The Journal of Special Education
http://sed.sagepub.com/

Comparison of Overlap Methods for Quantitatively Synthesizing Single-Subject Data

Mark Wolery, Matthew Busick, Brian Reichow and Erin E. Barton

J Spec Educ 2010 44: 18 originally published online 31 December 2008
DOI: 10.1177/0022466908328009

The online version of this article can be found at:
http://sed.sagepub.com/content/44/1/18

Published by:
Hammill Institute on Disabilities

HAMMILL INSTITUTE ON DISABILITIES

and

SAGE

http://www.sagepublications.com

The Journal of Special Education
Volume 44 Number 1
May 2010 18-28
© 2010 Hammill Institute on Disabilities
10.1177/0022466908328009
http://journalofspecialeducation.sagepub.com
hosted at http://online.sagepub.com

Comparison of Overlap Methods for Quantitatively Synthesizing Single-Subject Data

Mark Wolery
Matthew Busick
Vanderbilt University
Brian Reichow
Yale University Child Study Center
Erin E. Barton
University of Oregon

Four overlap methods for quantitatively synthesizing single-subject data were compared to visual analysts' judgments. The overlap methods were percentage of nonoverlapping data, pairwise data overlap squared, percentage of data exceeding the median, and percentage of data exceeding a median trend. Visual analysts made judgments about 160 A-B data sets selected randomly from the *Journal of Applied Behavior Analysis*. The four overlap methods were compared for data sets in which all visual analysts agreed a change in data occurred or a change did not occur across conditions. Each overlap method had unacceptably high levels of errors. Given the findings and weaknesses of the overlap methods, their use should be abandoned. The desirable characteristics of a quantitative synthesis method are described.

Keywords: *quantitative synthesis; single-subject designs; percentage of nonoverlapping data; percentage of data exceeding the median; pairwise data overlap squared*

Two assumptions underpinning the evidence-based practice movement are that research can be used to identify and evaluate effective practices, and multiple studies are needed to be confident in the evidence supporting a practice (Dunst, Trivette, & Cutspec, 2002; Horner et al., 2005; Odom et al., 2005). As a result, a need exists for aggregating findings from multiple studies to evaluate the strength of the evidence. Methods for addressing this need include critical literature reviews (Cooper, 1998), best-evidence syntheses (Slavin, 1987a, 1987b), and meta-analyses (Lipsey & Wilson, 2001). Meta-analysis is a method for aggregating findings, quantifying the magnitude of those findings, and conducting analyses of the moderator variables to account for differences in magnitude across studies (Lipsey & Wilson, 2001). Meta-analytic methods for group experimental research are well established (Dunst, Hambey, & Trivette, 2004; Lipsey & Wilson, 2001). For single-subject research, consensus does not exist on whether meta-analysis is applicable (Salzberg, Strain, & Baer, 1987; Scruggs, Mastropieri, & Casto, 1987) and, if applicable, how the effect sizes should be calculated (Campbell, 2004; Ma, 2006; Parker & Hagen-Burke, 2007).

A number of computational methods have been proposed for calculating effect sizes for single-subject research. These include various ways of calculating mean difference effect sizes (Busk & Serlin, 1992), regression-based effect sizes (Allison & Gorman, 1993; Center, Skiba, & Casey, 1985–1986; Faith, Allison, & Gorman, 1996), and the percentage of nonoverlapping data (PND; Scruggs et al., 1987; Scruggs & Mastropieri, 1998). The mean difference and regression-based effect size methods assume the data are independent. However, in single-subject studies, this assumption is not tenable because the

Authors' Note: Address correspondence to Mark Wolery, Department of Special Education, Box 228 Peabody College, Vanderbilt University, Nashville, TN 37203; e-mail: mark.wolery@vanderbilt.edu.

data are collected on an individual over time in the same setting, under the same conditions, with the same response definitions, and with the same recording procedures. Thus, the data are likely serially dependent rather than independent (Kazdin, 1984; Parker & Hagan-Burke, 2007). Autocorrelation can be used to identify and then account for the model of serial dependency; however, the limited number of data points in each condition often compromises identifying the model. Failure to identify the model of serial dependency because of the limited number of data points does not mean serial dependency is absent (Suen, 1987). Furthermore, the regression method appears to produce unreliable estimates of effect sizes (Busick, 2008). The methods for calculating mean difference and regression-based effect sizes should be avoided because of these problems. The use of the Pearson R (Parker & Hagan-Burke, 2007) and Pearson R^2 (Parker & Vannest, 2007) also assume independence and are subject to the same difficulties as the regression methods, thus should be avoided. The Kruskal-Wallis W^2 also has been proposed but "requires constant variance, symmetry of data distribution, and absence of outliers" (Parker & Hagan-Burke, 2007, p. 922). These assumptions make it undesirable with many single-subject data series. Thus, these indices are limited and may well lead to inaccurate and unsupported findings.

The PND was the first method proposed for quantitatively synthesizing single-subject data (Scruggs et al., 1987). The PND approach is not compromised by serial dependency in the data. It is, however, compromised by three data characteristics: (a) variability in the baseline condition (Ma, 2006), (b) a baseline datum point at the therapeutic ceiling or floor (Ma, 2006), and (c) the presence of trends in the data (White, 1987). Recently, several other indices of nonoverlapping data have been described: the percentage of data exceeding the median (PEM; Ma, 2006), the improvement rate difference (IRD; Parker & Hagan-Burke, 2007), the percentage of all nonoverlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007), the pairwise data overlap squared (PDO^2; Parker & Vannest, 2007), and the percentage of data exceeding a median trend (PEM-T; Wolery, Busick, Reichow, & Barton, 2008). A major criterion for judging the utility and rigor of these methods is to determine the extent to which they agree with judgments of visual analysts.

The purpose of this study was to determine the extent to which four overlap methods for quantitatively synthesizing single-subject data agree with visual analysts' judgments about differences in data patterns in two adjacent conditions. The compared methods were the PND, PDO^2, PEM, and PEM-T. The PND was selected because it is the oldest, most widely known, and most frequently used overlap method. The PDO^2 was selected because Parker and Vannest (2007) found it superior to the PND and PEM in agreeing with visual analysts' judgments. However, their comparison to visual analysts' judgments was based on a three-level rating system of the amount of improvement (*little to no improvement*, *moderate improvement*, and *strong or major improvement*); this method is different from that used in the current study. The PEM was chosen because it eliminates one of the confounding conditions of the PND; specifically, a ceiling or floor datum point in the first condition does not confound the PEM. The PEM-T was included because it is the only overlap method considering the trend in the data of the first condition. Three research questions were asked: (a) To what extent do these overlap methods (PND, PDO^2, PEM, and PEM-T) agree with visual analysts' judgments that a change occurred in the data patterns across two conditions? (b) To what extent do these overlap methods (PND, PDO^2, PEM, and PEM-T) agree with visual analysts' judgments that a change did not occur in the data patterns across two conditions? and (c) What is the total error rate for each overlap method (PND, PDO^2, PEM, and PEM-T) when visual analysts' judgments are used as the criterion?

Method

Selection and Preparation of Stimuli (Graphs)

A total of 160 graphs were selected randomly from the *Journal of Applied Behavior Analysis*, one graph from each issue for Volumes 1 through 40. Selected graphs had two adjacent conditions, with each condition having at least three data points. The following steps were used to select the graphs. First, for each issue of Volumes 1 through 40, a random number table was used to select one article. Second, when more than one figure existed in the selected article, a figure was selected randomly. If only one figure existed in the article, the selection process moved to Step 3. If the selected article did not have a figure with two adjacent conditions (each with at least three data points), another article from the same issue was selected. If the selected figure did not have two adjacent conditions that each had at least three data points, another figure was selected randomly from the same article. If none of the figures in an article had two adjacent conditions with each containing at least three data points, another article from the same issue was selected and the process was repeated. Third, for the selected figure, all adjacent condition changes for each data series on the figure were numbered—often, multiple data paths were graphed on the same figure, and each was numbered separately. Fourth, a random number table was used to select the adjacent conditions and data series. The adjacent conditions were selected without regard for whether they were baseline or intervention conditions; they were selected if they were adjacent and had at least three data points in each condition. This rule was based on the fact that all condition changes are evaluated when attempting to draw conclusions from single-subject research studies. The conditions were selected regardless of whether the purpose of the study was to accelerate or decelerate the measured behaviors; however, the intent of the condition change (i.e., to produce an increase or decrease in data) was identified and labeled.

When adjacent conditions were selected for each issue of Volumes 1 through 40, the value of each datum point was identified. The values were determined by using a transparent straightedge containing a grid of vertical and horizontal lines. A vertical line on the straightedge was placed over the ordinate (y-axis) of the graph, and the straightedge was moved up or down until a horizontal line intersected with a datum point. The value of the horizontal line intersecting with the ordinate was estimated, and the estimated values were tabled.

Graphs were constructed in Microsoft PowerPoint by using the procedures described by Barton, Reichow, and Wolery (2007). Based on the original figures, the ordinates were labeled as percentage or number of responses. The two adjacent conditions were labeled as Condition 1 and Condition 2, with Condition 1 on the left and Condition 2 on the right of the figure. The abscissa was labeled as sessions. The number of data points in each condition and in the graph was dependent on the number in the original figure. Each resulting graph was placed on a single page and numbered by the volume and issue from which it was selected. No information was presented about the behavior measured, goal of the study, or the components and parameters of the actual conditions. The selected graphs had a mean of 14.5 data points (range 6–24, median 14). The first condition had a mean of 6.8 data points (range 3–16, median 6). The second condition had a mean of 7.9 data points (range 3–19, median 7).

Procedures for Visual Analysts' Judgments

To determine whether the selected graphs contained data patterns that would be judged as different across conditions, the four authors independently viewed each of the 160 graphs and answered the following question: "Did a change exist in the data from Condition 1 to Condition 2?" No instructions were given to define change or no change. A binary (yes–no) choice was provided. After the four judges rated all graphs, their records were compared and the graphs were sorted into three groups: (a) all judges agreed a change occurred, (b) all judges agreed a change did not occur, and (c) judges disagreed about whether a change occurred.

Calculations of Overlap Methods

The four overlap methods were calculated for the graphs for which the four judges agreed a change in the data occurred across conditions and those for which they all agreed no change existed. A graph depicting how each overlap method was calculated is shown in Figure 1. For analysis purposes, the proportions were calculated rather than the percentages.

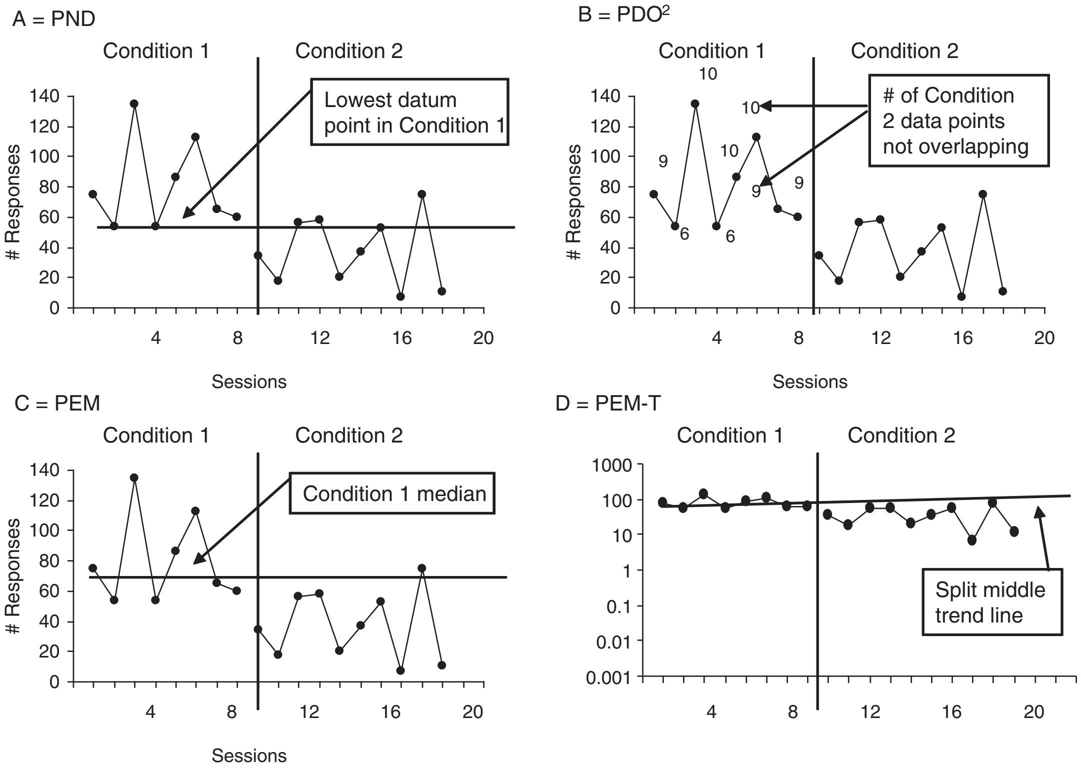
For PND, the steps were as follows:

1. Identify the intended change (increase or decrease) in the data from Condition 1 to Condition 2.
2. Identify the greatest or least datum point in Condition 1 (greatest if the intended change was to produce an increase in data and least if the intended change was to produce a decrease).
3. Draw a horizontal line through the data at the value of the datum point identified in Step 2.
4. Count the number of data points in Condition 2 above or below the line (on the therapeutic side) drawn in Step 3.
5. Divide the count from Step 4 by the total number of data points in Condition 2.

For PDO^2, the steps were as follows:

1. Identify the intended change.
2. For each datum point in Condition 1, count the number of data points in Condition 2 above or below (on the therapeutic side) its value (pairwise comparison).

Figure 1
Graphic Depiction of the Four Overlap Methods



Note: Graph A represents the calculation of the percentage of nonoverlapping data (PND), Graph B represents the calculation of the pairwise data overlap squared (PDO^2), Graph C represents the calculation of the percentage of data exceeding the median (PEM), and Graph D represents the calculation of percentage of data exceeding a median trend (PEM-T).

3. Sum all counts for all data points in Step 2.
4. Count the number of data points in Condition 1.
5. Count the number of data points in Condition 2.
6. Multiply the two counts (Steps 4 and 5) to determine the total number of pairwise comparisons.
7. Divide the sum from Step 3 by the product from Step 4.
8. Square the quotient.

For PEM, the steps were as follows:

1. Identify the intended change.
2. Identify the median of the data in Condition 1.
3. Draw a horizontal line through all the data at the Condition 1 median.
4. Count the number of data points in Condition 2 above or below the line (on the therapeutic side) drawn in Step 3.
5. Divide the count from Step 4 by the total number of data points in Condition 2.

For PEM-T, the steps were as follows:

1. Identify the intended change.
2. Graph the data on a semi-logarithmic chart.
3. Calculate and draw the split middle line of trend estimation through the Condition 1 data (White & Haring, 1980) and extend it through Condition 2.
4. Count the number of data points in Condition 2 above or below the trend line (on the therapeutic side) drawn in Step 3.
5. Divide the count from Step 4 by the total number of data points in Condition 2.

Results

Visual Analysis

The four judges used visual analysis and all agreed that 94 (58.8%) graphs showed a change in the data pattern from Condition 1 to 2. They all agreed that 27

Table 1
Number and Percentage of Graphs ($N = 94$) Judged as Having a Change in Data Patterns by Levels of Effect for Each Overlap Method

	Level of Effect (Scruggs & Mastropieri, 1998)		
	Questionable or Not an Effective Treatment (<0.70)	Effective Treatment (0.70–0.89)	Very Effective Treatment (0.90–1.0)
Overlap Method	Number (%)	Number (%)	Number (%)
PND	21 (22.3)	22 (23.4)	51 (54.3)
PDO^2	20 (21.3)	22 (23.4)	52 (55.3)
PEM	4 (4.3)	12 (12.8)	78 (83.0)
PEM-T	8 (8.5)	8 (8.5)	78 (83.0)

Note: PND = percentage of nonoverlapping data; PDO^2 = pairwise data overlap squared; PEM = percentage of data exceeding the median; PEM-T = percentage of data exceeding a median trend.

(16.9%) graphs showed no change in data patterns. Thus, the four judges agreed on 121 (75.6%) of the 160 graphs. For 13 (8.1%) graphs, three of four judges agreed a change in data patterns existed; for 13 (8.1%) graphs, three of four judges agreed a change did not exist across the two conditions; and for the remaining 13 (8.1%) graphs, two judges indicated a change occurred and two indicated no change occurred. The disagreements were not because of a single rater; rather, each rater disagreed with the others on some graphs. The 121 graphs on which the four judges agreed were used in subsequent analyses.

Agreement Between Overlap Methods and Visual Analysts' Judgments

Scruggs and Mastropieri (1998) provided the following guidelines to categorize effects using the PND: 0.90 to 1.0 is a very effective treatment, 0.70 to 0.89 is an effective treatment, and a proportion of less than 0.70 is questionable or not an effective treatment. These guidelines were used to evaluate the four overlap methods, because these are the only published standards for evaluating the overlap methods. For the graphs the four judges rated as having a change in data patterns, the number and percentage of graphs for each overlap method within these three categories are shown in Table 1. The PEM and PEM-T reported the highest percentage of graphs as having very effective treatments (83.0%). The PEM had the most graphs when the effective and very effective treatments categories were summed (95.7%); this was followed by the PEM-T, which had 91.5% for the two categories. The PDO^2 and PND had lower percentages of very effective treatments (55.3% and 54.3%, respectively) and effective plus very effective treatments (78.7% and 77.7%, respectively). For these graphs, the PEM had the lowest percentage of graphs as not having an effect (4.3%), followed by the PEM-T (8.5%). The PDO^2 and PND had much higher percentages, 21.3 and 22.3, respectively.

The number and percentage of graphs judged by all visual analysts as not having a change in the data patterns are shown for the four overlap methods in Table 2. A proportion of less than .70 was used as the criterion for not being effective. The PND had the highest agreement with the visual analysts (92.6%), which was followed by PDO^2 (74.1%), PEM-T (70.4%), and PEM (40.7%). Of these graphs, those rated by the overlap metrics as having effective or very effective treatments were PND (7.4%), PDO^2 (25.9%), PEM-T (29.6%), and PEM (59.3%).

No overlap metric had the highest agreement with visual analysts for the two types of data patterns, graphs with and without a change. Thus, the total error percentages were calculated for each overlap method and are shown in Table 3; the percentages in Table 3 represent the percentage of errors (disagreement with visual analysts' judgments). For graphs rated by visual analysts as having a change in the data patterns, an error was considered any proportion below .70; and for graphs rated by visual analysts as having no change in data patterns, any proportion of .70 or greater was considered an error. PDO^2 had the highest error percentage (22.3%) of the four overlap methods—slightly more than one fifth of the graphs resulted in errors. The PND had a slightly lower error percentage (19%), but nearly one fifth of the graphs resulted in errors. The PEM had an error percentage of 16.5, and the PEM-T had the lowest error percentage, at 13.2.

Table 2
Number and Percentage of Graphs (*N* = 27) Judged as Not Having a Change in Data Patterns by Levels of Effect for Each Overlap Method

	Level of Effect (Scruggs & Mastropieri, 1998)		
	Questionable or Not an Effective Treatment (<0.70)	Effective Treatment (0.70–0.89)	Very Effective Treatment (0.90–1.0)
Overlap Method	Number (%)	Number (%)	Number (%)
PND	25 (92.6)	1 (3.7)	1 (3.7)
PDO^2	20 (74.1)	5 (18.5)	2 (7.4)
PEM	11 (40.7)	8 (29.6)	8 (29.6)
PEM-T	19 (70.4)	2 (7.4)	6 (22.2)

Note: PND = percentage of nonoverlapping data; PDO^2 = pairwise data overlap squared; PEM = percentage of data exceeding the median; PEM-T = percentage of data exceeding a median trend.

Table 3
Total Error Percentage for Overlap Methods on Graphs (*N* = 121) Judged by All Visual Analysts as Having and Not Having a Change in Data Patterns

		Visual Analysts' Judgments		
Overlap Metric	Size of Effect	Change in Data Pattern	No Change in Data Pattern	Total Error Percentage
PND	0.70–1.0	73	2 (7.4%)	19.0
	<0.70	21 (22.3%)	25	
PDO^2	0.70–1.0	74	7 (25.9%)	22.3
	<0.70	20 (21.3%)	20	
PEM	0.70–1.0	90	16 (59.3%)	16.5
	<0.70	4 (4.3%)	11	
PEM-T	0.70–1.0	86	8 (29.6%)	13.2
	<0.70	8 (8.5%)	19	

Note: PND = percentage of nonoverlapping data; PDO^2 = pairwise data overlap squared; PEM = percentage of data exceeding the median; PEM-T = percentage of data exceeding a median trend.

Discussion

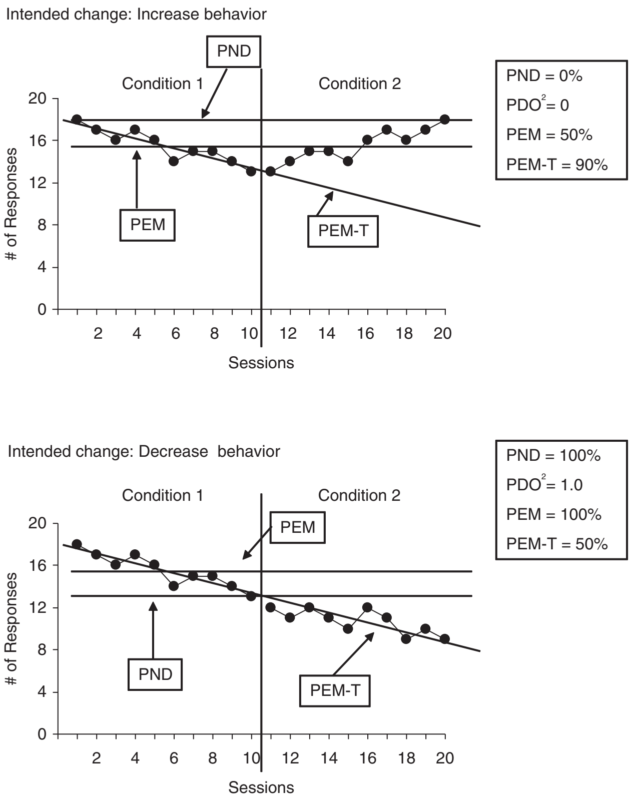
The purpose of this study was to determine the extent to which four overlap methods for conducting quantitative syntheses of single-subject data agreed with the judgments of visual analysts. The four judges using visual analysis all agreed a change occurred or did not occur on 121 of 160 graphs. For the graphs on which all judges agreed a change occurred, the PEM and PEM-T had higher agreement than did the PND and PDO^2. This finding replicates Ma's (2006) study showing the PEM was in greater agreement with authors' conclusions than was the PND. However, the PND and PDO^2 had more agreement than the PEM and PEM-T for the graphs visual analysts judged as not having a change in data patterns. This finding seems to replicate the findings of Parker and Hagen-Burke (2007) and Parker and Vannest (2007), who indicated that the PEM does not appear to discriminate graphs well and likely overestimates effects. The PEM-T had the lowest total error percentage, yet the PEM-T had errors on nearly an eighth of the graphs. Thus, for a quantitative synthesis, one of eight judgments about data change in two adjacent conditions is likely to result in errors. If each study had a minimum of three condition changes about which judgments were made (many studies will have more), an error is likely to occur once in every three studies—given the lowest error rate found in this study. Taken together, the data indicate these four overlap methods had error percentages exceeding acceptable limits. Thus, although these methods are not compromised by the serial dependency in the data, as mean difference effect sizes and regression-based effect sizes are, they do not discriminate well between graphs with and without changes in data patterns.

The overlap methods, however, have additional weaknesses, and four of these are discussed below. First, the overlap methods are not based on all of the data's characteristics. Time series data are characterized by their level, trend, and stability or variability. Furthermore, changes from one condition to another can occur in each of these characteristics separately or in combination (Kazdin, 1984). The overlap methods, by their calculations, are designed to detect changes in level across conditions. The PEM-T also can detect changes in trend across conditions, and this may account for why it had the lowest error percentage. The PND, PDO^2, and PEM do not detect changes in trend. As shown in Figure 2, failure to detect changes, or lack thereof, in trend across conditions can lead to inaccurate conclusions. None of the overlap methods is designed to detect changes in variability, which is a critical element in analyzing time series data and can be a desired outcome. Failures to detect trend changes and to account for variability are serious weaknesses in most of the overlap methods, making them prone to errors.

Second, the number of data points in the second condition influences the percentage of overlap. For example, if one datum point from Condition 2 overlaps with the Condition 1 data, then there must be at least 10 data points in Condition 2 to obtain 90% nonoverlap and thus a strong effect. The conditions selected for this study were substantially shorter (means of 6.8 and 7.9) than 10 data points. Thus, the shorter the second condition, the greater the influence of one overlapping datum point in the percentage calculation. Investigators with an effective independent variable can simply continue to collect data (even after an effect has been demonstrated) and thereby increase the size of the effect as calculated by the overlap methods.

Third, the overlap methods are not an estimate of the magnitude of the effects between conditions, although they are thought to represent magnitude. Effect sizes for group experimental research are estimates of magnitude. However, as shown in Figure 3, even when there is 100% nonoverlapping data between conditions, the magnitude of the data (size of the effect) can be quite different. This failure to estimate magnitude results in two problems. The estimates from the overlap methods cannot be integrated with effect sizes from group research, because they measure different elements. In addition, the meaning of the percentage of overlap is limited. Specifically, it reports only the proportion of overlap across the conditions—nothing more, nothing less—and it is not an estimate of the magnitude of effects.

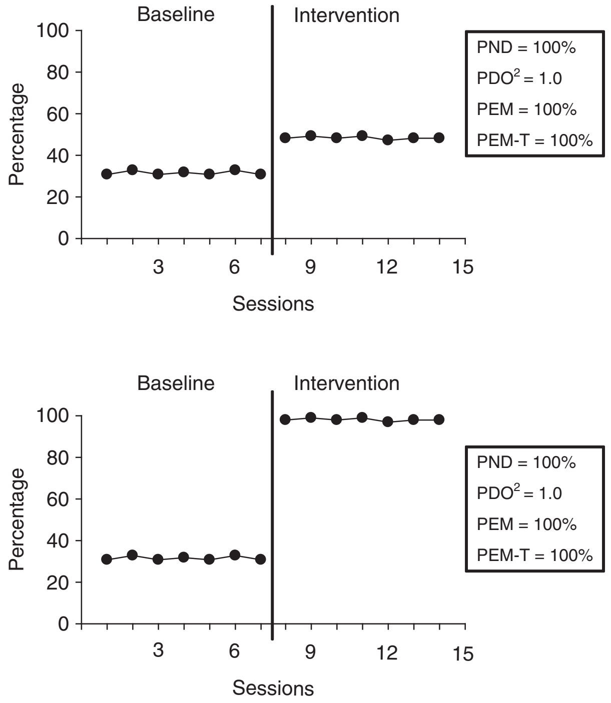
Figure 2
Overlap Methods and Trend



Note: The top panel shows a graph from a study intended to increase the behavior in the second condition, and a change in trend from decelerating to accelerating occurred when the conditions changed. The percentage of nonoverlapping data (PND), pairwise data overlap squared (PDO^2), and percentage of data exceeding the median (PEM) indicated no change occurred; the percentage of data exceeding a median trend (PEM-T) indicated a strong change occurred. The bottom panel shows a graph from a study intended to decrease the behavior in the second condition, and no change occurred when the conditions changed (trend from Condition 1 continued in Condition 2). The PND, PDO^2, and PEM indicated a strong change occurred; the PEM-T indicated no change occurred.

Fourth, the overlap methods fail to use the replication logic inherent in single-subject experimental research. A defining feature of single-subject research is that evidence that an effect exists is established through replication (Edgar & Billingsley, 1974;

Figure 3
Failure of Overlap Methods to Assess Magnitude



Note: Both graphs by visual analysis show no overlapping data between baseline and intervention conditions, and all overlap methods indicate strong effects. However, the magnitude of change in the two graphs is quite different, with the lower panel having a larger magnitude of change than the top panel. PND = percentage of nonoverlapping data; PDO^2 = pairwise data overlap squared; PEM = percentage of data exceeding the median; PEM-T = percentage of data exceeding a median trend.

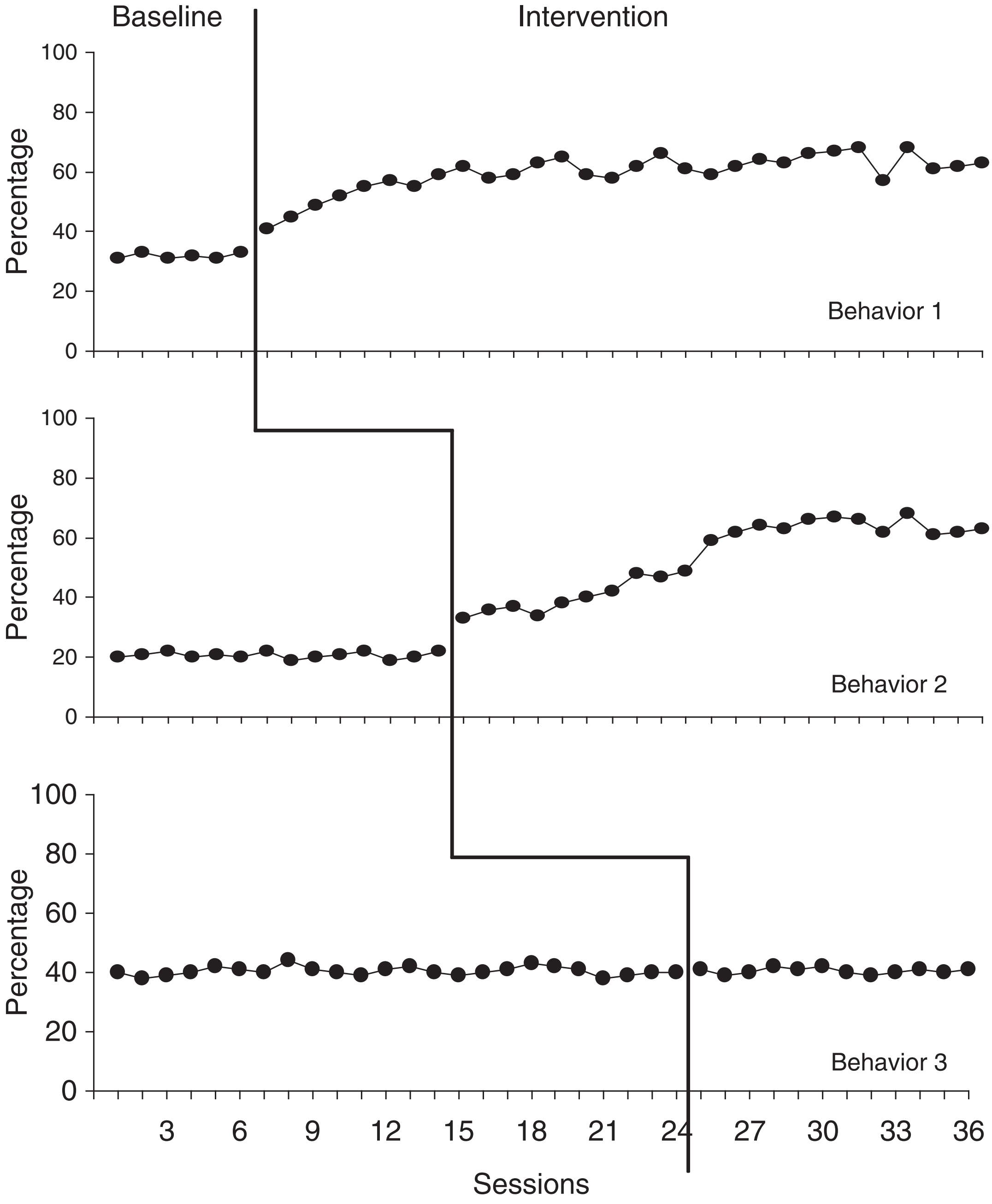
Kennedy, 2005; Tawney & Gast, 1984). Thus, a consistent change in the data pattern with each intra- and intersubject replication of a study is required to document that a functional relation exists between the independent and dependent variables. Ideally, those consistent changes also would be large, but the consistency of change is necessary to draw a conclusion about the likely presence of a functional relation. In Figure 4, a multiple-baseline design across three behaviors is shown. Because a change in the data pattern occurred for only the first two behaviors but not the third, the investigator cannot conclude that a functional relation is present. However, the percentage of nonoverlap was 100, 100, and 50, which would produce a mean percentage of nonoverlap of 83.3, which is classified as a moderate effect. Failure to attend to the consistency of change across the replications in a study, which is the central logic of single-subject studies, invalidates the overlap methods.

This study has limitations. First, the four visual analysts agreed on only 121 of 160 graphs. The high percentage of disagreement may have been due, in part, to the binary judgment they were asked to make (i.e., change or no change). When attempting to make this decision, our actual reaction to the graphs was “I am not sure” or “There is not enough data to tell.” Second, the visual analysts looked at only two adjacent conditions; thus, this level of agreement should not be equated with the types of judgments visual analysts make about complete figures containing all of the replications. We suspect the percentages of agreement would be higher among visual analysts if they inspected complete figures and made judgments about whether a functional relation was present rather than whether change occurred across two conditions. The overlap methods also were applied to only two adjacent conditions rather than the entire figure. It may be interesting to study the agreement between visual analysts’ judgments of functional relations and the overlap methods when the entire figure is used. Third, in this study we used the terms *questionable or not effective treatment*, *effective treatment*, and *very effective treatment*, following the lead of Scruggs and Mastropieri (1998). However, the overlap methods were applied only to two adjacent conditions and not to the entire study data; thus, the proportions presented do not represent the actual effectiveness or lack thereof of the variables used in the actual studies. Fourth, although the point is irrelevant in our view, the patterns shown by the PND and PDO^2 versus the PEM and PEM-T are likely because of using extreme datum points (PND and PDO^2) or the median (PEM and PEM-T).

In summary, the overlap methods fail to detect all of the characteristics of time series data (i.e., trend and variability), are compromised by the number of data points in the second condition, are not an estimate of the magnitude of effects, and do not use the replication logic inherent to single-subject research. These weaknesses taken together with the major finding of this study, which indicates their error percentages are unacceptably high (Table 3), lead to an inescapable conclusion: The overlap methods are inappropriate for quantitatively summarizing single-subject data. Given the inappropriateness of mean difference and regression-based effect size calculations, the field does not currently have a suitable synthesis method.

Given the press to synthesize the single-subject literature quantitatively, the field needs to devise a new metric. Although the calculation for such a metric is

Figure 4
A Failure to Replicate



Note: A multiple-baseline design across three behaviors in which a change in data patterns occurred for Behaviors 1 and 2 but not for Behavior 3.

not known, the characteristics of such a method can be proposed. First, and perhaps most important, it should focus on the replication logic of single-subject designs. Thus, it should take into account the consistency of effects across replications in each study. Second, it should use all the data of the study. This includes changes from baseline to intervention conditions as well as those from intervention to baseline conditions. Third, it should be an estimate of the magnitude of the effects across those replications. Fourth, it should take into account all the characteristics of the data, including level, trend, and variability. Fifth, it should be in high agreement with careful visual analysis. Sixth, it should not violate the assumptions about the nature of the data, such as serial dependency. Seventh, it should have some method of allowing analyses of moderator variables. If a metric is devised with these characteristics, then other issues must be addressed. Reviewers must determine whether the baseline conditions of the reviewed research are sufficiently similar to be considered the same, and thus a legitimate, comparison condition. A closely related issue is whether baseline conditions are adequately described to allow sound judgments about whether they are similar across studies (Lane, Wolery, Reichow, & Rogers, 2007). Finally, the procedural variations that are common across systematic replications of independent variables in single-subject research must be judged to be similar or different from one another. Specifically, how much procedural variation can occur before an independent variable is judged to be different in one study from another.

In conclusion, quantitatively synthesizing single-subject research is a laudable goal, because it would increase the objectivity of syntheses, allow quantification of the potential effects, and allow analyses of moderator variables. However, at this time, the field does not have a suitable method for calculating effects for studies that would allow them to be synthesized appropriately. A need exists for such a method, and a method should be devised that considers the replications inherent in single-subject research, is not compromised by serial dependency, and accounts for variability and trend as well as level changes.

References

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

Barton, E. E., Reichow, B., & Wolery, M. (2007). Guidelines for graphing data with Microsoft® PowerPoint™. *Journal of Early Intervention*, *29*, 320–336.

Busick, M. (2008, February). *Estimating effect sizes for single subject research: An evaluation of the reliability of the regression method*. Paper presented at the Conference on Research Innovations in Early Intervention, San Diego, CA.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwil & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 133–157). Hillsdale, NJ: Lawrence Erlbaum.

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

Center, B. J., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, *19*, 387–400.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Dunst, C. J., Hamby, D. W., & Trivette, C. M. (2004). Guidelines for calculating effect sizes for practice-based research synthesis. *Centerscope*, *2*(2), 1–10.

Dunst, C. J., Trivette, C. M., & Cutspec, P. A. (2002). Toward an operational definition of evidence-based practices. *Centerscope*, *1*(1), 1–10.

Edgar, E. B., & Billingsley, F. F. (1974). Believability when N=1. *Psychological Record*, *24*, 147–160.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Hillsdale, NJ: Lawrence Erlbaum.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practices in special education. *Exceptional Children*, *71*, 165–179.

Kazdin, A. E. (1984). Statistical analysis for single-case experimental designs. In D. Barlow & M. Hersen (Eds.), *Single case experimental designs* (pp. 285–324). Boston: Allyn & Bacon.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston: Allyn & Bacon.

Lane, K., Wolery, M., Reichow, B., & Rogers, L. (2007). Describing baseline conditions: Suggestions for study reports. *Journal of Behavioral Education*, *16*, 224–234.

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

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

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*, 137–148.

Parker, R. I., & Hagen-Burke, S. (2007). Median-based overlap analysis for single case data: A second study. *Behavior Modification*, *31*, 919–936.

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

Parker, R. I., & Vannest, S., (2007). *Pairwise data overlap for single case research.* Unpublished manuscript.

Salzberg, C. L., Strain, P. S., & Baer, D. M. (1987). Meta-analysis for single-subject research: When does it clarify, when does it obscure? *Remedial and Special Education*, *8*, 43–48.

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

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.

Slavin, R. E. (1987a). Ability grouping student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, *57*, 293–336.

Slavin, R. E. (1987b). Best-evidence synthesis: An alternative to meta-analysis and traditional reviews. In W. R. Shadish & C. S. Reichardt (Eds.), *Evaluation studies: Review annual* (Vol. 12, pp. 667–673). Thousand Oaks, CA: Sage.

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment*, *9*, 113–124.

Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education.* Columbus, OH: Charles Merrill.

White, O. R. (1987). The quantitative synthesis of single-subject research: Methodology and validation. Comment. *Remedial and Special Education*, *8*, 34–39.

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

Wolery, M., Busick, M., Reichow, B., & Barton, E. (2008, February). *Quantitative synthesis of single subject research*. Paper presented at the Conference on Research Innovations in Early Intervention, San Diego, CA.

Mark Wolery, PhD, is a professor of special education at Peabody College, Vanderbilt University. His current interests include transfer of stimulus control and single-subject research methods.

Matthew Busick, MEd, is a doctorate student in special education at Peabody College, Vanderbilt University. His current interests include educational interventions for children with autism and single-subject research methods.

Brian Reichow, PhD, is a postdoctoral associate at the Yale University Child Study Center. His current interests include interventions for young children with autism and the translation of research to practice.

Erin E. Barton, PhD, is an assistant professor of special education at the University of Oregon. Her current interests include teaching play skills to children with disabilities and single-subject research methods.