Id: 1263893

Participatory Educational Research (PER)
Vol.10(5), pp. 41-63, September 2023
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.23.74.10.5

A Comparison of Covariates, Equating Designs, and Methods in Equating TIMSS 2019 Science Tests

Elif Sezer Başaran [*]
Assessment and Evaluation in Education, Bursa Uludağ University, Bursa, Türkiye
ORCID: 0000-0002-7302-2724

Ceren Mutluer
Assessment and Evaluation in Education, Bolu Abant İzzet Baysal University, Bolu, Türkiye
ORCID: 0000-0002-3935-336X

Mehtap Çakan
Assessment and Evaluation in Education, Gazi University, Ankara, Türkiye
ORCID: 0000-0001-6602-6180

Article history
Received:
12.03.2023

Received in revised form:
1.05.2023

Accepted:
11.07.2023

Key words:
Anchor test design; covariates design; equating methods based on classical test theory; kernel equating; TIMSS 2019 science test

This research aimed to compare the equated scores by the methods based on classical test theory (CTT) and kernel equating, using covariates design (NEC) and anchor test design (NEAT). TIMSS 2019 science test scores equated by both Tucker, Levine true score, Levine observed score, equipercentile equating (pre-smoothing and post-smoothing) methods in CTT, and linear and equipercentile methods in kernel equating. Additionally, the covariates in NEC design were “home resources for learning,” “student confidence in science and mathematics,” “like learning science,” “instructional clarity in science lessons,” “math achievement,” “sex,” and “speaking the language of the test at home”. The equating results in NEC were compared with those in NEAT and EG. The participants comprised 1699 4th-grade students who attended the e-TIMSS 2019 in Canada, Singapore, and Chile. Results were analyzed according to equating errors and differences between equated scores. The research concluded that math achievement and home resources for learning could be used as covariates in NEC to equate the science test in case equating could not be done in the NEAT. However, when the other variables were used as covariates in NEC, the equated scores were very similar to the EG. Also, Tucker (CTT) and post-stratification (kernel) yielded similar equated scores in linear equating, and these methods were similarly different from kernel linear equating in EG. In equipercentile equating, the equated scores obtained from the post-smoothing (CTT) and EG were close to each other but slightly differed from post-stratification.

[*] Correspondency: eliffszr@gmail.com

Introduction

Comparing test results from different applications and testing conditions is an important issue in measurement and assessment. In some tests, students are handled different test booklets with different questions. Sometimes, individuals may have to apply for the same position with the tests held at different times. In such cases, it is necessary to perform certain statistical analyses to compare the test scores and ensure fairness in decision-making based on the test results. Statistical techniques used to compare test results are called 'test equating' methods. According to Kolen and Brennan (2014, p. 2), test equating is "a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably". Also, they emphasize that equating adjusts the difference between tests of similar difficulty and content.

Test scores with similar difficulty and content are equated with different data collection designs and statistical estimation methods. Equating designs may change depending on the number of tests answered by student groups, the application order of different tests, and whether tests have anchor items. According to Kolen and Brennan (2014), there are three commonly used equating designs: random groups, single group with counterbalancing, and common-item in non-equivalent groups. In random groups or equivalent groups design (EG), one group takes the X test, and the other takes the Y test. In the equating process, X test scores are converted to Y scores, or Y test scores are converted to X scores. All items in the X and Y tests may be different from each other, or some items in the two tests may be the same (Livingston, 2014). However, in this design, there may be differences between groups based on ability, which may affect the equating scores (Kolen & Brennan, 2014; Lu & Guo, 2018; Lyren & Hambleton, 2011).

In the non-equivalent groups with anchor test design (NEAT), some items in the X and Y tests are the same (i.e., common items), and the test scores are equated by considering these common items. However, there may be problems such as bias on test items, translation errors, and item deletion. For instance, if the number of differential item functioning (DIF) items in the test is high, the equating error in this design may be high (Atalay Kabasakal & Kelecioğlu, 2015; Yurtcu & Guzeller, 2018). If some of the common items are DIF items and the magnitude of DIF is large, this may create a more important problem (Atar, Atalay Kabasakal & Kibrislioglu Uysal, 2023). Sometimes, it is not possible to use common items for reasons such as safety (such as the Academic Personnel and Postgraduate Education Entrance Exam). An alternative method is the non-equivalent groups with covariates design (NEC).

Differences across groups are corrected in NEC by using covariates (González & Wiberg, 2017; Sansivieri et al., 2017). Variables such as gender, socioeconomic status, course achievement, and affective characteristics can be selected as covariates in the NEC. In a study by Wiberg and Branberg (2015), the scores obtained from a different standardized test and school scores were selected as covariates using university entrance exam scores. In the study by Yurtçu, Kelecioğlu, and Boone (2021), test scores of PISA 2012 Canada and Italy were equated by selecting gender and mathematics self-efficacy scores in data as the covariates. Similarly, in Akın Arıkan's (2020) study on the 2016 Monitoring and Evaluation of Academic Skills Project, gender and socioeconomic level were chosen as covariates. In Altintas and Wallin's (2021) study on Ankara University Examination for Foreign Students, gender and age were selected as covariates.

In addition to the equating designs such as NEC and NEAT, the equating methods are also essential for test equating. For example, in kernel equating, discrete score distributions are continued by using kernel smoothing methods. The equating process includes five steps: pre-smoothing, estimation of scores probabilities, continuization, computation of equating transformation, and computation of accuracy measures. Kernel equating can be realized by linear and equipercentile equating (EQ) methods. Additionally, among the popular test equating methods, some are based on classical test theory (CTT). For example, Tucker, Levine observed score (LevineOS), and Levine true score method (LevineTS), chained linear equating, can be used for linear equating in NEAT. Similarly, the frequency estimation method and chained EQ methods can be used for EQ. Some findings in the literature suggest that kernel equating and the methods based on CTT are close to each other, but kernel equating results have relatively fewer errors (e.g., Akın Arıkan & Gelbal, 2018; Liu & Low, 2008; Mao et al., 2006; von Davier et al., 2006).

A literature review on comparing equating designs yielded contradictory results. In some studies, more accurate results were obtained in NEAT than in NEC (Branberg & Wiberg, 2011), while others revealed better results in NEC than in NEAT design (Akın-Arıkan, 2020; Yurtçu, Kelecioğlu & Boone, 2021; Wallin & Wiberg, 2019; Wiberg & Branberg, 2015). According to Lu and Guo (2018), if the number of items in the anchor test is low, more accurate results are obtained in the covariate design than in the anchor test design. Even if there is no anchor test in some studies, it is suggested that equating can be done in NEC (Akın-Arıkan, 2020; Altintas & Wallin, 2021). However, in a study by Wiberg and Branberg (2015), university entrance test scores were equated by kernel equating, and verbal achievement test scores were chosen as the covariate. This study found that NEC had less equating error than NEAT, and NEAT had less equating error than EG. A similar conclusion was reached in a study by Wallin and Wiberg (2019). In Akın Arıkan (2020)'s study, gender and socioeconomic status were selected as covariates, and mathematics test scores were equated by kernel equating. The research results revealed that the least equating error was in the NEC, and NEAT equating by post-stratification method (PSE), one of kernel equating methods. In general terms, it was inferred that equating could be done in the NEC even if there was no anchor item.

The number of studies on NEC is limited in the literature, and they use a limited number of covariates. One of the underlying reasons is the low number of student- and school-oriented features in the equated tests. In this sense, such studies should address large-scale tests with a variety of data with many student- and school-oriented variables. The present study will focus on TIMSS 2019.

The fourth-graders from 58 countries participated in the TIMSS 2019 conducted by the International Association for the Evaluation of Educational Assessment. TIMSS is a comprehensive worldwide assessment and assumes that basic science and mathematics knowledge and understanding contribute to making sound financial decisions, using practical problem-solving skills, and having a fruitful personal life (Mullis, 2013). The present study focused on science achievement in the TIMSS 2019, in which both students' science achievement scores are calculated, and proficiency level in science is determined. Additionally, surveys (for students, parents, teachers, and school administrators) are applied to detect the variables that may affect achievement.

TIMSS provides researchers with a comprehensive data set, including achievement-related factors such as student characteristics, school resources, teacher qualifications, teaching practices, and family characteristics (e.g., language spoken at home, home resources for learning) (House, 2006; Leung, 2002). In the literature, the variables that can affect science achievement involve student characteristics (Coşkun, 2021; Pajares, 2008; Sarıer, 2020), school and teacher qualifications (Aydoğan & Gelbal, 2022; Coşkun, 2021; Sarıer, 2020), and family characteristics (Özkan, 2018; Salaway, 2008; Soysal, 2019; Üstün, 2007). The present study equated the scores from different science achievement test booklets in TIMSS 2019 by using the relevant covariates.

Besides, both methods based on CTT and kernel equating in NEAT were performed in this study. NEC results were compared with NEAT and EG results. Many student characteristics related to science achievement were selected as covariates. These covariates are home resources for learning, students confident in science and mathematics, like learning science, instructional clarity in science lessons, math achievement, sex and speaking the language of the test at home. The research results will contribute to selecting an appropriate equating design with covariate, and especially to equate standardized science tests.

Study Goal

This study aimed to compare the results from methods based on CTT and kernel equating in NEC and NEAT using TIMSS 2019 science test scores. The research questions are as follows:

1. Among the equating methods based on CTT, such as Tucker, LevineTS, LevineOS and EQ equating (pre-smoothing and post-smoothing), which one has the least equating error in the equated scores in NEAT obtained from science tests?
2. Among the kernel linear and EQ methods, which one has the least equating error in the equated scores in NEC and NEAT obtained from TIMSS 2019 science tests?
3. How approximate do the kernel equating results in NEC to the equating results based on kernel and CTT in NEAT, and the kernel results in EG?

Method

Design

This study aimed to find the equating method that yielded the most minor error among several equating methods based on CTT and kernel equating in NEC and NEAT. It also examined the covariates that could be used to equate science test scores. This study used basic research as it compared results from different equating methods and contributed to the existing theory by providing information (Wiersma & Jurs, 2005).

Sample

TIMSS 2019 research was conducted using both paper-and-pencil, and computer versions. Some countries took the tests on paper-and-pencil, while others used the digital platform (e-TIMSS). Also, in TIMSS research, students' achievement scores were calculated and their proficiency levels were determined according to different score ranges: "advanced" for 625 and above; “high” for 625 and 550; “intermediate” for 550 and 475; “low” for 475 and 400; “below low” for scores less than 400. The test version, achievement scores and proficiency levels of the countries participating in the research were reported in detail by the TIMSS team (Mullis et al., 2020).

This research examined the fourth-graders from several countries who participated in the TIMSS 2019 science test. First of all, the countries that answered the paper-and-pencil test were determined and classified according to their average mean success scores (e.g., high-level countries, and middle-level countries). This study specifically focused on the countries with many students from various science proficiency levels in the e-TIMSS. However, since all countries with an average score of "below low" benchmark took the paper-and-pencil version of TIMSS 2019, their data were not included in this study. Accordingly, the data from Singapore (595, high), Canada (523, intermediate), and Chile (469, low) were analyzed. Table 1 shows the student distribution by science proficiency in TIMSS 2019.

Table 1. TIMSS 2019 Science Achievement and Sample Size of Countries

		*International Benchmarks of Science Achievement**					*Sample Size*	
Country	*Average Scale Score*	*Advanced (625)*	*High (550)*	*Intermediate (475)*	*Low (400)*	*Below Low*	TIMSS 2019	This Study
Singapore	595 (High level)	38	36	19	5	2	5986	724 (42.6%)
Canada	525 (Intermediate)	7	30	38	20	5	13653	710 (41.8%)
Chile	469 (Low level)	1	13	34	34	18	4174	265 (15.6%)

Note. Percentage of students per country at proficiency levels in TIMSS 2019 (Mullis et al., 2020).

A total of 23813 students from Canada, Chile, and Singapore participated in the e-TIMSS. This study used the data only from students who answered the first and second science test booklets (Booklet 1 and Booklet 2) and did not have missing data in covariates (such as home resources for learning, students confident in science and mathematics, like learning science, instructional clarity in science lessons, math achievement, sex and speaking the language of the test at home). Providing assumptions among the countries where e-TIMSS application was used, three countries were selected among the countries with high, intermediate, and low achievement levels among the countries where the number of people per booklet was not small and the missing data was the lowest. Three representative countries have been identified where these conditions are met. Accordingly, the sample comprised 1699 students from Canada, Chile, and Singapore, who participated in the e-TIMSS application in TIMSS 2019 at the fourth-grade level.

Instruments

The data of this study was obtained from the official website of TIMSS (International Association for the Evaluation of Educational Achievement, 2021). The instruments were science achievement test, student and school questionnaires. In the TIMSS 2019 science achievement test of fourth grade, 45% of the total test score was from life science, 35% was from physical science, and 20% was from Earth science. The science achievement test was administered in the form of 14 booklets. In this study, Booklet 1 and Booklet 2 of the TIMSS 2019 science test applied as e-TIMSS were used because the number of test items was close and open-ended item numbers was low. Consisting of two subtests of 12 and 18 items, these booklets include multiple choice and constructed response (some are partially scored) items. In this study, correct or completely correct answers were coded with "1"; partially correct or wrong answers are coded with “0”.

In addition, the covariant variables in this study were (i) “home resources for learning,” (ii) “students confident in science,” (iii) “students confident in mathematics,” (iv) “students like learning science,” (v) “instructional clarity in science lessons,” (vi-vii) “math achievement

(Math1 and Math5)," (viii) "sex of students" and (ix) "speaking the language of the test at home". In this study, the mathematics achievement data were obtained from the mathematics achievement test and the data for other covariates were obtained from the student questionnaire. Each covariate was calculated as follows (Yin & Fishbein, 2020).

Home resources for learning was a scale measured and categorized by TIMSS (ASDGHRL). This variable included the number of books and children's books in the home, number home study supports, highest level of education and occupation of either parents. The Cronbach alpha coefficient was .59 in Canada, .65 in Chile, and .65 in Singapore. Students were expressed in three categories by TIMSS as many resources, some resources and few resources.

Students confident in science (ASDGSCS) and students confident in mathematics (ASDGSCM) were scales measured by TIMSS. The science scale consisted of seven items such as "I usually do well in science" and "Science makes me confused". The Cronbach alpha coefficient was .84 in Canada, .75 in Chile and .85 in Singapore. The mathematics scale consisted of nine items such as "I usually do well in mathematics", "Mathematics makes me confused" and "Mathematics is harder for me than any other subject". The Cronbach alpha coefficient was .87 in Canada, .82 in Chile and .87 in Singapore. Both of the scales were the 4-point Likert-type scale (i.e., agree a lot, agree a little, disagree a little, disagree a lot). Students were expressed in three categories by TIMSS as very confident, somewhat confident and not confident.

Students like learning science was a scale measured by TIMSS (ASDGSLS). The 4-point Likert-type scale (i.e., agree a lot, agree a little, disagree a little, disagree a lot) consisted of nine items such as "I enjoy learning science" and "I like to do science experiments". The Cronbach alpha coefficient was .91 in Canada, .85 in Chile and .91 in Singapore. Students were expressed in three categories by TIMSS as very much like, somewhat like and do not like.

Instructional clarity in science lessons was a scale measured by TIMSS (ASDGICS). The 4-point Likert-type scale (i.e., agree a lot, agree a little, disagree a little, disagree a lot) consisted of six items such as "I know what my teacher expects me to do" and "My teacher explains a topic again when we don't understand". The Cronbach alpha coefficient was .83 in Canada, .80 in Chile and .87 in Singapore. Students were expressed in three categories by TIMSS as high clarity, moderate clarity and low clarity.

Students' mathematics scores were calculated as five plausible values and categorized according to international benchmarks. Thus, these scores were expressed as a categorical variable with values between 1 and 5 by TIMSS. This study used the first and fifth plausible values (ASMIBM01 and ASMIBM05).

Sex of students was an item in the student questionnaire (ITSEX). The female was coded with "1" and male with "2". In addition, speaking the language of the test at home was an item in the student questionnaire (ASBG03). The question was "How often do you speak <language of test> at home?". The sections were "always", "almost always", "sometimes" and "never". This study combined the options and coded as yes (always and almost always) and no (sometimes and never).

Data Analysis

This study equated science scores from Booklet 1 (X test, new form) to scores from Booklet 2 (Y test, old form) using TIMSS 2019 data. For the test equating process, it was tested whether the assumptions of symmetry, measuring the same specification, equal reliability, independence from the group, equality were provided. Since the assumptions were provided, equated scores were calculated according to the equating designs and methods in line with the purpose of the research. The equated scores were obtained using the kequate (Andersson, Branberg & Wiberg, 2013, 2022) and equate (Albano, 2016) packages in R program (for Windows 4.2.2), and Rage. Equating in NEC and NEAT designs were coded as "EQ" for equipercentile and "L" for linear. The NEC designs were home resources for learning (RESOURCE and RESOURCE.L), students confident in science (SCICONF and SCICONF.L), students confident in mathematics (MATHCONF and MATHCONF.L), students like learning science (LEARNSCI and LEARNSCI.L), instructional clarity in science lessons (CLARITY and CLARITY.L), math achievement (MATH1 and MATH1.L, MATH5 and MATH5.L), sex of students (SEX and SEX.L) and speaking the language of the test at home (LANG and LANG.L).

In NEAT, these tests were equated with the external anchor. In addition, P and Q populations were equally weighted to form the synthetic population ($w_P = w_Q = 0{,}5$). The equating methods based on CTT were Tucker, LevineOS and LevineTS for linear; pre-smoothing and post-smoothing for EQ. For kernel equating, the equating methods in NEAT were PSE and chained equating (CE), both linear and EQ; the methods in NEC and EG were linear and EQ.

In kernel equating, log linear models were used in the pre-smoothing step. Gauss kernel function was used for continuation. Bandwidth selection is very important in kernel equating. If the bandwidths (h parameter) is ideal, EQ are obtained, and if it is wide, linear equating functions are obtained. The bandwidths were selected by the R program (kequate package) in this study. There were the bandwidths in Appendix 1.

The equating results obtained by different methods can be compared with the equating errors. However, the estimation of the equating accuracy depends on the framework adopted (e.g., methods based on CTT, kernel equation, item response theory) (Wiberg & González, 2016). For example, percent relative error (PRE) and standard error of equating (SEE) values can be used to evaluate kernel equating results (von Davier et al., 2004). The observed and equated score distribution moments are compared with the PRE values. That is, the differences between the ten moments of both discrete distributions are calculated. SEE is calculated by considering the equating function, equating design, and pre-smoothing score distributions. PRE, and SEE equations are as follows, denoted by the Jacobian matrix of the equating function is "J_φ"; the Jacobian matrix of the equating design "J_{DF}" and asymptotic covariance matrix of the score distributions in the pre-smoothing "C" (von Davier et al., 2004):

$$PRE(p) = 100\,\frac{\mu_p(\varphi(X)) - \mu_p(Y)}{\mu_p(Y)} \qquad (1)$$

$$SEE = \left\| J_\varphi J_{DF} C \right\| \qquad (2)$$

In contrast, the difference that matters (DTM) between the equated scores and scale scores was examined in studies that performed equating methods based on CTT (Wiberg & González, 2016). Graphs can be drawn to express the difference between the equated score and the evaluating criterion for the DTM. DTM can be used to evaluate kernel equating results (Wiberg & González, 2016). According to the DTM proposed by Dorans and Feigenbaum (1994), the difference between the equating score and the criterion is interpreted by considering the score unit (as cited in Liu et al., 2014). In addition, Suh et al. (2009) also evaluated the marginal differences by evaluating the score differences according to .05. This study generated graphs of the equated difference for each score and examined the magnitude of the difference with DTM and marginal meaningful. Since the science test scores were obtained with the total number of correct answers, the score unit is one and the DTM is determined as .5, which is half of the score unit (1/2). Accordingly, the score differences were interpreted as follows (Suh et al., 2009):

- meaningful if the difference is greater than or equal to .5 and there is DTM,
- marginally meaningful if the difference is between .05 and .49; but there is no DTM,
- not meaningful if the difference is less than 0.05 and there is no DTM.

Also, root mean squared difference (RMSE) can be used to determine the differences in equated scores. The equation of RMSE is as follows, where “$\bar{d}$” is the mean of the differences and "sd” is the standard deviation of the differences (von Davier et al., 2006):

$$RMSE = \sqrt{\bar{d}^2 + sd_d^2} \qquad (3)$$

RMSD (root mean squared difference) and WMSE (weighted mean square error) can be used to determine the difference in equated and raw scores. According to Kim and Lu (2018), “w_i” is the relative distribution of scores in the new form (X form), and the RMSD is calculated as follows:

$$RMSD = \sqrt{\sum_{i=0}^{k} w_i [\hat{e}_i(x_i) - e_i(x_i)]^2} \qquad (4)$$

In the WMSE coefficient calculated by Equation 5, “k” is the number of items in the X test; “S_X^2” is the variance of the raw scores on the X test; “X_{crit}” is the raw score of i in the X test; “X_E” is the equated scores obtained by different equating methods; “f_i” is the raw score frequency of i in the X test (Skaggs & Lissitz, 1986).

$$WMSE = \frac{\sum_{i=1}^{k-1} f_i (X_E - X_{crit})^2}{\sum_{i=1}^{k} f_i S_X^2} \qquad (5)$$

In addition, Newton-Raphson's Delta method can be used to determine the errors of equating methods based on CTT. Kolen and Brennan (2014) presented the following equations for linear (equation 7) and EQ (equation 8). In these equations, “x_i” is the equated score; “μ(X)” is the mean of X test; “$\sigma_{(X)}$” is the standard deviation of X test; “$\sigma^2_{(Y)}$” is the variance of Y test; “$P(x_i)$” is the percentile rank of X test and “∅” is the ordinate of the standard normal density at the unit-normal score.

$$N_{Total=}N_X+N_Y \qquad (6)$$

$$\text{Delta } l_Y(xi) \cong \sqrt{\left(\frac{2\sigma^2(Y)}{NTotal}\left\{2 + \left[\!\left[\frac{x_i - \mu(X)}{\sigma(X)}\right]\!\right]\right\}\right)} \qquad (7)$$

$$\text{Delta } e_Y(xi) \cong \sqrt{\left(\frac{4\sigma^2(Y)}{NTotal}\left\{\frac{(P(x_i)/100)(1-\left(\frac{P(x_i)}{100}\right))}{\emptyset^2}\right\}\right)} \qquad (8)$$

Results

For the test equating, we first calculated the descriptive statistics of the X and Y tests. Then we equated the tests and reported the equating's results according to the sub-problems. The descriptive statistics of the tests are shown in Table 2.

Table 2. Descriptive Statistics of New and Old Forms

	New Form		*Old Form*	
Statistics	*X Test*	*Anchor Test*	*Y Test*	*Anchor Test*
N	846	846	853	853
Item numbers	12	18	12	18
Mean	7.36	12.53	7.14	12.91
Standard deviation	2.580	3.242	2.498	3.066
Skewness	-.388	-.372	-.268	-.482
Kurtosis	-.469	-.626	-.655	-.227
Min	0	2	0	3
Max	12	18	12	18
Reliability*	.686	.735	.667	.703
Correlation**	.674		.650	

Note. *Cronbach alpha.
**Pearson correlation coefficient between the main test score and the anchor test.

According to Table 2, the item numbers of the main tests were equal and 12. The arithmetic mean of the tests was close to each other. The level of correlation between the main test and the anchor test was close to each other. Fisher Z coefficients were also calculated in testing the assumption of equality of reliability (Z_X = .84; Z_Y= .81). It is also seen that the reliability coefficients are very close to each other and the equality assumption of the reliability is provided with this information. The skewness and kurtosis coefficients are between -1.00 and +1.00.

Additionally, polyserial and point-biserial correlations between covariates and test scores were calculated using the psych package (Revelle, 2022) in the R program. According to Guilford (1956), the level of the relationship between the variables is interpreted as follows: "very weak" when the correlation is lower than .20, "weak" when it is between .20-.39, "medium" when it is between .40-.69, "high" when it is between .70-.89, and "very high" above .90. The correlations between covariates and test scores are shown in Table 3.

Table 3. Correlation Between Covariates and Test Scores

Covariant Variable	*X Test*	*Y Test*	*Comment*
Math achievement (Math1)	.757	.730	High
Math achievement (Math5)	.762	.709	High
Home resources for learning (RESOURCE)	.283	.273	Weak
Students confident in science (SCICONF)	.089	.178	Very weak
Students confident in mathematics (MATHCONF)	.182	.184	Very weak
Students like learning science (LEARNSCI)	.060	.127	Very weak
Instructional clarity in science lessons (CLARITY)	.110	.021	Very weak
Sex of students (SEX)*	-.010	.109	Very weak
Speaking the language of the test at home (LANG)*	-.073	-.013	Very weak

Note. *The values are calculated by point-biserial correlation.

According to Table 3, there was a high correlation between test scores and mathematics achievement, a weak correlation with "home resources for learning," and a very weak correlation with other covariates.

Results Equated by Methods Based on CTT in NEAT

In the first research question, the equating results obtained by different methods based on CTT in NEAT were compared. The linear methods are Tucker, LevineOS, and LevineTS; and the EQ methods are pre-smoothing and post-smoothing. The Delta, WMSE, and RMSD values are shown in Table 4.

Table 4. Delta, WMSE and RMSD Values Obtained from the Equating Methods Based on CTT in NEAT

	Equating Method	*Delta*	*WMSE*	*RMSD*
Linear	Tucker	.181	.026	.177
	LevineOS	.264	.082	.546
	LevineTS	.184	.075	.502
Equipercentile	Pre-smoothing	.083	.030	.202
	Post-smoothing	.082	.009	.063

According to Table 4, the post-smoothing (EQ) had the slightest error among the all methods, while LevineOS (linear) had the highest error. Tucker had the least error among the linear methods and post-smoothing had the least error among the EQ methods. According to Delta, the order of the methods with the least error is as follows: post-smoothing, pre-smoothing, Tucker, LevineTS, and LevineOS. According to WMSE and RMSD, the order of methods with the slightest error is as follows: post-smoothing, Tucker, pre-smoothing, LevineTS, and LevineOS. In addition, descriptive statistics of equated scores are in Appendix 2. Accordingly, the values of all equated scores were close to the old form. Equated scores are presented in Table 5.

Table 5. Equated Scores Derived from Different Methods Based on CTT

Score	*Tucker*	*LevineOS*	*LevineTS*	*Pre-smoothing*	*Post-smoothing*
0	0	0	0	.54	.15
1	.63	0	0	1.29	1.37
2	1.62	.76	.98	2.00	2.19
3	2.61	1.86	2.04	2.80	3.02
4	3.61	2.96	3.10	3.63	3.89
5	4.60	4.06	4.16	4.49	4.79
6	5.59	5.16	5.22	5.43	5.72
7	6.58	6.26	6.28	6.46	6.70
8	7.57	7.36	7.34	7.53	7.73
9	8.57	8.46	8.41	8.59	8.76
10	9.56	9.56	9.47	9.62	9.74
11	10.55	10.66	10.53	10.60	10.72
12	11.54	11.76	11.59	11.60	11.74

According to Table 5, the equated scores obtained after the pre- and post-smoothing for the raw score of 0, 1 and 2 have a higher value than the raw score, while the Tucker, LevineOS, and LevineTS scores have lower equated scores than the raw score. It has been observed that for raw score 3, the equated score based on only the final smoothing has a more excellent value than the raw score. In other raw score values, it is seen that all equated scores have a lower value than the raw score.

Results of Kernel Equating in NEC, NEAT and EG

In the second research question, X and Y test scores were equated by kernel equating methods (linear and EQ) in NEC, NEAT, and EG. There were the equated scores in Appendix 3 and descriptive statistics in Appendix 2. Accordingly, the values of all equated scores were close to the old form.

According to the PRE values, kernel linear results had more absolute values for the three or higher-order moments than EQ results. The PRE values of linear equating in NEC and EG were between -7.91 and 0; the moments in EQ were between -1.50 and .02, indicating a good fit. Similarly, according to PRE values in NEAT, the moments of the results obtained by the EQ methods were between -3.54 and .18; and the moments in linear equating methods were between -9.6 and 2.05. It indicates that the EQ results in all designs had a good fit, while the linear equating results showed worse fit. However, the first three moments in all methods were less than 1%. The small PRE values (and range) indicate that the conversion of X scores to Y scores was quite good. For example, there is up to 1.43% disagreement in the 10th moment for SCICONF.

The SEE graphics for the kernel linear NEC, NEAT and EG were in Figure 1. The SEEs in NEAT ranged from .0905 (at Score 8) to .2309 (at Score 0) for PSE.L; from .1015 (at Score 8) to .2862 (at Score 0) for CE.L. It was also seen that PSE.L had smaller equating error than CE.L in all scores. In NEC, some covariates clustered and there were actually two groups. Firstly, MATH1 and MATH5 had close errors. Secondly, the errors of the NEC and EG designs (for other covariates) had close errors. The first group scores were closer to the PSE in NEAT and the second group scores were closer to the scores in EG. The SEEs for MATH1 ranged from .0850 (at Score 8) to .2862 (at Score 0). The SEEs for MATHCONF ranged from .1158 (at Scores 8 and 9) to .2650 (at Score 0).

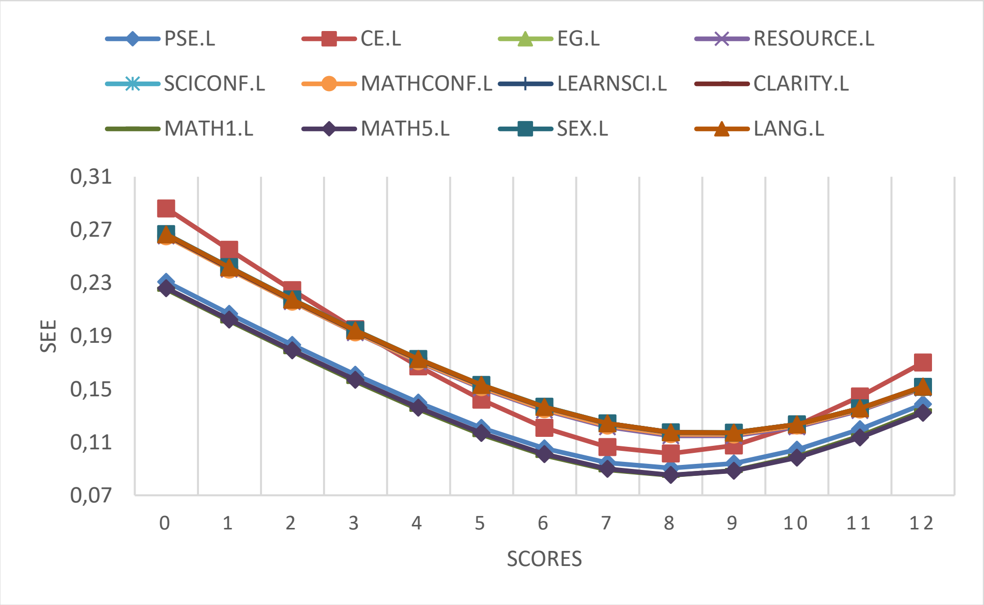


Figure 1. SEE for Kernel Linear Equating in NEC, NEAT and EG.

The SEE graphics for the kernel EQ (with optimal bandwidths) in NEC, NEAT and EG were in Figure 2. The SEEs ranged from .1107 (at Score 10) to .2540 (at Score 0) for PSE.EQ; from .1273 (at Score 10) to .2471 (at Score 0) for CE.EQ. It was also seen that CE.EQ had more equating errors than PSE.EQ in most scores (except Score 0). Similarly, some covariates in NEC clustered and there were actually two groups. Firstly, MATH1 and MATH5 had close errors. Secondly, the errors of the NEC and EG designs (for other covariates) had close errors. The SEEs for MATH1 ranged from .1014 (at Score 10) to .2252 (at Score 0). The SEEs for MATHCONF ranged from .1217 (at Score 10) to .2618 (at Score 0).

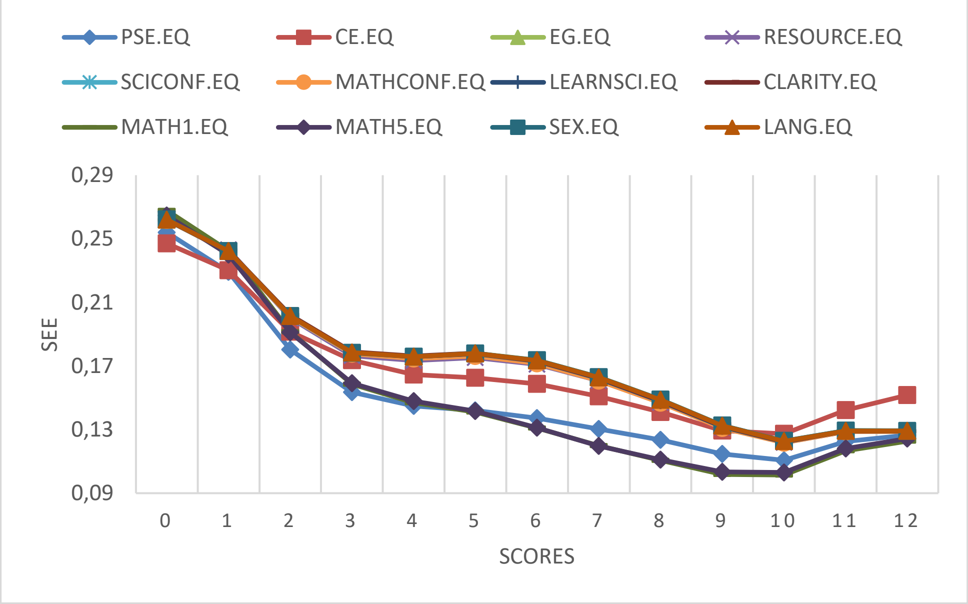


Figure 2. SEE for Kernel Equipercentile Equating in NEC, NEAT and EG.

According to Table 6 included WMSE and RMSD values, PSE was the method with the slightest error rate for linear and EQ. The errors in linear and EQ were very close to each other. There was little difference between the PSE in NEAT and EG, but EG had fewer errors. The errors in all covariates were similar in both linear and EQ in NEC. Nevertheless, MATH1 had fewer errors compared to others, while RESOURCE had more errors. The order of errors according to WMSE and RMSD values in both linear and EQ is as follows: MATH1, MATH5, LEARNSCI, CLARITY, MATHCONF, LANG, SEX, SCICONF, and RESOURCE.

Table 6. WMSE and RMSD Values Obtained from the Kernel Equating in NEC, NEAT and EG

		Linear		*Equipercentile*	
Equating Design		*WMSE*	*RMSD*	*WMSE*	*RMSD*
NEAT	PSE	.026	.179	.029	.195
	CE	.042	.281	.046	.311
EG	EG	.008	.055	.009	.062
NEC	MATH1	.002	.017	.003	.021
	MATH5	.005	.038	.006	.043
	LEARNSCI	.006	.046	.008	.053
	CLARITY	.007	.048	.008	.055
	MATHCONF	.007	.052	.009	.059
	LANG	.007	.053	.009	.060
	SEX	.008	.057	.010	.065
	SCICONF	.009	.062	.010	.071
	RESOURCE	.011	.074	.012	.083

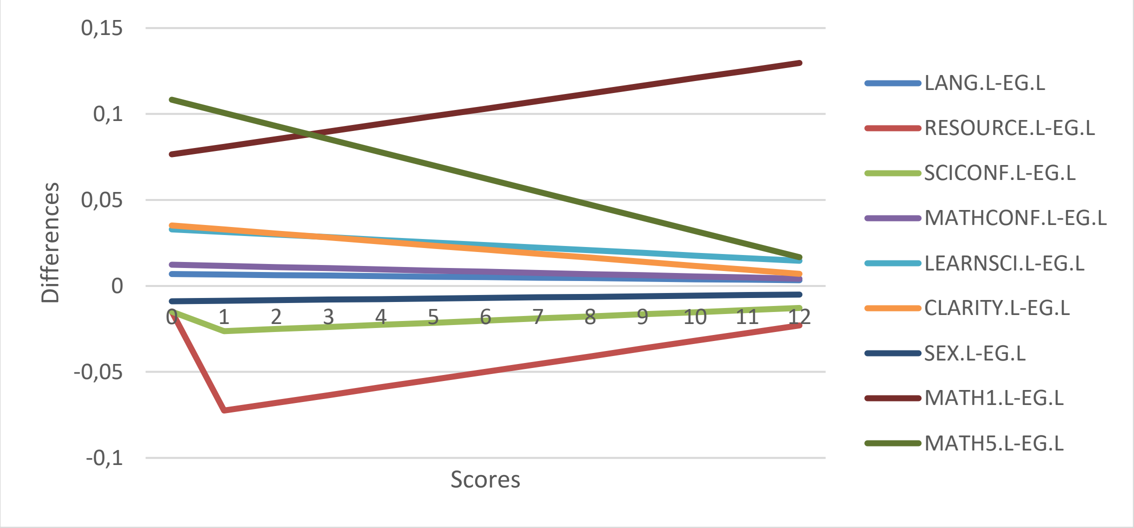
Comparison of NEC with NEAT and EG

In the third research question, the equating results in NEC using different covariates were compared with the kernel in EG and the method with the slightest error based on the kernel and CTT in NEAT. In the first and second research questions, the methods with the least equating errors in NEAT were Tucker (CTT, linear), post-smoothing (CTT, EQ), and PSE (kernel linear and EQ). The RMSE values of the equated score differences across methods are presented in Table 7. Accordingly, Tucker and PSE in linear equating had the smallest value (RMSE=.020). In EQ, post-smoothing and EG had the smallest value (RMSE=.077). It indicates that Tucker and PSE methods yielded similar equated scores. Similarly, EQ results indicate that the post-smoothing scores and EG scores are close to each other.

Table 7. RMSE Values of Differences Between NEAT and EG

	Linear		*Equipercentile*	
	EG	PSE	EG	PSE
PSE	.242	-	.213	-
Tucker	.228	.020	-	-
Post-smoothing	-	-	.077	.212

Figure 3 shows the comparison results regarding kernel linear equating in NEC with NEAT (PSE.L and Tucker) and EG. Accordingly, no point difference was higher than .5 and therefore there was no DTM. When NEC designs were compared with EG.L, the differences in MATH1 were between .05 and .5 at all score levels. Also the differences in RESOURCE and MATH5 ranged from .05 to .5 at some score levels. These results showed that the differences between MATH1 and EG.L were marginally meaningful, even within negligible limits. The same was true for RESOURCE and MATH5 for some points. However, the differences in LANG, SCICONF, MATHCONF, LEARNSCI, CLARITY and SEX are less than .05. In other words, the difference between these designs and EG was not meaningful. However, the equated score difference between NEC and PSE.L and Tucker was between .05 and .5 at all score levels. These results showed that the differences between all NEC designs, and PSE.L and Tucker methods were marginally meaningful, even within negligible limits.



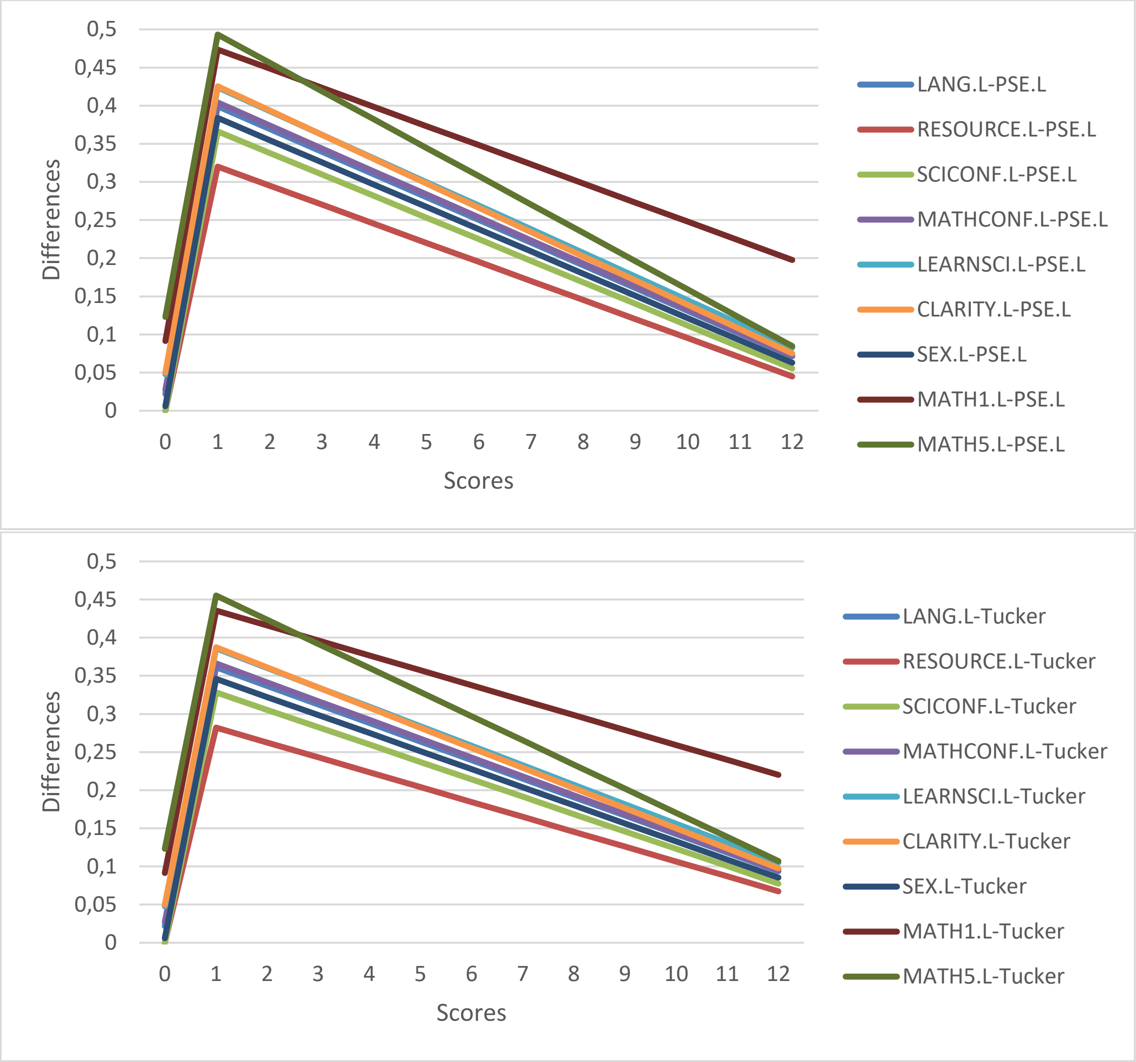


Figure 3. Comparison of Linear Equating in NEC with NEAT and EG.

Similarly, Figure 4 compares kernel EQ in NEC with NEAT (PSE.EQ and post-smoothing) and EG. Accordingly, no point difference was higher than .5 and therefore there was no DTM. When NEC designs were compared with EG.EQ and post-smoothing, the differences in MATH1, MATH5 and RESOURCE ranged from .05 to .5 at some score levels. These results showed that the differences were marginally meaningful, even within negligible limits. However, the differences in LANG, SCICONF, MATHCONF, LEARNSCI, CLARITY and SEX are less than .05. In other words, the differences between these designs and EG.EQ and post-smoothing were not meaningful. However, the equated score differences between NEC and PSE.EQ were between .05 and .5 at all score levels. The difference was found to be the highest at X=5. These results showed that the differences between all NEC designs, and PSE.EQ were marginally meaningful, even within negligible limits.

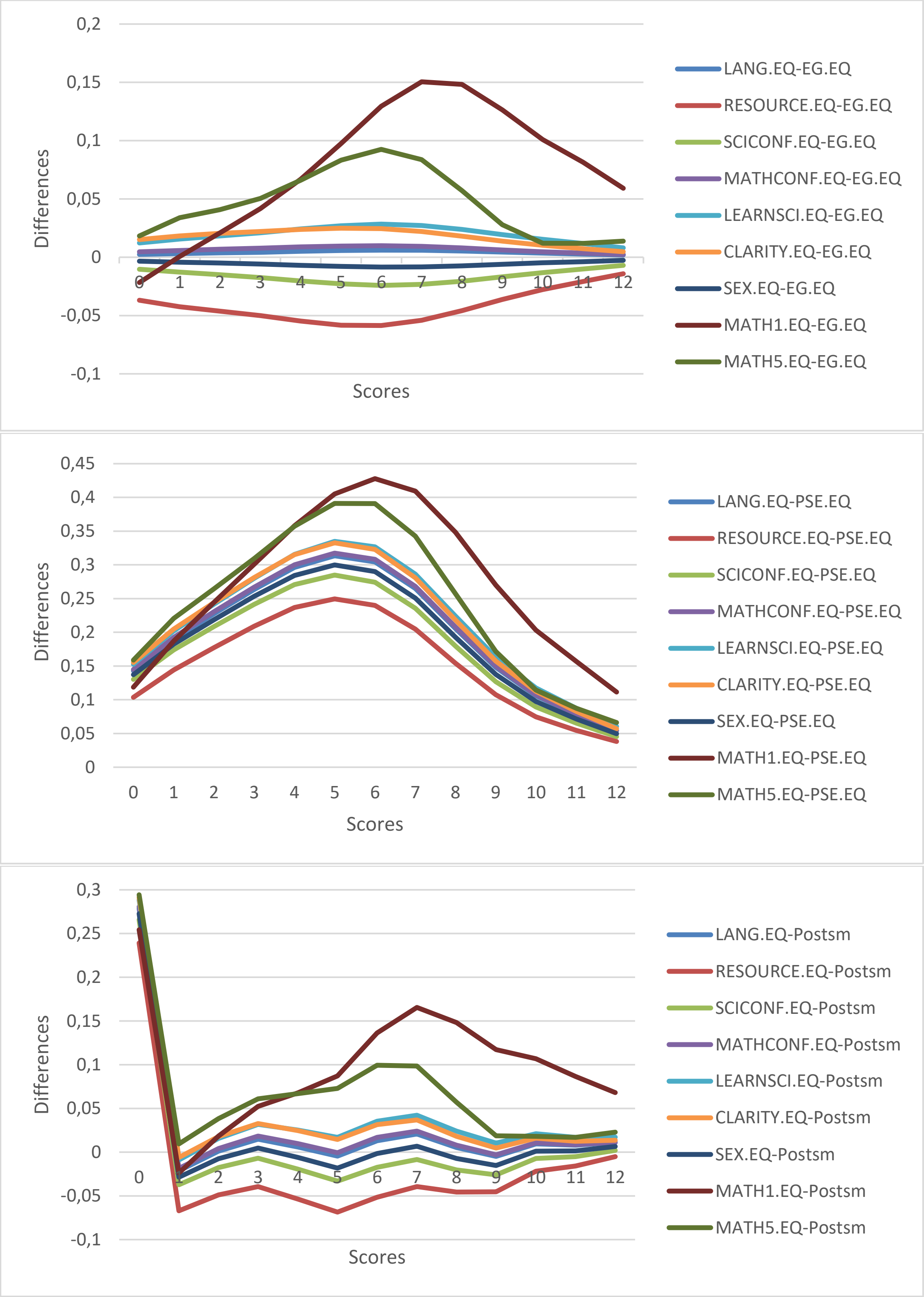


Figure 4. Comparison of Equipercentile Equating in NEC with NEAT and EG.

In addition, the RMSE values obtained by comparing NEC with NEAT and EG are shown in Table 8. Accordingly, the smallest RMSE for all covariates in both linear and EQ was obtained by comparing the NEC with the EG. It refers that the equated scores in NEC being close to those in EG. However, the highest RMSE value was obtained between the PSE (linear and EQ) and NEC, which indicates that the equated score differences between NEC and PSE are high.

Table 8. RMSE Values of Differences Between NEC and Other Designs

Covariate	*NEC-EG (L)*	*NEC-NEAT (PSE.L)*	*NEC-NEAT (Tucker)*	*NEC-EG (EQ)*	*NEC-NEAT (PSE.EQ)*	*NEC-NEAT (Postsm.)**
MATH1	.105	.334	.323	.094	.293	.120
MATH5	.069	.306	.292	.053	.265	.099
RESOURCE	.048	.194	.180	.044	.169	.080
LEARNSCI	.025	.265	.251	.020	.233	.083
CLARITY	.023	.263	.249	.019	.231	.083
MATHCONF	.009	.250	.236	.007	.220	.079
LANG	.005	.247	.233	.004	.217	.078
SEX	.007	.236	.222	.006	.207	.077
SCICONF	.020	.223	.209	.017	.196	.076

Note. *Post-smoothing

As seen in Table 8, the NEC with the covariates that yield the closest result with the EG in linear equating is LEARNSCI, CLARITY, MATHCONF, LANG, SEX, and SCICONF (RMSE<.03). The most distinctive result in EG was obtained in MATH1 (.1<RMSE<.2). The results between NEC and PSE in linear equating comparison were similar to the comparison NEC and Tucker. RESOURCE ($RMSE_{PSE}$=.19, $RMSE_{Tucker}$=.18) gave the closest result in PSE and Tucker. In contrast, the weakest result was obtained in MATH1 ($RMSE_{PSE}$=.33, $RMSE_{Tucker}$=.32). RMSE values in LEARNSCI, CLARITY, MATHCONF, LANG, SEX, and SCICONF were also close to each other (.22<$RMSE_{PSE}$<.27; .21<$RMSE_{Tucker}$<.25).

In EQ, the NEC results were close to the EG for all covariates (RMSE<.1). The NEC with the covariates that gave the closest scores to EG included LEARNSCI, CLARITY, MATHCONF, LANG, SEX, and SCICONF (RMSE<.02). The most distinctive result in EG was obtained in MATH1 (RMSE=.09). Similarly, when the post-smoothing and NEC were compared, the RMSE of all covariates were found to be close to each other (.08<RMSE<.12) and MATH1 yielded relatively different results (RMSE=.12). However, after a comparison of PSE and NEC, the lowest RMSE was found in RESOURCE. The highest RMSE was measured in MATH1. RMSE values of other covariates were also close to each other. Figure 3 and Figure 4 shows the equated score differences between NEC, and NEAT and EG.

Discussion, Conclusion, and Recommendations

In this study, TIMSS 2019 science test scores obtained from fourth-grade students were equated in NEC, NEAT, and EG with equating methods based on kernel and CTT. NEC designs used nine covariates with different correlations with test scores to address student characteristics. Accordingly, a high correlation was found between test scores and "math achievement" (MATH1 and MATH5) and a weak correlation between test scores and "home resources for learning" (RESOURCE). There was a very weak correlation between science achievement and the covariates of "student confidence in science (SCICONF) and mathematics (MATHCONF)," "like learning science" (LEARNSCI), "instructional clarity in science lessons" (CLARITY), "sex" (SEX) and "speaking the language of the test at home" (LANG). However, in the literature, there was a positive correlation between science achievement and other covariates except for language (Aydın, 2015; Aydoğan & Gelbal, 2022; Coşkun, 2021; Sarıer, 2020; Soysal, 2019; Üstün, 2007).

This study compared equating methods based on CTT in NEAT. Similar to our findings, the literature showed that the methods with the minor errors were Tucker (linear) and post-smoothing (EQ), while the method with the most error was LevineOS (Puhan, 2010). According to kernel equating results in NEAT, PSE had fewer errors than CE in both linear and EQ, which overlaps with the literature finding (Akın Arıkan, 2019, 2020). Also, in the present study, Tucker and PSE methods yielded similar equated scores in linear equating, and these methods were similarly different from kernel linear equating in EG (von Davier et al., 2004). In EQ, the equated scores obtained from the post-smoothing and EG were close to each other but slightly differed from PSE. These findings are consistent with the literature (Liu & Low, 2008; von Davier et al., 2006).

This study compared NEC results with NEAT (kernel and CTT) and EG (kernel). Both the RMSE values and the score differences indicated that NEC and EG results were closer to each other than NEAT. There were studies in the literature with similar findings (Branberg & Wiberg, 2011; Wiberg & Branberg, 2015).

This study also detailed the NEC results with different covariates and reached similar results for linear and EQ. The findings regarding NEC results in the literature vary depending on the covariate. For example, Akın Arıkan's (2020) study revealed that the results in NEC where gender and socioeconomic level were covariates differed. Similarly, the equating results in our study clustered according to the correlation level, and the equating errors were close to each other in each cluster. Basically, NEC designs were divided into two groups in which the SEEs were very close to each other. In the first group, there were MATH1 and MATH5 in which mathematics achievement was a covariate with a high correlation with science achievement. It was found that the first group scores were closer to the PSE in NEAT. The second group was covariates, which had weak and very weak correlation with science achievement (RESOURCE, SCICONF, MATHCONF, LEARNSCI, CLARITY, SEX, LANG). It was found that the second group scores were closer to the scores in EG.

In the present study, when some NEC designs were compared with EG, it was found that there was not meaningful difference for both linear and EQ. These NEC designs used covariates, which had very weak correlations with science achievement (LEARNSCI, SCICONF, MATHCONF, CLARITY, LANG, and SEX). Similar findings were also detected in the comparison of the aforementioned NEC designs and post-smoothing, which is one of the CTT methods in the NEAT. However, the equated score differences between some NEC designs (MATH1, MATH5, and RESOURCE), and EG (linear and EQ) and post-smoothing (CTT) were marginally meaningful, even within negligible limits.

In addition, the differences between all NEC designs, and PSE (kernel) and Tucker (CTT) methods in NEAT were found to be marginally meaningful, even within negligible limits. Similarly, there are findings in the literature suggesting that the results in the NEC and the PSE method in NEAT are close to each other (Wallin & Wiberg, 2016, 2019; Wiberg & Branberg, 2015), but there is no study comparing the results obtained from the NEC and methods based on CTT.

Additionally, it was determined that the most remarkable difference in both linear and EQ was between NEC, in which MATH1 was covariant, and the other designs (EG and NEAT). It did not overlap with the SEE results, which indicated that MATH1 and PSE were very close to each other yet different from EG. Since the SEE values in NEC, where MATH1 was a covariate, were less than .3, the differences between equation errors could stem from uncertain factors that might affect SEE (von Davier et al., 2006). SEE is calculated by the delta method and equating designs (von Davier et al., 2004), in which uncertain factors might lead to errors (von Davier et al., 2006). A potential inconsistency between SEE and other errors, e.g., WMSE, RMSD, and difference graphs) has also been revealed in several studies in the literature (Akın Arıkan, 2020; Wallin & Wiberg, 2019; Wiberg & Branberg, 2015).

Five math achievement scores are calculated in the TIMSS survey. This present study used the first (MATH1) and fifth (MATH5) group scores. It was determined that both were highly correlated with science achievement, but there was a partial difference in the equating results. MATH5 sometimes yielded close results to RESOURCE. For instance, the linear and EQ methods in EG and the RMSE values in the post-smoothing method were close to each other.

In conclusion, this study revealed that some NEC designs (MATH1, MAT5 and RESOURCE) while equating science tests were close to the NEAT and marginally meaningfully differentiated from the EG. Therefore, when there is no common item or the use of anchor items is not theoretically practical, the scores can be equated in NEC using math achievement and home resources for learning as covariates. However, in this study, it was seen that other NEC designs were not differ meaningfully with the EG. For this reason, when the science scores could not be equated in NEAT, some covariates should not be used in NEC, which were: student confidence in science, student confidence in mathematics, like learning science, instructional clarity in science lessons, sex and speaking the language of the test at home. In future studies, different covariates can be used in equating the science test scores. Also, it was found that the scores from the kernel EQ in EG and the post-smoothing were close to each other. In cases where an equating based on CTT cannot be performed, kernel EQ can be preferred. In addition, our study revealed some inconsistencies between SEE and other equation errors (RMSD, WMSE, difference graphs). This inconsistency may be due to differences in the calculation of equating errors. The SEE equation in kernel equating takes into account the equating design and delta method; however, equating design is not considered in calculating RMSD, WMSE and difference graphs. The difference between SEE and other equation errors can be investigated in future studies.

Note

A part of this study was presented as an oral presentation at 8th International Congress on Measurement and Evaluation in Education and Psychology (CMEEP 2022).

References

Akın Arıkan, C. (2019). A comparison of kernel equating methods based on neat design. *Eurasian Journal of Educational Research, 19*(82), 27-44. Retrieved from https://dergipark.org.tr/en/pub/ejer/issue/48089/608101

Akın-Arıkan, Ç. (2020). The impact of covariate variables on kernel equating under the non-equivalent groups. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 362-373. doi:10.21031/epod.706835

Akın Arıkan, Ç., & Gelbal, S. (2018). A comparison of traditional and kernel equating methods. *International Journal of Assessment Tools in Education, 5*(3), 417–427. doi:10.21449/ijate.409826

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1–36. doi: 10.18637/jss.v074.i08

Altintas, O., & Wallin, G. (2021). Equality of admission tests using kernel equating under the non-equivalent groups with covariates design. *International Journal of Assessment Tools in Education, 8*(4), 729–743. doi:10.21449/ijate.976660

Andersson, B., Branberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating using the R package kequate. *Journal of Statistical Software, 55*, 1-25.

Andersson, B., Branberg, K., & Wiberg, M. (2022). *Package 'kequate'.* Retrieved from https://cran.r-project.org/web/packages/kequate/kequate.pdf

Atalay Kabasakal, K., & Kelecioglu, H. (2015). Effect of differential item functioning on test equating. *Educational Sciences: Theory & Practice, 15*(5), 1229–1246. doi:10.12738/estp.2015.5.2505

Atar, B., Atalay Kabasakal, K., & Kibrislioglu Uysal, N. (2023). Comparability of TIMSS 2015 mathematics test scores across country subgroups. *The Journal of Experimental Education*, *91*(1), 82-100. doi:10.1080/00220973.2021.1913978

Aydın, M. (2015). *The effects of student-level and school-level factors on middle school students' mathematics achievement.* (Unpublished doctoral dissertation). Necmettin Erbakan University, Konya.

Aydoğan, İ., & Gelbal, S. (2022). Determination of the characteristics predicting science achievement through the classification and regression tree (cart) method: The case of TIMSS 2015 Turkey. *Education and Science, 47*(209), 239-259. doi:10.15390/EB.2022.9368

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

Coşkun, B. (2021). *The effects of student and school characteristics on TIMSS 2015 science and math achievement.* (Unpublished doctoral dissertation). Eskişehir Osmangazi University, Eskişehir.

González, J., & Wiberg, M. (2017). *Applying test equating methods using R.* Cham, Switzerland: Springer.

Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3th ed.). New York: McGraw-Hill.

House, J. D. (2006). The effects of classroom instructional strategies on science achievement of elementary-school students in Japan: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media*, *33*(2), 217-229.

International Association for the Evaluation of Educational Achievement. (2021). *TIMSS 2019 international database.* [Data set]. Retrieved from https://timss2019.org/international-database/?_gl=1*bf0qid*_ga*NDQyNzY0MjI0LjE2NDI2MjMxOTU.*_ga_L2FMXN42HR*MTY0MjYyMzE5NS4xLjAuMTY0MjYyMzE5NS4w

Kim, S., & Lu, R. (2018). *The pseudo-equivalent groups approach as an alternative to common-item equating* (Research Report No. ETS RR–18-02). ETS Research Report Series. doi:10.1002/ets2.12195

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

Leung, F. K. (2002). Behind the high achievement of East Asian students. *Educational Research and Evaluation, 8*(1), 87-108.

Liu, J., Guo, H., & Dorans, N. J. (2014). *A comparison of raw-to-scale conversion consistency between single- and multiple-linking using a nonequivalent groups anchor test design* (Research Report No. ETS RR–14-13). ETS Research Report Series. doi:10.1002/ets2.12014

Liu, J., & Low, A. C. (2008). A comparison of the kernel equating method with traditional equating methods using SAT data. *Journal of Educational Measurement*, *45*(4), 309-323.

Livingston, S. A. (2014). *Equating test scores (without IRT).* (2nd. ed.). Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/LIVINGSTON2ed.pdf

Lu. R.. & Guo. H. (2018). *A simulation study to compare nonequivalent groups with anchor testing and pseudo-equivalent group linking* (Research Report No. RR-18-08). Educational Testing Service. doi:10.1002/ets2.12196

Lyren, P. E., & Hambleton, R. K. (2011). Consequence of violated equating assumptions using the equivalent group design. *International Journal of Testing*, *11*(4), 308–323. doi:10.1080/15305058.2011.585535

Mao, X., von Davier, A. A., & Rupp, S. (2006). *Comparisons of the kernel equating method with the traditional equating methods on Praxis™ data* (Research Report No. RR-06-30). ETS Research Report Series. Retrieved from https://files.eric.ed.gov/fulltext/EJ1111483.pdf

Mullis, I. V. S. (2013). Introduction. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks*, (3-9). Boston College: TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/timss2015/frameworks.html

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* Boston College: TIMSS & PIRLS International Study Center. Retrieved from https://timss2019.org/reports/achievement/#science-4

Özkan, U. B. (2018). Comparative evaluation of TIMSS-2015 results in terms of educational resources at home. *Amasya Education Journal*, *7*(1), 98-120.

Pajares, F. (2008). Motivational role of self-efficacy beliefs in self-regulated learning. In D. H. Schunk, & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory and research and applications* (111-140). New York: Lawrence Erlbaum Associates.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement*, *47*(1), 54-75.

Revelle, W. (2022). *Package 'psych'*. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Salaway, L. J. (2008). *Efficacy of a direct instruction approach to promote early learning* (Unpublished doctoral dissertation). Duquesne University, Department of Counselling, Psychology and Special Education, Pittsburgh.

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A review of test equating methods with a special focus on irt-based approaches. *Statistica, 77*(4), 329–352. doi:10.6092/issn.1973-2201/7066

Sarıer, Y. (2020). Turkey's performance in TIMSS applications and variables predicting academic achievement. *Journal of Primary Education, 2*(2), 6-27.

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement,* *10*(3), 303–317. doi:10.1177/014662168601000308

Soysal, S. (2019). The effects of getting home learning resources and preschool education training on TIMSS 2015 mathematics and science performance. *Academy Journal of Educational Sciences*, *3*(2), 101-113.

Suh, Y., Mroch, A. A., Kane, M. T., & Ripkey, D. R. (2009). An empirical comparison of five linear equating methods for the NEAT design. *Measurement, 7,* 147–173. doi: 10.1080/15366360903418048

Üstün, E. (2007). *Okul öncesi çocuklarının okuma yazma becerilerinin gelişimi [Development of preschool children's literacy skills]*. İstanbul: Morpa.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data* (Research Report No. RR-06-02). ETS Research Report Series. doi:10.1002/j.2333-8504.2006.tb02008.x

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin., M. von Davier. & I. V. S. Mullis (Eds.). *Methods and procedures: TIMSS 2019 technical report* (16.1-16.331). Boston College: TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/timss2019/methods/chapter-16.html

Yurtcu, M., & Guzeller, C. O. (2018). Investigation of equating error in tests with differential item functioning. *International Journal of Assessment Tools in Education, 5*(1), 50–57. doi:10.21449/ijate.316420

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

Wallin, G. & Wiberg, M. (2016). *Nonequivalent groups with covariates design using propensity scores for kernel equating.* Paper presented in Quantitative Psychology The 81st Annual Meeting of the Psychometric Society, Asheville, North Carolina.

Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics, 44*(4), 390-414. doi:10.3102/1076998619838226

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

Wiberg, M., & González, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement, 53*(1), 106–125. http://www.jstor.org/stable/43940606

Wiersma, W., & Jurs, S. G. (2005). *Research methods in education: An introduction.* Boston: Pearson.

APPENDICES

Appendix 1. Descriptive Statistics of Equated Scores

The bandwidths in EQ for the optimal continuization were as follows:

- h_X=.546 and h_Y=.535 for NEAT.PSE;
- h_X=.545 and h_Y=.536 for NEAT.CE, EG and RESOURCE;
- h_X=.545 and h_Y=.537 for SCICONF and SEX;
- h_X=.546 and h_Y=.533 for MATH1;
- h_X=.542 and h_Y=.535 for MATH5;
- h_X=.545 and h_Y=.536 for MATHCONF, LEARNSCI, CLARITY and LANG.

The bandwiths in linear equating for the continuization were as follows:

- h_X=2538.895 and h_Y=2533.219 for NEAT.PSE;
- h_X=2580.024 and h_Y=2498.149 for NEAT.CE and EG;
- h_X=2575.277 and h_Y=2505.156 for RESOURCE.L;
- h_X=2578.673 and h_Y=2500.007 for SCICONF.L;
- h_X=2581.19 and h_Y=2497.522 for MATHCONF.L;
- h_X=2582.183 and h_Y=2496.311 for LEARNSCI.L;
- h_X= 2586.126 and h_Y=2497.995 for CLARITY.L;
- h_X=2577.147 and h_Y=2506.773 for MATH1.L;
- h_X=2593.167 and h_Y=2491.082 for MATH5.L;
- h_X=2580.151 and h_Y=2499.108 for SEX.L;
- h_X=2580.727 and h_Y=2498.053 for LANG.L.

Appendix 2. Descriptive Statistics of Equated Scores

	Equipercentile				*Linear*			
	Mean	*SD*	*Skewness*	*Kurtosis*	*Mean*	*SD*	*Skewness*	*Kurtosis*
X Test	7.36	2.58	-.39	-.48	7.36	2.58	-.39	-.48
Y Test	7.14	2.50	-.27	-.66	7.14	2.50	-.27	-.66
EG	7.14	2.50	-.28	-.66	7.14	2.50	-.39	-.48
PSE	6.94	2.56	-.23	-.74	6.94	2.57	-.39	-.48
CE	6.84	2.62	-.25	-.89	6.83	2.64	-.39	-.48
RESOURCE	7.10	2.51	-.27	-.67	7.10	2.51	-.39	-.48
SCICONF	7.12	2.50	-.27	-.66	7.12	2.50	-.39	-.48
MATHCONF	7.15	2.49	-.28	-.65	7.15	2.50	-.39	-.48
LEARNSCI	7.16	2.49	-.28	-.65	7.16	2.49	-.39	-.48
CLARITY	7.16	2.49	-.28	-.65	7.16	2.49	-.39	-.48
MATH1	7.25	2.51	-.32	-.63	7.25	2.51	-.39	-.48
MATH5	7.19	2.48	-.30	-.62	7.19	2.48	-.39	-.48
SEX	7.14	2.50	-.28	-.66	7.13	2.50	-.39	-.48
LANG	7.15	2.50	-.28	-.66	7.15	2.50	-.39	-.48
Tucker	-	-	-	-	6.94	2.56	-.39	-.48
LevineOS	-	-	-	-	6.66	2.84	-.39	-.48
LevineTS	-	-	-	-	6.66	2.74	-.39	-.48
Pre-smoothing	6.93	2.54	-.23	-.77	-	-	-	-
Post-smoothing	7.14	2.50	-.28	-.64	-	-	-	-

Appendix 3. Scores Equated by Kernel Equating Methods in NEC, NEAT and EG

	Score	*EG*	*Resource*	*SciConf*	*MathConf*	*LearnSci*	*Clarity*	*Math1*	*Math5*	*Sex*	*Lang*	*PSE*	*CE*
Equipercentile	0	.43	.39	.42	.43	.44	.44	.40	.44	.42	.43	.29	.64
	1	1.35	1.30	1.33	1.35	1.36	1.36	1.35	1.38	1.34	1.35	1.16	1.26
	2	2.19	2.15	2.18	2.20	2.21	2.21	2.21	2.23	2.19	2.20	1.97	1.87
	3	3.03	2.98	3.01	3.04	3.05	3.05	3.07	3.08	3.03	3.04	2.77	2.54
	4	3.89	3.83	3.87	3.90	3.91	3.91	3.95	3.95	3.88	3.89	3.60	3.30
	5	4.78	4.72	4.76	4.79	4.81	4.80	4.88	4.86	4.77	4.79	4.47	4.19
	6	5.72	5.66	5.70	5.73	5.75	5.75	5.85	5.82	5.71	5.73	5.42	5.23
	7	6.72	6.66	6.69	6.73	6.74	6.74	6.87	6.80	6.71	6.72	6.46	6.39
	8	7.73	7.69	7.71	7.74	7.76	7.75	7.88	7.79	7.73	7.74	7.53	7.56
	9	8.75	8.71	8.73	8.75	8.77	8.76	8.87	8.78	8.74	8.75	8.60	8.62
	10	9.74	9.72	9.73	9.75	9.76	9.75	9.84	9.75	9.74	9.75	9.64	9.58
	11	10.73	10.71	10.72	10.73	10.74	10.73	10.81	10.74	10.72	10.73	10.65	10.51
	12	11.75	11.74	11.74	11.75	11.76	11.75	11.81	11.76	11.75	11.75	11.70	11.56
Linear	0	.01	0	0	.03	.05	.05	.09	.12	.01	.02	0	0
	1	.98	.91	.96	.99	1.01	1.02	1.06	1.08	.97	.99	.59	.32
	2	1.95	1.88	1.93	1.96	1.98	1.98	2.04	2.04	1.94	1.96	1.59	1.34
	3	2.92	2.86	2.90	2.93	2.95	2.95	3.01	3.01	2.91	2.93	2.59	2.37
	4	3.89	3.83	3.87	3.90	3.91	3.91	3.98	3.97	3.88	3.89	3.58	3.39
	5	4.86	4.80	4.83	4.87	4.88	4.88	4.95	4.93	4.85	4.86	4.58	4.42
	6	5.82	5.77	5.80	5.83	5.85	5.85	5.93	5.89	5.82	5.83	5.58	5.44
	7	6.79	6.75	6.77	6.80	6.82	6.81	6.90	6.85	6.79	6.80	6.58	6.46
	8	7.76	7.72	7.74	7.77	7.78	7.78	7.87	7.81	7.75	7.77	7.57	7.49
	9	8.73	8.69	8.71	8.74	8.75	8.74	8.85	8.77	8.72	8.73	8.57	8.51
	10	9.70	9.67	9.68	9.70	9.72	9.71	9.82	9.73	9.69	9.70	9.57	9.53
	11	10.67	10.64	10.65	10.67	10.68	10.68	10.79	10.69	10.66	10.67	10.57	10.56
	12	11.63	11.61	11.62	11.64	11.65	11.64	11.76	11.65	11.63	11.64	11.57	11.58