ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);30-48

Calculation of Effect Size in Single-Subject Experimental Studies: Examination of Non-Regression-Based Methods*

Nihal ŞEN** Sedat ŞEN ***

Abstract
It is observed that meta-analysis studies have not been included in single-subject experimental studies as much as the experimental studies. In order to overcome this deficiency in the literature, regression-based and non-regression-based indexes that can be used as the effect size in single subject experimental studies have been recently developed. Since most of the regression-based indexes are affected by the serial dependency of single-subject experimental data, non-regression-based indexes that were less affected by this dependency and were preferred more than regression-based indexes were the main subject of this study. Although there are many indexes that are not based on regression in single-subject experimental studies, it is observed that most of the researchers prefer the percentage of non-overlapping data (PND) and percentage of zero data (PZD). There are many controversies in the literature especially on the use of the PND index. In the absence of a study describing the alternative indexes of PND and PZD in Turkey literature, the examination of non-regression-based indexes makes this study important. The aim of this study is to examine the non-regression methods used to calculate the effect size in single-subject experimental studies and to show how these methods will be applied in single-subject experimental research. In this aim, how to prepare the data, how to analyze it, how to synthesize it for more than one study and how to interpret the results are discussed. In this study, suggestions were made for the researchers based on 10 different indexes.

Key Words: Single-subject experimental research, meta-analysis, effect size, non-regression-based indexes.

INTRODUCTION

In recent years, new trends in studies in the field of education show that more emphasis is given to the synthesis of data from studies in a specific field. Hattie (2009) stated that practitioners and policy makers should summarize and compare various types of evidence obtained through meta-analysis in order to close the gap between research and practice in education (Kavale, 2001). Meta-analysis (Glass, 1976) is a method that provides a statistical summary of the studies conducted in a specific field. Meta-analysis mainly helps to calculate the overall mean value of the subject matter using the effect size values obtained from quantitative studies on a specific subject.

The meta-analysis method, which provides many benefits for researchers, can be applied in traditional meta-analysis studies with the effect size values such as standardized mean difference, correlation and risk ratio (Lipsey & Wilson, 2001). It is of great importance for the researchers to determine the effectiveness of the intervention in experimental studies. Besides the statistical significance, calculating the effect size value, which indicates the practical importance, gives the researcher information that can be interpreted (such as small effect, large effect) about the effectiveness of the intervention phase. Indeed, the American Psychological Association strongly recommends that every published work should report an effect size value (American Psychological Association, 2010).

Although single-subject experimental research (Kırcaali-İftar & Tekin, 1997) has a long-standing history, meta-analysis practices to synthesize studies in this area have begun in the late 1980s. Most of

* This study was presented at the 27th International Congress on Educational Sciences (ICES/UEBK-2018) (April 18-22, 2018, Antalya).

** Ph.D. student, Bolu Abant Izzet Baysal University, Institute of Educational Sciences, Special Education, Bolu, Turkey, e-mail: nihalllsennn@gmail.com, ORCID ID: orcid.org/0000-0002-9511-8401
*** Asst. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa, Turkey, e-mail: sedatsen@harran.edu.tr, ORCID ID: orcid.org/0000-0001-6962-4960

To cite this article:
Şen, N., & Şen, S. (2019). Calculation of effect size in single-subject experimental studies: Examination of non-regression-based methods. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 30-48. DOI: 10.21031/epod.419625

Received: 30.04.2018
Accepted: 16.12.2018

the researchers in this area failed to develop an acceptable effect size due to the problems arising from the data structure in single-subject experimental studies. Among these problems, single-subject experimental study data included repeated measurements on the same individual, hence the dependence on residual values (Huitema, 1985), the low number of measurements and the non-normal distribution of these measurements. These problems prevented the applicability of these statistics by providing some assumptions of parametric statistics in single subject experimental studies. In order to solve these problems, many indexes have been suggested for meta-analyses that can be applied in single-subject experimental studies (Beretvas & Chung, 2008). Among others, values based on percentage of non-overlapping data (Scruggs, Mastropieri & Casto, 1987) and the standardized average difference (Busk and Serlin, 1992) of the experimental and control groups (i.e., the baseline and intervention levels in single-subject experimental studies) are the major ones. In addition, a large number of indexes were proposed in the literature, mainly based on percentage calculations (Ma, 2006; Parker & Vannest, 2009; Parker, Vannest, & Brown, 2009; Parker, Vannest, & Davis, 2011). In addition to these nonparametric methods, many statistical methods based on regression analysis have been developed (Allison & Gorman, 1993; Huitema & McKean, 2000; Swaminathan, Rogers, Horner, Sugai, & Smolkowski, 2014; van den Noortgate & Onghena, 2003).

The methods proposed in the literature are suitable for different data designs to be obtained from experimental studies with single-subjects. Especially, the methods not based on regression have been widely accepted and still used in single-subject experimental studies (Aslan, Yalçın, & Özdemir, 2016; Aydın, 2017; Karasu, 2009a, 2009b, 2011; Korkmaz & Diken, 2010; Sönmez & Diken, 2010; Tavil & Karasu, 2013). It is also observed that the PND and PZD indexes are widely used by researchers in Turkey. Especially in order to find a solution to the criticisms of PND, many alternative effect size indexes which are not based on regression have been proposed in the literature. However, it is observed that other non-regression-based alternatives are rarely being used by researchers in Turkey (Bozkus-Genç, 2017; Kaya, 2015; Uysal, 2017). One of the reasons for little use of these methods by the researchers may be due to lack of methodological studies describing these indexes in Turkey. Many researchers abroad conducted studies describing and comparing the effect size values that can be used for single-subject experimental studies (Alresheed, Holt, & Bano, 2013; Campbell, 2004; Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014; Maggin, O'Keeffe, & Johnson, 2011; Maggin, Swaminathan et al., 2011; Manolov & Solanas, 2008; Manolov, Solanas, & Leiva, 2010; Olive & Franco, 2008; Parker et al., 2011; Wolery, Busick, Reichow, & Barton, 2010). Most studies in the literature are based on the comparison of non-regression-based indexes. Parker et al. (2011) compared nine indexes, but others generally compared a few indexes. The explanation and comparison of 10 indexes used in single-subject experimental research makes this study different from other studies in terms of the number of indexes discussed in a single study. According to the extensive literature review, the number of studies comparing or explaining these indexes in Turkey is almost negligible. The only example that can be given in this sense is Karasu (2009a)'s study of the effects of a group study selected from the treatment studies based on natural approaches to improve communication and social skills of children with autism. In that study, four different methods, percentage of non-overlapping data, percentage of zero data, Swanson model and ITSACORR (interrupted time-series analysis procedure), were compared to determine the most useful methods for calculating the effect-size in single-subject experimental studies. Correlation analyses was performed to determine the relationships between the effect size values obtained from four different methods and mean values were compared among the results of these methods.

The lack of a comprehensive study describing the indexes developed in the Turkish literature after PND and PZD makes this study important in which non-regression indexes are examined. The purpose of this study is to examine the non-regression methods of the meta-analysis of single-subject experimental studies and to show how these methods will be applied in single-subject experimental studies on sample data. In this purpose, how to prepare the data, how to analyze it, how to synthesize it for more than one study and how to interpret the results will be discussed. In this study, it was attempted to answer how the results obtained from sample data which are suitable for single-subject experimental structure differ between these indexes. All data used in this study are generated. Sample analysis was performed on the same data for all methods except PZD as PZD index was used to determine the effectiveness in

diminishing a behavior. Therefore, this can be considered a comparison study. The methods examined in this study do not include all indexes proposed in the literature. The most studied indexes were determined and included in this study by examining the comparison studies in single-subject experimental research literature (Alresheed et al., 2013; Campbell, 2004; Heyvaert et al., 2014; Maggin, O'Keeffe, & Johnson, 2011; Magin, Swaminathan et al., 2011; Manolov & Solanas, 2008; Manolov et al., 2010; Olive & Franco, 2008; Parker et al., 2011; Wolery et al., 2010).

Percentage of Non-overlapping Data (PND)

The percentage of non-overlapping data (PND) proposed by Scruggs et al. (1987) is a nonparametric index based on comparison of baseline and intervention levels. It is called percentage of non-overlapping data as it is based on taking into account the non-overlapping data between the baseline and intervention levels.

The calculation of this index is made by determining the ratio of the intervention level values that exceed the maximum value of the baseline level to the number of values obtained at the intervention level. The following steps are followed when calculating the PND value for an AB design:

1. Determine the maximum baseline value in the graph,
2. Draw a horizontal line from this determined maximum value to the right (intervention level),
3. Determine the intervention level values above this horizontal line,
4. Divide the number of data points obtained in the third step by the total number of data points at the intervention level,
5. The value obtained in the fourth step is multiplied by 100 to calculate the PND value.

In the intervention level situations where the target behavior is expected to increase, the intervention level values that exceed the highest value of the baseline level are taken into account when calculating the value of the PND. In cases where the target behavior is expected to decrease, the intervention level values below the lowest level of the baseline level are determined. The number of values obtained in these two cases is divided by the total number of data points at the intervention level. If a study involves more than one intervention, the PND indexes obtained from different interventions are combined to determine the mean value and this value is used to evaluate the overall effectiveness of all interventions. The PND value can also be calculated for the ABAB design, which is frequently applied in experimental studies with single subjects. In the ABAB design, the PND indexes are calculated for both of the AB designs. The mean of these two PND percentages is then calculated to find the overall value of the PND of the ABAB design.

The values of the baseline level (A) and the intervention level (B) of an AB design obtained from an individual's dependent variable (minimum score = 0, maximum score = 20) are presented in Figure 1. Ten measurements were assumed to be collected at both levels. This generated data set was assumed to represent an expected increase in the targeted behavior. When we look at the baseline level measurements in Figure 1, the highest baseline level value is 8. When a horizontal line is drawn over this value to the right, values above 8 are determined at the intervention level. According to the data in Figure 1, it can be seen that six data points (10, 9, 11, 10, 10, and 12) at the intervention level are above this line. If the number of data points exceeding the maximum value is divided by the total number of data points at the intervention level (10), a value of 0.6 is obtained. If we multiply this value by 100, the PND value is calculated as 60 (i.e., 0.6 × 100) based on the data presented in Figure 1.

According to Scruggs and Mastropieri (1998), the PND values should be over 90 in order to say that the intervention is "very effective". The values between 71 and 90 indicate "effective" intervention, while values between 50 and 70 indicate "questionable" or "moderate" effects. It indicates that the intervention "ineffective" where the percentage value is less than 50 (Scruggs, Mastropieri, Cook, & Escobar, 1986; Strain, Kohler, & Gresham, 1998). According to these criteria, the intervention in Figure 1 can be considered moderately effective.

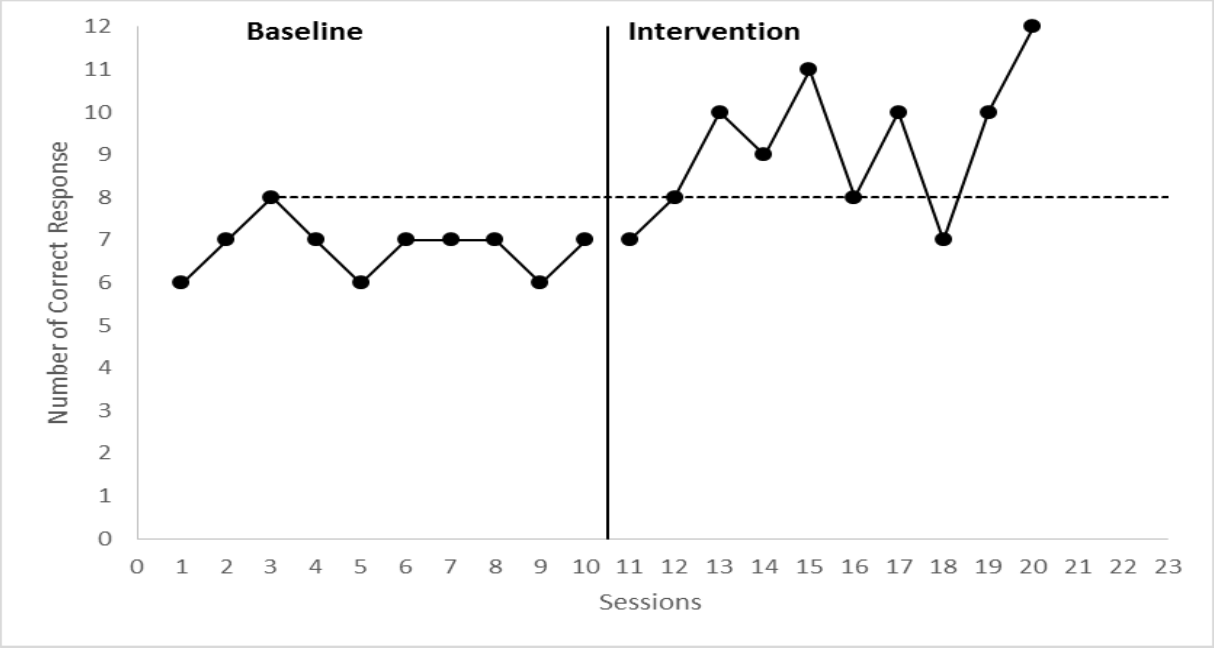

Figure 1. Calculation of Percentage of Non-overlapping Data in an AB Design.

Although PND index is not necessary to meet the assumptions of parametric statistics, it is often preferred by researchers as it is easy to compute and interpret (Ma, 2006). In addition to these advantages, the PND index has several limitations that have been criticized by researchers (Allison & Gorman, 1994; Strain et al., 1998). One of these criticisms is the failure of PND to behave as an effect size value. On the other hand, there are also explanations made by Scruggs and Mastropieri (2013) regarding the relationship between the size of the PND and the effect size. Another criticism of this value is that the percentage of non-overlapping data ignores all other values except for a single value in the baseline level. Another criticism directed at PND is that it does not take into account the trend stem from linear increase or decrease at the baseline level and does not detect the changes in slope. In the light of these criticisms, it should be noted that a reliable result cannot be obtained from the PND index alone (Scruggs & Mastropieri, 1998). This is because the PND index does not have a known sampling distribution and a *p* value cannot be calculated for this index (Parker, Hagan-Burke, & Vannest, 2007). Therefore, the fact that a *p* value cannot be calculated eliminates the possibility of making inferences based on this index. It is also criticized that the PND index value does not accurately reflect the effect of intervention level if 0 or 100 values are present at the baseline level (Ma, 2006). Scruggs and Mastropieri (1998) do not recommend calculating the PND in the case of a floor (0) or a ceiling (100) is present at among the baseline data points. Finally, as in other non-overlapping indexes, the increase in the value of the PND value as the number of intervention levels increases is seen as a limitation (Allison & Gorman, 1993).

Percentage of Zero Data (PZD)

Percentage of zero data (PZD) is an index developed by Scotti, Evans, Meyer, and Walker (1991) to show the effectiveness of the intervention level in single-subject experimental studies. PZD represents the degree of behavior suppression versus degree of behavior reduction and it is seen as a more stringent efficacy indicator (Campbell, 2004, p. 235). In other words, it is a measure that requires the targeted behavior to reach zero and remain at zero level.

The following steps are followed to obtain the PZD value in an AB design:

1. The first data point with zero at the intervention level is detected,
2. The number of intervention level data remaining at the zero point, including the zero point detected in the first step, is determined,

3. The number of zero values specified in the second step is divided by the number of data points after the first zero at the intervention level,
4. The value obtained in the third step is multiplied by 100 to obtain the PZD value.

The values of the baseline level (A) and the intervention level (B) of an AB design obtained from an individual's dependent variable (minimum score = 0, maximum score = 20) are presented in Figure 2. Ten measurements were assumed to be collected at both levels. This generated data set was assumed to represent an expected decrease in the target behavior.

When the intervention-level measurements are examined, it is the seventh data point where the participant gets first zero point. After this point, three more measurements were made on the participant and this participant scored two more zero points. Thus, the number of data points after this point, including the first zero, was four and the participant scored zero points in three of these four measurements. In order to calculate the PZD value, it is sufficient to divide the number of zero points by four. Thus, the PZD value is $3/4 \times 100 = 75$. In cases where there are multiple intervention levels (e.g., ABCD), some researchers (Reichle, 2007) have calculated the PZD value using only the last intervention level (e.g., D). In the multiple baseline designs across subjects, the PZD value is calculated for each subject separately (Schlosser & Koul, 2015). In their work on ABAB designs, Wehmeyer et al. (2006) stated that they calculated a PZD value for each pair of AB design and each calculated PZD was considered separate values.

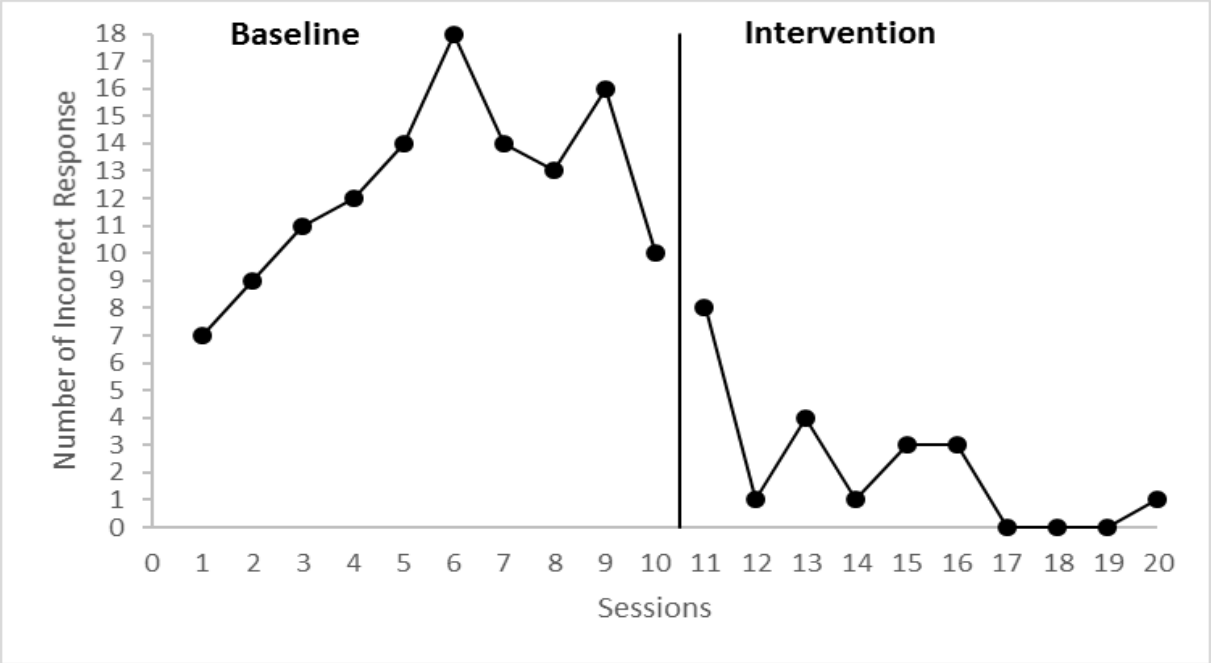

Figure 2. Calculation of Percentage of Zero Data in an AB Design.

The PZD index takes values from 0 to 100 and higher values indicate an effective intervention. The criteria suggested by Scotti et al. (1991) can be used in the interpretation of PZD. According to Scotti et al. (1991), values between 55-80% indicate "moderate" effect, while values above 80% indicate "high" effect. While PZD values below 18 percent imply "ineffective" intervention, values between 18-54% indicate "questionable" effect. According to the above criteria, the intervention in Figure 2 can be said to have "moderate" effect.

Since all data before the first zero point in the PZD index are ignored, data loss occurs in some cases when compared to PND. At the same time, non-zero values at the intervention level are not included in

the calculation of the PZD value. Besides, as in PND, the PZD index can also be affected easily by the trend arising from outliers and time (Allison & Gorman, 1993). Another disadvantage of the PZD index is that it only takes into account the data points at the intervention level. As it focuses only on the elimination of the target behavior, it is not suitable for every intervention situation.

Percentage of Data Points Exceeding the Median of Baseline Level (PEM)

As indicated by Ma (2006, p. 600), the calculation of PND index using the horizontal line in the presence of the 0 or 100 (ceiling and floor effect) value in the baseline level data makes it meaningless or zero. This is considered one of the weaknesses of this method. There may be a trend effect resulting from non-sudden decrease and increase in levels. This trend effect is ignored in the PND. For these reasons Ma (2006) argued that the risk of making Type II error was too high and, as an alternative to this method, the percentage of data points exceeding the median value of the baseline level (PEM) method was proposed.

When calculating the PEM value in an AB design, the following steps are followed:

1. The median value of the baseline level on the graph is determined,
2. A horizontal line is drawn from the median to the right.
3. The intervention level data points remaining above this horizontal line is determined,
4. The number of data points in the third step is divided by the total number of data points at the intervention level,
5. The value obtained in the fourth step is multiplied by 100 to calculate the PEM value.

If the intervention applied while calculating PEM value is expected to increase the target behavior, the intervention level data points above the median value of the baseline level are taken into account while the effect of the intervention in the intervention is expected to decrease the target behavior the intervention level data points below the median value are taken into account.

When we look at the baseline level data points in Figure 3, the median value of this level is 7. When a line is drawn over this value to the right side of the graph, values above 7 at the intervention level are determined. According to the data in Figure 3, eight data points (8, 10, 9, 11, 8, 10, 10, and 12) are seen above this line. At the intervention level, the value of the data exceeding the median value is divided by the total number of data points at the intervention level (10) to obtain a value of 0.8. If we multiply this value by 100, the PEM value in the data in Figure 3 is calculated as $0.8 \times 100 = 80$.

When the recommended criteria for the PEM index are considered, the values between 91% and 100% indicate a “very effective” intervention, while between 70% and 90% indicate “moderate” intervention effect. In studies with a PEM value below 70%, it can be said that the intervention effect is “questionable” or “ineffective” (Ma, 2006). According to Heyvaert et al. (2014), “questionable” intervention is between 50% and 70% and “ineffective” intervention occurs below 50%. In case of ineffective experiments, the data points show more or less continuous fluctuations around the median value. The PEM value obtained using the data in Figure 3 indicates a moderate effect according to the above criteria.

The PEM method has been found to be less affected by the autocorrelation in the data, and it has been found that it effectively differentiates the effective and ineffective interventions comparing to other indexes (Manolov, Solanas, & Leiva, 2010). The PEM index is used to calculate the change in the target behavior among AB levels, not the trend changes between the baseline and the intervention levels. In addition, this index does not take into account of the change and trend at the intervention level. The PEM value provides a partial solution to the inability to calculate the slope value in the PND when there is an orthogonal slope in the baseline- and intervention-level pairs after the first intervention level. According to Ma (2006), one of the limitations of PEM index is to ignore the magnitudes of the data points above the median in the calculation of this index, which means that this index is not sensitive to the data points above the median. In addition, how to calculate PEM in data cases other than AB designs was not mentioned in the original study.

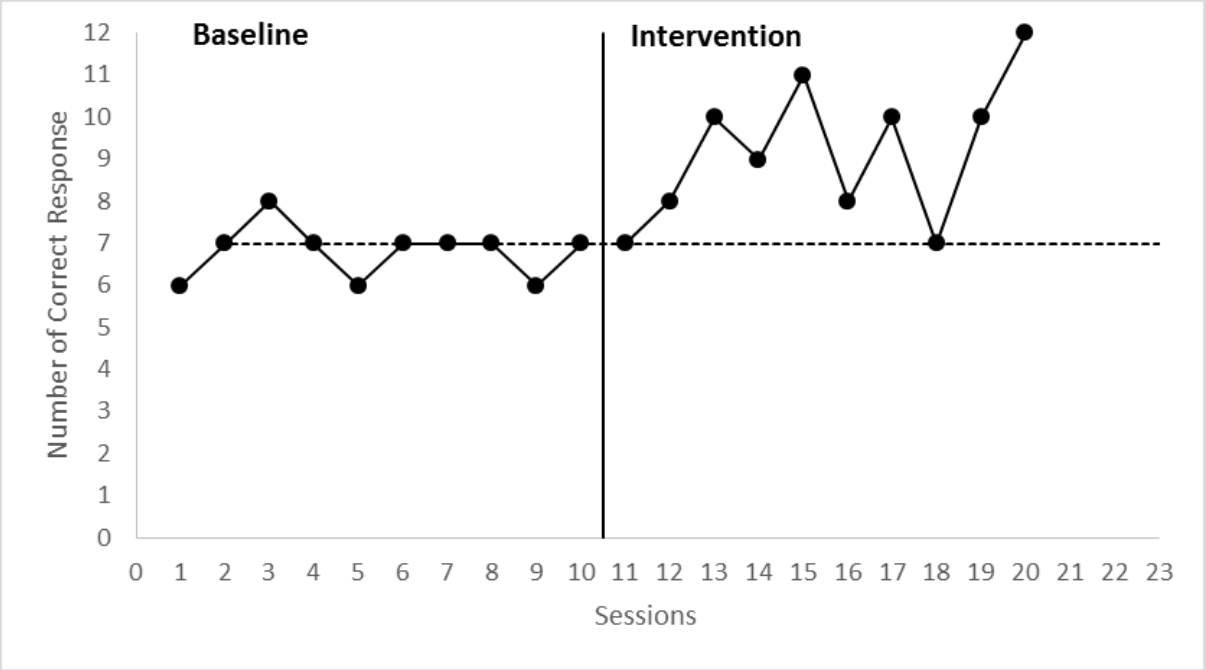

Figure 3. Calculation of Data Points Exceeding Median of Baseline Level in an AB Design.

Percentage of Data Exceeding a Median Trend of the Baseline Level (PEM-T)

Percentage of data exceeding the median trend of the baseline level (PEM-T) is an effect size index that takes into account the trend determined using the split-middle line technique (White & Haring, 1980) for single-subject experimental research. (Wolery et al., 2010). The PEM-T index, a version of PEM, is a non-parametric statistics used for the same purpose.

In the calculation of the PEM-T value in an AB design, the following steps are followed:

1. A split-middle line of trend is drawn on the graph from the baseline to the intervention level using the split-middle line technique as described in White and Haring (1980),
2. The intervention level data points above the median trend line are counted,
3. The number of data points in the second step is divided by the total number of data points at the intervention level,
4. The value obtained in the third step is multiplied by 100 to calculate the PEM-T value.

If we want to calculate the PEM-T value according to the baseline and intervention level data points presented in Figure 4, we need to create a median trend line using the data at the baseline level. To calculate PEM-T value in Figure 4, a median trend line was obtained using the R syntax created by Manolov, Sierra, Solanas and Botella (2014). As can be seen in Figure 4, eight data points at the intervention level remain above this trend line. Thus, we can calculate PEM-T as $8/10 \times 100 = 80$. In an experimental study with multiple AB designs, the PEM-T values are calculated for each AB design and the values obtained are averaged. The researchers who developed PEM-T did not make any suggestions on how to calculate this index for more complex designs.

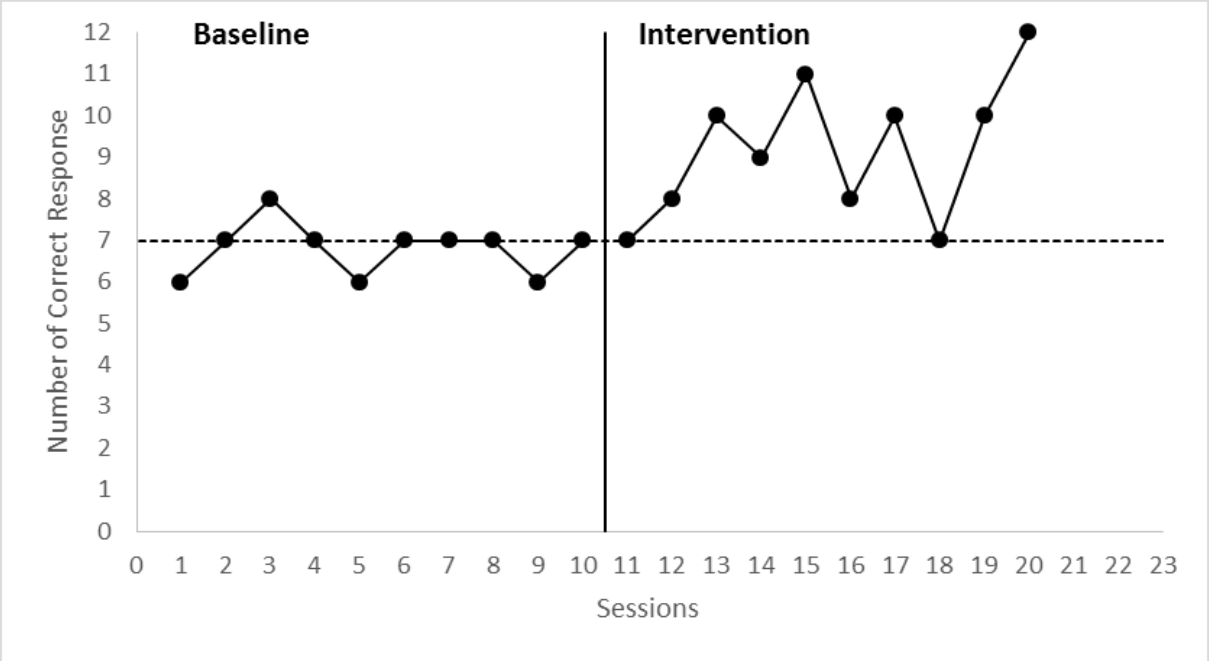

Figure 4. Calculation of the Percentage of Data Exceeding the Median Trend of the Baseline Level in an AB Design.

There are no clear criteria for interpreting PEM-T values in the literature. Some criteria presented in Ma (2006) can be used. According to these criteria, values of 90% and above indicate a "very effective" intervention and values between 70% and less than 90% indicate "moderate" intervention effect. Values between 50% and less than 70% refer to "questionable" or "mild" intervention. If PEM-T is below 50%, the intervention is considered "ineffective". These criteria need to be used more cautiously, as the effect size value may be smaller in graphs with a trend that begins at the baseline level. The PEM-T index also has some advantages and limitations. One of the main advantages of this index is that it is easy to interpret and can take account of the linear trend at the baseline level. The fact that the median trend line is affected by outlier values at the baseline level is also seen as a factor affecting the calculation of this index.

Percentage of All Non-Overlapping Data (PAND) and Phi Coefficient

The percentage of all non-overlapping data suggested by Parker et al., (2007) helps to find the percentage of non-overlapping data as in the other percentage of non-overlapping data indexes. The PAND value is defined as the percentage of remaining data points to total data points after eliminating the minimum number of data points that can remove the overlap between two levels (Parker et al., 2011).

In order to calculate the PAND value in an AB design, the following steps are followed:

1. Determine the data points that cause the overlap between the two levels,
2. Determine the minimum number of data points to eliminate the overlap between two levels,
3. Count the overlapping data points,
4. Calculate the percentage of overlapping data by dividing the number of overlapping data points determined in the third step by the total number of data points,
5. The percentage of overlapping data obtained in step 4 is subtracted from 100 to obtain the percentage of non-overlapping data (PAND).

Figure 5 shows the overlapping region and the data points that cause this overlapping area to calculate the PAND value according to the baseline and intervention level data. As shown in Figure 5, a line is drawn from the highest point of the baseline level and a line is drawn from the lowest point of the

intervention level and the overlapping region is determined between the two levels. A minimum number of data points are then decided to eliminate this overlap. If we remove a data point (8) at the baseline level and two data points at the intervention level (7 and 7) in the graph in Figure 5, we eliminate the overlap between these two levels and we get the number of all data points that do not overlap. Then, if we remove the number of data points (2 + 1 = 3) that overlap from the entire number of data points (20), we get the number of all the non-overlapping data points (17). Thus, we find the value of PAND (20-3)/20=0.85. If we multiply this value by 100 to represent it as a percentage, we get the percentage of all non-overlapping data.

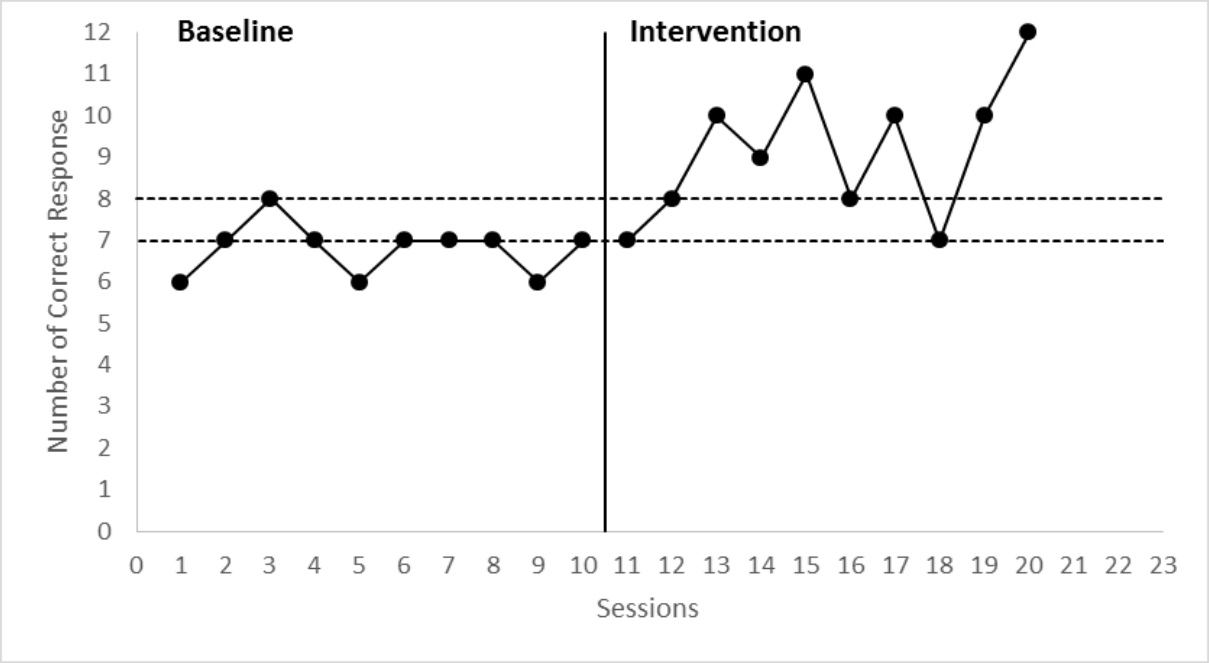

Figure 5. Calculation of the Percentage of All Non-Overlapping Data in an AB Design.

Parker et al. (2007) state that the PAND can be calculated only by looking at the graph, but it is difficult to find the correct results by inspecting the graph in cases where too many data points are present (complex single-subject experimental designs). PAND index, which is based on non-overlapping data values such as PND, is calculated using all data at the baseline level unlike the PND. Another advantage of the PAND index is that this index can be transformed into Pearson Phi correlation. Parker et al. (2007) state that the Phi value has a known sampling distribution, and the value of *p* for Phi can be obtained. It also allows for confidence intervals and power calculations. PAND index can be interpreted by converting it to Cohen's *d* value through the Phi value (Parker et al., 2007). Thanks to these features, the PAND value provides the opportunity to obtain a lot of information that cannot be obtained with the PND index. PAND, which can only measure average level changes, cannot control the positive baseline level trend (Parker et al., 2007, p. 196).

Nonoverlap of All Pairs (NAP)

The nonoverlap of all pairs (NAP) index developed by Parker and Vannest (2009) is a value calculated based on completely pair comparisons.

In an AB design, the following steps are followed to calculate the NAP value:

1. Each data point at the baseline level is compared with each data point at the intervention level. When we compare one of the baseline-level data points in an AB design to all of the intervention-level data points, there are three cases for each pair of data. The first of these cases

is the development of the data point taken at the baseline level towards the point of comparison at the intervention level (positive value). In other words, the data point at the intervention level is larger than that of baseline-level. Second, the level of baseline level data is no change (equal value or ties) from this level to the intervention level. Third, at the baseline level, the comparison of the data point at which we make the comparison is greater than the data points at the intervention level, i.e., it shows a decrease (negative).

2. A total score is calculated for each baseline level data point to be compared. When calculating this score, a value of 1 for positive ($A_n<B_n$) cases, a value of 0 for negative ($A_n> B_n$) and a value of 0.5 for cases with equal status ($A_n= B_n$) are given.
3. These values obtained in the second step are summed and the total development score is obtained for the baseline point subject to comparison.
4. We do the same for other data points at the baseline level, and we calculate the total development points.
5. By dividing this total development score obtained in the previous steps by the number of all data pairs that may be available (A × B), we obtain the NAP value.

The number of data pairs in an AB design is equal to product (A × B) of the number of data points at the baseline and intervention levels. The number of data pairs for the data in Figure 6 is 100 (i.e., 10 × 10). The data pairs that show increase (Pos), decline (Neg), and ties (Ties) are counted in Figure 6. Based on the data presented in Figure 6, development score is obtained as 10 (10 positive status) making the comparison of the first data point (6) at the baseline with each intervention level data point (6-7, 6-8, 6-10, 6-9, 6-11, 6-8, 6-10, 6-7, 6-10, 6-12). Similarly, for the second data point (i.e., 7), the development score is obtained as 9 (8 positive and 2 equal conditions). In the same way, the third data point (i.e., 8) of the baseline level is compared with each data point at the intervention level and shown in Figure 6 as an example. The development score for the third data point of the baseline is obtained as 7 (6 positive, 2 equal cases, 3 negative conditions) according to the comparison conditions in Figure 6. For each data point, the development scores are calculated as 10, 9, 7, 9, 10, 9, 9, 9, 10, and 9 for baseline sessions 1 to 10, respectively. When these values are summed, the total number of development data pairs equals to 91 and by dividing this number by the total number of data pairs (100), we calculate the NAP value as 91 (91%).

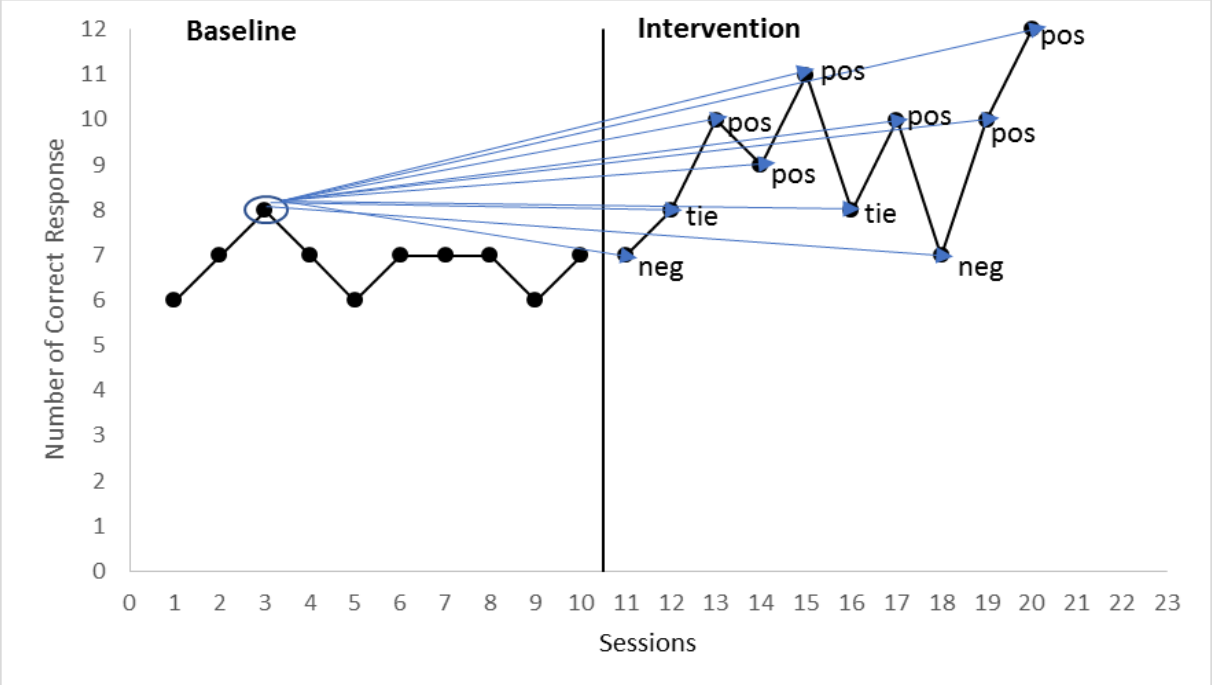

Figure 6. Showing the Data Pair Comparison Used for NAP Calculations for the 3rd Data Point (neg = negative; pos = positive; tie=equal).

The NAP index attempts to summarize the non-overlap between each baseline-level value and each intervention-level value. Briefly, it can be defined as the percentage of data showing development between A and B levels (Parker et al., 2011). Since the NAP takes into account all data pairs, it is a holistic non-overlapping data statistics and can be converted to the effect size Cohen's *d*. Like other non-overlapping data indexes, the NAP index can also be calculated using a graph, but it is not easy to compare all data pairs in large data sets. In addition to manual calculation from the graph, the NAP value can be obtained using receiver operating characteristics (ROC) or Mann Whitney U test analysis using familiar statistical programs such as SPSS. In the ROC analysis, Area Under the Curve (AUC) can be used to calculate the NAP index. If there are too many data points, or if there are multiple baseline and intervention levels, it may be difficult to calculate other non-overlapping data percentages from the graph. Even in these cases, the NAP value can be easily obtained by finding the area under the curve in the ROC analysis. In addition, the NAP value can be calculated by entering the data on a sheet via *http://www.singlecaseresearch.org/calculators/nap*. As mentioned in Parker and Vannest (2009, p. 359), the value under the curve (AUC value) is a probability value ranging from 0.5 to 1. According to the criteria given by Parker and Vannest (2009), the values less than 0.65 indicate "weak effect", the values between 0.66 and 0.92 indicate "moderate effect" and values above 0.92 indicate "very effective" intervention.

According to Parker and Vannest (2009), the NAP has five advantages compared to other non-overlapping data indexes. One of the main reasons for this is that they can better distinguish the results in the comparison of many published studies. The second advantage is that there is less error compared to other indexes which are calculated manually due to the fact that the scoring calculation can be performed with the help of commonly used computer programs such as SPSS. The third and fourth advantages are the high correlations between the data and the validity of the visual analysis and the conversion of the NAP to R^2 and thus to Cohen's *d* (Parker & Vannest, 2009). The fifth advantage given by Parker and Vannest (2009) is that the NAP value, which has low confidence intervals, offers more accurate results. Furthermore, in the case of autocorrelation in the longitudinal data, the NAP performs well (Manolov, Solanas, Sierra, & Evans, 2011).

Tau

Another effect size index developed by Parker, Vannest, Davis and Sauber (2011) is Tau (Kendall's Tau non-overlap). In the calculation of the tau index, non-overlapping data pairs are used as in the calculation of the NAP value. While the percentage of data pairs that do not overlap is found in the NAP, the percentage of overlap is excluded from the non-overlap percentage in the Tau index (Parker, Vannest, Davis, & Sauber, 2011). Given that PDP represents the number of data pairs that increases from the baseline level to the intervention level (positive) and the NDP represents the number of data pairs that are decreasing from the baseline level to the intervention level (negative) and TDP represents the total number of data pairs between the baseline and intervention levels. Tau index can be calculated as follows:

$$Tau = \frac{\text{PDP}-\text{NDP}}{\text{TDP}} \quad (1)$$

The number of all data pairs in an AB design, as in the PAND, is equal to the product number (A × B) of the baseline and intervention levels. Here, when we subtract the number of negative data pairs from the number of positive data pairs (the number of overlapping data pairs) and then dividing by the number of all data pairs, we get the Tau value.

In the data presented in Figure 6, by comparing each of the data pairs (100 data pairs) in a similar way to the calculations made in previous index, the number of positive pairs is 84, the number of negative pairs is 2 and the number of equal pairs is 14. Thus, the Tau value is calculated as (84-2) / 100 = 0.82. In order to convert this decimal value to a percentage scale, it is required to multiply by 100. The Tau value can also vary from 50% to 100%, as in the NAP index. As Parker et al. (2011) pointed out, the Tau value can be obtained using Kendall rank correlation or Mann-Whitney U test analysis using

conventional statistical programs (Parker, Vannest, Davis, & Sauber, 2011). The Tau value can be obtained by dividing the S value obtained in the Kendall rank correlation by the total number of data pairs (A × B).

In an AB design, the following steps are followed to obtain the Tau value from the Kendall rank correlation:

1. The Level variable is created: A Level variable is created by entering a value of 0 at the baseline level and a value of 1 at the intervention level (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
2. A Kendall rank correlation is calculated between the Level variable created in the first step and the raw data of baseline and intervention levels (6, 7, 8, 7, 6, 7, 7, 7, 6, 7, 7, 8, 10, 9, 11, 8, 10, 7, 10, 12),
3. S value is obtained from Kendall rank correlation (S = 82),
4. The S value in the third step is divided by the total number of data pairs (100) and the Tau value is obtained as .82. This value can be calculated by entering the raw data into a sheet via *http://www.singlecaseresearch.org/calculators/tau-u*.

Tau-U

Parker, Vannest, Davis, and Sauber (2011) proposed another Tau value (i.e., Tau-U) that can able to control undesirable positive baseline trend. The Tau-U index, as in Tau, does not only cover the non-overlapping data, but can also control the trend of the baseline level. In this way, the Tau-U index allows for the quantification of the increase in the intervention that occur only at the intervention stage, beyond the potential increase at baseline before the intervention (i.e., during the baseline level).

In an AB design, the following steps are followed to obtain the Tau-U value from the Kendall rank correlation:

1. A Level variable is created: Rank numbers are assigned to the data at the baseline level with the maximum value of 1 (3, 2, 1, 2, 3, 2, 2, 2, 3, 2) for (6, 7, 8, 7, 6, 7, 7, 7, 6, 7) and generated data are entered as Level A data. The last number in Level A is assigned to each element in Level B (7, 8, 10, 9, 11, 8, 10, 7, 10, 12) to continue (4, 4, 4, 4, 4, 4, 4, 4, 4, 4). A common Level variable is created by combining the data of these two levels (3, 2, 1, 2, 3, 2, 2, 2, 3, 2, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4),
2. A Kendall rank correlation is calculated between the Level variable created in the first step and the raw data of baseline and intervention levels (6, 7, 8, 7, 6, 7, 7, 7, 6, 7, 7, 8, 10, 9, 11, 8, 10, 7, 10, 12),
3. S value is obtained from Kendall rank correlation (S = 83),
4. The S value in the third step is divided by the total number of data pairs (100) and the Tau-U value is obtained as .83. We multiply by 100 to express this value on a scale of 100. This value can be calculated by entering the raw data into a sheet via *http://www.singlecaseresearch.org/calculators/tau-u*.

Parker et al. (2011) showed how the Tau-U value can be calculated for the AB and ABA designs. In this study, it was mentioned that only the first baseline (A) and intervention (B) levels are included in the calculation in more complex designs. Rakap (2015, p. 26) provided some explanations on how to use the Tau-U index, which is one of the most recent indexes developed in the literature, in other single-subject experimental designs. In designs with multiple baseline levels, Tau-U is calculated separately for each baseline-intervention (AB) comparison and the overall Tau-U value is obtained by taking the average of all Tau-U values obtained. When using the ABAB design, the Tau-U value is calculated for each baseline and intervention pairs (i.e., A_1B_1 and A_2B_2). The overall Tau-U value is also obtained by averaging the Tau-U values obtained from these pairs. In the ABCD design, according to Rakap (2015), Tau-U values should be obtained by comparing the intervention level with the baseline level for each level of intervention (e.g., AB, AC, AD given that A represents baseline level). The overall effect size of the full model should be calculated using the last intervention stage (AD comparison). In the case of

such design, the overall Tau-U value is the value obtained from the comparison of the baseline level and the final intervention level.

The criteria for the NAP suggested in Parker and Vannest (2009, p.364) can also be used for Tau-U. According to these criteria, percentages below 65 indicate "weak" or "small" effects, percentages between 66-92 indicate "moderate" effects, and percentages between 93 and 100 indicate "very effective" interventions. Since these are not specifically developed for Tau-U, caution should be noted in the use of these criteria. The Tau-U value of 82% using the data in Figure 6 indicates a moderate intervention.

The Tau-U has the flexibility to analyze trend and overlap separately or both at the same time. In addition to this flexibility, Tau-U index has more statistical power than other non-overlap indexes. The distinctiveness of the Tau-U index and its consistency with the visual analyses are among the strengths of this index. There is a decrease in the Tau-U index due to the attempt to control the trend. This decline is seen as a limitation of this index by Parker et al. (2011). The fact that the Tau-U index cannot be calculated in more complex single-subject experimental design than the AB and ABA designs is also seen as a limitation.

It should be noted that all of the non-overlap methods described so far are adapted from methods known as nonparametric dominance statistics in experimental studies. The dominance statistics can be defined as the probability of a randomly selected score from a group exceeding a score in a second group (Parker, Vannest, & Davis, 2011). The use of non-overlapping statistics, which are the equivalents of the dominance statistics in single-subject experimental studies has given more reliability to the publications in this area.

Mean Baseline Reduction (MBR)

Mean Baseline Reduction (MBR; Campbell, 2003) was developed by O'Brien and Repp (1990) to calculate the decrease in the level of intervention in the targeted behavior. The MBR index is also known as the percentage of decline from the baseline level and is applied to determine how much the behavior has decreased.

In order to calculate the MBR value, it is sufficient to calculate the means of the data points at the intervention level and baseline levels. The following steps can be followed when calculating MBR in an AB design:

1. Calculate the mean of the baseline level,
2. Calculate the mean of intervention level,
3. Subtract he mean of the intervention level from the mean of the baseline level,
4. Divide the value in the third step by the mean of the baseline level (Campbell, 2003),
5. By multiplying the value obtained in the fourth step by 100, the percentage of MBR is obtained. To put it another way; the MBR value is calculated with the following equation:

$$\text{MBR} = \frac{M_A - M_B}{M_A} \quad (2)$$

where M_A represents the mean of baseline level and M_B represents the mean of intervention level.

If the purpose of the intervention level is to increase rather than decrease a behavior, then the MBR value is calculated by changing the order of two terms (M_A and M_B) on the numerator part of Equation 2. In other words, the baseline level mean is subtracted from the intervention level mean and the resulting number should be divided by the mean of the baseline level. In order to be able to calculate the MBR over the data presented in Figure 6, we must first calculate the mean value for both levels. The mean values of the baseline level according to the data in Figure 6 is 6.8 and the mean value of the intervention level is calculated as 9.2. Thus, the percentage of MBR is calculated as ((9.2 - 6.8) /6.8) × 100 = 35.3.

When the MBR value was found to be 100%, the problematic behavior was completely eliminated and the 0% MBR value indicated that no change was observed according to the baseline level. Negative

MBR value shows that problematic behavior increases at intervention level. Some researchers recommend the use of the last three (Olive & Franco, 2008) or the last five (Lydon, Healy, O'Reilly, & McCoy, 2013) data points at the baseline level and intervention level when calculating the MBR value. If the values at the baseline level are zero, it is not possible to calculate the MBR value. This is a limitation of MBR. Although there are no clearly defined interpretation criteria for MBR, Bell, Skinner and Fisher (2009, p. 5) used small (.20), medium (.50) and high (.80) criteria for MBR values. In the literature, there are not many studies on how to calculate the MBR index in different single-subject experimental designs. Carr, Severtson, and Lepper (2009) used the last baseline and intervention-level data points to calculate the MBR for ABA designs. We also found Carr et al. (2009) used the baseline and intervention level data of each participant to calculate the MBR value for the multiple-baseline designs.

Standardized Mean Difference (SMD)

In the meta-analyses studies with multiple subjects, the standardized mean difference (Cohen's *d*, Hedges' *g* and Glass' delta) indexes are used to calculate the effect size. It has been suggested by researchers that a similar value to the standardized mean difference can be used in single-subject experimental studies (Busk & Serlin, 1992; Shadish, Hedges, & Pustejovsky, 2014). The following steps can be followed when calculating the SMD value (Busk & Serlin, 1992) to demonstrate the effectiveness of the experiment in single-subject experimental studies:

1. Calculate the mean of the baseline level,
2. Calculate the mean of the intervention level,
3. Calculate the standard deviation of the baseline level,
4. Subtract the mean of the baseline level from the mean of the intervention level,
5. The value in the fourth step is divided by the standard deviation of the baseline level (Campbell, 2003).

To put it another way; the SMD value is calculated with the following equation:

$$\text{SMD} = \frac{M_B - M_A}{S_A} \tag{3}$$

where M_A represents the mean of baseline level and M_B represents the mean of intervention level and S_A represents the standard deviation of the baseline level.

In order to calculate the SMD value using the data presented in Figure 6, we need to calculate mean values for both levels and only the standard deviation value for the baseline level. According to the data in Figure 6, the mean value of the baseline level is 6.8 and the mean value of the intervention level is calculated as 9.2. In addition, the standard deviation of the baseline level required to achieve the SMD value is 0.632. Thus, the SMD value is calculated as (9.2-6.8) /0.632 = 3.79. It is also shown by visual analysis and non-overlapping data-based indexes in which the effect of the intervention in Figure 6 is not very large. However, the result of the SMD causes the intervention to be seen as a very effective intervention. In the light of this example, it is likely that the SMD will provide misleading results in terms of its effectiveness.

In contrast to most other recommended indexes, the standardized mean difference (SMD) has a known sampling distribution, and allows statistics such as regression and ANOVA to be applied to the common scale. With the help of Binomial (Binomial) sign test, the confidence intervals for this index can be calculated (Busk & Serlin, 1992). According to Olive and Franco (2008, p. 8), there are many strengths of the SMD method. The first one is the use of mean value in the calculations and possibility to calculate this mean value in both increasing and decreasing intervention level effect situations. Unlike the approaches based on non-overlapping data percentage, no data value in the SMD method is ignored. Another strength of the SMD method is that it can be interpreted in the same way as the known effect size values (e.g., Cohen's *d*). According to Olive and Franco (2008, p. 7), the SMD value obtained in experimental studies with single subjects can be interpreted according to the criteria proposed by Cohen (1988). According to these criteria, d = 0.2, 0.5, and 0.8 are interpreted as small, medium, and large

effect sizes, respectively. Researchers are advised to be cautious when using these criteria. Olive and Smith (2005) state that there are many different single-subject experimental designs, and for each of these designs, the same way should be followed by using different data to calculate the SMD value. For example, the individual SMD values should be calculated for each subject in the multiple baseline designs that include different subjects. The original baseline level and the last applied intervention level should be used in the SMD calculation in the reversal designs. The SMD value calculation is also possible for an ABA'B' design. In these cases, the researcher should calculate the SMD values (SMD_1 and SMD_2) for both AB designs (AB and A'B') and obtain the mean of these two values. Among the criticized aspects of the SMD value is the possibility of a possible trend effect due to time and failure to account for the change in slope between levels. The standardized mean difference is considered to be a problematic effect size calculation method because of the effect of the autocorrelation and the resulting magnitude of the effect size is usually too large. In this context, many researchers in recent years are striving to develop this index. Although it is frequently used by many researchers, it is thought to be problematic to use them in the context of meta-analyses without overcoming the above-mentioned limitations. In addition, SMD statistics does not take into account the dependence on the longitudinal data structure as in most other non-regression approaches. This leads to incorrect estimation of standard error rates. Therefore, this may lead to an increase in the Type I or Type II error rates.

Reporting of Effect Size Indexes

The effect size values obtained from different studies must be gathered and interpreted to be used in a meta-analysis study. According to Maggin et al. (2011), the successful combination of effect sizes from different studies is useful to obtain generalizable results and to show to other researchers where the amount of effect at the intervention level is more or less. The summarization of the effect size values obtained from different studies is achieved by obtaining weighted average through traditional meta-analysis models in studies with multiple subjects. The proposed steps for traditional meta-analysis (literature review, selection of studies, calculation of effect sizes and summarization) also apply to the meta-analysis of single-subject experimental studies. The difference between multiple- and single-subject experimental studies is that how to summarize the effect size values obtained from different studies.

According to the research conducted by Maggin et al. (2011), researchers apply five different methods in experimental subjects. The most preferred of these methods is to summarize the effect size values obtained from all studies by calculating the mean effect size value. The second most commonly used summation method is to obtain an average value from all studies using a weighted average. In order to calculate the weighted average, it is necessary to obtain values such as inverse variance weights, confidence intervals or standard errors in the traditional meta-analysis. It is also possible to obtain weighted average with multilevel models. Other methods used to summarize effect sizes include reporting the median value of all studies, providing descriptive explanations, or showing the percentages of effective and ineffective studies.

In single-subject experimental studies, the researcher needs to pay attention to several points in obtaining the common effect size value using one of the methods described above (Vannest & Davis, 2013, pp. 107-108). First of all, the researcher who will make a meta-analysis should know the characteristics of single-subject experimental research. The researcher should be well aware of the levels in which a single subject will be compared between the levels of experimental research (Vannest & Davis, 2013). The calculation of the indexes presented above is shown for the AB design. Almost all of these indexes can also be applied to complex single-subject experimental designs. Even in these complex designs, most researchers obtain the index values described above by selecting a level A and a level B. What is important is to justify why the researcher has made such a decision. In addition, the researcher who is planning to make a meta-analysis in single-subject experimental studies should explain the method of calculating the effect size index of his/her choice and should provide a detailed description of the calculation method. WWC (Kratochwill et al., 2010) and many researchers recommend calculating and reporting the effect sizes using multiple indexes until a single effect size index is developed that is sensitive to all conditions specific to the data patterns obtained from single-subject experimental studies.

In addition to reporting the effect size value for each study included in the meta-analysis, the meta-analyst should also report details such as the number of participants in the research, the type and the number of behaviors examined, the number of studies and the type of weighting used (Vannest & Davis, 2013). Confidence intervals should also be provided when reporting effect size index values.

DISCUSSION and CONCLUSION

In this study, 10 indexes, which are used as effect size values in single subject experimental research, were examined. It was tried to explain how to calculate these indexes developed by different researchers on sample data. It is expected that this study will benefit the researchers with single-subject experimental studies. Firstly, PND and PZD indexes, which are the most preferred indexes in domestic and international literature, are explained and, the properties of other alternative indexes are presented. In this study, it was observed in the literature that the PND was preferred due to the fact that it is easily computable among the indexes compared, but the limitations were ignored. It has been underlined that some of the other indexes like PND (PAND, NAP, PEM and Tau) offer solutions to the limitations of PND but do not offer the advantages of traditional effect size values.

The fact that the effect size indexes are used in different behavior situations are among the distinguishing features of these indexes. While most of the indexes examined in our study are used in cases where target behavior is expected to increase or decrease, PZD index is used only in cases where the expected behavior is expected to disappear. If the researcher is doing an intervention on the elimination of a behavior in the participant, the index to be preferred may be the PZD. Another difference is the number of data points used in the calculation indexes. Some indexes only take into account data values at the intervention level (e.g., PZD), while some indexes (PND, PEM, PAND, etc.) take into account one or a few of the baseline data points. Therefore, it would be more accurate to use indexes that take into account all the data in both the baseline and intervention levels (NAP, Tau, Tau-U, SMD, and MBR). Statistics that take into account all of the values in the data are less influenced by outliers. In particular, the NAP, Tau and Tau-U indexes tend to produce more accurate results because they have more statistical power than other nonparametric indexes (Parker et al., 2011). Ignoring some data points can lead to misleading decisions about the effectiveness of the intervention level. In addition to taking into account all data points, indexes such as SMD and MBR are characterized by being able to be transformed into traditional effect size values or interpreted in the same way. In this context, it is observed that the values obtained in the SMD calculations are higher than the group design studies. Another index with this feature is the PAND index which can be converted to *d* value by obtaining Phi value. Most of the percentage indexes presented in this study do not produce values that can be interpreted as effect size values obtained from studies with multiple subjects. Another limitation of the indexes covered in this study is that most indexes cannot take into account the effect of autocorrelation and trend in single-subject experimental studies (as in time series). The autocorrelation effect is defined as the positive or negative correlation of repeated data collected from the same individual. This effect leads to incorrect calculation of the standard deviation values. The index values, which can take this negative effect into account (Tau-U), are found more ideal by the researchers. In cases of autocorrelation and data without trend, the use of PND and PEM-T may be recommended to practitioners to determine whether an intervention is effective in terms of ease. However, it would be more appropriate to use more advanced methods (e.g., Tau-U) in the context of meta-analyses.

REFERENCES

Alresheed, F., Hott, B. L., & Bano, C. (2013). Single subject research: A synthesis of analytic methods. *The Journal of Special Education Apprenticeship*, *2*(1), 1–18.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case∗. *Behaviour Research and Therapy*, *31*(6), 621–631.

Allison, D. B., & Gorman, B. S. (1994). "Make things as simple as possible, but no simpler." A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy*, *32*(8), 885–890.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Aslan, C., Yalçın, G. & Özdemir, S. (Mayıs, 2016). *Sosyal öykü tekniğinin etkililiği: Betimsel değerlendirme ve meta-analiz çalışması*. Teacher Education in Special Education, Vocational Training and Sports Konferansı (ELMIS), Konya, Türkiye.

Aydın, O. (2017). *Otizm spektrum bozukluğu olan bireylere matematik becerilerinin öğretimi: tek-denekli araştırmalarda betimsel ve meta analiz* (Yayımlanmamış yüksek lisans tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Bell, R. J., Skinner, C. H., & Fisher, L. A. (2009). Decreasing putting yips in accomplished golfers via solution-focused guided imagery: A single-subject research design. *Journal of Applied Sport Psychology, 21*(1), 1–14.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*, 129–141.

Bozkuş-Genç., G. (2017). *Otizm spektrum bozukluğu olan çocuklara soru sorarak iletişim başlatmanın kazandırılmasında temel tepki öğretiminin etkileri* (Yayımlanmamış doktora tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp.187–212). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24*, 120–138.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*(2), 234–246.

Carr, J. E., Severtson, J. M., & Lepper, T. L. (2009). Noncontingent reinforcement is an empirically supported treatment for problem behavior exhibited by individuals with developmental disabilities. *Research in Developmental Disabilities, 30*(1), 44–57.

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

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 35*(10), 2463–2476.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7*, 107–118.

Huitema, B. E., & Mckean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38–58.

Karasu, N. (2009a). Özel eğitimde delile dayalı yöntemlerin belirlenmesi: Tek denekli çalışma analizleri ve karşılaştırmaları. *Türk Eğitim Bilimleri Dergisi, 7*(1), 143–163.

Karasu, N. (2009b). Otizmden etkilenmiş bireylerde sosyal ve iletişim becerilerini arttıran yöntemlerin delile dayalı yöntem olarak belirlenmesi: Bir meta-analiz örneği. *Türk Eğitim Bilimleri Dergisi, 7*(3), 713–739.

Karasu, N. (2011). Otizmli bireylerin eğitiminde video ile model olma uygulamalarının değerlendirilmesi: Bir alanyazın derlemesi ve meta-analiz örneği. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 12*(2), 1–12.

Kavale, K. A. (2001). Decision making in special education: The function of meta-analysis. *Exceptionality, 9*, 245–268.

Kaya, F. (2015). *Otizm spektrum bozukluğu olan öğrencilere yiyecek-içecek hazırlama becerilerinin öğretiminde sesli anlatım içeren ve içermeyen video ipucunun karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Kırcaali-İftar, G., & Tekin, E. (1997). *Tek denekli araştırma yöntemleri*. Ankara: Türk Psikologlar Derneği Yayınları.

Korkmaz, Ö. T., & Diken, İ. H. (2010). Stereotipik davranışların azaltılmasında kullanılan yöntemlerin etkililiği: Betimsel ve meta analizi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 11(2)*, 1–12.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. *What works clearinghouse*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lydon, S., Healy, O., O'Reilly, M., & McCoy, A. (2013). A systematic review and evaluation of response redirection as a treatment for challenging behavior in individuals with developmental disabilities. *Research in Developmental Disabilities, 34*(10), 3148–3158.

Ma, H. (2006). An alternative method for quantifying synthesis of single-subject research: Percent of data points exceeding the median. *Behavior Modification, 30*, 598–617.

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality, 19*(2), 109–135.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*(3), 301–321.

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs: Quantitative proposals in the context of the evidence-based movement. *Behavior modification, 38*(6), 878–913.

Manolov, R., & Solanas, A. (2008). Comparing N=1 effect size indices in presence of autocorrelation. *Behavior Modification, 32*, 860–875.

Manolov, R., Solanas, A., & Leiva, D. (2010). Comparing "visual" effect size indices for single-case designs. *Methodology, 6*, 49–58.

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42*(3), 533–545.

O'Brien, S., & Repp, A. C. (1990). Reinforcement-based reductive procedures: A review of 20 years of their use with persons with severe or profound retardation. *Journal of the Association for Persons with Severe Handicaps, 15*(3), 148–159.

Olive, M. L., & Franco, J. H. (2008). (Effect) size matters: And so does the calculation. *The Behavior Analyst Today, 9*(1), 5–10.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*(2-3), 313–324.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40*, 194–204.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*(2), 135–150.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*(4), 303–322.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*(2), 284–299.

Rakap, S. (2015). Effect sizes as result interpretation aids in single-subject experimental research: description and application of four nonoverlap methods. *British Journal of Special Education, 42*(1), 11–33.

Reichle, J. (2007). Amongst methodologies of functional behavioral assessment, functional analysis yields more effective suppression outcomes. *Evidence-Based Communication Assessment and Intervention, 1*(4), 153–155.

Schlosser, R. W., & Koul, R. K. (2015). Speech output technologies in interventions for individuals with autism spectrum disorders: a scoping review. *Augmentative and Alternative Communication, 31*(4), 285–309.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation, 96*, 233–256.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*(3), 221–242.

Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education, 34*(1), 9–19.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24–33.

Scruggs, T. E., Mastropieri, M. A., Cook, S. B., & Escobar, C. (1986). Early intervention for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders, 11*, 260–71.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147.

Sönmez, M., & Diken, İ. H. (2010). Problem davranışların azaltılmasında işlevsel iletişim öğretiminin etkililiği: Betimsel ve meta-analiz çalışması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi*, *11*(1), 1–16.

Strain, P. S., Kohler, F. W., & Gresham, F. (1998). Problems in logic and interpretation with quantitative syntheses of single-case research: Mathur and colleagues (1998) as a case in point. *Behavioral Disorders, 24*, 74–85.

Swaminathan, H., Rogers, H. J., Horner, R. H., Sugai, G., & Smolkowski, K. (2014). Regression models and effect size measures for single case designs. *Neuropsychological Rehabilitation, 24*(3-4), 554–571.

Tavil, Y. Z. ve Karasu, N. (2013). Aile eğitim çalışmaları: Bir gözden geçirme ve meta-analiz örneği. *Eğitim ve Bilim, 38*(168), 85–95.

Uysal, H. (2017). *Zihin yetersizliği olan öğrencilere temel toplama işlemlerinde akıcılık kazandırmada iki farklı uygulamanın karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Van den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments & Computers, 35,* 1–10.

Vannest, K.J., & Davis, H. S. (2013). Synthesizing single case research to identify evidence based treatments. In B. G. Cook, M.Tankersley, & T. J. Landrum (Eds.), *Evidence-Based practices* (pp. 93–119). UK: Emerald Group Publishing.

Wehmeyer, M. L., Palmer, S. B., Smith, S. J., Parent, W., Davies, D. K., & Stock, S. (2006). Technology use by people with intellectual and developmental disabilities to support employment activities: A single-subject design meta analysis. *Journal of Vocational Rehabilitation, 24*(2), 81–86.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Merrill.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*(1), 18–28.