IJPES
www.ijpes.com
International Journal of Psychology and Educational Studies
ISSN: 2148-9378

# How Do Different Weighting Methods Affect the Overall Effect Size in Meta-Analysis? : An Example of Science Attitude in Türkiye Sample

**Yıldız YILDIRIM[1], Melek Gülşah ŞAHİN[2]**

[1] *Faculty of Education, Aydın Adnan Menderes University, Aydın, Türkiye* *0000-0001-8434-5062*

[2] *Faculty of Education, Gazi University, Ankara, Türkiye* *0000-0001-5139-9777*

ARTICLE INFO

*Article History*
Received 14.10.2022
Received in revised form 03.03.2023
Accepted 08.07.2023
Article Type: Research Article

ABSTRACT

There is increasing interest in meta-analysis in different fields due to the need to combine the results of primary research. One of the crucial concepts in combining results is weighting. This study examines how Hunter and Schmidt's method, weighting by sample size; Hedges and Vevea's method, weighting by inverse variance; and Osburn and Callender's method, unweighting, affect the overall effect size in meta-analysis. In this context, for meta-analysis, the search was done for studies examining the effects of alternative measurement and assessment techniques and methods in science education on science attitudes. The databases of the HEI National Thesis Center, Web of Science, ERIC, EBSCO, Google Scholar, and DergiPark were searched between 2010 and 2021. Eleven studies (with 14 effect sizes) that met the criteria were included in the meta-analysis. In line with the study's findings, it was observed that the overall effect sizes were significant and did not change much in the weighting methods. Besides, it was found that the method with the lowest standard error was unweighted. The weighting methods of Hunter and Schmidt and Hedges and Vevea gave similar results in terms of standard error. When the correlation coefficient between the weighting methods was examined, it was seen that all correlation coefficients were greater than 0.90.

Keywords:
Hedges-Vevea's method, Hunter-Schmidt's method, Weighting in meta analysis, Unweighting.

## 1. Introduction

Since the existence of human beings, new information has been obtained with the curiosity and needs of learning, and further information has been added to previous information, or old information has been replaced with new ones. There are subjective and objective processes, considering the process of obtaining information. Subjective processes can be regarded as the presence of senses, observations, transmissions, and authority. On the other hand, objective processes are the ones in which scientific processes are followed, and empirical methods are at the forefront. Objective processes are generally followed in science and the social sciences. However, due to the nature of the social sciences, similar or close results may not always be obtained compared to the sciences. According to Berliner (1992), educational research produces less reliable and inconclusive results compared to studies in the fields of physics, chemistry, geology, etc. For this reason, it is more difficult to conduct scientific research in the social sciences, especially educational research.

The same results may not be obtained in the studies carried out in the sub-fields of social sciences, especially on human behavior, since they are carried out on different samples and under different conditions. Because of these varying results in the social sciences, it is observed that researchers have carried out independent research for similar purposes. This situation leads to the need to obtain a general conclusion from the research conducted with the same purpose. One of the essential methods used to get a general result from these

[1]Corresponding author's address: Aydin Adnan Menderes University, Faculty of Education, Nu:125, Zafer, Efeler, Aydin, 09010/Türkiye
e-mail: yildizyldrm@gmail.com
**Citation:** Yıldırım, Y. & Şahin, M. G. (2023). How do different weighting methods affect the overall effect size in meta-analysis? : An example of science attitude in Türkiye sample. *International Journal of Psychology and Educational Studies*, *10*(3), 744-757. https://dx.doi.org/10.52380/ijpes.2023.10.3.1049

individual studies is meta-analysis. Situations such as the incremental increase in scientific knowledge, the ease of access to information, and the rise in the number of studies conducted under different conditions for similar purposes have led to an intensification of interest in meta-analysis.

The meta-analysis applies statistical methods to combine the findings of more than one primary research study (Glass, 1976). For this purpose, meta-analysis enables us to reach an overall effect size value by combining the effect size values of primary studies. The reason meta-analysis uses effect sizes to combine the findings of primary studies is because this statistic is standardized and addresses a common scale. At this point, we come across the concept of weighting, which is a crucial issue in meta-analysis. The meta-analysis, while calculating the overall effect size, uses the weighting methods found in the literature to weight the effect size values of the primary studies. Some of these methods are Hedges and Olkin's (1985) inverse variance method for the fixed-effects model, Hunter and Schmidt's (1990) weighting by sample size (HS) method, and Hedges and Vevea's (1998) inverse variance method (HV) for the random-effects model. These methods are frequently used in the literature. However, some studies have unweighted (UW) analyses. The concept of unweighting is essentially taking the arithmetic average of the effect sizes. When the literature is examined, it is seen that there are studies comparing different weighting methods (Bonett, 2008; Brannick et al., 2011; Englund et al., 1999; Field, 2005; Fuller & Hester, 1999; Manolov et al., 2014; Marin-Martinez & Sanchez-Meca, 2009; Scmidt et al., 2009; Shuster, 2010; van den Noortgate & Onghena, 2003). Englund et al. (1999) compared the results of unweighting and weighting with inverse variance using real data to test the unreliability of unweighting. Fuller and Hester (1999) examined the change in overall effect size and variance in the primary and moderator meta-analyses conducted with the UW and HS methods in their study using real data. In this direction, they handled seven different meta-analysis studies that were carried out earlier. Bonnett (2008), for correlation (r), compared Hedges and Olkin's, Shadish and Haddock's weighting methods for the fixed-effects model, the HV and HS weighting methods for the random-effects model, and a new fixed-effects model under different simulation conditions. In addition, they handled weighting and unweighting conditions comparatively well. Marin-Martinez and Sanchez-Meca (2009) compared the results of three methods: the HS weighting method for the random-effects model and the HV weighting method for both the fixed effect model and the random effect model, using simulation data under different conditions. Schmidt et al. (2009) also compared the HS and HV methods for the random-effects model. Brannick et al. (2011) compared the UW, HS, and HV procedures with the Monte-Carlo simulation study for effect sizes r and d.

When the studies above were examined, it was seen that different weighting methods were substantially compared based on the simulation data. It can be stated that comparisons based on real data are rare in the literature. In addition, it has been observed that studies comparing weighting methods in the literature are generally in the fields of psychology and management. It was also determined that the only research in the field of education was a simulation study. In addition to these studies, this study aimed to compare these three conditions with each other based on real data in the Turkish sample by both unweighting and weighting based on HV and HS methods in the field of education. The reason why HV (DerSimonian Laird) and HS methods were considered in the study is that weightings with these methods were frequently used in the literature, as stated before. In addition, it is necessary to shed light on the choice of the weighting method for researchers in educational sciences. Besides, examining the effects of weighting methods in educational sciences will improve the meta-analysis methodologically.

## 2. Methodology

### 2.1.Research Model

In line with the purpose of the study, it was examined how unweighting and weighting (UW) based on HS and HV methods affect the overall effect size in the meta-analysis. This research is a fundamental or basic study as it is concerned with the formulation of a meta-analysis theory (Kothari, 2004). Additionally, this research is a meta-analysis study because an overall effect size regarding the effect of alternative measurement and assessment techniques and methods in science education on science attitude is calculated by meta-analysis.

### 2.2. Data Collection Process

Within the research purpose, the effect of alternative measurement and assessment techniques and methods in science education on science attitudes has been handled as a meta-analysis subject. A search was done in the databases of the HEI National Thesis Center, Web of Science, ERIC, EBSCO, Google Scholar, and DergiPark between 2010 and 2021 to find primary research on the determined subject. These databases were searched between August 2021 and March 2022. Table 1 presents the keywords used for the search and the number of studies accessed.

**Table 1.** *Searched Databases, Keywords, and Number of Research Accessed*

| Databases | Keywords | Number of Research |
|---|---|---|
| HEI National Thesis Center | alternatif ölçme ve değerlendirme OR tamamlayıcı ölçme ve değerlendirme AND tutum AND fen AND deneysel | 102 |
| Web of Science | "alternative assessment" OR "authentic assessment" AND attitude AND science AND experimental AND Turkey | 507 |
| ERIC | "alternative assessment" "authentic assessment" attitude science experimental Turkey | 134 |
| EBSCO | "alternatif ölçme ve değerlendirme" OR "tamamlayıcı ölçme ve değerlendirme" AND "tutum" AND "fen" AND "deneysel" | 362 |
| Google Scholar | "alternatif ölçme ve değerlendirme" OR "tamamlayıcı ölçme ve değerlendirme" AND "tutum" AND "fen" AND "deneysel" | 340 |
| DergiPark | alternatif ölçme ve değerlendirme OR tamamlayıcı ölçme ve değerlendirme AND tutum AND fen AND deneysel | 255 |
| Total | | 1700 |

Table 1 shows that a total of 1700 studies were accessed. These studies were examined through their full texts, and the studies that met the inclusion criteria were included in the meta-analysis. The criteria for inclusion in the meta-analysis are listed as follows:

i) The primary study was published between 2010 and 2021,
ii) The primary study was done with Turkish students,
iii) At least one of the alternative measurement and assessment techniques and methods was used in the Science and Technology course.
iv) The primary study was designed with an experimental design (true, quasi, and poor),
v) The study group for the primary study was one of the 5th, 6th, 7th, and 8th grade levels,
vi) Science attitude was the dependent variable in the primary study,
vii) The studies reported the statistics required to calculate the Cohen d effect size, and the studies reported sample sizes.

Although the selection of primary studies was made according to the criteria above, the type of publication was not considered a criterion for primary studies. In this direction, the studies that make up the gray literature, such as papers, theses, and reports found as a result of the search, were included in the meta-analysis if they met the criteria. Studies that did not meet the criteria were excluded from the meta-analysis. Figure 1 illustrates the PRISMA flowchart for scanning and inclusion of primary studies (Liberati et al., 2009).

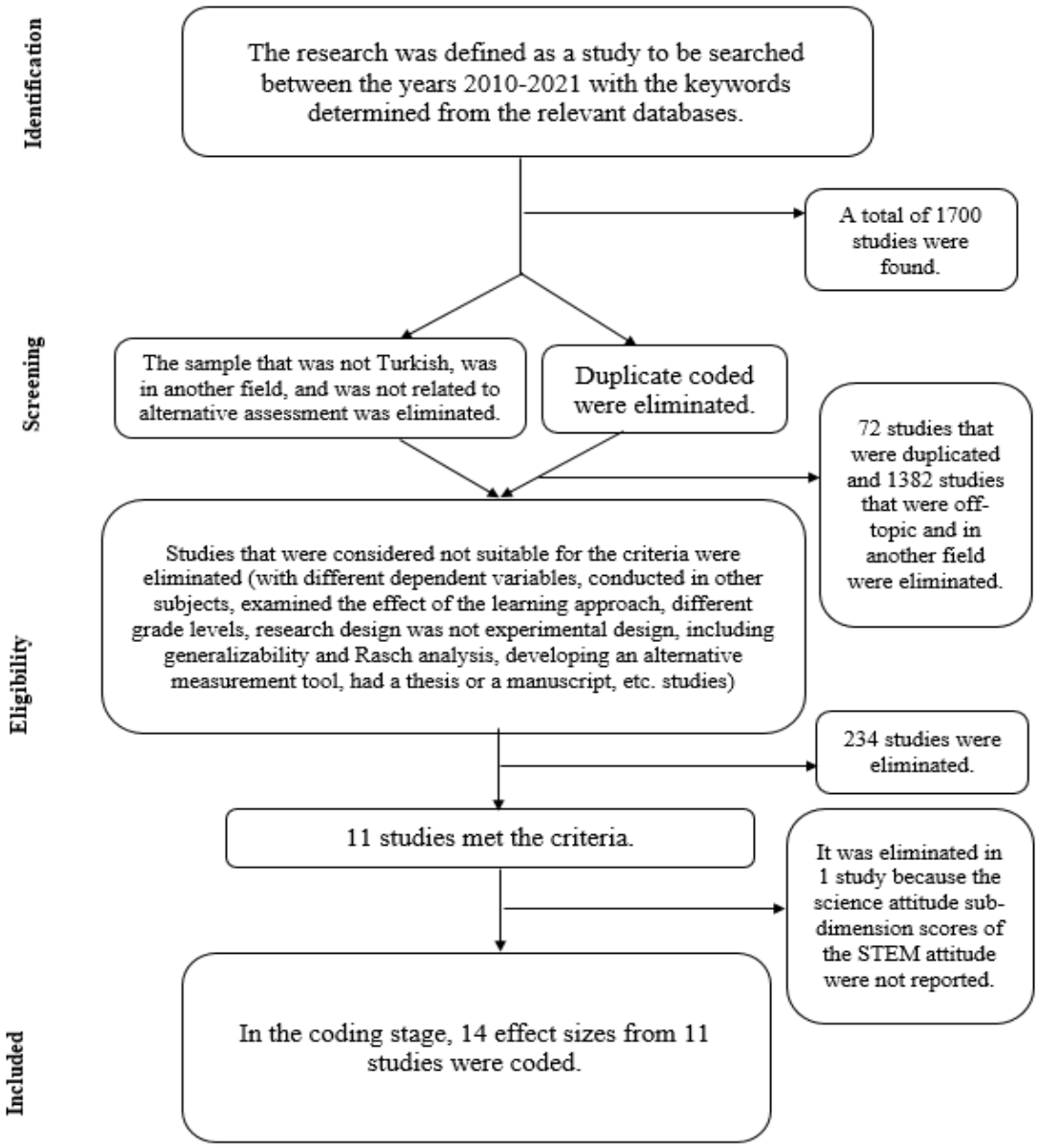



**Figure 1.** *PRISMA Flowchart*

As Figure 1 illustrates, primary studies were included in the meta-analysis by scanning them between 2010 and 2021. Since one of the inclusion criteria was that primary studies were supposed to be conducted in the last 10 years, a search was done between 2010 and 2021, and a total of 1700 studies were accessed when the databases were searched for the specified keywords. Of the 1700 studies, 1381 were excluded because they were conducted in another field (medicine, engineering, science, etc.) or because they did not include alternative assessments. Seventy-two of the studies were eliminated because they were duplicates. Of the remaining 247 studies, 234 were eliminated because they did not meet the meta-analysis inclusion criteria. The reasons for excluding these studies are as follows: the dependent variable was different; they were in a different field other than the science and technology course; the research was a qualitative, descriptive, or meta-analysis study; the grade level was different, etc. The score of science attitude was not reported in one of the remaining 13 studies, and therefore it was not included in the primary studies. In another study, since there was no information about which teaching method was used in the control group, we tried to make contact with the author via e-mail, but the author did not respond, and the study was excluded from the meta-analysis. It was determined that the remaining 11 studies met the inclusion criteria, and a total of 14 effect sizes were coded from these studies. Two researchers coded studies to examine intercoder reliability, and the percentage of agreement between coders was calculated with the formula of Miles and Huberman (1994, p. 64), reliability = number of agreements/(total number of agreements + disagreements).

The percentage of agreement was found to be 95.83%. When the reason for the inconsistency was examined, it was seen that the sample size of the experimental group belonging to the relevant effect size in the study was caused by incorrect coding. The reason for the wrong coding may originate from another experimental group in the related research. The researchers checked this sample size, and the correct sample size was determined, and the discrepancy was resolved.

In the coding of the primary studies, descriptive variables and the statistics given to calculate the effect size were used (See Appendix 1). Table 2 shows the descriptive features of 11 studies (14 effect sizes) included in the meta-analysis study.

**Table 2.** *Descriptive Features of Effect Sizes Obtained from Primary Studies*

| | Effect Sizes (f) | | | |
|---|---|---|---|---|
| Publication Language | Turkish<br>12 (%85.714) | | English<br>2 (%14.286) | |
| Publication Type | Manuscript<br>6 (%42.857) | | Thesis<br>8 (%57.143) | |
| Publication Year | 2010-2013<br>7 (%50) | 2014-2017<br>4 (%28.571) | 2018-2021<br>3 (%21.429) | |
| Research Design | True Experimental Design<br>1 (%7.143) | Quasi-Experimental Design<br>12 (%85.714) | Poor Experimental Design<br>1 (%7.143) | |
| Grade Level | 5th grade<br>0 (%0) | 6th grade<br>6 (%42.857) | 7th grade<br>7 (%50) | 8th grade<br>1 (%7.143) |

In primary studies, it was observed that alternative assessment was used in addition to the existing program in the experimental group and the existing program in the control group since the purpose of the researchers was to examine the impact of alternative assessment techniques and methods. However, a problem-based learning approach was used in one of the primary studies. Alternative measurement tools used in primary studies were as follows: performance task, self-assessment, peer evaluation, group evaluation, analytical rubric, concept map, meaning analysis, diagnostic branched tree, structured grid, puzzle, poster, diary, computer-assisted mind map, web-designed alternative measurement tools, drama, and predict-observe-explain. The learning areas and topics covered in primary studies are Light, Light and Sound, Matter and Heat, Systems in Our Body, Living Beings and Energy, and Electricity in Our Lives Attitude scales used in primary studies differed from one another, and eight different attitude scales were used.

### 2.3. Data Analysis

*The Cohend* effect size coefficient was calculated by using the mean and standard deviation of the post-test of the experimental group and the mean and standard deviation of the post-test of the control group to estimate the effect sizes of the primary studies based on the pretest-posttest control group design. In the primary study based on a single-group design, the Cohen *d* effect size was calculated by using both the mean and standard deviation of the post-test and the mean and standard deviation of the pre-test of the experimental group. Heterogeneity and publication bias were examined before conducting the meta-analysis. Within the context of heterogeneity, $Q$ statistics and the significance of $\tau^2$, $I^2$, $H^2$, and $R^2$ statistics were examined. Within the context of publication bias, the funnel plot, Rosenthal's fail-safe $N$ method, Kendall's tau, and Egger's regression tests were checked over. The results of the bias methods were obtained with the Jamovi 1.2.27.0 Major module. The Cohen *d* effect sizes of the primary studies were handled with the HV method based on a random effect model and the HS and UW methods based on a fixed effect model to get the overall effect size within the scope of the study. In the analyses based on these methods, calculations were made using the formulas given below. In obtaining the forest plots, Jamovi outputs were arranged based on these calculations.

**2.3.1. HS Weighting Method** Hunter and Schmidt (1990) suggested using sample sizes to weight effect sizes. This method is generally used when the correlation is used as the effect size coefficient. However, it can also be used for standardized mean differences. In this method, the weight is just the sample size, and Equation (1) is used to calculate the overall effect size.

$$\bar{d} = \frac{\sum[N_i d_i]}{\sum N_i} \quad (1)$$

In the Equation, $\bar{d}$, $d_i$ and $N_i$ represent the overall effect size, the effect sizes of primary studies, and the total sample size, respectively. Equations (2) and (3) are used to calculate the variance and standard error of the overall effect size (Schmidt et al., 2009).

$$s_{\bar{d}}^2 = \frac{\sum[N_i(d_i-\bar{d})^2]}{\sum N_i} \quad (2)$$

$$S_{e_{\bar{d}}} = \sqrt{\frac{s_{\bar{d}}^2}{k}} \quad (3)$$

**2.3.2. HV Weighting Method** Hedges and Vevea (1998) suggested that effect sizes should be weighted with the inverse of the sample error variance. For the random-effects model, the weight is expressed as $w_i^*$ and is the inverse of the sample error variance, as seen in Equation (4).

$$w_i^* = \frac{1}{v_i^*} \quad (4)$$

In the Equation, $v_i^*$ represents the sample error variance for the random-effects model. In the fixed-effects model, $v_i$ and $w_i$ are used for sample error variance and the weight of primary studies, respectively. Equation (5) is used to calculate the sample error variance, $v_i^*$, in Equation (4).

$$v_i^* = v_i + \tau^2 \quad (5)$$

In this Equation, $\tau^2$ represents the between-study variance. The calculation of $v_i$ (within-study variance) and $\tau^2$ (between-study variance) according to the DerSimonian and Laird method is given in Equations (6) and (7), respectively (Borenstein et al., 2009; Hedges & Olkin, 1985).

$$v_i = \frac{n^E+n^C}{n^E n^C} + \frac{d^2}{2(n^E+n^C)} \quad (6)$$

$$\tau^2 = \frac{Q-df}{\sum W_i - \frac{\sum W_i^2}{\sum W_i}} \quad (7)$$

In Equation (6), $n^E$ represents the sample size of the experimental group, and $n^C$ represents the sample size of the control group. Besides, In Equation (7), *df* represents the degree of freedom, that is, the number of primary studies minus one. $Q$ is a heterogeneity statistic and is the weighted sum of squares as given in Equation (8).

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i)^2}{\sum_{i=1}^{k} W_i} \quad (8)$$

**2.3.3. Unweighted (Unit Weighted) Condition** The UW method obtains an overall effect size by directly calculating the arithmetic mean of the effect sizes of primary studies without multiplying the effect sizes by any value. Accordingly, it can be said that each effect size is weighted by one unit, or the weighting value is 1. Based on this explanation, the formula used to obtain the overall effect size is given in Equation (9).

$$\bar{d} = \frac{\sum_{i=1}^{k} d_i}{K} \quad (9)$$

In Equation (9), $\bar{d}, d_i$ and $K$ represent the overall effect size, the effect sizes of primary studies, and the number of primary studies in the meta-analysis, respectively. In the UW method, the observed variances obtained from primary studies are weighted by one unit, summed, and divided by the square of the number of studies to calculate the overall effect size. The calculation of the mean-variance in the UW method is presented in Equations (10) and (11) (Bonnett, 2008; Osburn & Callender, 1992).

$$V_e = \frac{\sum V_{e_i}}{K} \quad (10)$$

$$V_{\hat{d}} = {V_e}/{K} \quad (11)$$

In Equation 10, $V_e$ and $K$ represent the mean of the sample error variance and the number of studies, respectively. $V_{e_i}$ represents the sample error variance of primary studies and is calculated as in Equation 6. In Equation 11, $V_{\hat{d}}$ is the sample error variance of the overall effect size is obtained by dividing the mean sample error variance by the number of studies. According to Osburn and Callender (1992), the reason for calculating the sample error variance of the overall effect size in this way is that the UW method gives more precise estimation results for the sample error variance.

In the data analysis, the correlation between the weighted effect sizes of the primary studies was calculated according to all three methods. The Spearman-Brown correlation coefficient was used to calculate the relationship between the weighted effect sizes obtained from the primary studies.

### 2.4. Ethical

The research is within the scope of the articles that do not require ethics committee approval as it is a meta-analysis study.

## 3. Findings

In accordance with the research problem, the overall effect sizes of the HV, HS, and UW, the standard error of the overall effect sizes, and the effect of the overall effect size on the significance level were examined. Data heterogeneity and publication bias were also examined before meta-analysis was performed according to different weighting methods. When Q and its significance were scrutinized within the context of heterogeneity, it was seen that Q = 81.564 ($p<.05$) and the significance of Q is an indicator of heterogeneity. Moreover, it can be said that heterogeneity is high because $I^2$ is 84.062%, and this value is higher than 75%. Since the value of the between-study variance ($\tau^2$) is 0.449 and also higher than 0, it is evidence of the existence of variability between studies and indicates heterogeneity. Finally, when $H^2$ and $R^2$ values were examined in terms of heterogeneity, they were found to be 6.274 and 6.356, respectively. The fact that these values are higher than 1 is an indicator of heterogeneity (Higgins & Thompson, 2002). In addition to heterogeneity, publication bias was also examined using the funnel plot, Rosenthal's fail-safe N, Kendall's tau, and Egger's regression test results. The funnel plot is given in Figure 2.

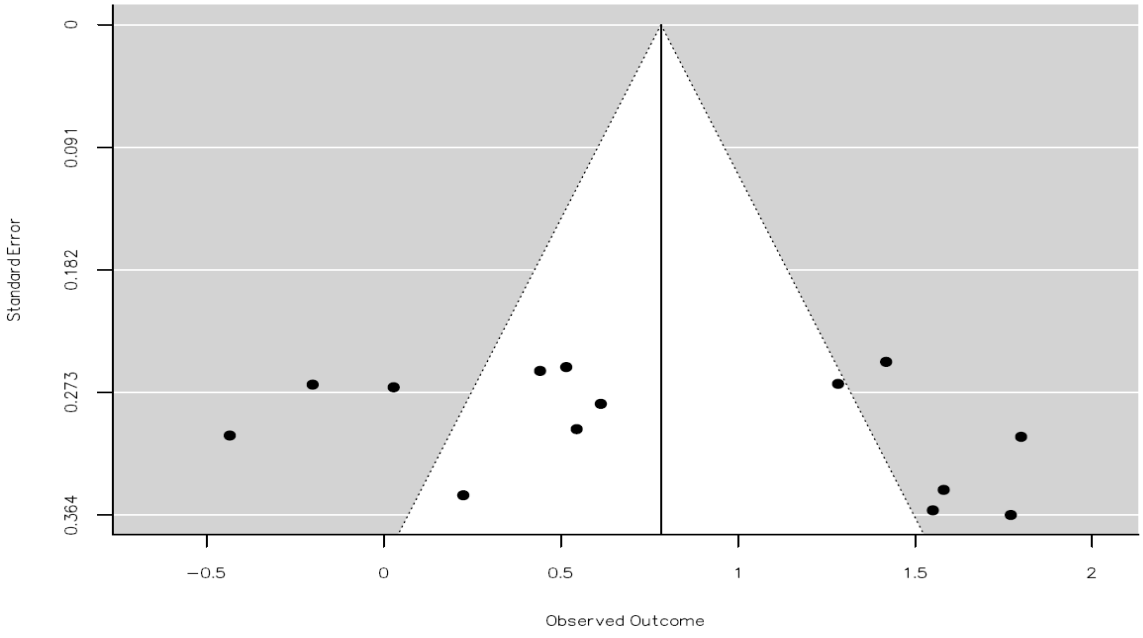



**Figure 2.** *Funnel Plot*

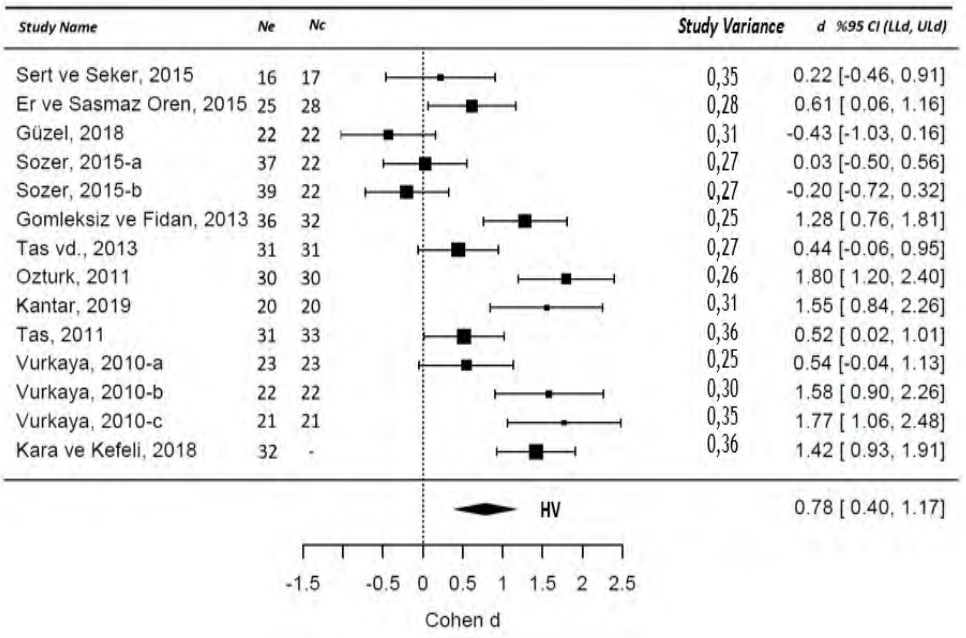
When the funnel plot given in Figure 2 was perused, it was seen that all studies had a high standard error, and therefore the points were gathered at the bottom of the plot. When the funnel plot was analyzed within the context of publication bias, it was seen that the points were almost distributed symmetrically around the reference line of the overall effect size. However, since interpreting the symmetry of the funnel plot is a subjective approach, Rosenthal's fail-safe N, Kendall's tau, and Egger's regression test results were also examined. According to Rosenthal's fail-safe N method, the number of studies that should be added to make the effect non-significant was found to be 476. There was no publication bias according to this method since the number of studies was much higher than 5k+10 ($k$ = number of primary studies included in the meta-analysis) (Rosenthal, 1979). In addition, Kendall's tau and Egger's regression intercept values were found to be 0.143 ($p$=0.518) and 1.389 ($p$=0.165), respectively. The fact that these statistics were not significant was also an indication of no publication bias.

The meta-analysis results of HV, HS, and UW methods are given in Table 3 for the situation where heterogeneity existed and publication bias did not exist;

**Table 3.** *Meta-analysis Results on Weighting Method*

| Weighting Method | $d$ | $V_d$ | Calculating $SE_d$ | $SE_d$ | $LL_d$ | $UL_d$ | $Z$ | $p$ |
|---|---|---|---|---|---|---|---|---|
| HV | 0.783 | 0.038 | $\sqrt{v_i^*}$ | 0.196 | 0.398 | 1.167 | 3.992 | 0.000 |
| HS | 0.766 | 0.520 | $S_{e_{\bar{d}}}$ | 0.193 | 0.388 | 1.143 | 3.973 | 0.000 |
| UW | 0.795 | 0.006 | $\sqrt{V}_{\hat{d}}$ | 0.022 | 0.753 | 0.837 | 36.988 | 0.000 |

Table 3 shows that the overall effect size did not vary much in the weighting methods in this study, where heterogeneity was high, and there was no publication bias. The largest overall effect size was found in the UW method with 0.795, while the lowest effect size was found in the HS method with 0.766. The largest variation among overall effect sizes is approximately 3%. Although the overall effect size did not vary much in the weighting methods, it was observed that the mean effect was significant ($p$<.05) in all methods. When the standard errors of the overall effect size were examined, it was seen that the lowest was in the UW method and was found to be 0.022. Another salient situation when the standard errors were examined was that the standard errors were similar for the HV and HS methods. The standard errors for these methods were 0.196 and 0.193, respectively. In this respect, it could be said that the standard error of the overall effect size in the HV and HS methods was similar. In parallel with the standard error, the narrowest confidence interval was obtained in the UW method, and the true value of the overall effect size was between 0.398 and 1.167. Similarly, in parallel with the standard error, the confidence intervals in the HV and HS methods were very close to each other and were in the range of 0.398-1.167 and 0.388-1.143, respectively. The forest plot for the HV, HS, and UW methods is given in Figure 3.



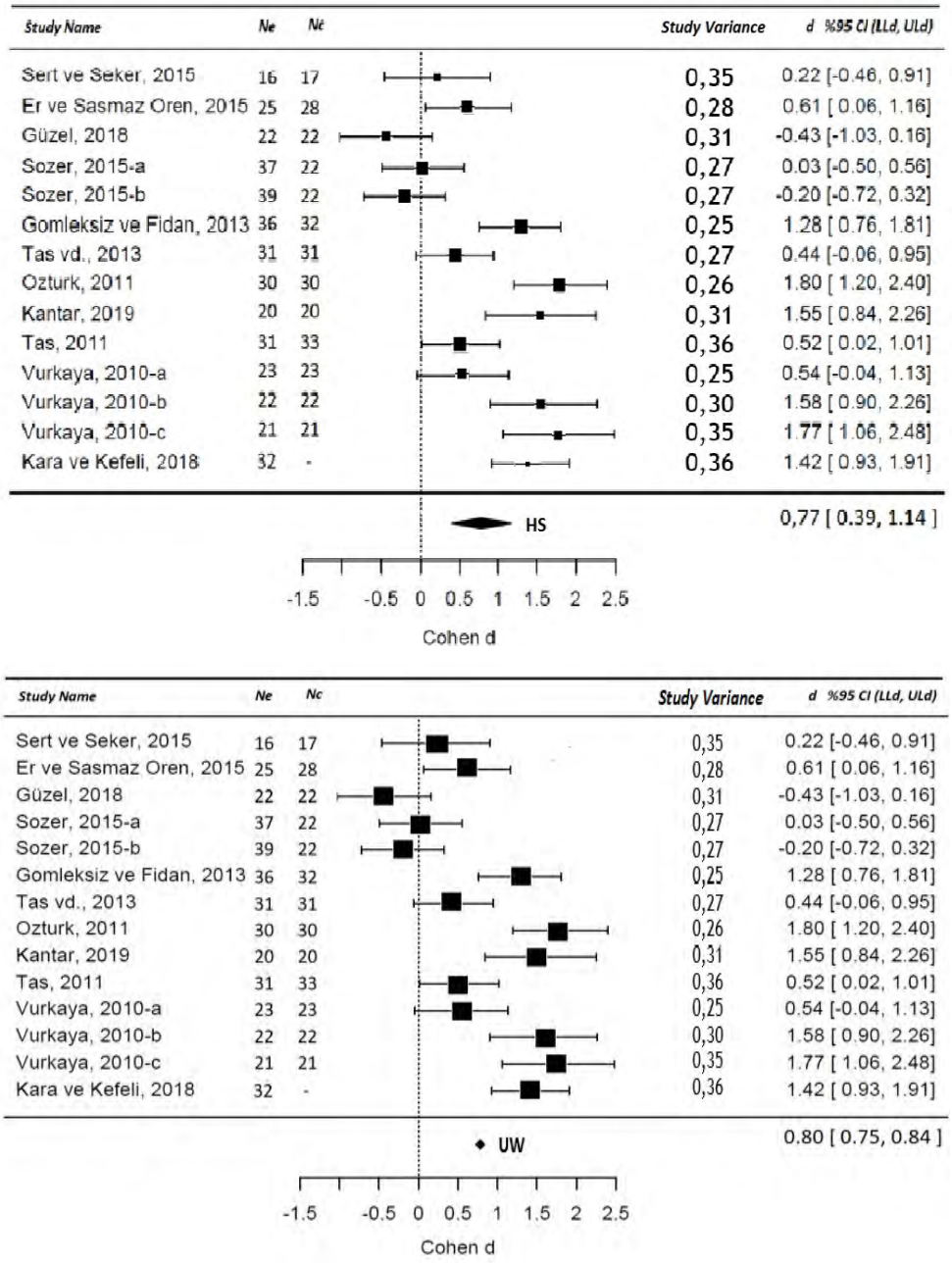



**Figure 3.** *Forest Plot of HV, HS, and UW Methods*

When the sizes of the squares of the primary studies in the forest plots in Figure 3 were examined, it was seen that the HV and HS methods were similar to one another. With these methods, it was seen that the study with different weights belonged to Kara and Kefeli (2018). Unlike these methods, since unit weighting is applied in UW, the sizes of the squares of the weights of the primary studies are equal to each other. It was observed that the lowest weight in the HV method was in the study of Vurkaya (2010-c), and in the HS method, it was in the study of Kara and Kefeli (2018). However, while the highest weight in the HV method was observed in the study of Kara and Kefeli (2018), the highest weight in the HS method was observed in the study of Gömleksiz and Fidan (2013). When the effect sizes and confidence intervals of the primary studies were examined, it was found that the effect sizes ranged from -0.43 to 1.80, while the lower bound for the confidence interval was -1.03, and the upper bound was 2.48. As the effect sizes of primary studies and the standard error (confidence interval) of the effect sizes were calculated independently of the weighting methods, they were the same in all three methods. In addition, among the primary studies, the study with the narrowest confidence interval belongs to Kara and Kefeli (2018), while the study with the widest confidence interval belongs to Vurkaya (2010-c). When the overall effect sizes were examined, it was seen that the diamonds were very similar to each other in terms of both location and size in the HV and HS methods. On the other hand, it could be stated that

the diamond in the UW method was narrower than in the HV and HS methods, and its location does not change significantly because the standard error of the overall effect was lower in the UW method than in the other methods, and therefore the confidence interval was narrower. The correlation between the weighted effect sizes of the primary studies obtained from all three methods was calculated. The calculated Spearman-Brown correlation coefficients and the significance of these correlation coefficients are given in Table 4.

**Table 4.** *Correlation Coefficients for Weighting Methods*

| Weighting Methods | HV | HS | UW |
|---|---|---|---|
| HV | 1 | | |
| HS | 0.930* | 1 | |
| UW | 0.996* | 0.934* | 1 |

*$p$<0.01 (2-tailed)

When Table 4 was examined, it was seen that the highest correlation (0.996) was between HV and UW methods, and the lowest correlation (0.930) was between HS and HV methods. Besides, the correlation between HS and UW methods was 0.934. It was observed that p<0.01 for all correlation coefficients obtained in the study. In this respect, it could be said that the relationships between all methods were significant. In addition, the fact that the correlation coefficients were approximately 0.90 indicates a high level of correlation between the methods. Especially the relationship between UW and HV methods was close to 1, which is an indication of an almost perfect relationship.

## 4. Conclusion, Discussion and Recommendations

In this study, the effects of different weighting methods on the meta-analysis results were examined. First of all, this study aimed to explore the impact of weighting and unweighting. Although there are many weighting methods in the literature, HV and HS methods, which are the most preferred weighting methods, were discussed in the current study. Based on the results of the study, it was seen that the overall effect size did not vary much in the weighting methods in situations where the heterogeneity was high, there were few primary studies, and there was no publication bias. The largest overall effect size was found in the UW method, and the lowest overall effect size was found in the HS method. When the literature was examined, Fuller and Hester (1999) also found similar overall effect sizes (correlation coefficients) in the HS and UW methods, which is similar to the findings obtained in this study.

As a result of the research, it was seen that the UW method had the lowest standard error. Osburn and Callender (1992) stated that in cases where the effect size values of primary studies were similar or equal and the sample size was highly variable, the standard error of the unweighted mean would be larger than the standard error of the weighted mean. They also stated that as the effect sizes of primary studies became more variable, the standard error of the unweighted mean tended to be lower than the standard error of the weighted mean. In this study, it was observed that the effect sizes of the primary studies differed from each other; however, the sample sizes did not vary much. Therefore, it is expected that the method with the lowest standard error would be the UW method. There are some findings in the literature that support the conclusion of this study, indicating that unweighting has a lower error than weighting (Bonett, 2008, 2009, 2010; Brannick, et al., 2019; Shuster, 2010). In meta-analysis research examining the effect of time with real data in the field of health, Shuster (2010) found that unweighting at some time points produced a lower standard error than the inverse variance weighting method (with the DerSimonian-Laird estimation) and the sample size weighting method. In the study of Fuller and Hester (1999), considering the confidence interval in primary and moderator meta-analyses, which is close to the number of primary studies in this study (10-14 studies), it was seen that the HS method in some studies and the UW method in some studies estimated with a lower error. In addition, researchers have always stated that when outliers are also included in the meta-analysis, there are no wider confidence intervals in the UW method than in the HS method. Hunter and Schmidt (1987) also discussed whether the weighted mean was better than the unweighted arithmetic mean. As a result, they pointed out that there were rare cases where unweighting would be better (Hunter & Schmidt, 2004). Hedges and Olkin (1985) also stated that in estimating the overall effect size, the weighting was less biased than the unweighting because primary studies with larger sample sizes had more weight. However, in this study, the sample sizes of the primary studies were quite close to each other. Therefore, this study exemplifies one of the rare cases where the UW method has a lower standard error than weighting. Finally, when the simulation

study conducted by Brannick et al. (2011), where the conditions were similar to the characteristics of our research, was examined, the highest estimation error was obtained in the UW method, unlike the results of our study.

As a result of the research, it was observed that the standard error was very similar in the HV and HS methods, but the error was slightly lower in the HS method. The fact that the overall effect size and standard error values were similar between the HV and HS methods might be due to the fact that both methods are based on sample size. While the HS method weights directly by using the total sample sizes, the HV method calculates the sample error variance using the sample sizes and weights the inverse variance. However, as the sample size of primary studies increases, the weight of the effect size also increases in both methods. Therefore, it is expected to give similar overall effect sizes and standard error values. In addition, the correlation between the weighted effect sizes obtained from the two methods was also found to be high. Bonett (2008) found in his research that the bounds of the confidence interval in the HS and HV methods were very close to each other, which is similar to the results of our study, under the conditions of k=10, r=0.60-0.90, N=20-100. In addition, as in our study, the bounds of the confidence interval in the HS method were slightly lower than the ones in the HV method. When the results of the research done by Marin-Martinez and Sanchez-Meca (2010), which had similar conditions to our study ($k$=10 and $k$=20, $\tau^2$=0.32, $N$=30), were examined, the error values in HS and HV methods were similar; however, unlike our study, the error of HV was lower. When the research done by Brannick et al. (2011), which also has similar conditions to our study, was examined, it was seen that the RMSR and AAE values of the HV and HS methods yielded very close results. In other words, according to the research results, the estimation errors of the HV and HS methods are similar to each other.

In this study, the relationship between weighting methods was also examined. As a result of the research, it was seen that the binary correlation coefficients between all methods were very high. The correlation coefficient between UW and HV was especially close to 1. In addition, the correlation coefficient between UW and HS was quite high. The reason why both correlation coefficients between UW and HV and between UW and HS methods were so high might stem from the fact that sample sizes in primary studies were very similar. Englund et al. (1999) examined the correlation between weighting and unweighting in a meta-analysis of 14 studies and found the correlation coefficient to be above 0.90, similar to the results of our research. In addition, the correlation coefficient between the HV and HS methods was also very high. It can also be said that this result is because both methods make an estimation based on sampling.

In conclusion, in this study, which was conducted with real data in education, it can be said that the unweighting had a lower standard error, and this method showed a high correlation with the weighting methods.

In this meta-analysis study conducted in the field of education, where heterogeneity is high, sample sizes of primary studies are close to each other, and there is no publication bias, unweighting has calculated a lower standard error than weighting methods. However, this does not indicate that unweighting or unit weighting is the most usable method. Because these findings are valid for the conditions of this study. Furthermore, weighting methods have strong statistical theories behind them. Accordingly, it is not appropriate to generalize under limited conditions where real data are used. In this study, the data for 14 effect sizes were discussed within the scope of a determined subject. Weighting methods can be compared in a meta-analysis study based on real data with a larger sample size. In order to generalize, it may be suggested to design post-hoc simulation studies based on real data in which different conditions (sample size, heterogeneity, bias, number of studies, estimation methods, etc.) may be effective in the results. In addition, the weighting methods used in this research or different weighting methods can be compared with each other in different studies to be conducted with real data.

## 5. References

References marked with an asterisk indicate studies included in the meta-analyses.

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. *Educational Researcher*, *31*(8), 18-20. https://doi.org/10.3102/0013189X031008018

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, 13, 173-181. https://doi.org/10.1037/a0012868

Bonett, D. G. (2009). Meta-analytic confidence intervals for standardized and unstandardized mean differences. *Psychological Methods*, 14, 225–238. https://doi.org/10.1037/a0016619

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, 15, 368-385. https://doi.org/10.1037/a0020142

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Brannick, M. T., Yang, L. Q., & Cafri, G. (2011). Comparison of weights for meta-analysis of r and d under realistic conditions. *Organizational Research Methods*, *14*(4), 587-607. https://doi.org/10.1177/1094428110368725

Brannick, M. T., Potter, S. M., Benitez, B., & Morris, S. B. (2019). Bias and precision of alternate estimators in meta-analysis: Benefits of blending Schmidt-Hunter and Hedges approaches. *Organizational Research Methods*, *22*(2), 490-514. https://doi.org/10.1177/1094428117741966

Englund, G., Sarnelle, O., & Cooper, S. D. (1999). The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology*, *80*(4), 1132-1141. https://doi.org/10.1890/0012-9658(1999)080[1132:TIODSC]2.0.CO;2

*Er, Ö., & Sasmaz-Oren, F. (2015). The effect of the alternative assessment approaches based education during the "light" unit of science and technology class in 7th grade on the academic achievements and attitudes of students. *Manisa Celal Bayar University Journal of Social Sciences*, *13*(4), 135-164. https://doi.org/10.18026/cbusos.17987

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*(4), 444–467. https://doi.org/10.1037/1082-989X.10.4.444

Fuller, J. B., & Hester, K. (1999). Comparing the sample-weighted and unweighted meta-analysis: An applied perspective. *Journal of Management*, *25*(6), 803-828. https://doi.org/10.1177/014920639902500602

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8. https://doi.org/10.2307/1174772

*Gömleksiz, M. N., & Fidan, E. K. (2013). The effect of computer assisted mind mapping on students' academic achievement, attitudes and retention in science and technology course. *Gaziantep University Journal of Social Sciences*, *12*(3), 403-426. Retrieved from https://dergipark.org.tr/tr/pub/jss/issue/24232/256876

*Güzel, Z. (2018). *The effects of problem based approach practiced through self and peer assessment on students' achievements and attitudes in science teaching (publication No. 506216)*. [Master's dissertation, Necmettin Erbakan University]. YOK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a metaanalysis. *Statistics in Medicine*, *21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Sage.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings (2. Ed.). Sage.

*Kantar, N. (2019). *The effect of alternative measurement and evaluation activities on 6th grade students' achievement in and attitudes towards science (publication No. 579261)*. [Master's dissertation, Ataturk University]. YOK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

*Kara, F., & Kefeli, N. (2018). The effect of using concept maps on student's success, logical thinking and attitudes towards science. *Necatibey Faculty of Education Electronic Journal of Science & Mathematics Education, 12*(2), 594-619. https://doi.org/10.17522/balikesirnef.506475

Kothari, C., R. (2004). *Research methodology methods & techniques*. New Age International Limited Publishers.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), 1-34. https://doi.org/10.1136/bmj.b2700

Manolov, R., Guilera, G., & Sierra, V. (2014). Weighting strategies in the meta-analysis of single-case studies. *Behavior Research Methods*, *46*(4), 1152-1166. https://doi.org/10.3758/s13428-013-0440-0

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*(1), 56-73. https://doi.org/10.1177/0013164409344534

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2. Ed.). Sage.

Osburn, H. G., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology, 77*(2), 115-122. https://doi.org/10.1037/0021-9010.77.2.115

*Öztürk, P. T. (2011). The effect of usage of concept maps, structured grid and diagnostic tree technics to teach the "living things and energy relations unit" on 8th grade of primary school students' attitudes towards science and technology lesson *(publication No. 294162)*. [Master's dissertation, Selcuk University]. YOK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. https://doi.org/10.1037/0033-2909.86.3.638

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97-128. https://doi.org/10.1348/000711007X255327

*Şeker, F., & Sert, H. (2015). The effect of complementary measurement and assessment approach on the attitude and success in science and technology course. *Mediterranean Journal of Humanities, 5*(2), 351-363. https://doi.org/10.13114/MJH.2015214577

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*, *29*(12), 1259-1265. https://doi.org/10.1002/sim.3607

*Sözer, E. (2015). *Influence of the multiple choice performance tasks on students' academic achievement, self-confidence and attitudes of the students towards course (publication No. 429373)*. [Master's dissertation, Gazi University]. YOK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

*Taş, E. (2011). A new web designed material approach on learning and assessment in science education. *Energy Education Science and Technology Part B: Social and Educational Studies, 3*(4), 567-578. Retrieved

*Taş, E., Çetinkaya, M., Karakaya, Ç., & Apaydın, Z. (2013). An ınvestigation on web designed alternative measurement and assessment approach. *Education and Science, 38*(167), 196-210. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1715/476

van den Noortgate, W., & Onghena, P. (2003). Estimating the mean effect size in meta-analysis: Bias, precision, and mean squared error of different weighting methods. *Behavior Research Methods, Instruments, & Computers, 35*(4), 504-511. https://doi.org/10.3758/BF03195529

*Vurkaya, G. (2010). *The effect of using alternative assessment activities on students? success and attitudes in science and technology course (publication No. 259467)*. [Master's dissertation, Kocaeli University]. YOK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

**Appendix 1.** *Coded Variables and Their Examples or Categories*

| | | Examples or Categories |
|---|---|---|
| Descriptive Variables | Publication Code | 101, 102, 103, 104-1, 104-2… |
| | Identity (Author Surnames, Year) | Surname, Year; Surname et al., Year |
| | Name of The Study | The effect of … etc. |
| | Publication Type | Manuscript/Thesis |
| | Publication Language | Turkish/English |
| | Publication Year | 2010, 2011, … |
| | Publication Place | Journal name or University (Instıtute) |
| | Volume-Number | e.g. 12(2) |
| | Author | Authors's names and surnames |
| | Database | Google Scholar/Dergipark/EBSCO/WOS/HEI |
| | Index | SSCI/SCI/TRDizin |
| | Teaching Model/Method | Alternative-Traditional etc. |
| | Approach in the Experimental Group | Alternative, Self and Peer Assessment etc. |
| | Approach in the Control Group | Traditional |
| | Measurement Tools Used in the Experimental Group and Control Group | Performance tasks, concept maps, Conductional communication grids, etc.-it did not applied etc. |
| | Subject (Sub-Topic) | Matter and Heat, Light and Sound, etc. |
| | Research Design | Quasi-experimental, Single group pre-test post-test experimental design, etc. |
| | Grade Level | 6. Grade / 7. Grade / 8. Grade |
| | The Developer/Adapter of The Attitude Scale | Taş (2006) etc. |
| | The Reliability Coefficient of The Attitude Scale | Reported / Not Reported |
| | The Number of Measurement Tools | 2, 4, 27 etc. |
| | The Application Course Hours | 16 hours, 32 hours etc. |
| | The Pilot Scheme of the Application | Yes / NA |
| | The Pilot Scheme for Alternative Measurement Tools | NA |
| | Data Analysis Method | ANCOVA, ANOVA, t-test etc. |
| | Application Time | 2008-2009 academic year etc. |
| Statistics for calculate the effect size | The Sample Size of the Control Group, | 17 – 33 |
| | The Sample Size of the Experimental Group, | 16 – 32 |
| | The Mean and Standard Deviation Scores of the Science Attitude Post-Test for the Experimental and Control Group Designs, | e.g. 4,22 (0,75) / 3,83 (0,68) |
| | The Mean and Standard Deviation Scores of the Science Attitude Pre-Test and Post-Test of the Experimental Group for the Single-Group Experimental Designs | e.g. 87,47 (10,22) / 73,72 (13,33) |