ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);149-164

The Effect of Item Weighting on Reliability and Validity*

Abdullah Faruk KILIÇ** Nuri DOĞAN ***

Abstract
The purpose of this study is to examine the effect of the item weighting method developed by researchers on the construct validity of the test. For this purpose, a Monte Carlo simulation study was carried out. Test length, average factor loadings, and sample size were considered as simulation conditions. Item weighting method was defined as follows: If average score of the individuals (calculated as individual's test score/the number of items) plus item difficulty index is 1 and over then item reliability index added to individual's item score (1 or 0); if not, then the item score of the individual (1 if 1, 0 if 0) is preserved. As a result of the research, it was observed that the weighting method contributes to the construct validity. According to the results of confirmatory factor analysis, the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) values were improved. According to the research findings, the weighting method used in this research can be recommended.

Key Words: item weighting, validity, reliability, EFA, CFA

INTRODUCTION

The validity of the scores obtained from tests used in the psychological field is among the most important subjects of the psychological measurement field. Validity is considered as a feature of the scores obtained from the applied tests (American Educational Research Association [AERA], American Educational Research Association [APA], & National Council on Measurement In Education [NCME], 2014) and can be collected under the construct validity as an umbrella term (Messick, 1995). In the process of collecting evidence for construct validity, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are frequently used (Nunnally & Bernstein, 1994). EFA is based on covariance structures and is a technique for obtaining fewer latent variables (factors) than the covariance matrix between observed variables (Daniel, 1989). In EFA, the aim is to reveal the factor structure of the clusters formed by the variables. In CFA, the theoretical construct is tested, and the structural properties of the variables measured before the analysis are known. For this reason, the purpose of CFA is to try to verify the predicted factor structure based on the measurements obtained from the measurement instrument (Stevens, 2009). In the process of collecting evidence for construct validity, both analyses have high importance. Nunnally (1978) emphasizes the importance of factor analysis by saying that it is at the heart of the measurement of psychological constructs.

Different measures may be taken to increase the validity of the scores obtained from the test. Following the scale development procedure during the development phase of the measurement instrument, knowing the theoretical subset well and reflecting it on the measurement instrument, and taking some measures in the implementation phase of the tests are examples. However, it has been thought that some weighting operations on the scores obtained after the test application can increase the validity of the scores obtained from the test, and studies on item weighting have been conducted accordingly (Erkuş, 2014; Ghiselli, 1964; Gulliksen, 1950; Rotou, Headrick, & Elmore, 2002).

* This study was conducted as an oral presentation during the Second International Social Sciences Symposium held in Antalya on May 18–20, 2017.

** Research Assistant, Hacettepe University, Graduate School of Educational Science, Ankara-Turkey, e-mail: abdullahfarukkilic@gmail.com, ORCID ID: 0000-0003-3129-1763
*** Professor Dr., Hacettepe University, Faculty of Educational Science, Ankara-Turkey, e-mail: nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Kılıç, A. F., & Doğan, N. (2019). The effect of item weighting on reliability and validity. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 149-164. DOI: 10.21031/epod.516057

Geliş Tarihi: 22.01.2019
Kabul Tarihi: 26.05.2019

When studies on item weighting are examined, it is seen that they were conducted mostly in the second half of the 20th century (Burt, 1950; Dick, 1965; Ghiselli, 1964; Guilford, 1954). Among the recommended methods related to item weighting in addition to the use of methods such as assigning values by multiple regression, assigning piecewise regression coefficients (Guilford, 1954), weighting with item discrimination indices (Birnbaum, 1968), and using factor analysis (Burt, 1950), some authors suggest methods like item weighting using test variance or item variance (Dick, 1965), weighting the items related to more important topics, taking into account the context in which the items are linked (Ghiselli, 1964), and weighting item clusters instead of individual items (Gulliksen, 1950). In a more recent study, Rotou, Headrick, and Elmore (2002) proposed weighting items by using multidimensional item response theory parameters and a hybrid weighting method based on the use of total score calculation in the classical test theory to calculate individual scores. When the results of item weighting studies are examined in general, it is observed that different weighting of the items for shorter tests is more efficient, item weighting has little effect when the number of items is between 10 and 20 (Ghiselli, 1964), the best weighting method for long tests is to make weights 1 of all items. When the average of item correlations is low, item weighting gives better results (Guilford, 1954), and when the number of components in the test decreases, scoring items differently is more effective in ranking individuals compared to the total points obtained by equally weighting (Ghiselli, 1964) (e.g., scoring for the correct answer, giving 0 for the wrong answer, and collecting the items that are correctly answered).

Research on weighting in Turkey has been carried out, but the emphasis was usually on the option weighting for multiple choice items (Akkuş & Baykul, 2001; Erdem, Ertuna, & Doğan, 2016; Gözen-Çıtak, 2010; Özdemir, 2004), and it has been seen that the research conducted on item weighting is limited. In the research carried out by Yurdugül (2010), the evaluation was based on the total scores of the individuals. The limitation in this research was that it focused on the total scores and rankings of individuals. In the current study, research was carried out on the construct validity. In addition, in contrast to other studies, a new weighting method was developed in current research.

When the methods have generally been evaluated, it has been asserted that the effort for item weighting will not be worth it (Guilford, 1954; Phillips, 1943). However, besides increasing the validity and reliability of the results obtained from the item weighting test, as it should maximize the difference between individuals (Horst, 1936), item weighting will help to better discriminate the individuals. Since today's highly developed computer technology also reduces the labor required for item weighting, even if it does not make an excessive contribution to validity and reliability, item weighting may be recommended due to piecewise contributions.

Although item weighting improves the validity and reliability of the results obtained from the test, and for this reason it is clearly important, it is surprisingly used in a small number of studies (Burt, 1950). Nowadays, due to the limited research conducted on the effects of item weighting, the item weighting method developed in this research was examined under different conditions. In the study, the following sub-problems were asked in order to search for answers to the question, "What is the effect of item weighting on the validity and reliability of the test?"

1. How does the explained variance ratio change that is described as a result of EFA, which is based on the matrix of converted item scores obtained by the proposed item weighting method?

2. How do the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and chi-square values change, which are described as a result of CFA, which is based on the matrix of converted item scores obtained by the proposed item weighting method?

3. How do the Cronbach's alpha reliability coefficient values change, which are described as a result of reliability analysis, which is based on the matrix of converted item scores obtained by the proposed item weighting method?

METHOD

This research is a Monte Carlo simulation study conducted to examine the effect of the weighting method on the validity and reliability of the test, which is proposed by researchers. Research data are limited by 1–0 scoring because of the multiplicity of categorical data types and with the idea that they should be studied separately. Another limitation is the generation of data in unidimensional construct in the case of multidimensional data, as it is difficult to deal with many conditions such as data type, number of dimensions, inter-dimensional relations, and number of items in dimensions.

As simulation conditions sample size (250, 1000, and 3000), the number of items (20, 30, and 40), and an average factor loading (0.5 and 0.7) are examined.

Regarding to sample size, small (250), medium (1000), and large (3000) samples were formed. Rhemtulla, Brosseau-Liard, & Savalei (2012) stated that 200 sample sizes are common in psychology literature. But in this study, exploratory and confirmatory factor analysis was conducted. 250, 1000 and 3000 sample sizes were chosen to avoid of the sample size requirement of the factor analysis (Comrey, 1988; Floyd & Widaman, 1995; Gorsuch, 1974; Guadagnoli & Velicer, 1988).

Because of unidimensional constructs were examined in this research, tests consist of 20, 30 and 40 items were formed as simulation condition. Tests consist of 20 items was used commonly in practice (MEB, 2013, 2019). So, the condition of 20 items was added simulation. 30 and 40 items were added to examine the effect of the number of items on the weighting method.

Factor loadings were between 0.30 and 0.50 could be considered low and above 0.70 could be considered high (Trierweiler, 2009). Thus, in this research, 0.50 (low) and 0.70 (high) was used for average factor loadings condition.

When all conditions were considered, a total of 18 simulated conditions were researched, and 1000 replications were made for each condition. The simulation conditions are presented in Table 1.

Table 1. Simulation Factors and Conditions

Fixed Conditions		Simulation Conditions		
Data Type	Number of Factors	Sample Size	Test Length	Average Factor Loading
		250	20	0.50
1-0	Unidimensional	1000	30	0.70
		3000	40	

When the conditions in Table 1 are considered together, 1 (data type) x 1 (number of factors) x 3 (sample size) x 3 (test length) x 2 (average factor loading) = 18 conditions are obtained. Since 1000 replications were made for each condition, the research was carried out on 18000 data files. EFA, CFA, and Cronbach's alpha reliability coefficient calculations were performed separately on 1000 replicated data files produced for each condition and the averages of the obtained values were calculated, and these values were compared with each other.

The averages of the descriptive statistics of replicated data generated according to the simulation conditions are presented in Table 2.

When Table 2 is examined, average values of data sets obtained from 1000 replications are seen. When the data sets are examined according to the simulation conditions, the mean skewness for the items are 0 and the mean kurtosis values are around 1. It can be stated that the average discrimination values change according to the average factor loading condition. The skewness values of the total score are around 0, and the kurtosis values are about 2 in the case where the average factor loading is 0.5 and 1.8 in the case of 0.7. When the psych package (Revelle, 2016) is used in two categorical data production, a cut-off score is entered for the skewness value. According to this cut-off point, the skewness of the data is negative, positive, or around zero. However, according to the cut-off point entered, the kurtosis is automatically adjusted according to the skewness. For example, in skewed data, the kurtosis values may be even higher.

Table 2. Descriptive Statistics of Simulation Data

Simulation Conditions		Item Statistics			Total Score Statistics After Conducting Item Weighting Procedure								
Sample Size	Number of Items	Factor Loading	Mean of Kurtosis Values of Items	Mean of Skewness Values of Items	Mode	Median	Arithmetic Mean	Maximum	Minimum	Skewness	Kurtosis	Mean Difficulty	Mean Discrimination
250	20	0.5	1.016	0.002	9.958	9.992	9.992	19.771	0.190	0.001	2.257	0.500	0.505
		0.7	1.017	0.001	10.172	9.986	9.993	20.000	0.000	0.001	1.822	0.500	0.690
	30	0.5	1.016	0.002	15.010	14.969	14.983	29.313	0.670	0.004	2.260	0.499	0.488
		0.7	1.017	0.002	15.076	14.950	14.985	29.999	0.000	0.004	1.823	0.500	0.680
	40	0.5	1.016	0.001	20.130	20.028	19.993	38.775	1.166	-0.003	2.261	0.500	0.479
		0.7	1.017	0.004	19.431	19.951	19.960	39.995	0.004	0.002	1.825	0.499	0.673
1000	20	0.5	1.004	-0.001	10.032	10.003	10.005	19.998	0.002	0.002	2.252	0.500	0.506
		0.7	1.004	-0.001	9.963	10.008	10.004	20.000	0.000	0.000	1.811	0.500	0.692
	30	0.5	1.004	0.001	14.843	14.993	14.994	29.910	0.099	0.000	2.260	0.500	0.488
		0.7	1.004	-0.002	14.892	15.020	15.012	30.000	0.000	-0.002	1.815	0.500	0.681
	40	0.5	1.004	0.000	20.063	19.995	20.003	39.695	0.305	0.002	2.260	0.500	0.480
		0.7	1.004	-0.003	20.439	20.043	20.028	40.000	0.000	-0.004	1.817	0.501	0.675
3000	20	0.5	1.001	0.000	9.977	10.000	9.998	20.000	0.000	0.001	2.250	0.500	0.506
		0.7	1.001	0.001	9.914	9.998	9.996	20.000	0.000	0.000	1.811	0.500	0.692
	30	0.5	1.001	0.000	14.965	14.997	14.998	29.999	0.001	0.001	2.258	0.500	0.489
		0.7	1.001	-0.001	15.212	15.000	15.004	30.000	0.000	-0.001	1.814	0.500	0.681
	40	0.5	1.001	0.000	20.102	20.001	20.003	39.974	0.037	0.000	2.258	0.500	0.480
		0.7	1.001	0.000	19.985	19.995	19.998	40.000	0.000	0.000	1.815	0.500	0.675

Weighting Method Used

Test scores are open to having random errors during the development, implementation, and scoring of the tests. Since the amount and direction of random errors are not known, it is not possible to eliminate them from the measurement results. Random errors can be estimated only on a group basis using statistical methods, not individual-base. Reliability values obtained for statistical tests or items are the values that help in the estimation.

With this in mind, the item difficulty index (pj) of each item and the average score (Ii) of the individual from the test were calculated. It was checked whether the sum of these two variables was greater than 1. If this sum was greater than 1, the item reliability index was added to the answer of the individual. If less than 1, the item score of the individual was unchanged. In this way, a new matrix of item scores was established.

This weighting method was developed by researchers based on the following explanations. In the study, item scores were first produced, and then a weighting process was carried out through an item scores matrix. For this purpose,

$$f(x_{ij}) = \begin{cases} x_{ij} + item\ reliability\ index, & p_i + I_j \geq 1 \\ x_{ij}, & p_i + I_j < 1 \end{cases} \quad (1)$$

function is used. Where, p_i represents the difficulty of item, and I_j represents the average score of the individual j, which can be expressed as follows:

$$I_j = \sum_{i=1}^{n} \frac{x_i}{n} \quad (2)$$

Where, xi refers to the score of the individual taken from the item i (0 or 1) and n refers to the total number of items. Thus, the average score is calculated for each individual. Accordingly, the average score of the individual will be between 0 and 1.

The weighting function is defined as a piecewise function. When the function is examined, xij expresses the answer of the individual j to the item i. According to this, the answer to the item i given by the individual j can take a value of 1 or 0. Regarding the piecewise function, if the sum of the average score of the individual and the item difficulty indices is 1 or more; the item reliability index is added to the item score of the individual (0 or 1). If this sum is less than 1, the item score of the individual is kept the same (0 or 1).

The purpose of defining the item weighting function as specified in Equation 1 is to try to correct the random error involved in the measurement results. When the function is examined, it is based on the principle of correcting the answer given by a successful individual to an easy question carelessly or due to different random error sources. Likewise, a minimally successful individual can also receive correction scores for the item difficulty that he or she can answer. This situation is shown schematically in Figure 1.

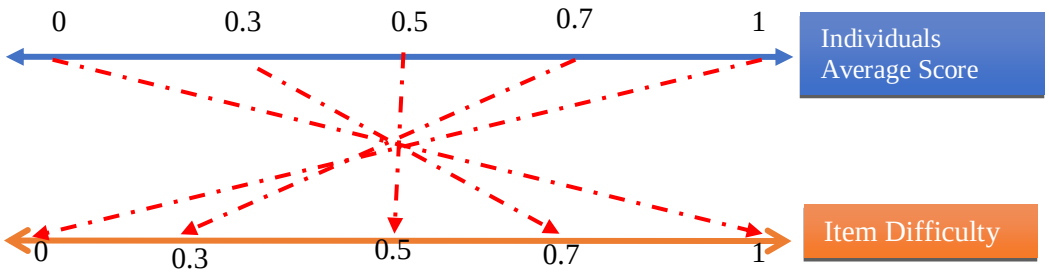

Figure 1. Item Weighting Function Chart

Figure 1 shows that when the average score of the individual is 0.3, for the individual to get correction points (1 + item reliability if the item is answered correctly, and 0 + item reliability if answered wrong), the item difficulty must be 0.7 or above, so the item must be easy. As the average score of the individual increases, the success of the individual increases, and the difficulty index of the item is also decreases, as it is getting more difficult. In this case, the weighting function can also work for an individual with low success. The important point in the function is that the individual has an average that he or she can respond correctly to that item. In Figure 1, a match between item difficulty and individual average score is presented for clarification of item weighting procedure. For example, if the average of the individual is 0.8, then the item difficulty is 0.20, and for the items above (0.20 and easier items), the weighting function will work.

Here is an example to explain this function: Assume that the average test score (total score / number of items) of a student is 0.62. In this case, an item's item difficulty index must be 0.38 (1-0.62) or above for the weighting of this student's score. The students' average scores for the items will not be weighted unless the item difficulty index 0.37 or lower. However, the students' average scores for the items will be weighted for the items with difficulty index 0.38 or higher. For more clarification, additional examples were given in Table 3.

When Table 3 was examined, it could be seen that if a student's average score + item difficulty index ≥ 1, the item's score will be weighted. But if it is smaller than 1, the item's score will not be weighted.

Table 3. Item Weighting Function Examples

Student Average Score	Item Difficulty Index	Whether Item Score Will Weight
0.20	0.80	Yes
0.20	0.85	Yes
0.20	0.79	No
0.20	0.50	No
0.50	0.50	Yes
0.50	0.60	Yes
0.50	0.30	No
0.50	0.20	No
0.70	0.30	Yes
0.70	0.40	Yes
0.70	0.20	No

To understand rationale of the method, let us suppose that the average of the individual is 0.9 and the difficulty of the item is 0.5. Then it is natural to expect that an individual who answers correctly to 90% of the items can also answer correctly to item with an average difficult. Similarly, an individual with an average of 0.20 may be expected to answer correctly to an item with an item difficulty index of 0.95. For these cases, weighting is performed by adding item reliability to the item score of the individual. If the individual gave the wrong answer to this item, an item reliability index is added to the item score to prevent it from getting 0 from that item. If the individual has already answered correctly that item, then the item score of the individual rises to the item reliability index. The reason for using the item reliability index here is that both item discrimination and the standard deviation of the item can be achieved at the same time. Thus, with the defined function, item scores of the individual are corrected by combining the item difficulty, item discrimination, and standard deviation of the item.

Process

In the study, data sets for each simulation condition were first generated using the psych package (Revelle, 2016) found in the R program (R Core Team, 2018). The aim was that the skewness value in data production is close to 0. For this reason, the cut-off score for the data produced in the dichotomously was taken as 0 (Revelle, 2016). The kurtosis values were based on the cut-off point.

After the data sets (1000 data sets for each condition) were generated according to the simulation conditions (18 conditions), the weighting process was applied to the data sets. The function presented in Equation 1 was used for weighting. The code was written by the researchers in the R program with the purpose of applying weighting to all data sets. Thus, a matrix of weighted item scores was generated from the generated data sets (1–0 form).

EFA, CFA, and reliability analyses were performed on all data sets before and after weighting (1000 replicative data sets of 18 conditions scored 1–0, and 1000 replicative data sets of 18 conditions). The psych package was used for EFA (Revelle, 2016). Since the weighted item scores matrix for EFA consisted of continuous data, Pearson correlation matrix was used for both non-weighted and weighted data sets for comparison. The Mplus software was used for CFA, but the Mplus Automation package in the R program was also utilized (Hallquist & Wiley, 2017). To conduct CFA, maximum likelihood (ML) estimation method was used for both non-weighted and weighted data sets for comparison.

The reported explained variance ratio calculated for the EFA in the study is the average explained variance ratio disclosed, which was obtained from 1000 replications for each condition. The average factor loading was obtained from 1000 replications for each condition, and then the average factor loading of the test was calculated according to the number of the items. CFI, RMSEA, and chi-square values calculated for CFA were obtained as an average as a result of the analysis of the data

sets obtained from 1000 replications. However, since the research was a simulation study, the average expression was not used. The average factor loading expression was used in the tables because the factors were calculated by taking the factor loading average (since the test shows the average factor loading).

It was examined that whether the explained variance ratio was normally distributed via Shapiro-Wilk test. Because they were normally distributed, t-test was used for comparison of explained variance ratio for original (1-0 data set) and weighted data set. To compare average factor loadings for original (1-0 data set) and weighted data set, Fisher's z-test was used. While Cohen's d was used for effect size to compare average of explained variance ratio, Cohen's q was used for effect size to compare average explained variance as two correlation coefficients. While Cohen's d interpreted as 0.2 small, 0.5 medium and 0.8 large, Cohen's q interpreted as 0.1 small, 0.3 medium and 0.5 large (Cohen, 1988, 1992).

RESULTS

Results are presented in the order of sub-problems.

Ratio of Variance Obtained as EFA Result and Findings for Average Factor Loadings

The explained variance ratios obtained by the simulation conditions as a result of the research and the average factor loadings are presented in Table 4.

The explained variance ratios are between 16.1% and 32.8% in the non-weighted data sets for all simulation conditions, and they range from 40.0% to 57.1% in weighted data sets. When the differences between the explained variance ratios that correspond to deviance before and after weighting are taken into account (explained variance ratio from after weighting minus explained variance ratio for binary scores), it is observed that they changed from 10.2% to 17.7% and the average was 13.8%. It is additional to note that these ratio values are derived by subtracting binary scores from the explained variance ratio for weighted scores.

For the simulation condition in which the average factor loading is 0.5, the increase in the explained variance ratio was less, and for simulation conditions with an average factor loading of 0.7, the explained variance ratio increased more. For simulation conditions with average factor loadings of 0.5 and 0.7, the increase in the sample size also increased the difference in the explained variance ratio. When the sample size was 3000, the increase of the number of items increased the difference in the explained variance ratio between before and after weighting. However, as the number of items decreased in the 250- and 1000-person samples, the explained variance ratio increased. As a result of the t-test performed to compare the explained variance ratios, it was observed that all differences were statistically significant ($\alpha<0.05$). When the effect size values were examined, it was observed that differences of the explained variance ratio had large effect size.

When the differences of the average factor loadings between before and after weighting were examined, it was observed that the differences varied between 0.096 and 0.138 and increased by an average of 0.119. In the case where the sample sizes were 250 and 1000, it was observed that with the reduction of the number of items the average factor loading difference before and after weighting was increased. On the other hand, when the average factor loading used in the simulation condition was 0.7 for a sample of 3000 individuals, the average factor loading difference between before and after weighting was increased. In addition, the difference of EVR increases with increasing the AFL in the samples of 250 and 1000 people. However, the difference in EVR decreases as the number of items increases. In the other hand, the difference for EVR was same as the number of items increased for 3000 sample size. According to these results, it could be said that the increase in the sample size, the effect of the increase in the number of items on the EVR decreases. As a result of the Fisher's z-test performed to compare the average factor loadings, it was observed that all

differences were statistically significant ($\alpha<0.05$). When the effect size values were examined, it was observed that differences of the average factor loadings had a small and medium effect size.

Table 4. Explained Variance Ratios and Average Factor Loadings

Simulation Conditions			Results Scoring 1–0		Results After Weighting		Difference		Cohen d for EVR	Cohen q for AFL
Sample Size	Number of Items	AFL	EVR	AFL	EVR	AFL	EVR	AFL		
250	20	0.5	0.164	0.401	0.277	0.522	0.113*	0.122†	5.86 (L)	0.15 (S)
	20	0.7	0.328	0.571	0.498	0.704	0.170*	0.133†	7.07 (L)	0.23 (M)
	30	0.5	0.164	0.400	0.269	0.506	0.105*	0.105†	3.85 (L)	0.13(S)
	30	0.7	0.328	0.571	0.483	0.682	0.154*	0.110†	3.61 (L)	0.18(S)
	40	0.5	0.163	0.400	0.266	0.496	0.102*	0.096†	3.23 (L)	0.12(S)
	40	0.7	0.327	0.570	0.474	0.667	0.147*	0.097†	2.73 (L)	0.16(S)
1000	20	0.5	0.162	0.401	0.278	0.526	0.116*	0.125†	11.60 (L)	0.16(S)
	20	0.7	0.327	0.571	0.500	0.707	0.173*	0.136†	14.60 (L)	0.23 (M)
	30	0.5	0.161	0.401	0.270	0.510	0.109**	0.109†	4.83 (L)	0.14(S)
	30	0.7	0.327	0.571	0.485	0.684	0.158*	0.113†	4.09 (L)	0.19(S)
	40	0.5	0.162	0.401	0.267	0.502	0.105*	0.101†	3.52 (L)	0.13(S)
	40	0.7	0.326	0.571	0.477	0.668	0.151*	0.097†	2.87 (L)	0.16(S)
3000	20	0.5	0.161	0.401	0.278	0.527	0.116**	0.125†	20.44 (L)	0.16(S)
	20	0.7	0.326	0.571	0.500	0.707	0.174*	0.136†	25.03 (L)	0.23 (M)
	30	0.5	0.161	0.401	0.280	0.529	0.119*	0.128†	22.15 (L)	0.16(S)
	30	0.7	0.326	0.571	0.502	0.709	0.176*	0.137†	27.29 (L)	0.24 (M)
	40	0.5	0.161	0.401	0.281	0.530	0.120*	0.129†	24.67 (L)	0.17 (S)
	40	0.7	0.326	0.571	0.503	0.709	0.177*	0.138†	27.28 (L)	0.24 (M)
Mean							0.138	0.119		

EVR: explained variance ratio; AFL: average factor loading,
*in terms of t-test result it is statistically significant ($\alpha<0.05$)
†in terms of Fisher's z-test it is statistically significant ($\alpha<0.05$)
(S): Small, (M): Medium, (L): Large

The impact of weighting on CFA results was examined before and after weighing in CFA as well as EFA.

Findings for Chi-Square Values and Results Obtained from CFA Fit Indexes

The CFI, RMSEA, and chi-square values obtained as a result of the CFA performed before (1–0 item score matrix) and after the weighting process applied to the item scores matrix are presented in Table 5.

When Table 5 is examined, it is seen that CFI values generally improve after weighting, and RMSEA and chi-square values tend to decrease. When the differences between CFI values between before and after weighting are examined, a change is observed between -0.003 (12th row) and 0.311 (1st row). The CFI values increased by an average of 0.112 for all simulation conditions after weighting. The CFI value was decreased when there were only 1000 subjects, 40 items, and an average factor loading of 0.7. There seems to be some improvement for all other conditions. When the average factor loading was 0.5, the improvement in the CFI was higher than the conditions when it was 0.7. When the sample size and number of item conditions are examined, when the sample size decreases and the number of items increases it can be said that the weighting tends to increase the CFI index.

Table 5. CFA, RMSEA, and Chi-Square Values Obtained from CFA

	Simulation Conditions			1–0 Scored Results			Weighting after Results			Difference		
Row Number	Sample Size	Number of Items	Average Factor Loading	CFI	RMSEA	Chi-Square	CFI	RMSEA	Chi-Square	CFI	RMSEA	Chi-Square
1	250	20	0.5	0.498	0.076	416.092	0.809	0.069	376.754	0.311	-0.006	-39.338
2		20	0.7	0.824	0.076	419.620	0.932	0.065	354.892	0.108	-0.011	-64.728
3		30	0.5	0.685	0.052	679.066	0.843	0.051	679.724	0.158	-0.001	0.658
4		30	0.7	0.869	0.055	718.176	0.915	0.056	763.311	0.046	0.001	45.135
5		40	0.5	0.751	0.042	1063.197	0.846	0.043	1124.488	0.095	0.002	61.291
6		40	0.7	0.877	0.047	1156.833	0.896	0.052	1351.126	0.019	0.005	194.293
7	1000	20	0.5	0.518	0.074	1102.929	0.822	0.067	941.487	0.304	-0.007	-161.442
8		20	0.7	0.844	0.071	1039.637	0.944	0.059	775.492	0.100	-0.012	-264.145
9		30	0.5	0.715	0.049	1362.145	0.862	0.047	1363.485	0.147	-0.001	1.340
10		30	0.7	0.900	0.048	1325.547	0.933	0.047	1513.180	0.033	0.000	187.633
11		40	0.5	0.799	0.036	1723.846	0.871	0.038	1989.171	0.071	0.002	265.325
12		40	0.7	0.923	0.037	1735.951	0.920	0.041	2550.180	-0.003	0.004	814.229
13	3000	20	0.5	0.522	0.074	2946.655	0.824	0.067	2462.226	0.302	-0.007	-484.428
14		20	0.7	0.847	0.070	2710.133	0.946	0.058	1917.261	0.099	-0.012	-792.872
15		30	0.5	0.721	0.048	3216.713	0.895	0.043	2709.393	0.174	-0.005	-507.320
16		30	0.7	0.905	0.046	2996.236	0.966	0.038	2192.891	0.061	-0.008	-803.345
17		40	0.5	0.806	0.036	3583.580	0.926	0.032	3064.073	0.120	-0.003	-519.507
18		40	0.7	0.931	0.035	3401.912	0.975	0.029	2596.216	0.044	-0.006	-805.696
	Mean									0.112	-0.004	-159.606

When the differences of the RMSEA index between before and after weighting are examined, these differences are seen to have changed between -0.012 (14th row) and 0.005 (6th row), and the average value is 0.004. Most of the simulation conditions result in a decrease in the RMSEA value. The biggest decrease was in the 20-item test with 3000 individuals and with an average 0.7 factor loading. When the number of samples decreases, and the number of items increases, the difference in RMSEA values also decreases. While the increase in the number of items for 250 and 1000 sample sizes resulted in a decrease in RMSEA, RMSEA values was improved for all conditions for 3000 sample size.

When the differences between before and after weighting of the chi-square value are examined, it is seen that these differences changed between -805.696 (18th row) and 814.229 (12th row). Chi-square values decreased after weighting in all conditions for the sample sizes of 3000. When the sample sizes were 250 and 1000, the weighting process caused a decrease in the chi-square values in 20-item tests. However, when the number of items were 30 and 40, chi-square values increased.

Findings for Coefficient of Reliability

The findings for the Cronbach's alpha reliability coefficient obtained as a result of the reliability analysis conducted before the weighting was applied to item matrix scores (1–0 item matrix scores) and after applying the weighting process are presented in Table 6.

Table 6. Cronbach's Alpha Values Obtained as a Result of Reliability Analysis

Simulation Conditions			1–0 Scored Results	Results after Weighting	Difference	Cohen's q for Cronbach's Alpha
Sample Size	Number of Items	Average Factor Loading	Cronbach's Alpha	Cronbach's Alpha	Cronbach's Alpha	
250	20	0.5	0.792	0.882	0.091[†]	0.31 (M)
	20	0.7	0.906	0.952	0.046[†]	0.35 (M)
	30	0.5	0.851	0.952	0.101[†]	0.59 (L)
	30	0.7	0.935	0.962	0.027[†]	0.28 (M)
	40	0.5	0.884	0.930	0.047[†]	0.26 (M)
	40	0.7	0.950	0.970	0.020[†]	0.26 (M)
1000	20	0.5	0.793	0.884	0.091[†]	0.31 (M)
	20	0.7	0.906	0.952	0.046[†]	0.35 (M)
	30	0.5	0.851	0.912	0.061[†]	0.28 (M)
	30	0.7	0.935	0.963	0.027[†]	0.29 (M)
	40	0.5	0.885	0.932	0.047[†]	0.28 (M)
	40	0.7	0.951	0.971	0.020[†]	0.27 (M)
3000	20	0.5	0.793	0.885	0.091[†]	0.32 (M)
	20	0.7	0.906	0.952	0.046[†]	0.35 (M)
	30	0.5	0.852	0.921	0.069[†]	0.33 (M)
	30	0.7	0.936	0.968	0.032[†]	0.35 (M)
	40	0.5	0.885	0.940	0.055[†]	0.34 (M)
	40	0.7	0.951	0.976	0.025[†]	0.36 (M)
Mean					0.052	

[†]in terms of Fisher's z-test it is statistically significant ($\alpha<0.05$)
(S): Small, (M): Medium, (L): Large

When Table 6 is examined, it is seen that the Cronbach's alpha coefficient changes between 0.792 and 0.951 for the non-weighted data sets and between 0.882 and 0.976 for the weighted data sets. When the differences between before and after weighing are examined, it is observed that they show changes between 0.020 (6th row) and 0.101 (3rd row). The Cronbach's alpha coefficients increased by 0.052 on average for all simulation conditions after weighting. For simulation conditions with an average factor loading of 0.7, the increase in the Cronbach's alpha coefficient was observed to be lower than the increase for the simulation conditions with an average factor loading of 0.5. As a result of the increase of the number of items in the simulation condition, the effect of the weighting process on the Cronbach's alpha coefficient is decreased. It has been observed that the increase in sample size generally results in a further increase in the confidence coefficient obtained after the weighting process. When the sample size and item number were evaluated together, it was observed that in the cases where the sample size increased, and the number of items decreased, the weighting process increased the Cronbach's alpha coefficient more. As a result of the Fisher's z-test performed to compare the average factor loadings, it was observed that all differences were statistically significant ($\alpha<0.05$). When the effect size values were examined, it was observed that differences of the average factor loadings had a small and medium effect size.

DISCUSSION and CONCLUSION

As a result of the research, it was observed that the proposed weighting method increased the explained variance ratio by 13.8%. As the average factor loading increases, the effect of the weighting process on the explained variance increases. Accordingly, the effect of the weighting process increases when the relationship between the items increases. This differs from the result obtained by the nominal weighting method used by Ghiselli (1964). The nominal weighting method decreases the average correlation between the components, and it has been reported that weighted

scores are more effective in ranking individuals (Ghiselli, 1964). In the current study, as the average factor loading of the items increased, the explained variance ratio also increased. Accordingly, it is recommended to use the weighting method used in the current research in tests with high relation between items.

When the effect of the weighting process on the explained variance ratio is examined, it can be said that the explained variance ratio also increases when the factor loading, number of items, and sample increase. When the averages obtained are examined, we do not agree with Guilford (1954) and Phillips (1943), who argued that item weighting would not be worth the effort. Increasing the ratio of explained variance in a psychological construct by around 13% is an important gain, and with the help of computer programs to operate weighting it is not difficult.

When the results of CFA are examined, it can be said that in general the weighting process improves CFI and RMSEA values. When chi-square values are examined, although there is no improvement for some models, it is observed that there was a decrease in chi-square values when evaluated as average. CFA estimation method can cause that result. To provide comparison of weighted and non-weighted analysis result, ML estimation method was used for CFA. On the other hand, when the change in the chi-square values which is close to zero, it can be said that the change in CFI values are quite high. So, it can be stated that there is a better fit in terms of CFI values.

When the reliability analysis results are examined, it is observed that the reliability coefficient on average increased to the 0.05 level. This result is similar to the research findings of Guilford, Lovell, and Williams (1942). However, when both EFA and CFA results are evaluated together, it is believed to be sufficient when the reliability coefficient is not reduced, because the increase in the number of items also increases the reliability coefficient. All calculated reliability coefficients show that the weighting results can be used. It is estimated that it may be sufficient for the weighting process not to have a lowering effect.

According to the results of the research, the weighting method recommended by researchers can be used by both researchers and policy practitioners. This weighting method contributes to the construct, but it should not be overlooked that it is being investigated for one-dimensional constructs. It may be advisable to researchers to investigate how the proposed weighting method produces results for two-dimensional tests or greater.

REFERENCES

American Educational Research Association (AERA), American Educational Research Association (APA), & National Council on Measurement In Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Akkuş, O., & Baykul, Y. (2001). Çoktan seçmeli test maddelerinin puanlamada, seçenekleri farklı biçimde ağırlıklandırmanın madde ve test istatistiklerine olan etkisinin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 20,* 9–15.

Birnbaum, A. (1968). Some latent models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Addison-Wesley: Reading, MA.

Burt, C. (1950). The influence of differential weighting. *British Journal of Statistical Psychology, 3*(2), 105–125. https://doi.org/10.1111/j.2044-8317.1950.tb00288.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*(5), 754–761. https://doi.org/10.1037/0022-006X.56.5.754

Daniel, L. G. (1989). Comparisons of exploratory and confirmatory factor analysis. In *Mid-South Educational Research Association*. Little Rock, AR: (ERIC Document Reproduction Service No. ED314447).

Dick, W. (1965). *Item weighting: Test parameter effects and comparison of the efficiency of various weighting methods.* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3564813).

Erdem, B., Ertuna, L., & Doğan, N. (2016, September). Çoktan seçmeli testlerde seçenek ağırlıklandırma yöntemlerinin testin faktör yapısına etkisinin incelenmesi (pp. 148–149). Paper presented at the Fifth International Congress on Measurement And Evaluation in Education And Psychology, Antalya, TR. Abstract retrieved from http://epod2016.akdeniz.edu.tr/_dinamik/333/53.pdf.

Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme-I: Temel kavramlar ve işlemler* (2nd ed.). Ankara: Pegem Akademi.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286–299. https://doi.org/10.1037/1040-3590.7.3.286

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York: McGraw-Hill.

Gorsuch, R. L. (1974). *Factor analysis*. Toronto: W. B. Saunders.

Gözen-Çıtak, G. (2010). Klasik Test ve Madde Tepki Kuramlarına göre çoktan seçmeli testlerde farklı puanlama yöntemlerinin karşılaştırılması. *İlköğretim Online, 9*(1), 170–187.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265–275.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Guilford, J. P., Lovell, C., & Williams, R. M. (1942). Completely weighted versus unweighted scoring in an achievement examination. *Educational and Psychological Measurement, 2*, 15–21.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hallquist, M., & Wiley, J. (2017). MplusAutomation: Automating Mplus Model Estimation and Interpretation. Retrieved from https://cran.r-project.org/package=MplusAutomation

Horst, P. (1936). Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika, 1*(1), 53–60. https://doi.org/10.1007/BF02287924

MEB. (2013). Temel Eğitimden Ortaöğretime Geçiş. Retrieved March 24, 2015, from http://oges.meb.gov.tr/docs2104/sunum.pdf

MEB. (2019). Sınavla öğrenci alacak ortaöğretim kurumlarına ilişkin merkezî sınav başvuru ve uygulama kılavuzu. Retrieved May 15, 2019, from https://www.meb.gov.tr/meb_iys_dosyalar/2019_04/03134315_Kilavuz2019.pdf

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037//0003-066X.50.9.741

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd. ed.). New York, NY: McGraw-Hill.

Özdemir, D. (2004). Çoktan seçmeli testlerin Klasik Test Teorisi ve Örtük Özellikler Teorisine göre hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlanması yönünden karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 26*, 117–123.

Phillips, A. J. (1943). Further evidence regarding weighted versus unweighted scoring of examinations. *Educational and Psychological Measurement, 3*(1), 151–155. https://doi.org/10.1177/001316444300300114

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/.

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research. Evanston, Illinois. Retrieved from https://cran.r-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373.

https://doi.org/10.1037/a0029315
Rotou, O., Headrick, T. C., & Elmore, P. B. (2002). An investigation of difference scores for a grade-level testing program. *International Journal of Testing, 2*(2), 83–105. https://doi.org/10.1207/S15327574IJT0202
Stevens, J. P. (2009). *Applied multivariate statistics for the social science* (5th ed.). London: Routledge.
Trierweiler, T. (2009). *An evaluation of estimation methods in confirmatory factor analytic models with ordered categorical data in LISREL*. (Doctoral dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3416004).
Yurdugül, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(1), 1–8.

Madde Ağırlıklandırmanın Güvenirlik ve Geçerliğe Etkisi

Giriş

Psikolojik alanda kullanılan testlerden elde edilen puanların geçerliği, psikolojik ölçme alanının en önemli konuları arasında yer almaktadır. Geçerlik, uygulanan testlerden elde edilen puanların bir özelliği olarak düşünülmekte ve şemsiye bir kavram olarak yapı geçerliği altında toplanabilmektedir.

Yapı geçerliğine yönelik kanıt toplama sürecinde genellikle açımlayıcı ve doğrulayıcı faktör analizinden yararlanılmaktadır. Açımlayıcı faktör analizi (AFA) kovaryans yapıları üzerine kurulmuş olup gözlenen değişkenler arasındaki kovaryans matrisinden daha az sayıda gizil değişkenler (faktörler) elde etmeye yarayan bir tekniktir. AFA'da amaç, değişkenlerin oluşturduğu kümenin faktör yapısını ortaya çıkarmaktır. DFA'da ise teorik olarak ortaya konulan kuramsal yapının test edilmekte ve analiz öncesinde ölçülen değişkenin yapısal özellikeri bilinmektedir. Bu nedenle DFA'da amaç ölçme aracından elde edilen ölçümlere dayanarak öngörülen faktör yapısının doğrulanmaya çalışılmasıdır. Yapı geçerliğine yönelik kanıt toplama sürecinde her iki analizin de önemi yüksektir. Nunnally (1978), faktör analizi psikolojik yapıların ölçümünün kalbinde yer almaktadır diyerek faktör analizinin önemi vurgulamaktadır.

Testten elde edilen puanların geçerliğini artırmak amacıyla farklı önlemler alınabilir. Buna örnek olarak ölçme aracının geliştirilme aşamasında ölçek geliştirme prosedürünü takip etmek, kuramsal alt yapıyı iyi bir şekilde bilmek ve ölçme aracına yansıtabilmek, testlerin uygulanma aşamasında bazı önlemlerin alınması gösterilebilir. Ancak test uygulandıktan sonra elde edilen puanlar üzerinde bazı ağırlıklandırma işlemleri ile de testten elde edilen puanların geçerliğinin artırılabileceği düşünülmüş ve madde ağırlıklıklandırmasına yönelik araştırmalar yürütülmüştür.

Madde ağırlıklandırmaya yönelik araştırmalar incelendiğinde çoğunlukla 1900'lü yılların ilk yarısında yer aldığı görülmektedir. Madde ağırlıklandırmayla ilgili olarak önerilen yöntemler arasında çoklu regresyon yoluyla değer atama, kısmı regresyon katsayılarını atama Guilford (1954), madde ayırıcılık indeksleri ile ağırlıklandırma Birnbaum (1968), faktör analizini kullanma (Burt, 1950) gibi yöntemlerin kullanılmasının yanında test varyansını ya da madde varyansını kullanarak madde ağırlıklandırma (Dick, 1965), maddelerin bağlantılı olduğu içerik dikkate alınarak daha önemli konularla ilişkili olan maddeleri ağırlıklandırma (Ghiselli, 1964) tek tek maddeler yerine madde kümelerinin ağırlıklandırma şeklinde yöntemler öneren yazarlarda bulunmaktadır (Gulliksen, 1950). Günümüze daha yakın bir çalışmada ise Rotou, Headrick ve Elmore (2002) çok boyutlu madde tepki kuramı parametreleri kullanarak maddeleri ağırlıklandırmayı ve bireylerin puanlarını hesaplamak için klasik test kuramındaki toplam puan hesaplamasını kullanmaya dayanan bir hibrit ağırlıklandırma yöntemi önermiştir.

Türkiye'de ağırlıklandırmaya yönelik araştırmalar yürütülmüş ancak genellikle çoktan seçmeli maddeler için seçenek ağırlıklandırma üzerinde durulmuş (Akkuş & Baykul, 2001; Erdem, Ertuna, & Doğan, 2016; Gözen-Çıtak, 2010; Özdemir, 2004) madde ağırlıklandırma üzerinde yürütülen araştırmaların sınırlı olduğu görülmüştür (Yurdugül, 2010). Yurdugül (2010) tarafından yürütülen

araştırmada bireylerin toplam puanları üzerinden değerlendirme yapılmıştır. Mevcut araştırmada ise yapı geçerliğine yönelik araştırmada bulunulmuştur.

Yöntemler genel olarak değerlendirildiğinde, Guilford (1954) ve Phillips (1943) madde ağırlıklandırma için harcanan emeğe değmeyeceğini belirtmiştir. Ancak madde ağırlıklandırmayla testten elde edilen sonuçların geçerliği ve güvenirliğinin artırılmasının yanında bireyler arasındaki farkı da maksimize etmesi gerektiğinden (Horst, 1936) bireyleri daha iyi ayırmayı sağlayacaktır. Günümüzde bilgisayar teknolojisinin oldukça gelişmiş olması madde ağırlıklandırma işlemine harcanacak emeği de azaltacağından geçerlik ve güvenirliğe aşırı katkı yapmasa bile kısmi katkıları sebebiyle kullanılması önerilebilir.

Madde ağırlıklandırmanın sonuçları genel olarak incelendiğinde, kısa testler için maddelerin farklı ağırlıklandırmanın daha verimli olduğu, madde sayısının 10 ila 20 arasında olduğunda madde ağırlıklandırmanın çok az etkili olduğu (Ghiselli, 1964), uzun testler içinse en iyi ağırlıklandırmanın tüm maddeler için 1 olarak seçilmesi olduğu belirtilmektedir. Maddeler arası korelasyonların ortalaması düşük olduğunda madde ağırlıklandırmanın daha iyi sonuçlar verdiği (Guilford, 1954) ve testteki bileşen sayısı azaldıkça maddeleri farklı puanlamanın eşit ağırlıklandırma yoluyla elde edilen toplam puana (Örneğin, doğru cevap için 1, yanlış cevap için 0 puan vermek ve doğru cevap verilen maddeleri toplamak gibi.) göre bireyleri sıralama üzerinde daha etkili olduğu ifade edilmiştir (Ghiselli, 1964).

Madde ağırlıklandırmanın testten elde edilen sonuçların geçerliğini ve güvenirliğini artırıcı etki yapması ve bu nedenle de açıkça önem arz etmesine rağmen çok az sayıda çalışmada kullanılması Burt (1950) tarafından da şaşırtıcı olarak ifade edilmiştir. Günümüzde madde ağırlıklandırmanın etkilerine yönelik olarak yürütülen araştırmaların sınırlı olması nedeniyle bu araştırmada araştırmacılar tarafından geliştirilen madde ağırlıklandırma yöntemi farklı koşullar altında incelenmiştir. Araştırmada "Madde ağırlıklandırmanın testin geçerlik ve güvenirliğine etkisi nasıldır?" sorusuna yanıt aramak amacıyla i) önerilen madde ağırlıklandırma yöntemiyle elde edilen dönüştürülmüş madde puanları matrisi üzerinden yürütülen AFA sonucunda açıklanan varyans oranı nasıl değişmektedir?, ii) önerilen madde ağırlıklandırma yöntemiyle elde edilen dönüştürülmüş madde puanları matrisi üzerinden yürütülen DFA sonucunda CFI, RMSEA ve ki-kare değerleri nasıl değişmektedir?, iii) önerilen madde ağırlıklandırma yöntemiyle elde edilen dönüştürülmüş madde puanları matrisi üzerinden yürütülen güvenirlik analizi sonucunda Cronbach Alfa güvenirlik katsayısı değerleri nasıl değişmektedir? Sorularına yanıt aranmıştır.

Yöntem

Araştırmacılar tarafından önerilen ağırlıklandırma yönteminin testin geçerlik ve güvenirliği üzerindeki etkisinin incelenmesi amacıyla Monte Carlo simülasyon çalışması yürütülmüştür. Araştırmanın verileri, kategorik veri türlerinin çokluğu ve ayrı çalışılması gerektiği düşüncesiyle 1-0 puanlamayla sınırlandırılmıştır. Diğer bir sınırlılık ise veri setlerinin tek boyutlu üretilmesidir. Bunun nedeni ise çok boyutlu verilerde veri türü, boyut sayısı, boyutlar arası ilişkiler, boyutlardaki madde sayıları vb. gibi birçok koşulu bir arada ele almanın çalışmayı amacından uzaklaştırabileceği düşüncesidir.

Simülasyon koşulu olarak örneklem büyüklüğü (250, 1000 ve 3000), madde sayısı (20, 30 ve 40) ve ortalama faktör yükü (0.5 ve 0.7) ele alınmıştır. Örneklem büyüklüğü olarak küçük, orta ve büyük olacak şekilde örneklemler oluşturulmuştur. Ağırlıklandırma sonrasında madde sayısının, faktör analizi ve puanların güvenirliğine etkisini incelemek için madde sayısı da simülasyon çalışmasına koşul olarak eklenmiştir. Faktör yükleri ortalaması da tek boyutluluğun güçlü ya da zayıf olması durumunda ağırlandırmanın etkisini görmeyi sağlayacağı düşüncesiyle ele alınmıştır. Bütün koşullar ele alındığında toplamda 18 simülasyon koşulu araştırılmış ve her bir koşul için 1000 replikasyon yapılmıştır.

Araştırmada kullanılan ağırlıklandırma yönteminde,

$$f(x_{ij}) = \begin{cases} x_{ij} + madde\ güvenirlik\ indeksi, & p_i + I_j \geq 1 \\ x_{ij}, & p_i + I_j < 1 \end{cases} \quad (1)$$

fonksiyonu kullanılmıştır. Burada p_i, i maddesinin madde güçlüğünü, I_j ise j. bireyin ortalama puanını ifade etmektedir. Yani;

$$I_j = \sum_{i=1}^{n} \frac{x_i}{n} \quad (2)$$

şeklinde ifade edilebilir. Burada x_i, bireyin i. maddeden aldığı puanı (0 ya da 1), n ise toplam madde sayısını ifade etmektedir. Böylece her bireyin ortalama puanı hesaplanmaktadır. Buna göre bireyin ortalama puanının 0 ile 1 arasında değer alacağı söylenebilir.

Ağırlıklandırma fonksiyonu parçalı fonksiyon olarak tanımlanmıştır. Fonksiyon incelendiğinde x_{ij}, j bireyinin i maddesine verdiği yanıtı ifade etmektedir. Buna göre j bireyinin i maddesine verdiği yanıt 1 ya da 0 değerinin alabilir. Parçalı fonksiyonun kuralları incelendiğinde eğer bireyin testten aldığı ortalama puan yani I_j ile i maddesinin madde güçlük indeksinin toplamı 1 ve daha büyükse o zaman bireyin madde puanına (0 ya da 1) i maddesinin madde güvenirlik indeksi eklenmektedir. Eğer bu toplam 1'den küçükse bu durumda bireyin madde puanı aynen korunmaktadır.

Madde ağırlıklandırma fonksiyonunun Denklem 1'de belirtilen şekilde tanımlanmasının amacı, ölçme sonuçlarına karışan tesadüfi hatayı düzeltmeye çalışmaktır. Fonksiyon incelendiğinde başarılı bir bireyin kolay bir soruya dikkatsizlikle veya farklı tesadüfi hata kaynakları nedeniyle verdiği cevabın düzeltilmesi esasına dayanmaktadır. Aynı şekilde düşük başarılı bir birey de kendi cevaplayabileceği madde güçlüğü için düzeltme puanı alabilmektedir.

Sonuç ve Tartışma

Araştırma sonucunda önerilen madde ağırlıklandırma yönteminin açıklanan varyans oranını ortalama %13.8 artırdığı gözlenmiştir. Ortalama faktör yükü arttıkça ağırlıklandırma işleminin açıklanan varyans üzerindeki etkisi artmıştır. Buna göre maddeler arasındaki ilişki arttıkça ağırlıklandırma işleminin etkisinin arttığı söylenebilir. Bu sonuç Ghiselli (1964) tarafından belirtilen nominal ağırlıklandırma yönteminden elde edilen sonuçla farklılaşmaktadır. Nominal ağırlıklandırma yönteminde bileşenlere farklı ağırlıklıklar atanmaktadır. Ghiselli (1964) tarafından belirtilen nominal ağırlıklandırma yöntemiyle bileşenler arası ortalama korelasyon azaldıkça ağırlıklandırılmış puanların bireylerin sıralanmasında daha fazla etkili olduğu raporlanmıştır. Mevcut araştırmada ise maddelerin ortalama faktör yükü arttıkça açıklanan varyans oranının da arttığı gözlenmiştir. Buna göre maddeleri arasındaki ilişkileri yüksek olan testlerde mevcut araştırmada kullanılan ağırlıklandırma yönteminin kullanılması önerilebilir.

Ağırlıklandırma işleminin açıklanan varyans oranına etkisi incelendiğinde faktör yükü, madde sayısı ve örneklem arttıkça açıklanan varyans oranın da arttığı söylenebilir. Guilford (1954) ve Phillips (1943) harcanan emeğe nazaran elde edilen iyileşmenin önemsiz olduğunu vurgulamıştır. Ancak mevcut araştırmadan elde edilen ortalamalar incelendiğinde madde ağırlıklandırma yönteminin kullanılmasının harcanan efora değecek sonuçlar ürettiği düşünülmektedir. Diğer bir deyişle kullanışlık açısından araştırmada önerilen ağırlıklandırma yönteminin önerilebileceği söylenebilir. Bir psikolojik özellikteki açıklanan varyans oranını %13 civarında arttırmak önemli bir kazançtır ve artık bilgisayar programlarının da yardımıyla ağırlıklandırma yapmak çok da zor olmamaktadır.

Doğrulayıcı faktör analizi sonuçları incelendiğinde ise genel olarak ağırlıklandırma işleminin CFI ve RMSEA değerlerinde iyileşme sağladığı söylenebilir. Ki-Kare değerleri incelendiğinde bazı modeller için iyileşme olmadığı gözlense de ortalama olarak değerlendirildiğinde ki-kare değerlerinde de bir düşme olduğu gözlenmiştir.

Güvenirlik analizi sonuçları incelendiğinde ise ortalama olarak 0.05 düzeyinde güvenirlik katsayısının yükseldiği gözlenmiştir. Bu sonuç Guilford, Lovell, & Williams (1942) araştırma bulgularıyla da benzerdir. Ancak AFA ve DFA sonuçları birlikte değerlendirildiğinde güvenirlik katsayısının azalmamasının yeterli olabileceği düşünülmektedir. Çünkü madde sayısının artması güvenirlik katsayısını da arttırmaktadır. Hesaplanan tüm güvenirlik katsayıları ağırlıklandırma sonuçlarının kullanılabileceğini göstermektedir. Ağırlıklandırma işleminin güvenirliği düşürücü bir etki yapmamasının yeterli olabileceği değerlendirilmektedir.

Araştırma sonuçlarına göre araştırmacılar tarafından önerilen ağırlıklandırma yönteminin kullanılması hem araştırmacılara hem de politika uygulayıcılarına önerilebilir. Yapı geçerliğine katkı sunan bu ağırlıklandırma yönteminin tek boyutlu yapılar için araştırıldığı gözden kaçırılmamalıdır. Araştırmacılara iki yada daha çok boyutlu testler için önerilen ağırlıklandırma yönteminin nasıl sonuçlar ürettiğini araştırılması önerilebilir.