ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 336-347

Test Equating with the Rasch Model to Compare Pre-test and Post-test Measurements

Zeynep UZUN * Tuncay ÖĞRETMEN **

Abstract
The purpose of this study is to prove the equitability of pre and post-tests with the Rasch Model and to provide the observability of individual and interindividual ability changes by evaluating the equated tests with stack analysis within the scope of the Rasch Measurement Theory. The pre-test and post-test data that are applied in this study were derived from the project named A Model Proposal to Increase Turkey's Success in the field of Mathematics in International Large-Scale Exams: Effectiveness of the Cognitive Diagnosis based Monitoring Model No. 115K531, which started on 15/11/2015 and was supported by the TUBİTAK SOBAG 3501 program. The tests were analyzed with the Rasch model, and the fit of the data to the Rasch model was evaluated, and then the Rasch Model and the Separate estimation-Common person method were applied for equating process. Lastly, individual and interindividual ability changes were observed by applying the stack analysis method with the Rasch model. As a result of the analysis of pre and post-tests with the Rasch model, it was concluded that they meet the requirements of the model. As a consequence of the equating process, the equitability of pre-test and post-test was proved, and it was observed that the individual and interindividual ability change could be evaluated by analyzing the pre-test and post-test data with the stack analysis method.

Key Words: Rasch model, test equating, stack analysis.

INTRODUCTION

When there is a need to compare tests, the first thing to be examined is whether the tests in question are comparable or not. For this purpose, the tests are equated with the equating methods, and, in the result of success, the comparability of the tests is proven.

Equating is defined as adjusting one test form's unit system to another test form's unit system (Angoff, 1971) and is a statistical process (Kolen & Brennan, 2004). It is applied with two methods: horizontal and vertical equating.

The horizontal equating is used at comparable difficulty levels and in need of equating the test forms in which the ability distributions of the candidates who take the exam are similar, while the vertical equalization is used at different difficulty levels and in need of equating the test forms in which the ability distributions of the candidates who take the exam are different (Hambleton & Swaminathan, 1985).

For instance, while the horizontal equating is used when the application of different forms of the test is required, the vertical equating can be used for the purposes such as; evaluating a student who performs well above her/his class with a test a few levels ahead, tracking learning development of a student with exams, evaluating multiple groups at different levels with a single scale (Hambelton & Swaminathan, 1985), working with standardized tests (Crocker & Algina, 1986), analyzing the effect of intervention as an individual by proving the comparability of scores, considering the possibility that in pre-test/post-test applications, which is also examined in this study, items may not function in the same way for all those who took the test (Anselmi, Vidotto, Bettinardi, & Bertolotti, 2015).

* Graduate Student, Ege University, Faculty of Education, Izmir-Turkey, zuzun2204@gmail.com, ORCID ID: 0000-0003-4681-0044
** Ph.D, Ege University, Faculty of Education, Izmir-Turkey, tuncay.ogretmen@ege.edu.tr, ORCID ID: 0000-0001-7783-1409

To cite this article:
Uzun, Z., & Öğretmen, T. (2021). Test equating with the Rasch model to compare pre-test and post-test measurements. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 336-347. doi: 10.21031/epod.957614

Received: 25.06.2021
Accepted: 29.11.2021

Equating can be applied based on two approaches: Item Response Theory or Classical Methods (Hambelton & Swaminathan, 1985). Classical Methods are divided into two as Equipercentile equating and Linear equating (Angof, 1971).

In the Equipercentile equating method, the scores in the two tests to be compared are accepted as equivalent if the frequency distributions are the same for a particular sample. The method ensures that the converted score distributions are the same. However, using raw scores causes problems in meeting the requirements, which were defined by Hambleton and Swaminathan (1985), such as subject independence, unidimensionality, and symmetry. Therefore, the Equipercentile equating method is considered group-dependent (Hambleton & Swaminathan, 1985).

In Linear equating, scores corresponding to the same standard score are considered to be equal (Angoff, 1971). Just like Equipercentile equating, it is group-dependent and does not meet the equating requirements (Hambleton & Swaminathan, 1985). In addition to that, in equating with classical methods in the comparison of pre-test and post-test scores, the statistical significance and magnitude of the difference between the average scores obtained from both tests can be mentioned. Examining the individual development of the students or the individual development differences among the students is out of the question.

Item Response Theory (IRT), on the other hand, is advantageous compared to classical methods since item and ability estimations can be made independently from the sample. If the item response model fits the data, the requirements in the classical method will be met due to its equality, symmetry, and invariance features (Kolen, 1981).

When the pre/post-test applications are compared by equating the test forms with the IRT, it is possible to examine not only the change in the average scores of the sample but also the individual development of the students. Two things can be achieved by measuring change at the individual level; first, characteristics that can separate students based on whether or not they have shown any development, which can be used in future applications, and secondly, the degree of change in ability seen in cases where the effect desired to be evaluated differs due to individual differences of students (Anselmi et al., 2015).

In the Item Response Theory, true score and observed score equating methods are recommended for equating (Kolen & Brennan, 2004). In the true score equating method, the tests are equated at the θ ability levels. Therefore, for equating Concurrent Estimation (Lord, 1980) and Separate Estimation methods are used.

In the observed score equating method, the score distributions of the tests are estimated with the selected IRT model, and the scores are equated with the Equipercentile equating method (Kolen & Brennan, 2004). When the Item Response Theory approach is preferred to equate the tests, it is necessary to decide which IRT model will be used for data analysis before choosing the equating method.

Purpose of the Study

The purpose of this study is to equate the pre and post-tests that are prepared and applied within the scope of the project named A Model Proposal to Increase Turkey's Success in the field of Mathematics in International Large-Scale Exams: Effectiveness of the Cognitive Diagnosis based Monitoring Model No. 115K531 (Başokçu et al., 2018), which started on 15/11/2015 and was supported by the TUBİTAK SOBAG 3501 program using the Rasch model based on the Item Response Theory and the Separate Estimation-Common Person Equating. Tests evaluate with stack analysis to ensure the observability of individual and interindividual ability changes.

Subproblems

For pre-test analysis: Do the pre-test data fit the Rasch model? Is the pre-test unidimensional? Does the pre-test have sufficient distinctiveness?

For post-test analysis: Do the post-test data fit the Rasch model? Is the post-test unidimensional? Does the post-test have sufficient distinctiveness?

For equating procedure: Can pre-test and post-test scores be compared? Can pre-test and post-test scores be converted into each other? Can the individual and interindividual change of the effect be analysed by evaluating the pre-test and post-test data on the same scale?

METHOD

In this study, all data analyses were conducted with the Dichotomous Rasch model. The Separate estimation-Common person method was applied to evaluate whether the pre-test and post-test measures were comparable. Pre-test and post-test were compared with stack analysis.

Instrument and Sample

The first identical 29 items of pre-test/post-test exam, which consists of 30 multiple choice items, prepared within the scope of project A Model Proposal to Increase Turkey's Success in the field of Mathematics in International Large-Scale Exams: Effectiveness of the Cognitive Diagnosis based Monitoring Model, constitute the measuring instrument of the research while a total of 1225 six-graders in 42 classes of 5 different schools in Izmir province constitute the sample of the exam.

Procedure

In this study, analyses are carried out in three steps. In the first step, the data obtained from the students' pre-test and post-test applications are analyzed separately with the Dichotomous Rasch model, and the fit of the data with the model and the statistical characteristics of the tests are evaluated. In the second step, the equating process is applied between the pre-test and post-test with the common person Separate estimation-Common person method, the pre-test as the reference. And in the third step, the observability of individual and inter-individual ability changes are evaluated with the Dichotomous Rasch model and with the pre-test and post-test data, which are proven to be equitable with each other by stack analysis method.

Rasch Analysis

Rasch analysis is a single parameter IRT model that estimates test items' parameters and the characteristics that are intended to be measured according to the possible answers for the items. The ability and parameter estimations are independent of the sample to which the test is applied. In Rasch analysis, knowledge is a function of the difference between person ability and item difficulty. As with the Guttman scale, it is assumed that the person will answer all items up to her/his ability level correctly. In the Rasch model, individuals and items can be positioned on the same scale, and using the Equation 1, which was used in a one-parameter logistic model, the probability of a person with θ ability level to correctly answer item i in a b_i difficulty is calculated in the model.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}},\ i = 1,2,\dots,n \qquad (1)$$

The Rasch model is applied by choosing the appropriate one among the three based on the number of item categories and weighting: Dichotomous, Andrich Rating Scale Model (RSM) or Master Partial Credit Models. In this study, analyses are carried out with the Winsteps 3.92.1 program using the Dichotomous Rasch model. The Rasch model requires unidimensionality, local independence, monotonic rising, and non-intersecting item response functions. In order to obtain test and item statistics and to evaluate the validity and reliability of the test, Rasch-based item-response threshold ordering, fit of the data to the model, item difficulty and person ability, unidimensionality and local independence, differential item functioning, scatter, and reliability analyses should be performed.

Item-category average measures

In the Rasch model, the values of the ability means corresponding to the categories are examined in order to evaluate the valid discriminability of the categories and to reveal if the item is understood correctly by the test takers. For a Dichotomous model, if the ability means of category 0 of an item is lower than the ability of category 1, it is established that the item was correctly understood by the test takers. The difference between the means indicates the power of discrimination.

Model fit tests

By evaluating the concordance statistics, it is determined to what extent the items, the individuals, and the test fit the Rasch model. In the analyses performed with the WINSTEPS program, item fit and person fit are evaluated with INFIT and OUTFIT MNSQ (mean square) sizes and unit standard deviation values. If the INFIT and OUTFIT MNSQ (mean square) sizes are between 0.5 and 1.5, it indicates that the scale is unidimensional and the sample size is sufficient. (Linacre, 2016)

In the present study, Among the 1225 students who took the pre-test and post-test, responses of 10 were excluded from the analysis due to missing data, as with the responses of a total of 4 more students, 2 in the pre-test and 2 in the post-test, were excluded from the analysis as well because their MNSQ values were higher than 4.0.

Item difficulty and person ability

In the Rasch model, item difficulty and person ability are expressed in logit. The difficulty of items refers to the corresponding level of ability. An item has a 50% probability of being answered correctly at the corresponding level of ability. A higher logit represents a more difficult item and a person with greater abilities.

Unidimensionality and local independence

The Rasch model requires meeting the unidimensionality assumption (Chang, Wang, Tang, Cheng, & Lin, 2014). Meeting the unidimensionality assumption is also an indication of local independence. Item parameters may be estimated biased in the state of unachieved unidimensionality under the unidimensionality assumption.

Differential item functioning

In order to evaluate whether the items show bias or not in the Rasch analysis, the size of the DIF contrast and the statistical significance of the Mantel Hanzel Chi-square value are examined. DIF contrast should be between -0.50 and 0.50 logit values, and Mantel Hanzel Chi-square value should be statistically insignificant ($p \geq .05$). A negative DIF contrast value indicates that the item is easy for the subject, while a positive DIF contrast value indicates that it is difficult for the subject (Linacre, 2016). Item bias can be seen as uniform item bias, where bias is seen at the same rate at all levels of ability, or as non-uniform item bias, where it occurs at specific or varying ability ranges. In this study, uniform item bias is evaluated based on the gender variable.

Separation and reliability

In the Rasch analysis, reliability is evaluated through personal reliability, person separation index, item reliability and item separation index. In the case of the measurement error getting smaller, the reliability values become insensitive and cannot exceed the upper limit of 1. At this point, the separation indices provide this congestion to be stated (Wright, 1996a).

Person separation is used in the classification of individuals, and its low value (< 2 and person reliability < .8) shows that the measuring tool is not sensitive enough to distinguish individuals at lower and upper-performance levels (Linacre, 2016).

Item separation, on the other hand, enables the evaluation of the concordance of item hierarchy to the expectations. Low item separation (< 3 = high, medium, low item difficulty, and item reliability < .9) indicates that the sample is not large enough to evaluate the rate of concordance between item hierarchy and expectations (Linacre, 2016).

Equating Treatment

When evaluating a certain effect with pre/post-test, to observe the change in individuals or the functioning of the items, pre and post-test should be equated. For this purpose, in the study, it was evaluated whether the data were suitable for stack analysis with separate estimation common person equating. Ability measurements obtained by analyzing the pre-test and post-test separately using the Dichotomous Rasch Model in Winsteps 3.92.1 program were used in the equating process. The process consists of three steps:

1. Using the ability measures obtained by the Dichotomous Rasch model, a trend line is obtained by placing the pre-test ability measures of a small portion of the sample on the x-axis and the post-test ability measures on the y-axis. If the line angle is 45 degrees to the x-axis, it is considered that the pre-test and post-test measures are convertible to each other, and the data are considered to be suitable for both stack analysis that allows the examination of change on an individual basis and rack analysis that allows examination of change on an item basis. If the trend line cannot provide the 45-degree angle,
2. Empirical intercept and slope values of the trend line are used to capture the slope. These values convert y-axis measures to x-axis measures or the reverse. If the intervention method whose effect is to be examined is to be evaluated, the post-test data is shifted by using the coordinate where the trend line cuts the x-axis and the slope value (pre-test parameters), and the trend line is obtained again. If the aim is to make a decision about the result of the desired intervention method, the pre-test should be shifted with the post-test parameters and the trend line is obtained again. If the new trend line cannot provide the 45-degree angle, it is assumed that equating is not possible between the two tests. If it provides the 45-degree angle, it is considered that the equating procedure is successful for the part taken from the sample, and to examine whether the equating will be valid for the whole sample,
3. Equating analysis is applied to entire sample data with the same coordinate and slope value. With the trend line obtained by using the whole sample, making an angle of 45 degrees with the x-axis, it is determined that the equating process is successful and the two test data are suitable for stack analysis.

In the present study, the ability measures of the first 68 students in the data set were used to evaluate the first step of the equating process. In the second step, the post-test data were shifted using the pre-test parameters.

Stack Analysis

To evaluate a specific effect applied and to evaluate the effect on an individual and group basis in the selected sample, where pre-test and post-test are applied, stack analysis is suggested (Wright, 1996b, 2003). Stack analysis is the analysis of the sample by combining the pre-test and post-test data. In this combination, the post-test data are added under the pre-test data as if different people took this exam. More specifically, stack analysis is Rasch analysis by arranging data. By adding the post-test data below the pre-test data, the data is stacked. Stack analysis is performed by applying Rasch analysis using stacked data. While the number of items does not change in the stacked analysis, the person sample doubles, and the difficulty of the items is kept constant between two-time points. To perform

stack analysis, equating pre and post-tests must be successful. In this study, stack analysis was applied with Winsteps 3.92.1 program using the Dichotomous Rasch model.

RESULTS

Rasch Analysis (First Step Results)

Item-category mean order

To examine if the pre-test and post-test item response categories were understood correctly by the students, the averages of the students who chose the 0 and 1 categories of each item were compared, and it was seen that the average ability values of the students who chose category 1 for each item were higher than the students who chose the category 0 of the item. The fact that the average ability values of the students who chose category 1 of items for the pre-test and post-test were higher than the students who chose category 0 of the item shows that the categories were correctly distinguishable, and the items were correctly understood. In other words, it was proven that the students can choose the categories that fit the purpose of the test.

Model fit tests

It was observed that the data of 1211 students included in the analysis fit the Rasch model (with an MNSQ value lower than 4.0). Students' mean infit and outfit values and standard deviations were revealed as follows: for the pre-test; mean infit 1.00 and SD 0.16, mean outfit 0.97 and SD 0.36, and for the post-test; mean infit 1.00 and SD 0.14, mean outfit 1.01 and SD 0.33. Since the infit and outfit values found were close to 1, it was stated that the sample fit the Rasch model.

When the infit and outfit values of the items were evaluated to evaluate the model fit, it was seen that the values in the pre-test and post-test were between 1.50 and 0.50, and the items were found to be fit with the model. The mean infit, outfit, and standard deviation values of the tests are revealed as follows: for the pre-test; mean infit 1.00 and SD 0.08, mean outfit 0.97 and SD 0.19, and for the post-test; mean infit 0.99 and SD 0.12, mean outfit 1.01 and SD 0.22. Since the infit and outfit values found were close to 1, it is stated that the test is compatible with the Rasch model.

Item difficulty and person ability

For the pre-test, the ability measurements ranged from 1.90 to -3.80 logit, and the average ability measurements were -1.38 (SD: 0.81) logit. The ability measurements for the post-test ranged from 5.01 to -3.67 logit, and the average ability measurement was -0.95 (SD: 1.11) logit. The item difficulty average measure for the pre-test was found to be 0.0, and the item difficulty average measure for the post-test was also found to be 0.0. The difficulty levels of the items as a result of the pre-test and post-test separate analyses and the difficulty values of the items found as a result of the equating process are given in Table 1 in The Stack Analysis section. In Figures 1 and 2, item difficulty and person ability distributions for pre-test and post-test are presented.

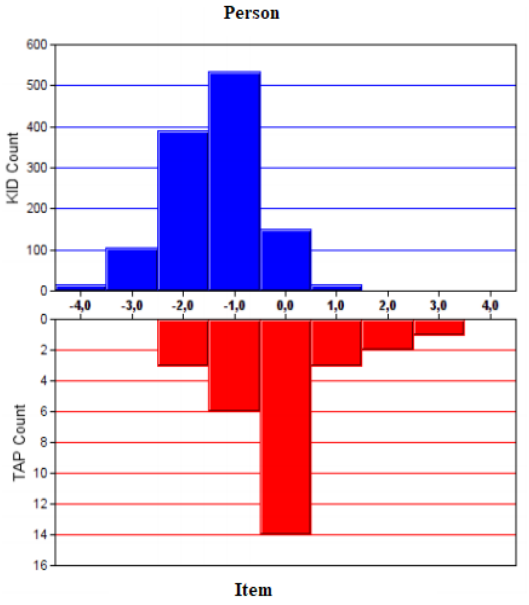

Figure 1. Pre-test İtem Difficulty and Person Ability Distribution

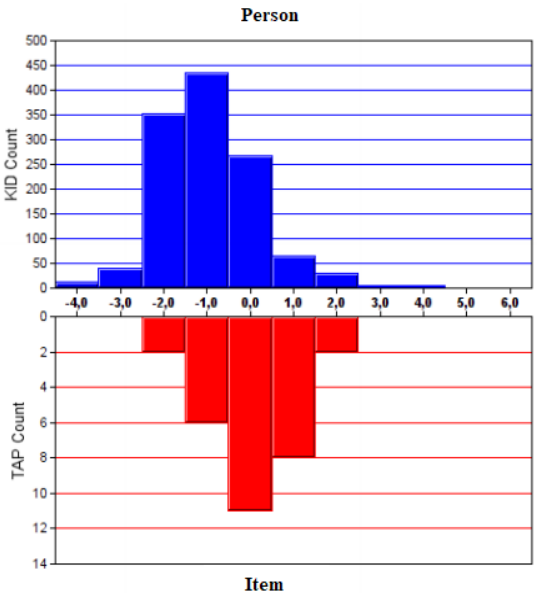

Figure 2. Post-test İtem Difficulty and Person Ability Distiribution

In Figure 1, the item range is between 3.03 and -1.87, while the skill range is between 1.90 and -3.80. In the pre-test, there is no item suitable for the ability level of the students evaluated on the logit 2 and 3 ability measure. As a result, it was determined that the pre-test could not differentiate the sample sufficiently.

Unidimensionality and local independence

For pre-test in the evaluation of multidimensionality, it was revealed that the percentage of observed explained variance (21.9%) and the percentage of unexplained observed variance (78.1%) were equal to the expected percentages; whereas the unexplained variance in the 1st contrast values was calculated as 1.78, revealing that the corresponding unexplained observed variance percentage was lower than the expected unexplained variance percentage. It can be said that there is no such doubt for the pre-test since it was considered as a multidimensionality possibility when the unexplained variance in the 1st contrast value is higher than 2.

As the results of the main component analysis, the Winsteps program divides the items into 3 clusters based on their 1st contrast loads and checks whether they measure the same thing by comparing them. The disattenuated correlation value difference between the item clusters being lower than 0.7 and the Pearson correlation value being lower than 0.3 indicates a second dimension. In the case of a second dimension, person measurements become biased.

It was determined that, in this study, for pre-test, the disattenuated correlation between item clusters 1 and 3 was lower than 0.7, and the Pearson correlation values between 1-3rd and 2-3rd clusters were lower than 0.3. The eigenvalue of significant contrasts between the items is greater than 2 (Linacre, 2016). It was determined that the multidimensionality effect did not create a significant difference since the unexplained variance contrast value for the pre-test was lower than 2.

For the post-test, the following was noted: observed variance percentage (26.3%) was higher than the expected value, the unexplained observed variance percentage (73.7%) was very close to the expected value (74%), and the unexplained variance in the 1st contrast value was calculated as 1.81, whose corresponding unexplained observed variance percentage was lower than expected unexplained variance percentage. It can be said that there is no such doubt for the post-test since it was considered as a multidimensionality possibility when the unexplained variance in the 1st contrast value is higher than 2. And since it was determined that the disattenuated correlations between item clusters were

greater than 0.7 and the Pearson correlation values were greater than 0.3, no doubt would suggest multidimensionality.

Differential item functioning

As a result of the DIF analysis performed to understand whether the items show gender bias, only the DIF contrast of the 24th item for the pre-test was found to be higher than 0.50, and the Mantel Hanzel Chi-square values ($p \geq .05$) were not statistically significant revealing that there was no gender bias. Since there was no DIF contrast higher than 0.50 for the post-test, likewise, it also means that the post-test did not show gender bias as well.

Separation and reliability

Based on the results, the Cronbach Alpha reliability coefficient of the pre-test analysis was .64, the reliability (Model) value of the individuals was .62, and the separation coefficient was 1.27. If the reliability value is below 0.80, it indicates that people are clustered in groups 1 or 2 (Linacre, 2016). When the individual reliability value is evaluated with the separation coefficient, the individual reliability and discriminability of the pre-test were found insufficient. Cronbach Alpha value was also found to be at an insufficient level. The item reliability coefficient and discrimination index of the pre-test were determined as .99 and 12.34, respectively, and the reliability and discrimination of the items were stated as quite good.

Based on the results, the Cronbach Alpha reliability coefficient of the post-test analysis was 0.83, the reliability (Model) value of the individuals was 0.80, and the separation coefficient was 2.02. The person reliability value between 0.80 and 0.90 indicates that the sample can be divided into 2 or 3 groups (Linacre, 2016). When the person reliability value was evaluated with the separation coefficient, the person reliability of the post-test was found sufficient. Moreover, Cronbach Alpha value was found to be at an insufficient level. The item reliability coefficient and discrimination index of the post-test were determined as .99 and 12.06, respectively, and the item reliability was stated as quite good. It was determined that the pre and post-tests meet the requirements and can be equated with the Rasch model approach.

Separate Estimation Common Person Analysis (Second Step Results)

The equating procedure with separate estimation common person analysis was applied to compare the pre-test and post-test data on the same metric. Considering that the intervention method that is being examined within the scope of the research will be developed, the equating method was applied by using pre-test parameters. Based on the Dichotomous Rasch model, the measurements of the first 68 people, starting from the highest ability level, were drawn with the pre-test measurement values on the x-axis and the post-test person measurement values on the y-axis. Furthermore, it was observed that the measurements were not parallel. In this case, the rack analysis was found inappropriate to apply. A correction was performed in the post-test using the coordinate -1.53, which is the point where the line obtained intersects the x-axis, and 0.73, which is the slope value of the line, to ensure the equating of the measurements. It was seen that the new line slope obtained was .997, and the equating process was found to be successful for 68 people. To examine whether the equating obtained as a result of the correction process will be valid for the whole sample, analysis was once more applied, this time considering the whole sample. The line slope was calculated as .996 and the equating between the two tests was still valid. The correlation value between tests was 0.52, and the common variance was 27%. The correlation value, free of measurement error, was calculated as 0.74. With this determination, it was proved that pre-test and post-test can be evaluated on the same metric with stack analysis and test scores can be converted to each other.

Equation 2 and Equation 3 can be used for conversion:

$$\text{Pre-test Score (x-coordinate)} * \text{slope} + \text{y-coordinate} = \text{Estimated Post-test Score} \quad (2)$$

Post-test Score (y-coordinate) / slope + x-coordinate = Estimated Pre-test Score (3)

In the scope of this study pre and post-test scores can be converted into one another using the Equation 4 and Equation 5:

Pre-test Score (x-coordinate) * 1.37 + 2.1 = Estimated Post-test Score (4)

Post-test Score (y coordinate) / 1.37 - 1.53 = Estimated Pre-test Score (5)

Stack Analysis (Third Step Results)

Stack analysis is the analysis of the sample by combining the data of the pre-test and post-test. In this combination, the post-test data are added under the pre-test data as if different people took this exam. While the number of items does not change, the person sample doubles and the difficulty of the items is kept constant between two-time points. For the stack analysis applied, it is evaluated whether the item categories are understood according to the purpose of the test. When the average of the response categories of the items was examined, it was determined that the average ability value of the students who preferred the category 1 of the item, for all items, was higher than the students who chose the category 0. This situation means students can choose the categories according to the purpose of the test. In Table 1, stack Analysis, pre-test, post-test item difficulty measurements, and mathematical general and domain-specific ability areas are given.

Table 1. Pre-test, Post-test, and Stack Analysis Item Measurements

Item	Content Domain	Mathematical Abilities				Pre-test	Stack Analysis	Post-test	Level Change of Difficulty
		Mathematization	Using symbolic and technical language	Reasoning and developing a strategy	Communication and association				
1	Number	X				-1.87	-1.69	-1.55	Harder
2	Number		X			-1.03	-1	-1.02	Same
3	Number	X		X		-1.06	-0.96	-0.9	Harder
4	Number		X		X	-0.66	-0.45	-0.29	Harder
5	Number	X	X			0.09	0.15	0.16	Harder
6	Number	X	X			0.4	0.53	0.62	Harder
7	Number	X		X		-1.23	-1.03	-0.86	Harder
8	Number		X			-1.79	-1.69	-1.64	Harder
9	Number	X	X			-0.45	-0.4	-0.39	Harder
10	Number	X	X		X	0.99	1.08	1.12	Harder
11	Number	X				1.78	0.35	-0.39	Easier
12	Number			X	X	0.14	0.08	-0.01	Easier
13	Number	X				0.39	0.24	0.09	Easier
14	Number	X	X			0.07	-0.25	-0.57	Easier
15	Geometry	X		X	X	0.1	0.39	0.63	Harder
16	Geometry	X		X	X	0.9	0.75	0.6	Easier
17	Number	X			X	-0.49	-0.11	0.24	Harder
18	Number	X				-0.52	-0.37	-0.27	Harder
19	Number	X		X		0.46	0.56	0.61	Harder
20	Number	X		X	X	0.98	0.99	0.97	Same
21	Number			X	X	-0.18	0.05	0.23	Harder
22	Number			X	X	0.15	-0.13	-0.41	Easier
23	Number			X		-1.61	-1.39	-1.2	Harder
24	Number	X	X	X	X	3.03	2.65	2.45	Easier
25	Number		X		X	0.46	0.6	0.69	Harder
26	Number	X	X			-1.04	-0.96	-0.94	Harder
27	Number	X		X		1.87	1.67	1.52	Easier
28	Number	X		X		-0.32	-0.21	-0.14	Harder
29	Number	X		X		0.41	0.56	0.65	Harder

In Table 1, item difficulties are expressed as logit, and changes in item difficulty in the post-test compared to the pre-test are indicated in the difficulty level change column in order to make it easier

to notice. When the items were examined, it was found that the item difficulty values of the items 1, 3, 4, 5, 6, 7, 8, 9, 10, 15, 17, 18, 19, 21, 23, 25, 26, 28, and 29 increased in the post-test compared to the pretest. The pretest difficulty for the item 2 was -1.03 logit, and the post-test difficulty was -1.02 logit, and for the item 20 the pre-test difficulty was 0.98 logit, and in the post-test 0.97 logit and it was determined that the item difficulty values were very close. On the other hand, it is seen that the difficulty values of the other items decreased in the post-test compared to the pretest; in other words, they were easier for the students. In order to inform about the content of the items, the general and domain-specific mathematical abilities of the items, which are prerequisites in mathematical literacy (Başokçu et al., 2018), are also included in Table 1.

Item difficulty values can't change between pre-test and post-test in stack analysis. Items get a fixed value for two tests. The cross graph of the pre-test and post-test ability measurements of the stack analysis, which shows the individual ability change when the item difficulties are kept constant at two-time points, is given in Figure 3.

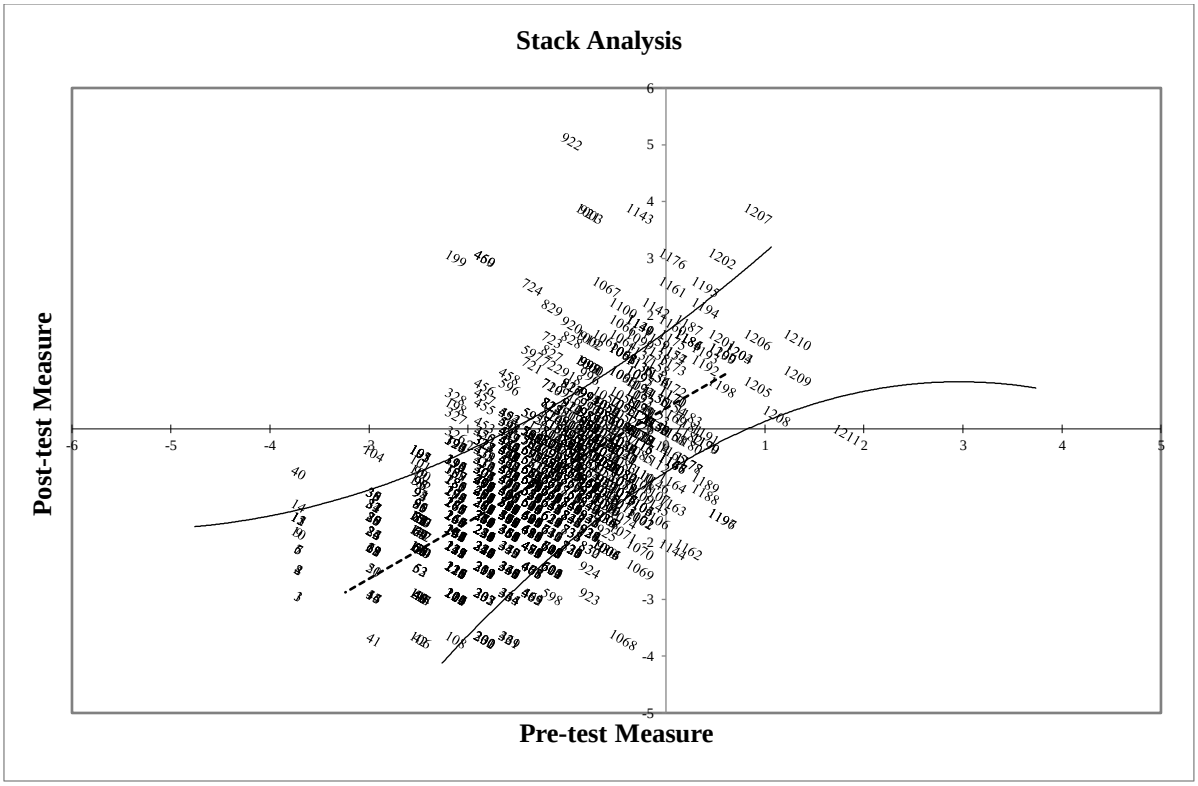

Figure 3. Cross Chart of Pre-test and Post-test Ability Measurements of Stack Analysis

In Figure 3, the pre-test measures of the students are shown in the x-axis and the post-test measures in the y-axis are shown as logit. There is a statistically significant difference ($t_{(1210)} = 12.79$, $p < .05$) in favor of the post-test between the pre-test and post-test measures results of the students. The correlation between the pre- and post-test measures of the stack analysis was 0.52. The correlation was found to be moderate. This shows that the effect applied between the pre-test and the post-test leads to different levels of change among students. Students above the identity line performed better in the post-test than the pretest. For a student who falls below the identity line, the situation is the opposite. In the graph, it is seen that the success of students numbered 922, 199, 459, and 460 increased more than other students. The ability measurement of student number 922 was calculated as -0.95 logit in the pre-test and 5.06 logit in the post-test, and an ability increase of 6.01 logit was observed. This value is 5.12 for the student numbered 199, 4.83 for the student numbered 459 and 460. Students numbered 1068, 1162, and 1069, which are below the identity line and at the farthest point, were negatively affected by the

effect between the pre-test and the post-test, and a decrease in the ability of 3.23, 2.35 and 2.22 logits was observed in the students, respectively.

DISCUSSION and CONCLUSION

When the results of item bias, multidimensionality, and discrimination analysis of the post-test are compared with the pre-test results, it is seen that the pre-test results were much weaker than the post-test results. The reason for this situation may be that students encounter types of questions that they did not encounter before in the pre-test application, while this effect disappeared in the post-test since they grasped the structure of the item better through the follow-up tests. One of the project findings from which the data was collected supports this view. This finding is that the problem situations that students encounter in the skill area they want to gain affect their success. Exposing students to problem situations similar to those in the tests aimed at increasing the level of success will increase success (Başokçu et al., 2018).

In the equating procedure, it has been proven that the pre-test and post-test ability measures are convertible to each other, as the slope of the trend line obtained with all sample ability measures in the pre-test and post-test to the x-axis is .996. Within the scope of this study, there was no need to transform ability measures with conversion formulas. The convertibility has only been demonstrated, as the ability for stack analysis needs to be convertible between pre- and post-testing of the ability measures. It has been observed that the ability measures obtained from the pre-test and post-test are comparable and can be evaluated on the same scale as the stack analysis.

In accordance with the previous study (Başokçu et al., 2018), it was observed that there was a significant difference in favor of the post-test between the pre-test and post-test ability measurements obtained within the scope of the stack analysis. As a result of the comparison of the pre-test and post-test ability measures obtained with the stack analysis with the help of graphics, students who were differently affected by the effect applied between the pre-test and the post-test could be determined. The change in students' ability levels can be compared. Thus, it has been seen that the level of individual and inter-personal ability change can be evaluated.

In the pre- and post-test evaluation, there are studies in which common item equating is used (e.g. Fujita & Mayekawa, 2011) or only the stack analysis method is used without using the equating procedure (e.g. Cunningham & Bradley, 2010; Herrmann-Abell, Flanagan, & Roseman 2012; Ling, Pang & Ompok, 2018). Common person equating was preferred to prove the equivalence of equivalent forms structure (e.g. Cavanagh, 2012; Popp & Jackson, 2009; Taylor & McPherson, 2007). It was stated by Masters (1985) that the same results can be achieved with both equatings in the Rasch Model, and the common person equating tests unidimensionality more clearly. For this reason, common person equating and stack analysis are used in this study to show that the intervention effect can be evaluated on an individual basis. Compared to previous studies, a stricter 1st contrast value was taken as the criterion in this study compared to Ling et al. (2018) study. Compared to the studies of Cunningham and Bradley (2010) and Anselmi et al. (2015), it is presented with a better percentage of person who fit the model. The results are generally in accordance with previous studies, and no feature has been identified that can make the procedure specific or hinder the implementation of equating or stack analysis.

In this study, the factors that affected the students who benefited more from the effect applied between the pre-test and the post-test or could not benefit from it were not researched. The factors that increase or decrease the success of the student can be determined by interviewing the students who are affected differently individually or by applying tests on possible factors to these students. Thus, the method of intervention can be developed, individualized, or differentiated for the groups to be determined. In the field of education, each student's unique and individual talent is an investment for the future of society. From this point of view, it is considered that the equating steps with the Rasch model are a suitable choice for studies that evaluate the intervention methods (the effect of the use of materials, the effect of the teaching model, etc.) whose effects are desired to be investigated with the pre-post test application.

REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Anselmi, P., Vidotto, G., Bettinardi, O., & Bertolotti, G. (2015). Measurement of change in health status with Rasch models. *Health and Quality of Life Outcomes, 13*(1), 1-7. doi: 10.1186/s12955-014-0197-x

Başokçu, O. T., Bardakçı, V., Çakıroğlu, E., Öğretmen, T., Yurdakul, B., & Akyüz, G. (2018). *Uluslararası geniş ölçekli sınavlarda Türkiye'nin matematik başarısını arttırabilmek için bir model önerisi: Bilişsel tanıya dayalı izleme modelinin etkililiği* (Proje No. SOBAG 3501). Retrieved from https://open.metu.edu.tr/bitstream/handle/11511/50310/TVRnMU56SXk.pdf

Cavanagh, R. F. (2012, December). *Engagement in classroom learning: Ascertaining the proportion of students who have a balance between what they can do and what they are expected to do.* Paper presented at the 2012 Annual International Conference of the Australian Association for Research in Education, Sydney, Australia.

Chang, K. C., Wang, J. D., Tang, H. P., Cheng, C. M., & Lin, C. Y. (2014). Psychometric evaluation, using Rasch analysis, of the WHOQOL-BREF in heroin-dependent people undergoing methadone maintenance treatment: Further item validation. *Health and Quality of Life Outcomes*, *12*(1), 1-9. Retrieved from https://link.springer.com/article/10.1186/s12955-014-0148-6

Crocker, L., & Algina J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cunningham, J. D., & Bradley, K. D. (2010, May). *Applying the Rasch model to measure change in student performance over time*. In American Educational Research Association Annual Meeting, Denver, CO.

Fujita, T., & Mayekawa, S. I. (2011). A comparison between common item equating with pre-and post-reading and listening tests. In C Ho, M.-F. G. Lin (Eds.), *E-Learn: World conference on e-learning in corporate, government, healthcare, and higher education* (pp. 626-631). Association for the Advancement of Computing in Education (AACE).

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Academic Publishers Group.

Herrmann-Abell, C. F., Flanagan, J. C., & Roseman, J. E. (2012, March). Results *from a pilot study of a curriculum unit designed to help middle school students understand chemical reactions in living systems* (Online Submission). Paper presented at the NARST Annual International Conference, Indianapolis.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*(1), 1-11. Retrieved from http://www.jstor.org/stable/1434813

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York: Springer.

Linacre, J. M. (2016). *Winsteps® (Version 3.92. 0)* [Computer Software]. Winsteps, Beaverton, OR, USA. Retrieved from www.winsteps.com

Ling, M. T., Pang, V., & Ompok, C. C. (2018). Measuring change in early mathematics ability of children who learn using games: Stacked analysis in Rasch measurement. In Q. Zhang (Ed.), *Pacific rim objective measurement symposium (proms) 2016 conference proceedings* (pp. 215-226). Singapore: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Masters, G. N. (1985). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82. doi: 10.1177/014662168500900107

Popp, S. E. O., & Jackson, J. C. (2009, April*). Can assessment of student conceptions of force be enhanced through linguistic simplification? A rasch model common person equating of the FCI and the SFCI*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Taylor, W. J., & McPherson, K. M. (2007), Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. *Arthritis Care & Research, 57*(5), 723-729. doi: 10.1002/art.22770

Wright, B. D. (1996a). Reliability and separation. *Rasch Measurement Transactions*, *9*(4). Retrieved from https://www.rasch.org/rmt/rmt94n.htm

Wright, B. D. (1996b). Time 1 to time 2 (pre-test post-test) comparisons and equating: Racking and stacking. *Rasch Measurement Transactions, 10*(1). Retrieved from https://www.rasch.org/rmt/rmt101f.htm

Wright, B. D. (2003). Rack and Stack: Time 1 vs. time 2 pre-post. *Rasch Measurement Transactions, 17*(1), 905-906. Retrieved from https://www.rasch.org/rmt/rmt171a.htm