots, u_{n(X)}\}$ , and

$$941$$

$$942 A_{k,p(X)} = ((k - 1)/p(X), k/p(X)).$$

$$943$$

944 Likewise, given  $p(X)$ ,  $\{u_{1(Y)}, \dots, u_{n(Y)}\}$  and  $\{z_1, \dots, z_r\}$ , the full conditional posterior distribu-  
 945 tion of  $\mathbf{w}_{p(Y)}$  (i.e., of  $G_Y$ ) is Dirichlet( $\mathbf{w}_{p(Y)} | \alpha_{1,p(Y)}, \dots, \alpha_{p(Y),p(Y)}$ ).

946 Given  $\{u_{1(X)}, \dots, u_{n(X)}\}$ , the full conditional posterior distribution of  $p(X)$  is proportional  
 947 to

$$948$$

$$949 \pi(p(X)) \prod_{i=1}^{n(X)} \beta(x_i | \theta(u_{i(X)}, p(X)), p(X) - \theta(u_{i(X)}, p(X)) + 1),$$

$$950$$

PSYCHOMETRIKA

951 and given  $\{u_{1(Y)}, \dots, u_{n(Y)}\}$ , the full conditional posterior distribution of  $p(X)$  is defined simi-  
 952 larly. Thus, a straightforward Gibbs sampler can be used for  $p(X)$  and for  $p(Y)$ .

953 Given  $p(X)$  and  $\{z_1, \dots, z_r\}$ , the sampling of each latent variable  $u_{i(X)}$  ( $i = 1, \dots, n(X)$ )  
 954 from its conditional posterior distribution, proceeds as follows. With probability

955 
$$\beta(x_l | \theta(u_{l(X)}, p(X)), p(X) - \theta(u_{l(X)}, p(X)) + 1) / (\kappa_1 + \kappa_2),$$

956  
 957  
 958  $l \neq i$ , set  $u_{i(X)}$  equal to  $u_{l(X)}$ , where

959  
 960 
$$\kappa_1 = \sum_{k=1}^{p(X)} \{ (mG_0(A_{k,p}) + r\widehat{F}_r(A_{k,p})) \beta(x_i | k, p(X) - k + 1) \},$$

961  
 962 
$$\kappa_2 = \sum_{q \neq i} \beta(x_q | \theta(u_{q(X)}, p(X)), p(X) - \theta(u_{q(X)}, p(X)) + 1).$$

963  
 964  
 965  
 966 Otherwise, with probability  $\kappa_1 / (\kappa_1 + \kappa_2)$ , draw  $u_{i(X)}$  from the mixture distribution:

967  
 968 
$$\pi G_0(A_{k^*, p(X)}) + (1 - \pi) \text{Uniform}\{z_k \in (A_{k^*, p(X)})\},$$

969  
 970 with mixture probability defined by  $\pi = m / (m + r\widehat{F}_r((k^* - 1) / p(X), k^* / p(X)))$ , where  $k^*$  is a  
 971 random draw from a distribution on  $k = 1, 2, \dots$  that is proportional to

972  
 973 
$$\{ mG_0(A_{k^*, p}) + r\widehat{F}_r(A_{k^*, p}) \} \beta(x_i | k, p(X) - k + 1).$$

974  
 975 Also,  $\text{Unif}\{z \in (a, b]\}$  denotes a uniform distribution on a discrete set of values falling in a  
 976 set  $(a, b]$ . Each latent auxiliary variable  $u_{i(Y)}$  ( $i = 1, \dots, n(Y)$ ) is drawn from its conditional  
 977 posterior distribution in a similar manner (replace  $p(X)$  with  $p(Y)$ , the  $u_{i(X)}$ s with  $u_{i(Y)}$ s, and  
 978 the  $x_i$ s with  $y_i$ s).

979 Furthermore, up to a constant of proportionality, the full conditional posterior density of the  
 980 variables  $\{z_1, \dots, z_r\}$  is given by

981  
 982 
$$(z_1, \dots, z_r | \mathbf{w}_{p(X)}, \mathbf{w}_{p(Y)}, p(X), p(Y))$$
  
 983 
$$\propto \pi(z_1, \dots, z_r) \text{dir}(\mathbf{w}_{p(X)} | \alpha_{1,p(X)}, \dots, \alpha_{p(X),p(X)}) \text{dir}(\mathbf{w}_{p(Y)} | \alpha_{1,p(Y)}, \dots, \alpha_{p(Y),p(Y)}),$$

984  
 985 where  $\text{dir}(\mathbf{w} | \alpha_{1,p}, \dots, \alpha_{p,p})$  denotes the density function for the Dirichlet distribution, and

986  
 987 
$$\alpha_{k,p(X)} = mG_0(A_{k,p}) + r\widehat{F}_r(A_{k,p}) + n(X)\widehat{F}_u(A_{k,p}), \quad k = 1, \dots, p(X).$$

988  
 989 Also,  $\pi(z_1, \dots, z_r)$  denotes the density function for the first  $r$  samples from a Pólya-urn scheme  
 990 with parameters  $(m, G_0)$ ; see, for example, Blackwell and MacQueen (1973).

991 In particular, to generate a sample of each latent  $z_j$  ( $j = 1, \dots, r$ ) from its full conditional  
 992 posterior distribution, the following steps are taken. Let

993  
 994 
$$A_{1,t} = (c_1 = 0, c_2], \quad A_{2,t} = (c_2, c_3], \quad \dots,$$
  
 995  
 996 
$$A_{l,t} = (c_{l-1}, c_l], \quad \dots, \quad A_{t,t} = (c_{t-1}, c_t = 1],$$

997  
 998 denote sets formed by taking the union defined by

999  
 1000 
$$\{0, c_2, \dots, c_{t-1}, 1\} = \{k/p(X) : k = 1, \dots, p(X)\} \cup \{k/p(Y) : k = 1, \dots, p(Y)\},$$

1001 and let  $\widehat{F}_{r-1}$  denote the empirical distribution of  $\{z_d : d \neq j\}$ . Then a sample of each latent  $z_j$  is  
 1002 generated from the mixture distribution

$$1003 \pi_z G_0(A_{l^*,t}) + (1 - \pi_z) \text{Uniform}\{z_d \in A_{l^*,t}\},$$

1004 where  $\pi_z = m/(m + (r - 1)\widehat{F}_{r-1}(A_{l^*,t}))$  is the weight of the mixing distribution, and  $l^*$  is a  
 1005 random draw from a distribution on  $l = 1, \dots, t$  defined by

$$1006 \text{dir}(\mathbf{w}_{p(X)} | \alpha_{1,l,p(X)}, \dots, \alpha_{p(X),l,p(X)}) \text{dir}(\mathbf{w}_{p(Y)} | \alpha_{1,l,p(Y)}, \dots, \alpha_{p(Y),l,p(Y)}),$$

1007 with

$$1008 \alpha_{k,l,p(X)} = G_0(A_{k,p(X)}) + (r - 1)\widehat{F}_{r-1}(A_{k,p(X)}) + \mathbf{1}(A_{k,p(X)} \cap A_{l,t}),$$

$$1009 \alpha_{k,l,p(Y)} = G_0(A_{k,p(Y)}) + (r - 1)\widehat{F}_{r-1}(A_{k,p(Y)}) + \mathbf{1}(A_{k,p(Y)} \cap A_{l,t})$$

1010 for  $k = 1, \dots, p(X)$ , and for  $k = 1, \dots, p(Y)$ , respectively.

### 1011 Appendix III. The Loglinear Model Used for the Third Data Example

1012 As an Appendix to Section 3.3, here we provide details about the log-linear methods used for  
 1013 the kernel equating model, in the analysis of test scores arising from the non-equivalent groups  
 1014 design. As mentioned in that section, independent log-linear models were fit only to cells without  
 1015 structural zeros. These models are jointly defined by

$$1016 \log \widehat{g}_{X,V_1}(x_k^*, v_l^*) = \alpha_X + \sum_{t=1}^{T(X)=4} \widehat{\lambda}_X(x_k^*)^t + \sum_{t=1}^{T(V_1)=2} \widehat{\lambda}_{V_1}(v_l^*)^t$$

$$1017 + \sum_{t=1}^{I(X)=3} \sum_{t'=1}^{I(V_1)=2} \widehat{\lambda}_{X,V_1,t,t'}(x_k^*)^t (v_l^*)^{t'},$$

$$1018 \log \widehat{g}_{Y,V_2}(y_k^*, v_l^*) = \alpha_Y + \sum_{t=1}^{T(Y)=4} \widehat{\lambda}_Y(y_k^*)^t + \sum_{t=1}^{T(V_2)=4} \widehat{\lambda}_{V_2}(v_l^*)^t$$

$$1019 + \sum_{t=1}^{I(Y)=3} \sum_{t'=1}^{I(V_2)=1} \widehat{\lambda}_{Y,V_2,t,t'}(y_k^*)^t (v_l^*)^{t'}.$$

1020 The first fitted model presented above was selected after using Akaike's Information Criterion  
 1021 (AIC; Akaike, 1973) to compare the goodness-of-fit of 160 log-linear models described by com-  
 1022 binations of the vector of values for  $(T(X), T(V_1), I(X), I(V_1))$ , for  $T(X), T(V_1) \in \{1, \dots, 4\}$   
 1023 and  $I(X), I(V_1) \in \{0, 1, 2, 3\}$  (and  $I(X) = I(V_1)$  whenever  $I(X)$  or  $I(V_1)$  is zero), and similarly  
 1024 for the second fitted model presented above. The AIC was 256.06 and 229.14 for the selected log-  
 1025 linear models for the first and second contingency tables, respectively. Then conditional on the  
 1026 maximum likelihood estimates of  $(\widehat{G}_{X,V_1}, \widehat{G}_{Y,V_2})$  obtained via these best models, the estimate of  
 1027 the discrete test score distributions  $(\widehat{G}_X, \widehat{G}_{V_1}, \widehat{G}_Y, \widehat{G}_{V_2})$  are obtained via marginalization. Then  
 1028 the bandwidth estimates  $(\widehat{h}_X = .58, \widehat{h}_Y = .61, \widehat{h}_{V_1} = .55, \widehat{h}_{V_2} = .59)$  were obtained using their  
 1029 least-squares minimization method described in Section 1. Finally, a reviewer points out that in-  
 1030 stead, more elaborate log-linear models can be used for the data. These models would include up  
 1031 to four cross-product moments, as described in von Davier et al. (2004, Chap. 10).

PSYCHOMETRIKA

References

1051  
 1052  
 1053 Akaïke, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csaki  
 (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.  
 1054 Angoff, W. (1971). Scales, norms, and equivalent scores. In R. Thorndike (Ed.), *Educational measurement* (2nd ed.,  
 1055 pp. 508–600). Washington: American Council on Education.  
 1056 Blackwell, D., & MacQueen, J. (1973). Ferguson distributions via Polya urn schemes. *Annals of Statistics*, *1*, 353–355.  
 1057 Dalal, S., & Hall, W. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statis-  
 tical Society, Series B*, *45*, 278–286.  
 1058 De Iorio, M., Müller, P., Rosner, G., & MacEachern, S. (2004). An ANOVA model for dependent random measures.  
 1059 *Journal of the American Statistical Association*, *99*, 205–215.  
 1060 Diaconis, P., & Ylvisaker, D. (1985). Conjugate priors for exponential families. *Annals of Statistics*, *7*, 269–281.  
 1061 Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.  
 1062 Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209–230.  
 1063 Hjort, N., & Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In V. Nair (Ed.), *Advances  
 in statistical modeling and inference: essays in honor of Kjell Doksum*. Singapore: World Scientific. Chap. 23.  
 1064 Holland, P., & Thayer, D. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal  
 of Educational and Behavioral Statistics*, *25*, 133–183.  
 1065 Joe, H. (1997). *Multivariate models and dependence concepts* (2nd ed.). Boca Raton: Chapman and Hall/CRC.  
 1066 Karabatsos, G., & Walker, S. (2007). *A Bayesian nonparametric approach to test equating*. Presented at the national  
 council of measurement in education, Chicago, April 2007.  
 1067 Karabatsos, G., & Walker, S. (2009, to appear). Coherent psychometric modeling with Bayesian nonparametrics. *British  
 Journal of Mathematical and Statistical Psychology*.  
 1068 Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York:  
 1069 Springer.  
 1070 Lorentz, G. (1953). *Bernstein polynomials*. Toronto: University of Toronto Press.  
 1071 Müller, P., & Quintana, F. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*, 95–110.  
 1072 Müller, P., Quintana, F.A., & Rosner, G. (2004). A method for combining inference across related nonparametric  
 Bayesian models. *Journal of the Royal Statistical Society, Series B*, *66*, 735–749.  
 1073 Petrone, S. (1999). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, *26*, 373–393.  
 1074 Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.  
 1075 Von Davier, A., Holland, P., & Thayer, D. (2004). *The kernel method of test equating*. New York: Springer.  
 1076 Walker, S. (2004). New approaches to Bayesian consistency. *Annals of Statistics*, *32*, 2028–2043.  
 1077 Walker, S., & Muliere, P. (2003). A bivariate Dirichlet process. *Statistics and Probability Letters*, *64*, 1–7.  
 1078 Walker, S., Damien, P., Laud, P., & Smith, A. (1999). Bayesian nonparametric inference for random distributions and  
 related functions. *Journal of the Royal Statistical Society, Series B*, *61*, 485–527.  
 1079 Walker, S., Lijoi, A., & Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional  
 models. *Annals of Statistics*, *35*, 738–746.  
 1080  
 1081  
 1082  
 1083  
 1084  
 1085  
 1086  
 1087  
 1088  
 1089  
 1090  
 1091  
 1092  
 1093  
 1094  
 1095  
 1096  
 1097  
 1098  
 1099  
 1100