The comparison of the scores obtained by Bayesian nonparametric model and classical test theory methods*

Meltem Yurtcu¹, Hülya Kelecioglu² and Edward L Boone³
¹Division of Measurement and Evaluation in Education, Department of Educational Sciences, İnönü University, Malatya, Turkey
²Division of Measurement and Evaluation in Education, Department of Educational Sciences, Hacettepe University, Ankara, Turkey
³Department of Statistical Science and Operations Research, Virginia Commonwealth University, Richmond, VA, USA

Abstract
Bayesian Nonparametric (BNP) modelling can be used to obtain more detailed information in test equating studies and to increase the accuracy of equating by accounting for covariates. In this study, two covariates are included in the equating under the Bayes nonparametric model, one is continuous, and the other is discrete. Scores equated with this model were obtained for a single group design for a small group in the study. The equated scores obtained with the model were compared with the mean and linear equating methods in the Classical Test Theory. Considering the equated scores obtained from three different methods, it was found that the equated scores obtained with the BNP model produced a distribution closer to the target test. Even the classical methods will give a good result with the smallest error when using a small sample, making equating studies valuable. The inclusion of the covariates in the model in the classical test equating process is based on some assumptions and cannot be achieved especially using small groups. The BNP model will be more beneficial than using frequentist methods, regardless of this limitation. Information about booklets and variables can be obtained from the distributors and equated scores that obtained with the BNP model. In this case, it makes it possible to compare sub-

*Some of this article was presented as oral presentation at 6th International Congress on Measurement and Evaluation in Education and Psychology, Prizren, Kosovo in 5–8 September 2018.

Corresponding author:
Meltem Yurtcu, Division of Measurement and Evaluation in Education, Department of Educational Sciences, İnönü University, Malatya 44280, Turkey.
Email: meltem.yurtcu@gmail.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
categories. This can be expressed as indicating the presence of differential item functioning (DIF). Therefore, the BNP model can be used actively in test equating studies, and it provides an opportunity to examine the characteristics of the individual participants at the same time. Thus, it allows test equating even in a small sample and offers the opportunity to reach a value closer to the scores in the target test.

**Keywords**

Single group design, Bayesian nonparametric model, test equating, covariates, equated scores

**Introduction**

Tests assessing content areas are frequently used in the education systems to evaluate students. Although the tests appear to cover the same content, they do not necessarily allow for a direct comparison, as the questions may differ according to the testing tools. Test equating studies allow a comparison to be made as they can quantify directly comparable student performance. Test equating methods are based on different theories such as the Classical Test Theory (CTT) and the Item Response Theory (IRT).\(^1\)\(^-\)\(^4\) These two theories differ from one another according to the level of providing assumptions they depend on. The IRT considers the probability of a correct answer given to a test item according to an ability parameter and item statistics.\(^5\) In the CTT, the test score, real score and error score parameters are taken into consideration.\(^6\) Both models have some limitations. These limitations are the dependence on the group and the test in CTT. In IRT, the model is considered as a complex structure, and it is difficult to apply in practice. Measurement precision is the same in CTT, for all individuals regardless of individual characteristics. In IRT, it is calculated depending on the latent properties.\(^6\) Therefore, many researchers argue that the IRT method is superior to the CTT method.\(^6\)\(^-\)\(^10\) According to the results obtained from test equating studies, there are studies supporting that IRT-based methods give better results than CTT-based equating methods.\(^11\),\(^12\) However, the good results of IRT approaches depend on the sample size.\(^6\),\(^13\) In the literature, it is possible to see studies showing that test equating based on CTT gives better results than IRT in small samples.\(^14\) Even among the CTT equating methods, it has been observed mean and linear equating methods give better results in small samples compared to equal percentage equating methods.\(^15\) The Bayesian model, which is based on probability, gives better results than the estimates made according to these theories and provides more detailed information about the research.\(^16\)\(^-\)\(^22\) The Bayes\(^2\) statistical method has a flexible structure and will contribute to the development of educational research. Recently, an equated scores approach applying Bayesian modelling has been used.\(^16\),\(^23\),\(^14\) Bayesian and traditional (classical) approaches both deal with unknown parameters in a statistical model.\(^25\),\(^26\) In the analysis of the traditional approach, only one parameter is considered as unknown, while in the Bayes approach, all unknown parameters are treated as ambiguous.\(^21\) Therefore, in the approach to the concept of parameters according to classical methods, the parameters are considered to be constant. According to the Bayesian approach, each of
the parameters is accepted as a probability variable and is based on obtaining inferential information using previous information. Unlike classical methods, the Bayesian approach gives more relevant posterior inferences by including prior knowledge in the study to quantify certain uncertainties. The Bayesian approach provides many advantages, and a few are mentioned below:

1. can be used in complex models that cannot be estimated with classical approaches
2. to calculate probability within the confidence interval
3. to include historical information in the analysis when necessary
4. to analyse cases where there is not a large sample

In addition to these significant advantages, the necessity of meeting the assumptions required for some analyses (such as sample size for many analyses, continuity or smoothing for equating) is eliminated with the Bayesian approach. However, the Bayesian method also has some disadvantages. The most important of these is the selection of the a priori in the model. Choosing an a priori that gives false information can mislead the estimation results. In addition, the presence of too many parameters in the Bayesian model causes confusion and increases calculation costs.

The test equating results can also be influenced by variables that belong to individual personalities and situations. These variables constitute most of the information known as a priori in Bayesian statistics. Bayesian approaches are well-defined probability models that allow for observed data and unknown values. Parametric and semi-parametric equating methods are used to convert two or more tests into one result. Since the Bayes parametric model uses a limited number of parameters, it has some limitations. The basis of Bayes nonparametric models lies in the flexible use of the number of parameters in the models. Thus, the Bayesian nonparametric model could allow the use of covariates to estimate functions for equating transformations. With the inclusion of these variables in the process, changes occur in the transformation functions, and equating between the categories of variables becomes possible. According to the BNP, the equivalent of the score from the first test in the second test does not have the bijection feature as in the previous equating methods. In the BNP approach, there is considerable difficulty in choosing prior distributions that will ensure the interdependence between response data and covariates. It is very important to select the correct covariates with the BNP model as they can help reduce bias in any model as long as they explain the differences between groups and relate to test scores. Age, gender, educational status, type of school, amount of time the tests are administered, average school performance, and grade are the variables generally used as covariates in the literature. Covariates should be carefully chosen to have a significant effect on the selected feature. As the number of categorical variables increases, the number of individuals falling into each category decreases. Therefore, using a less categorical variable
or reducing the number of categories may be needed. Covariates in a linear equat-
ing method were applied in the classical methods. The observed variables were
combined with the past variables of the individual using the linear regression
model to create a relationship between the two test scores. To create this model
based on the maximum likelihood estimation method, test scores are expected to
provide assumptions, such as normality. However, it is very difficult to make this
assumption in small samples. Here, using the BNP model could be more useful
than using frequentist methods.

In BNP models, covariates are treated as dependent random probability vari-
ables. The covariates to be included in the process a priori are the Bernoulli distri-
bution if it is dichotomous, or the Dirichlet distribution can be used if it is
continuous and has a range. The Dirichlet process is formed as an extension of
the Dirichlet distribution in continuous space. Dirichlet Process (DP) Models play
a central role in nonparametric Bayes inferences. One of the biggest reasons for
this central role is that there are many variations of the Dirichlet processes. One of
the ways of displaying the Dirichlet distribution is Sethuraman’s Dirichlet process
definition. The main point in these Sethuraman representations is the definition of
a discrete variable related to the location of each function. The effect of the cov-
ariates in the distribution of test scores will reveal any dependencies that can be
explained using the Dependent Dirichlet Process (DDP), which is an extension of
the Dependent Process (DP) model. In later studies, Barrientos, Jara and Quintana expanded the model and built the Dependent Bernstein polynomial
process (DBPP) model. Barrientos et al. discuss two specific cases of DBPP. A
dependent stick-breaking process with common weights and predictor-dependent
support points was used in this study, known as a single weight DBPP (wDBPP).
This model provides a covariate-dependent equation transformation. Thus, the
equated test score can be obtained by combining the scores of the individuals with
the covariates.

The credible interval used in the Bayes method carries more information than
point estimation. Credible intervals for posterior inferences produced by the BNP
method can be summarized with values such as average and 95% credible interval
(highest density interval, HDI). The points within this range are more accurate than
the points that fall outside.

Test equating studies are generally carried out without covariates in classical test
equating methods. However, this leads to the conclusion that the individuals who
were getting the same score fall in the same category of variables, although they
have unique characteristics. For example, for the tests to be equated, the scores of
the two students who got 60 points from the first test may differ according to their
gender or their attitude to the course. In addition, the scores of a female student
who likes mathematics very much and a male student who likes mathematics at a
moderate level can be compared. Thus, instead of evaluating only on equated
scores as in classical methods, comments can be made on the scores obtained
according to the characteristics of individuals. For this reason, including covariates
in the analysis can help to reduce measurement errors. Test equating is very
important to compare the scores of individuals who took different tests at different times and be able to use the scores interchangeably. Different equating methods are proposed to achieve this. Every equating method has assumptions to meet. The accuracy of the estimates increases as long as the assumptions are provided. Reaching correct estimates may allow a fair comparison to be made for equating, especially in small samples. This research is based on comparing equated scores obtained by the BNP method with the equated scores obtained from linear and mean equating methods in classical test theory in small groups. In the BNP model of this research, two covariate variables were used. The study was carried out using the single group design, in which two or more test papers are applied to the same group. To prevent the bias that may arise from the sample size in the comparison of the methods, mean and linear equating methods, which are less sensitive to sample size, are preferred to the CTT models. The combination of these factors makes this study different from other studies. There is no need to make an assumption about sample size in the BNP model. The 2011 booklet is taken as the equated test, which is also a rescaled test. The difference between the scores obtained by three test equating methods and the target test scores is considered the comparison criterion. Also, in this study, the credible intervals of the equated scores of the BNP model and the differences between the groups were examined as an advantage of the BNP model.

Method

This study compares two approaches, and the aim is to compare the equated scores obtained from the BNP model with the outcomes of classical test theory equating methods. Two different (mean and linear) equating methods are discussed for the classical theory. The distance between the equated scores obtained from mean, linear equating methods and the BNP model to the target test score, were then analysed.

Data used in research

Twenty questions from the mathematics section of the Placement Test (PT, the Turkish acronym is SBS) carried out by the Ministry of National Education of Turkey (MNE) for 8th graders in 2011 and 2012 were used for data collection. The data were obtained with the permission of the ministry. These questions were given to 102 8th grade students in a public school in Samsun in the spring semester of the 2015–2016 academic year. Data were collected using convenience sampling, which is based on ease of accessibility. The sample consisted of 51 female and 51 male students.

In the BNP model, two covariates were used to obtain the equated score: gender and attitude scale towards mathematics. The Attitude Scale towards Mathematics, developed by Önal is comprised of 22 items, consisting of 11 positive and 11 negative questions. The scale is the 5-Likert, where ‘Totally Agree’, ‘Agree’,
Undecided’, ‘Disagree’ and ‘Strongly Disagree’ are used and the scale consists of four factors: interest, anxiety, work and necessity. The internal consistency coefficients of the scale’s factors are 0.89, 0.74, 0.69 and 0.70, respectively. Negative items are inverted, and the total score obtained from the scale is used in the study. High scores indicate individuals show positive attitudes and vice versa.39

In the literature, students’ attitudes towards mathematics are generally evaluated as positive and negative.40–43 Added to these categories, a neutral attitude towards mathematics is also discussed.44 In this study, attitude towards mathematics is considered as a continuous variable in the analysis process. Representational situations for three different mathematical attitude levels are also included to facilitate understanding, comparison and interpretation of the model.

Data analysis. The equation conditions on the 8th grade’s PT mathematics subtests used in the study are tested for the participants. Equated scores are obtained according to Classical Test Theory equating methods and the BNP model. Total scores are used in equating PT mathematics subtests consisting of 20 questions. Test equation conditions, unidimensionality, the difference between variances and average difficulties, and the difference between reliability were tested, and then the equated scores were obtained. To obtain equated scores in the BNP model, codes were created based on the DBPP model and test equating process in the R.3.2.1 programme using the model proposed by Gonzalez et al.23,24 The equations of the process are:

\[ G \text{ represents DBPP, for } G = \{G_x : x \in X\}; \]

\[ g_x(\cdot) = \sum_{j=1}^{\infty} w_j(x) \beta\left(\cdot \mid [k \theta_j(x)], k - [k \theta_j(x)] + 1\right), \]

\[ \lceil \cdot \rceil \text{ represents the ceiling function} \]

In this formulation

\[ \theta_j(x) = h_x(z_j(x)), H = \{H_z : z \in Z\} \]

\[ w_j(x) = \nu_x \{n_j(x)\} \prod_{l<j} [1 - \nu_x \{n_j(x)\}] \]

and \( k \) is a control parameter for the prior distribution. \( V = \{v_x : x \in X\} \) and \( H = \{h_x : x \in X\} \) are bijective continuous functions for every \( x \in X \). \( H = \{h_z : z \in Z\} \) as so \( v_1, v_2, \ldots \) and \( \alpha > 0 \) are independent and defined random variables defined by \( \beta(1, \alpha) \) distribution; \( k \) is a discrete random variable with distribution indexed by a finite-dimensional parameter \( \lambda \); \( \theta_j(z) = h_z(r_j(z)), r_1, r_2, \ldots \) and \( \eta_j = \{\eta_j(x) : x \in X\} \) for \( j \geq 1 \) are independent and identically distributed real-valued stochastic processes with law indexed by the parameter \( \psi \).

As \( Z \) represents covariate space, \( F_z \) represents covariate-dependent random probability distributions. For \( z \in Z, \{F_z : z \in Z\} \), the formulation of \( wDBPP \) is below:
This model, representing an infinite set of beta distributions, states that the test scores have covariate-dependent sample densities. This model can be demonstrable as $\{F_z; z \in Z\} \sim \text{wDBPP}(\alpha, \lambda, \psi, H)$, providing a covariate-dependent equation transformation. $^24$

Prior distribution specification. The application of the wDBPP model is based on the Makov Chain Monte Carlo (MCMC) methods. For wDBPP, $h_z(\cdot) = \exp\{\cdot\}/(1 + \exp\{\cdot\})$, $r_j(z) = z^T\gamma_j$ and $\gamma_j \sim N_p(\mu, S), j = 1, 2, \ldots$.

Here

$$v_j | \alpha \sim \beta(1, \alpha), k | \lambda \sim \text{Poisson}(\lambda) \quad (k > 1),$$

$$\mu | m_0, S_0 \sim N_p(m_0, S_0), S | \nu, \psi \sim IW_p(v, \psi)$$

$IW_p(v, A)$; scale matrix A, represents the $p$-dimensional inverted-Wishart distribution with $\nu$ degrees of freedom. This study used the fitted values of the model found in the studies of Gonzalez et al. $^{23,24}$ These values are $\lambda = 25$, $m_0 = 0_p$, $S_0 = 2.25 \times I_p$, $\nu = p + 2$ and $\alpha = 1$. To explain the posterior distribution of the wDBPP model and obtain the posterior distribution samples of all model parameters, the MCMC algorithm was applied.

Posterior inference. All computations were coded and performed in the R 3.2.1 statistical programme. $^45$ The posterior distribution is given by:

$$p(v, k, w, \gamma | y, z)$$

$$\propto \prod_{i=1}^{n} \sum_{j=1}^{10} w_j \beta \left( \frac{e^{z^T\gamma_j}}{1 + e^{z^T\gamma_j}}, k - \frac{e^{z^T\gamma_j}}{1 + e^{z^T\gamma_j}} + 1 \right) \prod_{j=1}^{10} \beta(v_j | 1, 1)$$

$$\times \left[ \frac{25^k e^{-25}}{k!(1 - e^{-25})} \right] \left\{ \prod_{j=1}^{10} (2\pi)^{-\frac{1}{2}} e^{-0.5(\gamma_j - \mu)^T S^{-1}(\gamma_j - \mu)} (2\pi)^{-\frac{1}{2}} e^{-0.5 m_0^T S_0^{-1} m_0} \right\}$$

$$\times \frac{|\psi|^2}{2^{\frac{5}{2}} \Gamma_\nu(2)} |S|^\frac{\nu}{2} e^{-\frac{1}{2} \psi S^{-1}}$$

To create a posterior predictive distribution the following was used:

$$p(T | y, z) = \int p(v, k, w, \gamma | y, z) L(T | v, k, w, \gamma) dv \, dk \, dw \, d\gamma$$

where
\[ L(T|\nu, k, w, \gamma) = \sum_{j=1}^{10} w_j \beta \left( T \left[ k \frac{e^{z_j^T \gamma}}{1 + e^{z_j^T \gamma}} \right], k - \left[ k \frac{e^{z_j^T \gamma}}{1 + e^{z_j^T \gamma}} + 1 \right] \right). \]

\[
\begin{align*}
\mathbf{t}_x &= F^{-1}_x(\cdot) \\
\mathbf{t}_y &= F^{-1}_y(\cdot) \\
\mathbf{t}_z &= F^{-1}_z(\cdot) \\
\mathbf{t}_w &= F^{-1}_w(\cdot) \\
\mathbf{t}_k &= F^{-1}_k(\cdot) \\
\mathbf{t}_g &= F^{-1}_g(\cdot)
\end{align*}
\]

This shows the sum obtained for the determined distributions.

To show the T score distributions; \( t_x \) scores obtained from the \( X \) test on \( T \), \( t_y \) scores obtained from \( Y \) test on \( T \) and \( z \) are considered as covariates. The transformation functions are as follows:

For the two-year exams, the Metropolis-Hastings sampler was used to obtain MCMC samples from the posterior distribution. Initially, 5,000 MCMC samples were taken to assess the sampler’s convergence and tune the sampler to improve acceptance rates. These samples were discarded as burn-in samples. This was followed by taking 110,000 samples from the tuned sampler. All inferences were made using these samples. Summaries of the parameters are given in Supplemental Appendix 2. After sampling, equated scores were obtained using the cumulative posterior predictive distributions. The codes used in the process for one of the booklets are given in Supplemental Appendix 3. In the study, credible intervals were determined, and the intervals for the predicted scores were examined. The credible intervals were calculated using the 2.5%, 50% and 97.5% quantiles from the posterior instead of the HDI.

**Results**

The average scores of the test applied in 2011 and 2012 were 59.07 and 57.02, respectively.

**Equations based on the classical test theory**

According to classical test theory, unidimensionality, difference between variance, average difficulty and reliability of the tests were examined in the equating performed according to classical test theory.

**Unidimensionality.** As a result of the factor analysis performed for unidimensionality, the Kaiser-Meyer-Olkin (KMO) value obtained from the mathematics subtest is 0.822 for 2011 and 0.772 for 2012. These KMO values indicate that the data is suitable for factor analysis. Chi-square statistics obtained from the Barlett test are significant (\( p = 0.00 < 0.05 \)) for the subtests of both years. This result supports KMO values and is another indicator of the suitability of the data for factor analysis.
Testing the difference between variances and average difficulties. Levene’s test was used to examine the variance of the tests, and there is no significant difference between the variances of the subtests. To test whether there is a significant difference between the average difficulties of the subtests, the t statistic was used. The difficulty levels of the two subtests were different at a 0.05 significance level.

Testing the difference between the reliability of subtests. KR-20 reliability coefficients for the 2011 and 2012 subtests were computed. The Fisher Z transformation correlation coefficient was converted to the normal distribution’s Z variable, and the difference between the correlations was examined. It was found the subtests for each year had the same reliability.

Examined equation conditions were provided for the classical test theory equating methods. For mean and linear equating methods, the difference between the equated score and score on the target test is shown in Supplemental Appendix 1.

Test equating with BNP model

This article uses the BNP model for the test equation, allowing the use of covariates to predict the score distribution, controlling group invariance. Thus, the BNP model for equating does not require the same checking of conditions found in other equating methods. The of the student distributions were obtained according to the covariates used in the model. The points distribution for the mathematics subtests for female students is given in Figure 1, representing the three different mathematical attitude levels. Other distributions can be obtained according to the characteristics of each participant. Figure 1 includes only representative graphs of the three levels. Since the distributions of individuals at the mentioned level will be similar universally, the changes in their distribution in different booklets will reveal the difference between the booklets.

The scores of female students’ attitude levels are concentrated in the middle scores, and the intensity varies according to the year. The score distribution of female students who have a neutral attitude towards mathematics and took the exam in 2011 is more pronounced than the distribution at other levels. Although the distribution of female students who have positive or negative attitudes towards mathematics is similar, the students with a positive attitude towards mathematics are observed to have higher densities in high scores.

In 2012, the density of average scores increased in the distribution of female students according to their mathematics attitude compared to 2011. The density of high scores also increased among female students having a positive attitude towards mathematics. In contrast, the density of low scores increases among those with a negative attitude in the 2012 booklet. Since the score distributions of female students vary according to their attitudes towards mathematics, it can be said that attitude towards mathematics explains mathematical achievement. It should be noted that the score distribution of those who have a neutral attitude towards mathematics remained the same. Since the individuals with the same characteristics showed different distributions in different years, it is possible to comment on the
difficulty or discrimination levels of the booklets. In this case, it can be said that the booklet of 2012 contains more distinctive and more straightforward questions for female students.

Credible intervals of the female students’ score distribution varied according to their attitude towards mathematics in both years, but the intervals are generally wide. A wide credible interval indicates uncertainties in the estimates for the values falling in these ranges. These uncertainties can be explained because the

Figure 1. Posterior mean density and 95% posterior credible density intervals for female students showing neutral, positive and negative attitude towards mathematics according to the booklets.
participants with specific profiles are not present in the sample or the questions failed to distinguish these individuals.

Equated scores were computed using an equating function after combining the covariates and the scores obtained from the tests. The score distribution was obtained for each sub-group to observe the difference between the booklets. The graphs of female students’ equated scores according to their attitude towards mathematics, representing three levels, are given in Figure 2. This shows that female students who showed negative attitudes towards mathematics in equating among the booklets would get high scores in the target test towards the middle scores and low scores towards the high scores. If they have a positive attitude towards mathematics, their scores in the target test would be higher than the scores from the equated test. However, if they showed a negative attitude, they would show low scores. Considering the distributions and equated scores, the scores obtained in the 2012 booklet were low even if there was a slight difference from the scores in 2011. Therefore, it is possible to state that the booklet in 2012 contained slightly more complex questions than the 2011 booklet. In the 2012 booklet, it was observed that female students who showed positive attitudes towards mathematics

\textbf{Figure 2.} Equating of female students with the BNP model according to their neutral, positive and negative attitudes towards mathematics.

A = Equated test (2011).

B = Equated scores; The expected score from the target test (2012).
increased their probability of getting high scores, while those who showed negative attitudes towards mathematics increased the likelihood of getting low scores. Therefore, we can say that the booklet in 2012 contains more distinct questions in determining the attitudes of female students towards mathematics.

The distribution of male students is shown in Figure 3.

Increased density of high scores was shown by the distributions obtained for male students in the 2011 booklet. When attitudes towards mathematics were positive, this density was higher than for the other booklet. It was observed that male students who have negative attitudes towards mathematics are more likely to have lower scores in the 2012 booklet than in the other exam. Here, it is possible to state that the questions in the 2012 booklet were more difficult for male students.

Figure 3. Posterior mean density and 95% posterior credible density intervals for male students showing neutral, positive and negative attitude towards mathematics according to the booklets.
compared to 2011. In both booklets, when the attitude towards mathematics was examined separately according to the levels, it can be stated that this variable explains the mathematics achievement as the students’ scores changes. In addition, it is possible to say that this variable can reveal the difference between the exams by creating different distributions according to booklets. When the credible intervals of male students were examined, the graphs showed that the widest range had a neutral attitude towards mathematics. Credible intervals increased at the highest scores for the students who showed positive attitudes towards mathematics and at the lowest scores for the students who showed negative attitudes towards mathematics. It is possible to make similar comments to the deductions made for the female students as to uncertainty due to the width of these confidence intervals. After the distributions with the BNP model were obtained, many different comparisons could be made. Graphs of equated scores obtained by equating between booklets for male students are shown in Figure 4.

The expected scores in the target test were generally lower than the equated test for the male students. In particular, it was possible to see this difference in the high scores for male students who had a neutral and positive attitude towards mathematics.
mathematics. This difference was clear between the middle and low scores for the students who exhibited a negative attitude towards mathematics.

When the distributions and equated scores are taken into consideration, it is possible to say that the exam that was conducted in 2011 was both easier and contained more distinct questions for male students. According to attitude levels, male students have different distributions in both booklets. This shows that attitude towards mathematics was a variable that can explain mathematical success. In addition, it can be stated that this variable partially explains the difference between these booklets because the distributions of male students differed at all levels in the booklets.

The differences between the booklets and groups were obtained by considering the comparability feature of the BNP model according to sub-categories. Based on this information, equated scores were obtained for each individual with different levels of mathematics attitude and gender. In addition, equated scores were derived according to the mean and linear equating methods in the classical test. The distributions of the students’ equated scores by each method are given in Figure 5.

Regarding the densities of equated scores, the standard deviations of the equated scores are: $sd_{mean} = 5.3057$; $sd_{linear} = 4.8215$ and $sd_{Bayes} = 4.7998$. The distribution of the scores obtained from the target test is shown in black. The distribution of the equated scores obtained from the BNP model appears to be closest to the target test. In addition, equated scores obtained from these three methods are examined according to their distance from the target test. These differences are given in Supplemental Appendix 1.

Figure 5. Distribution of equated scores by mean, linear equating methods, using the BNP model and distribution of the target test.
Conclusion and discussion

More than one form of a test with the same content and statistical features can be used in large-scale studies, mainly to ensure reliability. To make correct decisions about individuals and compare them, test equating was made between these examinations. However, although these tests are called parallel, they differ slightly from each other. In addition, test equating is used to ensure the interchangeability of different examinations applied to the same individuals.\textsuperscript{46} Even if individuals with the same characteristics completed these test papers that measure the same structure, a slight difference in the psychometric properties of the tests might cause significant differences in test results. Individuals do not always reflect their capabilities, and the differences in the difficulties of the tests affect these scores. This leads to misinterpretation in the comparison of test results. Therefore, test equating studies are carried out to eliminate comparison errors. Individual characteristics that cause individuals to obtain different scores can be considered as a covariate. The BNP model helps to provide convenient explanations of these differences. The requirement of obtaining booklets with the same level of difficulty and level of discrimination was challenging. The difference between the scores of the individuals is thought to have been caused by the differences in the booklets. Past variables of the individuals can show tolerance of the difference between booklets and make it easier for individuals to compare their scores. Therefore, the second test score could be estimated according to the points that individuals could get in the first test. Thus, the differences between the score obtained from the second and predicted scores can be examined. The estimation methods used in classical methods can also be affected by the characteristics of the group. In this respect, using a model that can make more accurate predictions and provide detailed information by including the characteristics of individuals in the analysis process will be an important step in the test equating process.

In this study, equated scores were obtained using the BNP model, a new approach that facilitates the inclusion of variables in the test equating process. Two different variables were used to describe students' characteristics. Gender was a discrete variable, and the attitudes of individuals towards mathematics were used as a continuous variable during the analysis. Representational situations were submitted to facilitate understanding and comparing three different mathematical attitude levels—positive, negative and neutral. To make a comparison, linear equating methods and mean equating methods were used from the classical test theory, in which various assumptions are checked before computing equated scores. The BNP model with covariates for the test equation controls group invariance. Thus, the BNP model for equating does not require the same conditions that need to be checked in other equating methods.\textsuperscript{24} This research examines the comparison of the BNP model and CTT in test equating in small groups. The combination of these factors makes this study different from other studies. Because of the comparisons, the best results are the equated scores obtained from the BNP model. This shows that the Bayesian model provided better results than classical tests in a small sample.
Studies that state that the Bayes method makes better predictions than classical methods and provides more helpful information to support this result.\textsuperscript{16,17,21}

Two advantages of using the BNP model in test equating analyses are shown in this study. First, the BNP model could make robust predictions when assumptions were not met or were missing. Thus, it allowed test equating even in a small sample and provided an opportunity to reach a value close to the target test scores. The second advantage was the simplicity of including the characteristics of the individuals in the model as a covariate. There are many studies in the literature showing that using covariates in the Bayesian approach gives better results and reduces the equating error in cases where the assumptions of the classical theory are not fully met, or the sample size is not sufficient.\textsuperscript{20,47} In addition, unlike the classical model, using the correct covariates in the BNP model allows the analysis of the characteristics of included variables according to their subcategories.\textsuperscript{19,20,22,25,34,35}

The density of high scores in both booklets was shown to increase among students who had a positive attitude towards mathematics, and the density of low scores increased among students who showed a negative attitude. The high score density of male students was higher than female students, which can be expressed as male students had better scores than female.

Figure 1 shows female students were more successful in 2012, whereas Figure 3 shows that male students were more successful in 2011. This outcome indicates that there may be gender-biased questions in the booklets. Comments can be made about the direction of DIF regarding the distributions. Therefore, the presence of DIF in the booklets can be tested as a result of the test equating performed by the BNP model. The DIF studies in the literature identified unfair questions in 2011\textsuperscript{48,49} and 2012 booklets\textsuperscript{50} of the mathematics subtest. Regarding psychometric properties of the booklets, it can be said that the exam conducted in 2011 was more straightforward and the booklet applied in 2012 contained more distinct questions for female and male students.

In conclusion, the results obtained using the BNP model are closest to the target test. More detailed analyses can be performed by including gender and attitude towards mathematics as covariates. This eliminates the negativity arising from accepting individuals with different characteristics similar to occurrences in classical methods. Researchers who want to obtain information about subgroups and research cases where assumptions are not met can carry out equating studies using the Bayesian approach. Many characteristics can be added to the model as covariates. This model allows for the comparison of features such as difficulty and discrimination between booklets.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Meltem Yurtcu https://orcid.org/0000-0003-3303-5093

Supplemental material
Supplemental material for this article is available online.

References
1. Ryan J and Brockmann F. A practitioner’s introduction to equating with primers on classical test theory and item response theory. Washington, DC: CCSSO, 2009.
2. von Davier A. Statistical models for test equating, scaling, and linking. New York, NY: Springer Verlag, 2011.
3. Kolen MJ and Hendrickson AB. Scaling, norming, and equating. In: Secolsky C and Denison BD (eds) Handbook on measurement, assessment, and evaluation in higher education. New York, NY: Routledge, 2013, pp.257–276.
4. Hambleton RK and Swaminathan H. Item response theory: principles and applications. Baston: Kuluwer-Nijhoff Publishing, 1985.
5. Gonzalez J. SNSequate: standard and nonstandard statistical models and methods for test equating. J Stat Softw 2014; 59(7): 1–30.
6. Hambleton RK and Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educ Meas Issues Pract 1993; 12(3): 3847.
7. Fan X. Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educ Psychol Meas 1998; 58(3): 357–381.
8. Jabrayilov R, Emons WHM and Sijtsma K. Comparison of classical test theory and item response theory in individual change assessment. Appl Psychol Meas 2016; 40(8): 559–572.
9. Reise SP and Haviland MG. Item response theory and the measurement of clinical change. J Pers Assess 2005; 84: 228–238.
10. Prieler JA. So wrong for so long: changing our approach to change. Psychologist 2007; 20: 730–732.
11. Yang WL. The effects of content homogeneity and equating method on the accuracy of common Item test equating. Unpublished doctoral dissertation, Michigan State University, Michigan, 1997.
12. Peterson NS, Cook LL and Stocking ML. IRT versus conventional equating methods: a comparative study of scale stability. J Educ Stat 1983; 8(2): 137–156.
13. Dorans NJ, Pommerich M and Holland PW. Linking and aligning scores and scales. In: Dorans NJ, Pommerich M and Holland PW (eds) Statistics for social and behavioral sciences. New York, NY: Springer, 2007.
14. Mead AD and Meade AW. Test construction using CTT and IRT with unrepresentative samples. In: Paper presented at the annual meeting of the society for industrial and organizational psychology, Atlanta, GA, USA, 8–10 April 2010, p.56. SIOP.
15. Tsai TH. Estimating minimum sample sizes in random groups equating. In: Paper presented at the annual meeting of the national council on measurement in education, Chicago, IL, USA, 24–27 March 1997. NCME.
16. Karabatsos G and Walker SG. A Bayesian nonparametric approach to test equating. *Psychometrika* 2009; 74(2): 211–232.
17. Kruschke JK, Aguinis H and Joo H. The time has come: Bayesian methods for data analysis in the organizational sciences. *Organ Res Methods* 2012; 15(4): 722–752.
18. MacEachern SN. Dependent Dirichlet processes. *Department of Statistics, The Ohio State University*. 2000. https://people.eecs.berkeley.edu/~russell/classes/cs294/f05/papers/maceachern-1999.pdf (accessed 12 July 2017).
19. Kim S, Livingston SA and Lewis C. Effectiveness of collateral information for improving equating in small samples. Princeton, NJ: Educational Testing Service, 2009.
20. Kim S, Livingston SA and Lewis C. Collateral information for equating in small samples: a preliminary investigation. *Appl Meas Educ* 2011; 24: 302–323.
21. van de Schoot R, Kaplan D, Denissen J, et al. A gentle introduction to bayesian analysis: applications to developmental research. *Child Dev* 2013; 85(3): 1–19.
22. Oh HJ, Guo H and Walker ME. Improved reability estimates for small samples using empirical Bayes techniques. Research report, RR-09-46, 2009. Princeton, NJ: ETS. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2009.tb02203.x (accessed 5 July 2017).
23. Gonzalez J, Barrientos AF and Quintana FA. Bayesian nonparametric estimation of test equating functions with covariates. *Comput Stat Data Anal* 2015; 89: 222–244.
24. Gonzalez J, Barrientos AF and Quintana FA. A dependent Bayesian nonparametric model for test equating. In: Millsap RE, Bolt DM, van der Ark LA, et al. (eds) *Quantitative psychology research*. Cham: Springer, 2015, pp.213–226.
25. Wiberg M and Branberg K. Kernel equating under the non-equivalent groups with covariates design. *Appl Psychol Meas* 2015; 39(5): 349–361.
26. Kaplan D and Depaoli S. *Handbook of structural equation modeling*. Hoyle RH (ed.). New York, NY: The Guilford Press, 2012.
27. Boone EL, Ye K and Smith EP. Using data augmentation via the Gibbs Sampler to incorporate missing covariate structure in linear models for ecological assessments. *Environ Ecol Stat* 2009; 16: 75–87.
28. van de Schoot R and Depaoli S. Bayesian analyses: where to start and what to report. *Eur Health Psychol* 2014; 16(2): 75–84.
29. Kruschke JK. Bayesian estimation supersedes the t test. *J Exp Psychol Gen* 2013; 142(2): 573–603.
30. De Iorio M, Müller P, Rosner GL, et al. An ANOVA model for dependent random measures. *J Am Stat Assoc* 2004; 99(465): 205–215.
31. Müller P and Quintana FA. Nonparametric bayesian data analysis. *Stat Sci* 2004; 19(1): 95–110.
32. Orbanz P and Teh YW. Bayesian nonparametric models. In: Sammut C and Webb GI (eds) *Encyclopedia of machine learning*. Boston, MA: Springer, 2010, pp.81–89. https://www.stats.ox.ac.uk/˜teh/research/npbayes/OrbTeh2010a.pdf (accessed 7 February 2017).
33. Shah A and Ghahramani Z. Determinantal clustering process-A non-parametric Bayesian approach to kernel based semi-supervised clustering. In: *Proceedings of the Twenty Ninth conference on uncertainty in artificial intelligence*, Bellevue, WA, USA,
11–15 August 2013. AUAI Press. http://auai.org/auai2013/prints/papers/200.pdf (accessed 22 January 2017).

34. Branberg K and Wiberg M. Observed score linear equating with covariates. *J Educ Meas* 2011; 48(4): 419–440.

35. Liou M, Cheng PE and Li M. Estimating comparable scores using surrogate variables. *Appl Psychol Meas* 2001; 25: 197–207.

36. MacEachern SN. Dependent nonparametric processes. In: *ASA proceedings of the section on Bayesian statistical science*, 1999. Alexandria, VA: American Statistical Association. https://people.eecs.berkeley.edu/~russell/classes/cs294/f05/papers/maceachern-1999.pdf (accessed 12 January 2017).

37. Barrientos AF, Jara A and Quintana F. Fully nonparametric regression for bounded data using Bernstein polynomials. *J Am Stat Asso* 2017; 112: 806–825.

38. Kruschke JK. Bayesian data analysis. *Wiley Interdiscip Rev Cogn Sci* 2010; 1(5): 658–676.

39. Önal N. A study un the development of a middle school students’ attitudes towards mathematics scale. *Elementary Educ Online* 2013; 12(4): 938–948.

40. Lee K and Anderson J. Gender differences in mathematics attitudes in coeducational and single sex secondary education. In: *Mathematics education in the margins (Proceedings of the 38th annual conference of the Mathematics Education Research Group of Australasia)* (eds Marshman M, Geiger V and Bennison S). Sunshine Coast: MERGA, 13–18 June 2015, pp.357–364.

41. Collis B. Sex differences in the association between secondary school students’ attitudes toward mathematics and toward computers. *J Res Math Educ* 1987; 18(5): 394–402.

42. Batool T, Akhter S and Kalsoom T. Exploring gender differences in attitude towards mathematics at secondary level in Pakistan. *J Bus Soc Rev Emerg Econ* 2020; 6(2): 587–596.

43. Zan R and Martino P. Attitude toward mathematics: overcoming the positive/negative dichotomy. In: Sriraman B (ed.). *Beliefs and mathematics. The Montana mathematics enthusiast: monograph series in mathematics education*. Charlotte, NC: Age Publishing & The Montana Council of Teachers of Mathematics, 2008, pp.197–214.

44. Gee BC. Attitudes toward mathematics and basic mathematical understanding of prospective elementary school teachers at Brigham Young University. 1965. Corvallis, OR: Oregon State University. https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/2r36v198m?locale=en (accessed 19 January 2021).

45. R Development Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013. http://www.R-project.org (accessed 15 April 2016).

46. Öztürk N and Anıl D. A study on equating academic staff and graduate education entrance examination scores. *Educ Sci* 2012; 37(165): 180–193.

47. Livingston SA and Lewis C. *Small-sample equating with prior information*. ETS Research Rep. No. RR-09-25, 2009. Princeton, NJ: ETS. https://files.eric.ed.gov/fulltext/ED507811.pdf (accessed 5 July 2017).

48. Kan A, Sünbüloğlu and Ömür S. 6. - 8. Sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Univ J Fac Educ* 2013; 9(2): 207–222.

49. Terzi R and Yakar L. Differential item and differential distractor functioning analyses on Turkish high school entrance exam. *J Meas Eval Educ Psychol* 2018; 9(2): 136–149.
50. Yıldırım H. *An investigation of item bias of mathematics subtest in 2012 year level determination exam*. Unpublished Master Thesis, Gazi University, Graduate School of Educational Sciences, Ankara, 2015.

**Author biographies**

Meltem Yurtcu received her doctoral degree from Hacettepe University in Division of Educational Measurement and Evaluation in 2018. She is assistant professor in Inonu University at Department of Measurement and Evaluation in Education. Her research focused on Test equating, Differential item functioning, Bayesian approach.

Hülya Kelecioglu is a Professor at Department of Measurement and Evaluation in Education in Hacettepe University. Her research focused on Test equating, Item response theory, Classical test theory, Differential item functioning.

Edward L Boone is a Professor at Department of Statistical Sciences and Operations Research. His research focused on Bayesian statistical methods with applications in the environment, health care and national security, Statistical consulting.