ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 192-211

The Comparison of the Equated Tests Scores by Various Covariates using Bayesian Nonparametric Model *

Meltem YURTÇU ** Hülya KELECİOĞLU *** Edward L. BOONE ****

Abstract
This research is based on obtaining equated scores by using covariates in the Bayesian nonparametric model. As covariates in the study, gender, mathematics self-efficacy scores, and common item scores were used. The distributions were obtained for all score groups. Hellinger Distance was calculated to obtain the distances between the distributions of equated scores by using covariates and the distribution of the target test scores. These distances were compared with the distributions of equated scores obtained from methods based on Item Response Theory. The study was conducted on Canadian and Italian samples of Programme for International Student Assessment (PISA) 2012. PARSCALE and IRTEQ were used for classical methods, and R was used for Bayesian nonparametric model. When gender, mathematics self-efficacy scores, and common item scores were used as covariates in the model, distance values of obtained equated scores to target test scores were close to each other, but their distributions were different. The closest distribution to target test scores was achieved when gender and mathematics self-efficacy scores were used together as covariates in the model, and the farthest distributions were obtained from item response theory methods. As a result of the research, it was determined that the model is more informative than the classical methods.

Key Words: Test equating, Bayesian nonparametric model, covariates, equated scores, score distribution.

INTRODUCTION

It is very important to compare the scores of the individuals evaluated by the tests. Equating is used to compare the scores obtained from different test forms that serve the same purpose. One of the most important steps of equating is the selection of the equating method, which differs regarding the use of common items or common individuals. The methods involving common individuals can be classified as single group design, counterbalanced design, and equivalent group design, whereas the method involving common items in non-equivalent groups is named as Non-Equivalent groups with Anchor Test (NEAT) (Branberg & Wiberg, 2011). NEAT is used when there is no chance of applying another questionnaire and the data required to reveal the difference between the groups were obtained from common items/tests (Liou, Cheng, & Li, 2001; Moses, Deng, & Zhang, 2010). The selection of the common tests is crucial in the design, and the selected test should have a similar mean and item difficulty with the tests in question and should represent this test in terms of content (Dorans, Moses, & Eignor, 2010; Kolen, 1988; Kolen & Brennan, 2014; Mittelhaeuser, Beguin, & Sijtsma, 2011; Sinharay & Holland, 2006; Wei, 2010; Wiberg & von Davier, 2017). The common test should be one-dimensional, should have a high correlation with the scores of the other tests to be equated, and should reflect the exact structure of the test forms (Wallin & Wiberg, 2017). In addition, the use of common tests that address the trends over time in NEAT design may be appropriate only for certain individuals, which may create a bias for equating. If the common tests/items fail to satisfy these conditions, the

* This study is based on Meltem Yurtçu's doctoral thesis titled "The Comparison of the Equated Tests Scores by Using Various Covariates using Bayesian Nonparametric Model".

** Assist. Prof, Inonu University, Faculty of Education, Malatya-Turkey, meltem.yurtcu@gmail.com, ORCID ID: 0000-0003-3303-5093
*** Prof. Ph.D., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934
**** Prof. Ph.D., Virginia Commonwealth University, Department of Statistical Science and Operations Research, Richmond-USA, ELBoone@vcu.edu, ORCID ID: 0000-0003-0755-6899

To cite this article:
Yurtçu, M., Kelecioğlu, H., & Boone, E. L. (2021). The comparison of the equated tests scores by various covariates using bayesian nonparametric model. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 192-211. doi: 10.21031/epod.864744

Received: 19.01.2021
Accepted: 17.06.2021

reliability of equating and other processes associated with common tests/items will be negatively affected (Wei, 2010; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017). Moreover, the tests to be equated may not have any common items or tests. In this case, the bias and mean standard error can be reduced by adding variables associated with the test scores to the test equating process, which allows to explain the difference between the groups (Branberg & Wiberg, 2011; Liou et al., 2001; Oh, Guo, & Walker, 2009; Wiberg, 2015; Wiberg & Branberg, 2015), and to increase the accuracy of the estimation (Branberg & Wiberg, 2011; Kim, Livingston, & Lewis, 2009, 2011; Livingston & Lewis, 2009; Oh et al., 2009; Wiberg & Branberg, 2015). Wiberg and Branberg (2015) stated that using a single common variable that has a high correlation with the test scores could give results similar to a common test. Liou et al. (2001) also suggested that the variables selected from historical data of the individuals may give better results than common tests.

In recent years, Non-Equivalent Groups with Covariates (NEC) design, which uses common variables/covariates in the absence of common items, has been added to the literature (Branberg & Wiberg, 2011; Wiberg & Branberg, 2015). The design involving the use of both common item/s and covariate/s is called NEATNEC (Wiberg & Branberg, 2015).

The most important assumption of NEC design is that covariates are able to explain the difference between groups. The most important step of this design is that the situational distributions of the test scores should be the same in both groups in terms of covariates categories. This is an indication that the achievement of individuals is evaluated according to their categorical characteristics. However, if the test scores to be equated were obtained at different time periods (i.e., equating a new test with an old test), this hypothesis may not be valid because the characteristics of the test scores and the covariates may have changed over time (Wiberg & Branberg, 2015).

Although many researchers have described covariates in different terms, they emphasized that these variables are related to test scores, and they can explain the difference between groups (Branberg & Wiberg, 2011; Kim et al., 2009; Liou, 1998; Liou et al., 2001; Wiberg & Branberg, 2015; Wright & Dorans,1993). In the literature, the variables such as age, gender, and educational status were observed to be included as covariates (Branberg & Wiberg, 2011; Gonzalez, Barrientos, & Quintana,2015a; Karabatsos & Walker, 2009; Liou et al., 2001; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017). The accuracy of the prediction may increase with the increase of the number of covariates added to the study, which makes the number of covariates added to the study important. Another important issue is the number of covariate categories. As the number of covariate categories increases, the number of individuals falling into each relevant category may decrease. Therefore, limiting the number of variable categories will give more appropriate results and will strengthen the prediction (Wallin & Wiberg, 2017; Wiberg & Branberg, 2015).

Equating methods are based on various theories and assumptions, which are classified in the literature as Classical Test Theory and Item Response Theory (IRT). However, in recent years, Bayesian approach has come to the fore in test-equating studies.

Bayesian Approach

In the classical approach, the *p*-value is used to test the significance of null hypotheses, which varies according to the sample and purpose of the researcher (Berger, Boukai, & Wang, 1997; Kruschke, 2010; Kruschke, Aguinis, & Joo, 2012; Lee & Boone, 2011; Rounder, Morey, Speckman, & Province, 2012). This can be considered as a disadvantage because point estimation affects the outputs in terms of reaching an accurate result. The confidence interval used in Bayesian approach carries more information than point estimation. The confidence intervals for posterior inferences generated by Bayesian approach can be expressed with the mean and 95% confidence interval (highest density interval/HDI). The points falling in this range are more accurate than the points that are outside (Kruschke, 2010).

Bayesian approach provides well-defined probabilistic models for observed data and unknown values. There are two types of Bayesian approaches. Parametric Bayesian approach uses a limited number of

parameters, but it has some limitations, whereas the flexible use of the number of parameters in the models constitutes the basis of Bayesian nonparametric approach (De Iorio, Müller, Rosner, & MacEachern, 2004, Müller & Quintana, 2004; Orbanz & Teh, 2010; Shah & Ghahramani, 2013). Dirichlet Process (DP) Model is one of the models that have a central role in Bayesian nonparametric approaches (De Iorio et al., 2004; Gonzalez et al., 2015a; Petrone, 1999a). This model allows the inclusion of the covariates in equating process. The randomness effect of the variables on the distribution of the test scores will appear as dependency, which is explained by the Dependent Dirichlet Process (DDP), an extension of the DP model (Barrientos, Jara, & Quintana, 2016; MacEachern, 1999, 2000). However, the selection of prior distributions in Bayes nonparametric approaches is usually very difficult. Petrone (1999a, 1999b) suggested using Bernstein-Dirichlet Prior (BDP) model to eliminate this limitation. In their studies, Barrientos et al. (2016) expanded the model further and developed Dependent Bernstein Polynomial Process (DBPP) model. Barrientos et al. (2012, 2016) discussed two specific types of DBPP. In this study, DBPP involving a dependent stick-breaking process with common weights and predictor-dependent support points was employed. This type is called single-weight DBPP (wDBPP). Z represents covariate space, and F_z represents covariate-dependent random probability distributions.

For $\forall z \in Z, \{F_z : z \epsilon Z\}$, wDBPP can be formulated as;

$$f_{(z)}(\cdot) = \sum_{j=1}^{\infty} w_j \beta\left(z | \lceil k\theta_j(\mathbf{z}) \rceil, k - \lceil k\theta_j(\mathbf{z}) \rceil + 1\right)$$

This model, which represents an infinite set of beta distributions, suggests that the test scores have covariate-dependent sample densities. This model can be shown as:

$\{Fz; z \in Z\} \sim$ wDBPP$(\alpha, \lambda, \psi, H)$.

Where $= \{\boldsymbol{h_z}; \boldsymbol{z} \in \boldsymbol{Z}\}$; $\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots\ldots, \alpha > 0$ are independent, random variables whose distribution is defined by $\beta(1, \alpha)$; k is a discrete random variable with a distribution indexed to a finite-dimensional parameter λ, $\boldsymbol{\theta_{j(z)}} = \boldsymbol{h_z}\left(\boldsymbol{r_j}(\boldsymbol{z})\right), \boldsymbol{r_1}, \boldsymbol{r_2}$ ...., are independent and identically distributed real-valued stochastic processes indexed by the parameter ψ. This model provides a covariate-dependent equating transformation (Gonzalez, Barrientos, & Quintana, 2015b).

In this study, the accuracy of the predictions and their contribution to the test equating process were analyzed by comparing the equated scores obtained from Bayesian Nonparametric Model (BNP) by using various covariates at NEC design.

METHOD

The research was conducted with real data. The distribution of equated scores obtained from the scaling methods based on IRT was compared with the distributions of equated scores obtained from the BNP model.

Sample

The data used in the research was obtained from PISA 2012. In order to carry out the equating process in non-equivalent groups, two countries with different success levels were selected. According to PISA 2012 math results, the data of Canada, which was ranked as 13th with an average score of 518, and Italy, which was ranked as 32nd with an average score of 485, were taken from the database published by OECD (http://www.oecd.org/pisa/data). The records with missing data were removed, and Italian data with a sample size of 908 and Canadian data with a sample size of 931 were used in the analysis.

Data Collection Tools

In PISA 2012, a cognitive test measuring students' mathematics literacy and a student questionnaire were used. The data of the research is comprised of the Italian students' responses to booklet 5 and Canadian students' responses to booklet 6 of the mathematics sub-test. Booklets 5 and 6 were selected to be used in the research because of the equal number of math questions and the high number of common items. There were 12 common items in the booklets.

Gender and mathematics self-efficacy score (MATHEFF) were used as covariates in the analysis, where gender is a two-category variable and MATHEFF is a continuous variable. In addition, the anchor item scores were taken as the covariate in the BNP model. The reason for using MATHEFF is that it is defined as the variable that explains the mathematics achievement (Ayotola & Adedeji, 2009; Hackett & Betz, 1989; Koğar, 2015; Schulz, 2005; Siegle & McCoach, 2007; Thien & Darmawan, 2016). This variable was derived from the sum of the item scores, where a higher score indicates lower self-efficacy. MATHEFF scores varied between 8-32. But, since the scores range between 0-1 in the model, MATHEFF scores were also converted into the 0-1 range, showing the change within one unit.

Another covariate used in the NEC design of the BNP model was gender. There are many studies in the literature using gender as covariate (Branberg & Wiberg, 2011; Gonzalez et al., 2015a, 2015b; González & Wiberg, 2017; Liou et al., 2001). In addition, in many studies, gender is considered as a variable that creates differentiation among groups (Martin, Mullis, Foy, & Stanco, 2012; Yıldırım, Yıldırım, Yetişir, & Ceylan, 2013).

Regarding the equating studies performed in non-equivalent groups, the number of common items in the tests should be equal to at least 20% of the number of questions to minimize the equating error (Angoff, 1971). The study was carried out with 24 items in NEC design, and the total score of the common items was used as the covariate. In NEAT design, 12 items were taken as external commonitems, and the study was carried out with 36 items. To avoid them from affecting the model as a different criterion, partially scored items in the booklets were converted into two category-scores.

Data Analysis

In the research, IRT-based scale conversion methods and the analyses using the BNP model were carried out separately. First of all, unidimensionality and local independence were tested for IRT. Factor 10.3 analysis software was used to test unidimensionality, which was analyzed over 36 items. The unidimensionality of 36-item in booklets was taken as the proof of the unidimensionality of the 24-item version. As a result of the factor analysis, Kaiser Mayer Olkin (KMO) value of booklet 5 was found to be .95, whereas Bartlett's value was 7086.60 (df = 630; $p < .001$). Regarding booklet 6, KMO value was .94 and Bartlett's value was 6427.00 (df = 630, $p < .001$). KMO values indicated the sufficiency of the sample sizes for the analysis, and Bartlett's value indicated the factorizability of the data set. Regarding these values, it can be said that the tests were unidimensional.

The unidimensionality of the booklets provided insight about local independence assumption. Moreover, in order to test the local independence assumption, the correlation between the items was calculated for the top and bottom 27% of the data (Kelley, 1939). The correlation between the top and bottom groups was found to be lower than the overall correlation; therefore it was concluded that the local independence assumption was met.

Parameter estimation

The two test forms to be scaled in the study are parallel. The parameters obtained from these forms were estimated from different individuals, and the mean and standard deviations of the groups were different; therefore the estimations were made using separate calibration methods.

Equating by NEAT design was performed using ability parameters. The -2loglikelihood values obtained for 2 parameter logistic model (PLM) and 3 PLM were tested by chi-square test and 3 PLM

model was found to be significant. Therefore, the parameters were estimated according to 3 PLM method. Parscale 4.1 program was used in the estimation of item parameters.

Scale conversion

Common items were taken as external common items in NEAT design to allow a comparison with NEC design. IRTEQ software was used to convert the parameters taken from the PARSCALE software to the same scale. Since IRT true-score equating is more accurate and precise (Li, Jiang, & von Davier, 2012), this process was carried out on true-score. In the study, booklet 6 was taken as the target test, whereas booklet 5 was taken as the basic test.

Test equating by Bayes nonparametric approach

In order to make accurate statistical predictions, Markov chain Monte Carlo (MCMC) sampling method was used to obtain a sample representing the universe (Kruschke, 2015; StataCorp, 2015). In this study, MCMC method was used to estimate population parameters (k, γ, w) of the BNP model. General information about the population can be obtained using covariates. MCMC processes were performed separately for Canada and Italy data sets. The covariates and parameters compatible with the data are combined in the files prepared in MCMC sampling by using DBPP.

The covariates used in the research were added to the model as anonymous priors. This fact prevented the bias that may arise from the effects of these variables on the posterior distributions of the scores and ensured a more objective evaluation.

Prior distribution specification: The distributions of wDBPP based on MCMC method were given as:

$\boldsymbol{h_z(\cdot) = \frac{exp\{\cdot\}}{1+exp\{\cdot\}},\ r_j(z) = z^T\gamma_j}$ and $\boldsymbol{\gamma_j|\ \mu, S \sim^{iid} N_p(\mu, S),\ j = 1, 2, \ldots.}$

Here; $\boldsymbol{v_j|\ \alpha \sim \beta(1, \alpha), k|\ \lambda \sim Poisson(\lambda)\ \|_{\{k>1\}}, \mu|\ m_0, S_0 \sim N_p(m_0, S_0), S|v,\ \psi \sim IW_p\ (v, \psi)}$. In equation I, $\boldsymbol{W_p\ (v, A)}$; scale matrix A represents *p*-dimensional inverted-Wishart distribution with degrees of freedom v. The values that Gonzalez et al. (2015a) found to be significant in their study, were also included in their study of 2015b, therefore the following values were used while generating the prior distribution $\boldsymbol{\lambda = 25, m_0 = 0_p, S_0 = 2.25 * I_p, v = p + 2}$, and $\boldsymbol{\alpha = 1}$. MCMC algorithm was run to explain the posterior distribution of wDBPP model and to obtain the posterior distribution samples of all model parameters.

Posterior inference: All computations were coded and performed in R 3.2.1 statistics software. The posterior probability distribution was given by:

$$
\begin{aligned}
& p(v, k, w, \gamma | y, z) \\
& \propto \prod_{i=1}^{n}\left[\sum_{j=1}^{10} w_j \beta\left(y_i \left| \left[k \frac{e^{Z_i^T \gamma_j}}{1+e^{Z_i^T \gamma_j}}\right], k - \left[k \frac{e^{Z_i^T \gamma_j}}{1+e^{Z_i^T \gamma_j}}\right] + 1\right.\right)\right]\left[\prod_{j=1}^{10} \beta(v_j|1,1)\right] \\
& \times \left[\frac{25^k e^{-25}}{k!\,(1-e^{-25})}\right]\left[\prod_{j=1}^{10}(2\pi)|S|^{-\frac{1}{2}} e^{-0.5(\gamma_j-\mu)^T S^{-1}(\gamma_j-\mu)}\right](2\pi)|S_0|^{-\frac{1}{2}} e^{-0.5(m_0)^T S_0^{-1}(m_0)} \\
& \times \frac{|\psi|^2}{2^2 \Gamma_2(2)}|S|^{\frac{7}{2}} e^{-\frac{1}{2}tr(\psi S^{-1})}
\end{aligned}
$$

The posterior predictive distribution was given as below:

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

Where

$$L(T|v,k,w,\gamma) = \sum_{j=1}^{10} w_j \beta \left(T| \left\lfloor k \frac{e^{Z_i^T \gamma_j}}{1+e^{Z_i^T \gamma_j}} \right\rfloor, k - \left\lfloor k \frac{e^{Z_i^T \gamma_j}}{1+e^{Z_i^T \gamma_j}} \right\rfloor + 1 \right)$$

shows the sum obtained for the identified distributions.

The number of iterations was first set as 5000 to test the parameters in the generated files. Then, MCMC number was set as 150 000, and the analyses were performed by repeating 10 times for each file in order to obtain a proper distribution. The analyses of the test forms were carried out simultaneously, which took around 10 hours and 23 minutes for each file.

The algorithm of Gibbs and Metropolis-Hastings sampling method was as follows. It was used to explain the posterior distribution obtained by gathering the covariables with the model in MCMC files:

An initial $v^* \sim p(v|v^{(i)})$ value is obtained by using Metropolis-Hastings ratio; if the initial value is reasonable, it is accepted; if not, it is rejected, and the process continued until the most appropriate value is obtained (there were 10 v values in the research).

An initial $\gamma^* \sim p(\gamma|\gamma^{(i)})$ value is obtained by using Metropolis-Hastings ratio; if the initial value is reasonable, it is accepted; if not, it is rejected, and the process continued until the most appropriate value is obtained (there were 20 γ values in the research).

An initial $k^* \sim p(k|k^{(i)})$ value is obtained by using Metropolis-Hastings ratio; if the initial value is reasonable, it is accepted; if not, it is rejected, and the process continued until the most appropriate value is obtained (there was 1 k value in the research).

After completing this stage, the equated scores were obtained using cumulative distributions of the test scores.

The transformation functions are as follows, where T is score distribution; $\boldsymbol{t_x}$ represents the scores obtained from test X, $\boldsymbol{t_y}$ represents the scores obtained from test Y, and z represents the covariates;

$$\boldsymbol{t_x} = \boldsymbol{F^{x^{-1}}}(\cdot)$$

$$\boldsymbol{t_y} = \boldsymbol{F^{y^{-1}}}(\cdot) \quad \Longrightarrow \quad \boldsymbol{t_y} = \boldsymbol{\varphi}(\boldsymbol{t_x}) = \boldsymbol{F^{y^{-1}}}(\boldsymbol{F^x}(\cdot))$$

$$\boldsymbol{t_{z_x}} = \boldsymbol{F^{z_x^{-1}}}(\cdot)$$

$$\boldsymbol{t_{z_y}} = \boldsymbol{F^{z_y^{-1}}}(\cdot) \quad \Longrightarrow \quad \boldsymbol{t_{z_y}} = \boldsymbol{\varphi}(\boldsymbol{t_{z_x}}) = \boldsymbol{F^{z_y^{-1}}}(\boldsymbol{F^{z_x}}(\cdot))$$

The analyses conducted to obtain equated scores were completed in 7 days and 6 hours. The equating process was completed by putting the generated profile distributions into the percentiles determined for covariate categories.

DBPP model defines continuous distribution functions in (0-1) range. Therefore, the score estimations were made in this range as Gonzalez et al. (2015b) have done in their study. After equating, the scores were converted to the scale-of-100 so that the highest score will be 100. This is considered as the best scaling method in equating studies involving the tests with different ranges (Livingston, 2004). Therefore, the continuous variables used in the distributions were converted and analyzed in (0-1) range, then the graphics and distributions obtained for equated scores were converted to the scale-of-100 and interpreted.

Comparison criteria

In traditional equating methods, standard criteria such as Root Mean Square Error (RMSE), Mean Square Error (MSE), bias, and standard errors (SE) are used to assess parameter estimation error.

However, it is difficult to compare the results obtained by the methods based on different models such as IRT and BNP (Wiberg & Gonzalez, 2016). Therefore, in this study, the comparison of the results using the criteria such as RMSE and MSE was not possible. Hellinger Distance, which provides statistical information, was used in this study to compare the equated scores obtained by BNP and IRT methods to target test's scores. This distance is the sum of the distances between the points of each distribution. There are many forms of Hellinger distance. Hellinger Distance used to compute the distance between two distributions f and g (Boone, Merrick, & Krachey, 2012) is formulated as;

$$\widehat{H}(f,g) = \left[\frac{1}{2}\int\left(\sqrt{\hat{f}(x)} - \sqrt{\hat{g}(x)}\right)^2 dx\right]^{1/2} \approx \left[\frac{1}{2}\sum_{l=1}^{k}\left(\sqrt{\hat{f}(x)} - \sqrt{\hat{g}(x)}\right)^2 (x_l - x_{l-1})\right]^{\frac{1}{2}}$$

The distances between the distributions of the scores were computed according to the method above, and the distributions are shown through graphics in the results part. One of the titles (participants, sample, or working group) should be used with respect to the group formation procedure used in the study. The information about the sampling procedure and the group should be given in this part.

RESULTS

In PISA 2012, the mean score and standard deviation of 908 Italian students, who answered booklet 5, was 51.51 and 20.72, respectively. Whereas the mean score and standard deviation of 931 Canadian students who answered booklet 6 was 52.27 and 22.06 respectively.

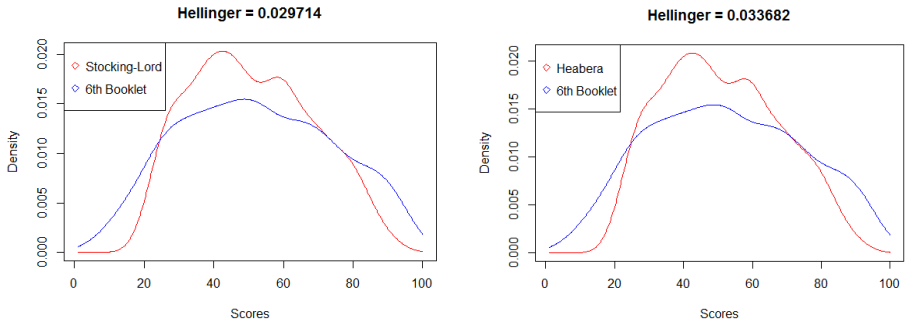
Equating errors occurred as a result of scaling according to IRT methods in the NEAT design were computed, and the score distributions obtained from various methods were analyzed.

In the two booklets, answers taken by two non-equivalent groups were used for scaling. RMSE values were calculated.

Table 1. RMSE Values Obtained According to IRT Methods

Mean – Mean	Mean-Sigma	Stocking-Lord	Heabera
0.149	0.13	0.20	0.18

The lowest error was obtained from Mean-Sigma method and the highest error from Stocking-Lord method. New ability parameters were computed, and item parameters of the target test were used for finding true scores. Probability density distributions of each method and their distance from the target test were calculated using Hellinger distances.



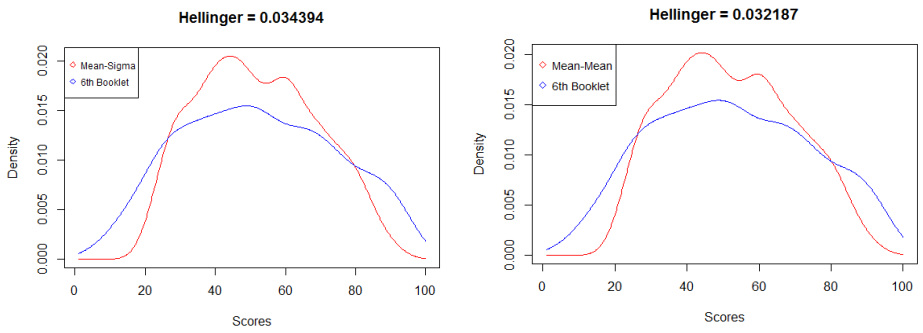
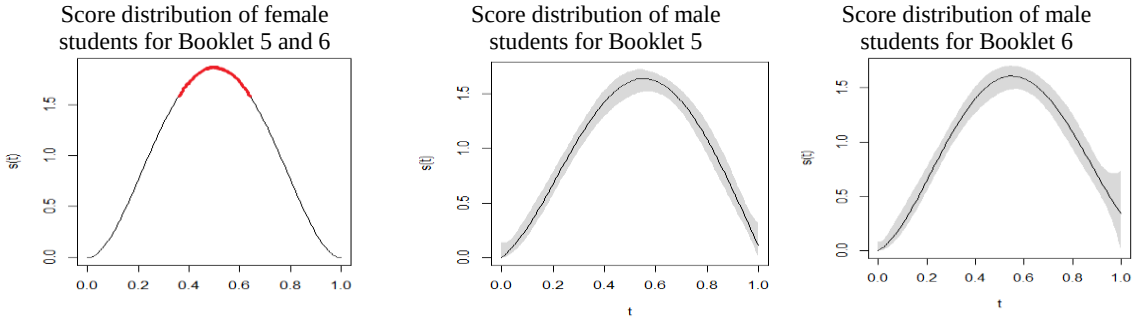


Figure 1. The Distribution of Obtained Scores and Their Distance from the Target Test's Scores

Regarding the probability density distributions of the predicted scores in Figure 1, the distributions of the scores were observed to be similar and to be at approximately similar distances to the target test's distribution according to the Hellinger distance. Although Mean-Sigma method gave the lowest RMSE, the distributions obtained from the characteristic curve methods were closer to the distribution of the target test. According to Hellinger distance, Stocking-Lord method was the closest distribution with 0.029714.

The distance between the distribution of equated scores obtained by using gender as covariate in the BNP model and the distribution of target test's scores

Gender was taken as covariate, and students' scores were gathered with this variable. Distributions were first examined according to the booklets. Figure 2 shows the distribution of the scores and confidence intervals that best reflect the population for each gender.



Note. The confidence interval is shown in red to female because it was very narrow.

Figure 2. Score Distributions and Confidence Intervals according to Gender for Booklets.

The distributions were observed to be similar. Especially, the distribution of female students was the same for both booklets. The accuracy of the score estimation was checked through confidence intervals. Confidence intervals of female students' score distributions were found to be quite narrow, whereas male students' confidence intervals were wide, which may indicate uncertainty in the estimation of these scores. The decrease in the accuracy may be due to the low number of students in the sample used for the estimation of scores, or due to the fact that the scores of the students having the same profile were distributed in a wide range.

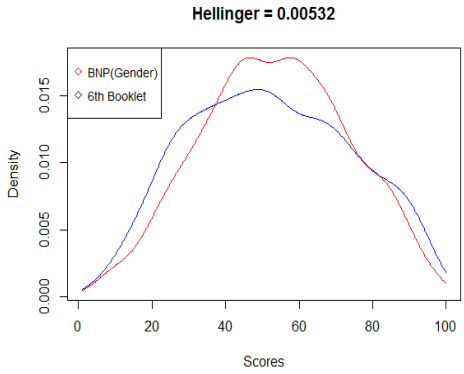


Figure 3. Distribution of the Target Test's Scores and the Scores Equated with Gender

The score equated with gender covariate was calculated for each student. The distributions of equated scores and target test's scores were compared. The distance between these distributions was calculated by Hellinger distance. As can be seen from Figure 3, the distribution of equated scores was observed to be sharper than the distribution of the target test's scores. The distance between these two curves was 0.00532, which was approximately one-fifth of the distance obtained by IRT methods.

The distance between the distribution of equated scores obtained by using MATHEFF as covariate in the BNP model and the distribution of target test's scores

MATHEFF was taken as the covariate, and students' scores were associated with this variable. The score distributions that best reflect the population according to MATHEFF levels were computed. The distributions of scores at different MATHEFF levels were analyzed according to booklets.

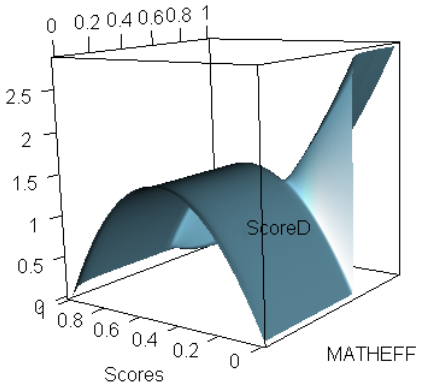


Figure 4. The Distributions of the Scores Equated with MATHEFF

Students at different MATHEFF levels had different profiles. The distribution of each profile was computed. Test score distributions of booklets 5 and 6 according to MATHEFF levels of the students were similar, therefore they are shown in a single graph in figure 4. As students' self-efficacy levels decrease (or for higher values of MATHEFF), the intensity of their scores decreases. Based on these distributions in each profile, it was also possible to see at which scores the students' distribution changed and how this change was affected for both booklets.

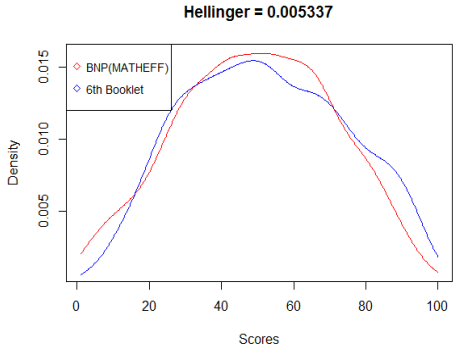


Figure 5. Distribution of the Target Test's Scores and the Scores Equated with MATHEFF Levels

In the BNP model, the distribution of equated scores was very close to the distribution of the target test' scores. Hellinger distance was calculated as 0.005337. This distance is significantly lower than the distance obtained from IRT methods and the distance of the model obtained using gender. Compared to the BNP model using gender, the distributions were observed to approach and differentiate from the target test at different points. In the model using MATHEFF, the distribution of equated scores moved away from the target test at the ends, whereas in the model using gender, the distribution of equated scores differed from the target test in average values.

The distance between the distribution of equated scores obtained by using both gender and MATHEFF as covariates in the BNP model and the distribution of target test's scores

Students' MATHEFF scores were examined according to gender. The distributions obtained for female students were similar to males for booklets 5 and 6, therefore, graphs are shown for both genders in figures. Figure 6 and 7 shows the distributions of the students for booklets 5 and 6.

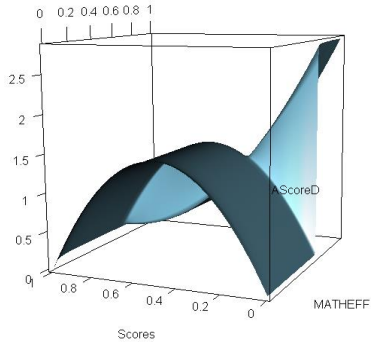


Figure 6. Distributions for Booklet 5

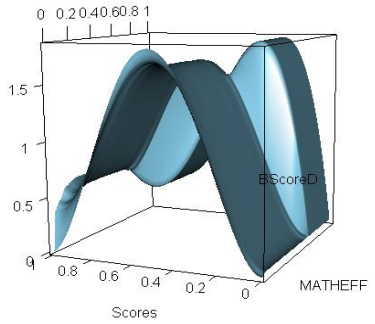


Figure 7. Distributions for Booklet 6

Regarding booklet 5, it was observed that the intensity of high scores of both genders' students with low mathematics self-efficacy decreased. In booklet 6, the students of both genders with low mathematics self-efficacy were observed to be clustered around 20. As can be seen from these distributions, students' intensity around high scores decreased as MATHEFF scores get higher, which indicates lower mathematics self-efficacy levels.

So, it can be concluded that booklet 6 was easier than booklet 5 for both female and male students. In addition, the differentiation of the distributions in booklets may indicate that using these two covariates was effective in revealing the differences between the booklets. Equated scores were obtained using the cumulative distributions of these distributions generated by combining covariates and individuals' scores. The probability distributions of equated scores and target tests were examined together in Figure 8.

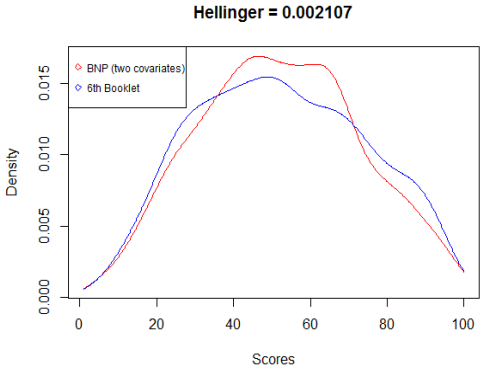


Figure 8. Distribution of the Target Test's Score and the Scores Equated with both Covariates

The distribution of equated scores is very close to the target test when both covariables were included in the model; Hellinger distance is also relatively small (0.002107) compared to other models. From Figure 8, it can be seen that equated scores obtained by using two covariates got closer to the target test. In particular, the approximation of distributions to the extreme values might indicate that the model could be used to tolerate the error in extreme values.

The distance between the distribution of equated scores obtained by using common items as covariate in the BNP model and the distribution of target test's scores

In the first part of the study, equated scores were obtained from common items according to IRT scaling methods. In this section, the scores obtained from the sum of common items were used as a covariate. The distributions obtained from the combination of student scores and covariates are shown in Figures 9 and 10.

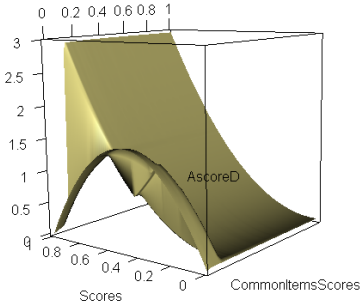


Figure 9. Distributions for Booklet 5

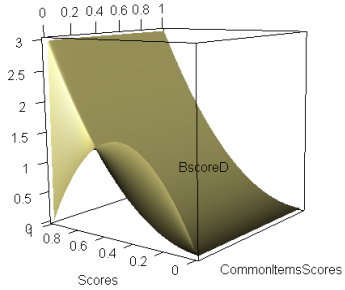


Figure 10. Distributions for Booklet 6

In order to check whether common items reflect the tests or not, the correlation between common test scores and test scores was examined. These correlations were found to be .79 for booklet 5 and .75 for booklet 6. Accordingly, it can be said that common items represent the tests statistically.

According to Figure 9, if common item scores were not included in the model as covariate or they contributed to the model with very low scores in booklet 5, the density of students was observed to increase on average scores and densities towards the end scores decreased. With the increase of common item scores, the shapes of distributions differed from first distributions, and it was observed that low score densities decreased and high score densities increased.

Regarding Figure 10, which shows the analysis results for booklet 6, if common item scores were not included in the model as a covariate or contributed to the model with very low scores, students are concentrated around the mean. The distributions of students were quite similar for other score levels. Therefore, regarding the individuals with other scores than low common item scores, the distributions are similar for both booklets. The differences in common item scores failed to explain the difference in the math achievement of the students. Booklet 6 was observed to be easier than booklet 5.

Equated scores were obtained according to common item scores of students. The probability distributions of these scores and target test were examined together, and their distributions are given in Figure 11.

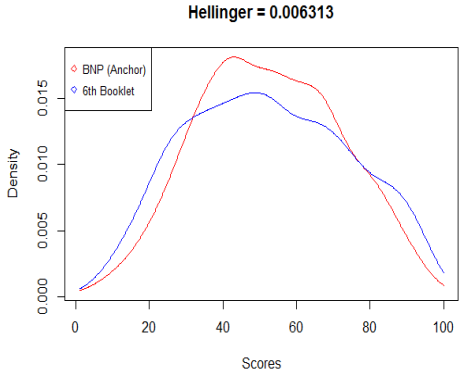


Figure 11. Distribution of the Target Test's Scores and The Scores Obtained from BNP Model with Common Items

Hellinger distance between the distribution of equated scores obtained by using common items as covariate and the distribution of target test scores was calculated as 0.006313. This distance was smaller than the one of the IRT methods, but it was greater than the values obtained from BNP models with other covariates. The distribution of equated scores obtained using common items is similar to the distribution of the equated scores obtained using gender. Both distributions diverged from target test's distribution at the ends. Although the numerical values of Hellinger distances were insufficient, their shapes supported the information given about these distributions.

DISCUSSION and CONCLUSION

In this study, equated scores were computed using the BNP model, bringing a different perspective than classical methods. Gender, mathematics self-efficacy scores, and the sum of common items scores were used as covariates. Equated scores were computed for different covariates, and the distances between these scores' distributions and the distribution of the target test's scores were examined. The explanation of mathematics achievement by the variables and the differences between booklets were interpreted using the BNP model. The results obtained from IRT and BNP models and their interpretation are given below.

The scores taken from common items were considered as the external common test in IRT equating methods; the minimum error was obtained from Mean-Sigma method, whereas the maximum error from the Stocking-Lord method. Therefore, it was concluded that external common items caused more error than moment methods in reducing the difference between items' characteristic curves; and the difference between the discriminant parameters obtained from common tests applied to the groups was less than the difference in characteristic curves. Regarding the distances between the distribution of true scores obtained by IRT scaling methods and distribution of target test's scores, the closest distribution was obtained from Stocking-Lord method. This fact can be expressed as that Stocking-Lord method produced closer values, even though it generated more erroneous predictions than other IRT methods.

In the BNP model, similar score distributions were obtained from female and male students for each booklet when gender was considered as the only covariate. Although gender was seen to be insufficient in showing the difference between the booklets, it was found that booklet 6 was comprised of easier questions than booklet 5. In spite of similar distributions, the confidence intervals of male and female students' distributions were different. Since the same distributions were obtained for the students of both genders, it was concluded that gender has no significant effect on mathematics

performance/achievement. There are various studies supporting this fact in the literature (Hall & Hoff, 1988; Lindberg, Hyde, Petersen, & Linn, 2010; Thien & Darmawan, 2016).

In the BNP model, when MATHEFF was taken as the covariate, the distributions of the students with medium and high scores were similar. The distributions of both booklets varied according to the MATHEFF level; therefore, it was found that MATHEFF was effective on mathematics achievement. Thus, it can be concluded that MATHEFF explains mathematics achievement. The literature contains studies showing that MATHEFF explains mathematics achievement (Ayotola & Adedeji, 2009; Ding, 2016; Hackett & Betz, 1989; Koğar, 2015; Schulz, 2005; Siegle & McCoach, 2007; Thien & Darmawan, 2016). In traditional equating, if the knowledge of individuals is not included, score distributions of each student group would be considered to be the same. In this study, the differentiation in the score distribution of the students in various sub-groups was kept under control, and equated scores of each sub-group were computed. Regarding the model in which MATHEFF was used, it was concluded that the distribution of equated scores approaches the distribution of target test's scores. The most important assumption of NEC design is that the distribution categories obtained from covariates should be the same for the sub-groups (Wiberg & Branberg, 2015). The differences between booklets can be observed using this assumption. Since MATHEFF distributions were similar in both booklets, it was concluded that either this variable could not fully explain the difference between booklets, or the booklets were very similar. However, even in this case, it could be said that booklet 5 contained more difficult questions than booklet 6.

When both MATHEFF and gender were used as covariates in the BNP model, the information obtained from the model was more detailed than the models with a single covariate. If two covariates are used in the model, it is possible to distinguish the variables affecting the distributions of students' mathematics achievement and the magnitude of this effect. The distributions in booklets were the same for both genders. In our case, different distributions were obtained for different booklets and MATHEFF levels. The use of these variables together revealed that they could explain both the difference between booklets and mathematics achievement levels. The distribution of equated scores obtained using two covariates was observed to approach the distribution of target test's scores more than other models.

When the sum of common item scores in the BNP model was used as a covariate, only the distributions of low-score students varied, and the range was quite small. Therefore, the distribution of medium- and high-score students was observed to remain the same. In other words, it was concluded that common items were at the same level and uniform; otherwise they would change the distribution of test scores directly. The same result was obtained for both booklets. The correlation of common item scores was higher for booklet 5 and caused more distributional variations for this booklet. This fact showed that common items were more similar to the questions in booklet 5 and made more distinctions between the sub-groups with different scores in this booklet. Since the distributions obtained from common item scores did not differ significantly according to the booklets, it was concluded that common items don't adequately explain mathematics achievement. The distance between the distribution of the scores equated with common item scores and the distribution of the target test's scores showed the effectiveness of the method but using two covariates in the model was more effective. There are studies supporting the use of covariates for achieving more positive results in equating process, in cases where common items do not possess the properties required for equating or the assumptions of test equating are not satisfied (Dorans & Holland, 2000; Liou et al., 2001; Wright & Dorans, 1993).

When only MATHEFF and only gender were used as a covariate, the distributions did not differ significantly according to booklets. In the model where two covariates were used, distribution differences were observed according to booklets. In the model where the common item scores were used, distribution differences were observed in the low-score student group. This result suggested that in BNP models, common item scores explained the difference between the booklets more than MATHEFF scores. Despite different covariate types used in BNP models, booklet 6 was observed to be easier than booklet 5. Likewise, it is possible to say that the questions in booklet 5 were more distinctive.

Regarding the distributions of equated scores and the distances of these distributions to target test, the comparison between IRT methods and BNP models was straightforward. The distributions of equated scores obtained from the BNP model were closer to the distributions of the target test. The distances between the distributions of equated scores using the BNP model and the distribution of target test's scores were smaller. The closest distance was obtained from the distribution of the BNP model using two covariates together. Therefore, it can be said that more precise estimations are obtained by using BNP model. There are many studies supporting that the Bayesian method makes better predictions than classical methods, and it can be used to obtain much useful information (Karabatsos & Walker, 2009; Kruschke et al., 2012; van de Schoot, et al., 2013).

It was very difficult to compare BNP models that use different covariates according to Hellinger distances. Even though the numerical values obtained from Hellinger distance between BNP models is not sufficient for decision making, the shape of the distributions supported the information about the distance to the target test. Since BNP model uses score distributions for equating, it doesn't require any limitation such as having a same number of individuals in the basic test and target test. Moreover, there is no need to limit the number of individuals in the sub-groups involved in the tests. In the study, the low number of individuals in some sub-groups and the inclusion of covariates to the model as missinformation caused large confidence intervals. However, in spite of large confidence intervals, BNP models would yield more useful and informative results.

As BNP model keeps group invariance under control, the irregularities and discontinuities of the distributions have been eliminated. For this reason, there is no need for pre-smoothing, the selection of the bandwidth parameter, and the derivation of the standard error of equating used in other equating methods (Gonzalez et al., 2015b). This is an indication of the importance of the model (Karabatsos & Walker, 2009).

In future research, researchers may use the model for test equating without using any covariate. When covariate is used in the model, the study can be carried out to determine the items with DIF (Differential Item Functioning) according to variable/s' categories. In the model, equated scores can be obtained using different continuous and discrete covariates such as socioeconomic status, age, etc.

REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Ayotola, A., & Adedeji, T. (2009). The relationship between mathematics self-efficacy and achievement in mathematics. *Procedia Social and Behavioral Science, 1*, 953-957. Retrieved from https://cyberleninka.org/article/n/1232855.pdf

Barrientos, A. F., Jara, A., & Quintana, F. (2012). On the support of MacEachern's dependent dirichlet processes and extensions. *Bayesian Analaysis*, 7(2), 277-310. Retrieved from https://projecteuclid.org/download/pdfview_1/euclid.ba/1339878889

Barrientos, A. F., Jara, A., & Quintana, F. (2016). *Fully nonparametric regression for bounded data using Bernstein polynomials.* Retrieved from http://www.mat.uc.cl/~ajara/Publications_files/DependentBernstein.pdf

Berger, J. O., Boukai, B., & Wang, Y. (1997). Unied frequentist and bayesian testing of a precise hypothesis. *Statistical Science*, *12*(3), 133-160. Retrieved from https://www2.stat.duke.edu/~berger/papers/statsci.pdf

Boone, E. L. Merrick, J. R. W., & Krachey, M. J. (2012). A Hellinger distance approach to MCMC diagnostics. *Journal of Statistical Computation and Simulation*, *84*(4), 833-849. doi: 10.1080/00949655.2012.729588

Branberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*(4), 419-440. doi: 10.1111/j.1745-3984.2011.00153.x

De Iorio, M., Müller, P., Rosner, G., L., & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. J*ournal of the American Statistical Association, 99*(465), 205-215. doi: 10.1198/016214504000000205

Ding, Y. (2016). *How do students' mathematics self-efficacy, mathematics self-concept and mathematics anxiety influence mathematical literacy?-A comparison between Shanghai-China and Sweden in PISA 2012* (Master thesis). University of Gothenburg, Faculty of Education, Gothenburg, Sweden.

Dorans, J. N., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306. doi: 10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS RR-10-29). New Jersey: ETS, Princeton.

González J., & Wiberg M. (2017) Recent developments in equating. In J. González & M. Wiberg (Eds.), *Applying test equating methods: Methodology of educational measurement and assessment* (pp. 157-178). Switzerland: Springer, Cham

Gonzalez, J., Barrientos, A. F., & Quintana, F. A. (2015a). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics and Data Analysis 89*, 222-244. doi: 10.1016/j.csda.2015.03.012

Gonzalez, J., Barrientos, A. F., & Quintana, F. A. (2015b). A dependent Bayesian nonparametric model for test equating. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W-C. Wang, (Eds.) *Quantitative psychology research* (pp. 213-226). New York: Springer Cham Heidelberg New York Dordrecht London.

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education, 20*(3), 261-273. doi: 10.2307/749515

Hall, C. W., & Hoff, C. (1988). Gender differences in mathematical performance. *Educational Studies in Mathematics 19*(1988) 395-401. Retrieved from https://link.springer.com/content/pdf/10.1007%2FBF00312455.pdf

Karabatsos, G., & Walker, S. G. (2009). A bayesian nonparametric approach to test equating. *Psychometrika, 74*(2), 211-232. doi: 10.1007/S11336-008-9096-6

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17-24.

Kim, S., Livingston, S. A., & Lewis, C. (2009). *Effectiveness of collateral information for improving equating in small samples*. New Jersey: ETS, Princeton.

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating in small samples: A preliminary investigation. *Applied Measurement in Education, 24*(4), 302-323. doi: 10.1080/08957347.2011.607057

Koğar, H. (2015). PISA 2012 matematik okuryazarlığını etkileyen faktörlerin aracılık modeli ile incelenmesi. *Eğitim ve Bilim, 40*(179), 45-55. doi: 10.15390/EB.2015.4445

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-36. doi: 10.1111/j.1745-3992.1988.tb00843.x

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3nd. ed.). New York: Springer.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews; Cognitive Science, 1*(5), 658-676, doi: 10.1002/wcs.72

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (Second Ed.)*: A tutorial with R, JAGS, and Stan.* Waltham, MA: Academic Press / Elsevier.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4) 722-752. doi: 10.1177/1094428112457829

Lee, A. H., & Boone, E. L. (2011). A frequentist assessment of Bayesian inclusion probabilities for screening predictors. *Journal of Statistical Computation and Simulation, 81*(9), 1111-1119. doi: 10.1080/00949651003702135

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Mesurment, 49*(2), 167-189. doi: 10.1111/j.1745-3984.2012.00167.x

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135. doi: 10.1037/a0021276

Liou, M. (1998). Establishing score comparability in heterogeneous populations. *Statistica Sinica, 8*, 669-690. Retrieved from http://www3.stat.sinica.edu.tw/statistica/oldpdf/A8n33.pdf

Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement, 25*(2), 197-207. doi: 10.1177/01466210122032000

Livingston, S. A. (2004). *Equating test scores (Without IRT).* Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

Livingston, S. A., & Lewis, C. (2009). Small-sample equating with prior information. (ETS Research Rep. No. RR-09-25). New Jersey: ETS, Princeton.

MacEachern, S. N. (1999). *Dependent nonparametric processes*. Retrieved from https://people.eecs.berkeley.edu/~russell/classes/cs294/f05/papers/maceachern-1999.pdf

MacEachern, S.N., (2000). *Dependent Dirichlet processes* (Tech. rep). Department of Statistics, The Ohio State University. Retrieved from https://people.eecs.berkeley.edu/~russell/classes/cs294/f05/papers/maceachern-1999.pdf

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Boston College, MA, USA: International Study Center.

Mittelhaeuser, M.-A., Beguin, A. A., & Sijtsma, K. (2011). *Comparing the effectiveness of different linking design: The internal anchor versus the external anchor and pre-test data* (Measurement and Research Department Reports, 1). Arnhem: Cito.

Moses, T., Deng, W., & Zhang, Y.-L. (2010). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating* (RR-10-23). New Jersey: ETS, Princeton.

Müller, P., & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*(1), 95-110. doi: 10.1214/088342304000000017

Oh, H. J., Guo, H., & Walker, M. E. (2009). *Impraved reability estimates for small samples using empirical Bayes teshniques* (RR-09-46). New Jersey: ETS, Princeton.

Orbanz, P., & Teh, Y. W.(2010). Bayesian nonparametric models. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning.* Boston, MA: Springer. doi: 10.1007/978-0-387-30164-8_66

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statistics 27*(Varsa sayı no) 105-126. Retrieved from https://www.jstor.org/stable/pdf/3315494.pdf?refreqid=excelsior%3A7e6e0614f5a5f181dfd25d2ad6947bc6

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics 26*, 373-393. Retrieved from https://www.jstor.org/stable/pdf/4616563.pdf?refreqid=excelsior%3A801798d1ac07988dafb6e83769c949b2

Rounder, J. N., Morey, R. D., Speckman, P. L., & Province, M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(2012), 356-374, doi: 10.1016/j.jmp.2012.08.001

Schulz, W. (2005, April). *Mathematics self-efficacy and student expectations: Result from PISA 2003*. Annual Meetings of the American Educational Research Association in Montreal. Retrieved from https://files.eric.ed.gov/fulltext/ED490044.pdf

Shah, A., & Ghahramani, Z. (2013, September). *Determinantal clustering process- A nonparametric bayesian approach to kernel based semi-supervised clustering.* Proceedings of the TwentyNinth Conference on Uncertainty in Artificial Intelligence. Retrieved from http://auai.org/uai2013/prints/papers/200.pdf

Siegle, D., & McCoach, D. B. (2007). Increasing student mathematics self-efficacy through teacher training. *Journal of Advanced Academics, 18*(2), 278-312. Retrieved from https://files.eric.ed.gov/fulltext/EJ767452.pdf

Sinharay, S., & Holland, P. W. (2006). *Choice of anchor test in equating* (RR-06-35). New Jersey: ETS, Princeton.

StataCorp. (2015). *Stata Bayesian analysis reference manual release 14*. College Station, TX: StataCorp LLC. https://www.stata.com/manuals14/bayes.pdf

Thien, L. R., & Darmawan, I. G. N. (2016). Factors associated with Malaysian mathematics Performance in PISA 2012. In L. M. Thien, N. A. Razak, J. Keeves, & I. G. N. Darmawan (Eds.), *What can PISA 2012 data tell us?: Performance and challenges in five participating Southeast Asian countries* (pp. 81-105). Rotterdam: Sense Publisher.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2013). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 1-19. doi: 10.1111/cdev.12169

Wallin, G., & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology – 81st annual meeting of the psychometric society, Asheville, North Carolina*. New York: Springer.

Wei, H. (2010, May). *Impact of non-representative anchor items on scale stability*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Denver, CO.

Wiberg, M. (2015). Anote on equating test scores with covariates. In E. Frackle-Fornius (Ed.), *Festschrift in honor of Hans Nyquist on the occasion of his 65th birthday* (pp. 96-99). Stockholm: Department of Statistics Stockholm University, Sweden.

Wiberg, M., & Gonzalez, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement. 53*(1), 106-125. Retrieved from: http://www.mat.uc.cl/~jorge.gonzalez/papers/TR/Assess_TR.pdf

Wiberg, M., & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing, 17*(2), 105-126. doi: 10.1080/15305058.2016.1277357

Wiberg, M., & Branberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349-361. doi: 10.1177/0146621614567939

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (RR-93–04). New Jersey: ETS, Princeton.

Yıldırım, H. H., Yıldırım, S., Yetişir , M. İ., & Ceylan, E. (2013). *PISA 2012 ulusal ön raporu*. Ankara: MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü (YeğiTek).

Parametrik Olmayan Bayes Yöntemiyle Ortak Değişkenlere Göre Yapılan Test Eşitlemelerinin Karşılaştırılması

Giriş

Denk olmayan gruplarda ortak test deseninde ortak testin seçimi oldukça önemli olup bu test, eşitlenecek olan testler ile benzer ortalama, madde zorluğuna sahip olmalı ve bu testleri içerik olarak temsil etmelidir (Dorans, Moses, & Eignor, 2010; Kolen, 1988; Mittelhaeuser, Beguin, & Sijtsma, 2011; Sinharay & Holland, 2006; Wei, 2010; Wiberg & von Davier, 2017). Ancak ortak testler bu tür özelikleri her zaman sağlayamayabilir. Ortak testlerin tek boyutlu olmaması, diğer testlerdeki puanlarla yüksek oranda ilişki vermemesi, test formlarındaki yapıyı tam olarak ölçmede yetersiz kalması (Wallin & Wiberg, 2017) veya uygulamasından kaynaklı hataların olması (Liou, Cheng, & Li, 2001) eşitlenmedeki güvenirliği ve ortak testlere bağlı diğer süreçleri etkilemektedir (Wiberg & von Davier, 2017; Wei, 2010). Bu durumlara ek olarak, sadece zaman içerisindeki eğilimleri ele alan ortak testlerin denk olmayan gruplarda ankor madde (NEAT) deseninde kullanılması, sadece belirli bireyler için uygun olabilir ki bu durumda eşitleme için bir yanlılık oluşturabilir. Bu da testlerin güvenirliklerini olumsuz yönde etkileyecektir (Wiberg & Branberg, 2015; Wiberg & von Davier, 2017; Wei, 2010). Ayrıca birçok büyük uygulamaları gerektiren sınavlarda ortak madde veya ortak test bulunmamaktadır. Bu durumda test puanları ile ilişkili ve gruplar arasındaki farkı açıklayabilen değişkenlerin kestirim sürecine ek bilgi olarak veya ortak testlerin yerine eklenmesi ile yanlılık ve ortalama standart hata azaltılabilir (Branberg & Wiberg, 2011; Liou ve diğerleri, 2001; Oh, Guo, & Walker, 2009; Wiberg, 2015; Wiberg & Branberg, 2015). Böylece kestirimin doğruluğunu arttırabileceği için eşitleme çalışmaları birçok yönden incelenebilecektir (Branberg & Wiberg, 2011; Kim, Livingston, & Lewis, 2009, 2011; Livingston & Lewis, 2009; Oh ve diğerleri, 2009; Wiberg & Branberg, 2015). Son yıllardaki çalışmalarda ortak maddelerin olmadığı durumda ortak değişkenlerin kullanılması ile Denk Olmayan Gruplarda Ortak değişken (Non-equivalent Groups with Covariates /NEC) (Branberg & Wiberg, 2011; Wiberg & Branberg, 2015) ve hem ortak madde hem de ortak değişkenlerin kullanılması ile NEATNEC deseni literatüre eklenmiştir (Wiberg & Branberg, 2015). Bu çalışma NEC deseni üzerinden yürütülmüştür.

NEC deseninin en önemli varsayımı, ortak değişkenlerin gruplar arasındaki farklılığı açıklayabildiğidir. Test puanlarının durumsal dağılımlarının, ortak değişkenlerin kategorilerine göre her iki grupta da aynı olması bu desen için en önemli adımdır (Wiberg & Branberg, 2015). Bu adımın en önemli parçası olan ortak değişkenlerin seçimi ise oldukça önemlidir. Birçok araştırmacı ortak değişkenleri farklı terimlerle ifade etmiş olsa da bu değişkenlerin test puanları ile ilişkili olması ve gruplar arasındaki farkı açıklayabilecek nitelikte olmasına vurgu yapmıştır (Branberg & Wiberg, 2011; Kim ve diğerleri, 2009; Liou, 1998; Liou ve diğerleri, 2001; Wiberg & Branberg, 2015; Wright & Dorans, 1993). Alanyazında ortak değişken olarak genellikle yaş, cinsiyet, eğitim durumu gibi değişkenlerin yer aldığı görülmektedir (Branberg & Wiberg, 2011; Gonzalez, Barrientos, & Quintana, 2015a; Karabatsos & Walker, 2009; Liou ve diğerleri, 2001; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017).

Ortak testin kullanımından daha iyi sonuç vermesi için ortak değişkenlerin sayısı arttırılabilir. Ancak ortak değişken sayısı arttıkça, bu değişkenlerin kategorilerine düşen birey sayısı azalacağından dolayı değişkenlere ait kategori sayılarının sınırlandırılması daha uygun sonuçlar verecektir (Wiberg & Branberg, 2015; Wallin & Wiberg, 2017).

Son yıllarda Bayes yaklaşımı da test eşitleme çalışmalarında öne çıkmaktadır. Özellikle Parametrik olmayan Bayes yaklaşımı (BNP) ortak değişkenlerin modele eklenmesini olası hale getirmektedir. Bu araştırmada iki farklı ortak değişken kullanılarak NEC deseninde BNP modeline göre elde edilen eşitlenmiş puanlar Madde Tepki Kuramı (MTK) yöntemleri ile karşılaştırılarak test eşitleme sürecine katkısı incelenmiştir.

Yöntem

Araştırmada ortak maddelerin bulunmadığı NEC deseninde farklı ortak değişkenler ile BNP modeli kullanılmıştır. Modellere göre elde edilmiş olan puan dağılımları ve eşitlenmiş puanların hedef teste olan uzaklığı Hellinger uzaklığı ile incelenmiştir. Araştırma gerçek veri üzerinde yürütülmüş olup BNP modeline göre elde edilen eşitlenmiş puanlara ait dağılımlar ile madde tepki kuramına dayalı olarak ölçekleme yöntemlerinden elde edilen eşitlenmiş puanların dağılımları karşılaştırılmıştır.

Araştırmanın evreni ve örneklemi

Denk olmayan gruplar arasında eşitleme yapmak için PISA 2012 verilerinden yararlanılmıştır. Kayıp ve eksik veriler temizlendikten sonra, 5. kitapçık için 908 kişilik İtalya verisi, 6. kitapçık için 931 kişilik Kanada verisi kullanılmıştır.

Veri toplama araçları

PISA 2012 kapsamında öğrencilere uygulanan matematik okuryazarlığını ölçen bilişsel testten ve öğrenci anketinden yararlanılmıştır. NEC deseni için cinsiyet, matematik öz yeterlik puanı (MATHEFF) ve ortak madde puanları ortak değişken olarak alınmış ve ortak değişkenlerin kullanılması ile elde edilen sonuçlar birbirleri ile karşılaştırılmıştır. Çalışma NEC deseninde 24 madde üzerinden yürütülmüş olup, ortak maddelerin toplam puanı ortak değişken olarak kullanılmıştır. NEAT deseninde ise 12 madde dış ortak madde olarak alınmış olup 36 madde üzerinden çalışma yürütülmüştür.

Verilerin analizi

Araştırmada MTK kuramına dayalı ölçek dönüştürme yöntemleri ve BNP modeli için analizler ayrı ayrı sürdürülmüştür. İlk olarak MTK varsayımlarından tek boyutluluk ve yerel bağımsızlık test edilmiş ve testlerin tek boyutlu olduğu sonucuna varılmıştır. Alt ve üst gruplardaki korelasyon ile toplam gruptaki korelasyon birlikte incelenerek yerel bağımsızlık varsayımı desteklenmiştir.

Parametre kestiriminde veri seti ile uyumlu model olarak 3 PLM anlamlı bulunmuş ve analizler bu yönteme göre kestirilmiştir. Madde parametrelerinin kestirimi için Parscale 4.1 programından yararlanılmıştır. Kalibre aşamasında Bayes modellerini temel alan modellerden Expected A Posteriori (EAP) yöntemi kullanılmıştır.

Ölçek Dönüşümü için NEC deseninde ortak değişkenler ortak madde yerine kullanılarak 24 madde üzerinden analizleri gerçekleştirilecektir. NEAT deseninde de ortak maddeler, NEC deseni ile karşılaştırmayı sağlayabilmek için, dış ortak madde olarak alınmıştır. IRTEQ programı ile ölçekleme yapılmıştır. Araştırmada 6. kitapçık hedef test olarak belirlenmiştir. 5. kitapçık temel test olarak alınmış ve gerçek puan hesaplanmıştır.

Parametrik olmayan bayes (bnp) yaklaşımına göre test eşitleme: BNP yöntemi kullanılarak yapılan eşitleme çalışmaları ile eski ve yeni test puanları arasında kurulabilecek ilişki ortak değişkenlerin sürece katılması ile şekillendirilmiştir. Modelde yer alan parametrelerin kestirimlerinde uygun

sonuçlar elde edebilmek için MCMC yöntemi kullanılmıştır. MCMC örnekleme süreci ile hazırlanan dosyalarda DBPP modeli kullanılarak veriye uygun parametreler ve ortak değişkenler birleştirilmektedir. Kanada ve İtalya veri setleri için ayrı ayrı MCMC süreçleri yürütülmüştür. Daha sonra ise eşitleme fonksiyonundan yararlanarak eşitlenmiş puanlar elde edilmiştir. Çalışmada elde edilen puan dağılımları ile birlikte güven aralıklarına da yer verilmiştir.

BNP modeli için, Gonzalez ve diğerlerinin (2015a, 2015b) çalışmalarında kullanmış olduğu formüllerden yararlanılarak R 3.2.1 programında kodlar oluşturularak analizler gerçekleştirilmiştir.

Karşılaştırma kriteri: Çalışmada, MTK yöntemleri ile BNP Yöntemi ile elde edilen eşitlenmiş puanları karşılaştırmak için istatistiksel bilgi veren ve eşitlenmiş puanlara ait dağılımların hedef teste olan uzaklıklarını inceleyen Hellinger Uzaklığı kullanılmıştır.

Sonuç ve Tartışma

Araştırmada ortak maddelerden elde edilen puanlar dış ortak test olarak alınmıştır. Ortak maddelerin parametreleri üzerinden yapılan ölçekleme sonucunda Stocking-Lord yönteminin diğer MTK yöntemlerine göre daha hatalı kestirim yapmış olsa dahi gerçek puan olarak hedef teste daha yakın değerler ürettiği şeklinde ifade edilebilir. Li, Jiang ve von Davier (2012) de araştırmasında MTK gerçek puan eşitleme ile elde edilen puanların daha doğru ve kesin olduğunu vurgulamaktadır.

BNP modelinde ortak değişken olarak sadece cinsiyet ele alındığında, kız ve erkek öğrenciler için kitapçıklarda benzer dağılımlar elde edilmiştir. Cinsiyet değişkenin kitapçıklar arasındaki farkı göstermede yetersiz olduğu sonucu görülse de 6.kitapçığın 5.kitapçıktan daha kolay sorular içerdiği sonucu elde edilmiştir. Ortak değişken olarak cinsiyetin kullanıldığı araştırmaları literatürde görmek mümkündür (Branberg & Wiberg, 2011; Gonzalez & Wiberg, 2017; Gonzalez ve diğerleri, 2015a, 2015b; Liou ve diğerleri, 2001). Aynı kitapçığı almış olan kız ve erkek öğrenciler için güven aralıkları farklılık gösterse de dağılımları oldukça benzer olup cinsiyetin matematik performansı üzerinde önemli bir etkisinin olmadığını göstermektedir. Literatürde bu durumu destekleyen benzer çalışmaların yer aldığını görmek mümkündür (Hall & Hoff, 1988; Lindberg, Hyde, Petersen, & Linn, 2010; Thien & Darmawan, 2016).

BNP modelinde ortak değişken olarak MATHEFF alındığında tüm düzeylerdeki bireylere yönelik üç boyutlu bir dağılım grafiğine yer verilmiştir. Orta ve yüksek puana sahip bireylere ait dağılımlar benzerlik göstermiş, düşük düzeydeki puana sahip bireylere ait dağılımlar ise farklılaşmıştır. Kitapçıkların her ikisi için de dağılımlar MATHEFF puan düzeyinde göre değişim gösterdiğinden, MATHEFF değişkeninin matematik performansında bireyler arasındaki farkı ortaya koyduğu sonucuna ulaşılmaktadır. Dolayısı ile MATHEFF ortak değişkeninin matematik başarısını açıkladığı sonucuna ulaşılabilir. Literatürde MATHEFF değişkeninin matematik başarısını açıkladığını gösteren çalışmalar yer almaktadır (Ayotola & Adedeji, 2009; Ding, 2016; Hackett & Betz, 1989; Koğar, 2015; Thien & Darmawan, 2016; Schulz, 2005; Siegle & McCoach, 2007). Geleneksel yöntemle yapılan eşitleme çalışmalarında bireylere ait önsel bilgilere yer verilmemesi durumunda her birey için eşitleme dağılımları aynı olarak alınacaktır. Bu çalışma ile bireylere ait puan dağılımlarının alt gruplarda farklılaşması kontrol altında tutularak, alt gruplara göre eşitlenmiş puanlar elde edilmiştir. MATHEFF değişkenin modelde kullanılması ile eşitlenmiş puanlardan elde edilen dağılımın, hedef testteki puanlara yaklaştığı sonucunu ortaya çıkarmaktadır. NEC deseninde ortak değişkenlerden elde edilen dağılımlara ait kategorilerin alt gruplar için aynı olması (Wiberg & Branberg, 2015) varsayımdan yararlanılarak kitapçıklar arasındaki farklar gözlenebilmektedir. MATHEFF değişkeninin her iki kitapçıkta da benzer dağılımlar vermiş olması ile kitapçıklar arasındaki farkı tam olarak açıklayamadığı veya kitapçıkların birbirlerine oldukça benzer oldukları söylenebilir. Fakat bu durumda dahi, bu alt problem için elde edilen sonuçlarda 5.kitapçığın, 6.kitapçığa kıyasla zor sorular içerdiği ifade edilebilir.

MATHEFF ve cinsiyet birlikte ortak değişken olarak BNP modelinde kullanıldığında daha önceki alt problemlere kıyasla modelde daha detaylı bilgiler elde edilmiştir. Bu alt problem ile hangi değişkenin bireylerin matematik başarısına ait dağılımlarını ne kadar değiştirdiğini görmek mümkündür. Bu iki değişken birlikte ele alındığında, her kitapçık ve MATHEFF değişkenindeki her puan düzeyi için farklı

dağılımlar oluşturduğundan, bu değişkenlerin birlikte hem kitapçıklar arasındaki farkı hem de matematik başarısını açıklayabildiği sonucunu ortaya koymuştur. İki ortak değişken kullanımı ile elde edilen eşitlenmiş puanların dağılımının hedef test puanlarına ait dağılıma yaklaştığı sonucu gözlemlenmiştir.

BNP modelinde ortak madde puanları ortak değişken olarak alındığında bireylere ait elde edilen puan dağılımları sadece düşük puanlarda ve çok az bir ranjda değişmektedir. Dolayısı ile ortak maddelerden yüksek puan alan bireyler ile düşük puan alan bireylerin puan dağılımları benzerlik göstermektedir. Bu da farklı düzey ortak madde puanına sahip öğrencilerin matematik başarıları arasında net bir ayrım yapılmadığını göstermektedir. Yani ortak maddelerin aynı düzey ve tek tip olduğu veya direkt test puanlarına etki ederek dağılımlarını değiştirdiği sonucunu ortaya çıkarmaktadır. İki kitapçık için de bu durum benzer şekildedir. Ancak ortak madde puanlarının 5.kitapçıkla daha yüksek korelasyon vermesi ve bu kitapçıktaki dağılımlarda daha çok değişim yapmış olması, ortak maddelerin 5.kitapçıktaki sorulara daha çok benzediği ve bu kitapçıktaki farklı puan almış alt gruplar arasında daha fazla ayrım yaptığını göstermektedir. Ortak madde puanlarından elde edilen dağılımların kitapçıklara göre büyük bir farklılık göstermemesi, ortak maddelerin matematik başarısını yeterli düzeyde açıklamadığı sonucunu ortaya çıkarmıştır. Ortak madde puanlarının kullanılması ile elde edilen eşitlenmiş puanlar ile hedef teste ait dağılım arasındaki uzaklık yöntemin etkili olduğunu ancak iki ortak değişken kullanılmasının ortak maddelerden daha etkili olduğu sonucunu ortaya çıkarmıştır. Ortak maddelerin eşitleme için gereken özellikleri taşımadığı veya test eşitleme için varsayımların ihlal edildiği durumlar için, ortak değişkenlerin kullanılmasının eşitleme sürecinde daha uygun sonuçlar vereceğini destekleyen çalışmalar literatürde yer almaktadır (Dorans & Holland, 2000; Liou ve diğerleri, 2001; Wright & Dorans,1993).

Sadece MATHEFF ve sadece cinsiyet değişkeni kullanıldığında dağılımlar kitapçıklara göre aşırı bir farklılık göstermemektedir. İki ortak değişkenin kullanıldığı modelde dağılımların kitapçıklara göre farklılıkları açık bir şekilde görülmekte; ortak madde puanlarının kullanıldığı modelde ise düşük ortak madde puanlarında kitapçıklara göre dağılımların farklılaştığı görülmektedir. Bu durum BNP modellerinde; ortak madde puanlarının kitapçıklar arasındaki farkı, sadece MATHEFF değişkeni kullanıldığı modelden daha çok açıkladığı sonucunu ortaya çıkarmaktadır.

Bütün BNP modellerinde farklı ortak değişkenler kullanılsa dahi 6.kitapçığın 5.kitapçıktan daha kolay olduğu ve bu kitapçıkta bireylerin yüksek puan olma yoğunluğunun fazla olduğu sonucu ortaya çıkmaktadır. Aynı şekilde yine her model için 5.kitapçıktaki soruların daha ayırıcı olduğunu söylemek mümkündür.

Eşitlenmiş puanlara ait dağılımlar ve bu dağılımların hedef teste uzaklıkları incelendiğinde MTK yöntemleri ve BNP modelleri arasında karşılaştırma yapmak kolaydır. BNP modeli ile elde edilmiş olan eşitlenmiş puanlar için hesaplanan Hellinger Uzaklığı, MTK ölçek dönüştürme yöntemlerine göre oldukça düşük olup, bu dağılımlar hedef teste daha yakındır. Bu dağılımlardan en yakın uzaklığı iki ortak değişkenin kullanıldığı BNP modeli vermiştir. Dolayısı ile eşitlenmiş puan-hedef teste ait dağılımların birbirlerine MTK yöntemlerine kıyasla yakınlaştığı ve bu model kullanılarak daha kesin kestirimler elde edildiği sonucuna ulaşılmıştır. Bayes yönteminin klasik yöntemlerden daha iyi kestirim yaptığını ve daha yararlı bilgiler için de kullanılabileceğini ifade eden çalışmalar bu sonucu desteklemektedir (Karabatsos & Walker, 2009; Kruschke, Aguinis, & Joo, 2012; van de Schoot ve diğerleri, 2013).

BNP modeli ile grup değişmezliği kontrol altında tutulduğu gibi dağılımların düzensizliği ve süreksizliği de giderilmiş olduğundan; diğer eşitleme yöntemlerinde kullanılan ön-düzgünleştirme, bant genişliği parametresinin seçimi ve eşitlemenin standart hatasının türetilmesine ihtiyaç duyulmamaktadır (Gonzalez ve diğerleri, 2015b). Bu durum ise modelin önemliliğinin bir göstergesidir (Karabatsos & Walker, 2009).