An Analysis of Effect Sizes for Single-Subject Research:
A Statistical Comparison of Five Judgmental Aids

Lee L. Mason
The University of Texas at San Antonio

Effect sizes for single-subject research were examined to determine to what extent they measure similar aspects of the effects of the treatment. Seventy-five articles on the reduction of problem behavior in children with autism were recharted on standard celeration charts. Pearson product-moment correlations were then conducted between two previously unexamined effect sizes, celeration and celeration change, as well as three more common statistics, the mean baseline reduction, the percentage of non-overlapping data, and the percentage of zero data. Significant correlations were found for both celeration and celeration change, suggesting that these and other effect sizes measure somewhat similar aspects of the effect of the treatment. These findings and limitations are discussed within the broader context of evidence-based practices in education.

The Use of Judgmental Aids in Single-Subject Research

Recent legislation, such as the No Child Left Behind Act of 2001 and the Individuals with Disabilities Education Improvement Act of 2004, calls for the use of evidence-based practices to make curricular and instructional decision in the classroom. Underlying this legislation is the assumption that educators will select interventions that would provide the strongest benefit for their student population. Evidence-based practices are research-validated instructional techniques that have met rigorous standards for research design, methodological quality, and the magnitude of the effect. Randomized controlled trials and meta-analyses, which rely on statistical evaluation, typically identify evidence-based practices by examining effect sizes that measure the magnitude of the effect of an intervention (Cohen, 2001). On the other hand, single-subject research relies on the use of visual analysis in “reaching a judgment about the reliability or consistency of intervention effects by visually examining graphed data” (Kazdin, 1982, p. 232). As a result, comparisons across studies become somewhat more subjective. Furthermore, rather than determining effect sizes across groups of participants, single-subject designs compare the effect of an intervention with an alternative treatment or an adjoining phase.

Parker, Vannest, and Brown (2009) note that even the best visual analyses are commonly supported by simple statistical heuristics. According to Michael (1974), who preferred the plain English term “judgmental aids” rather than “statistics,” these numbers are simply stimuli that more easily elicit responses from researchers and practitioners than raw data alone. For instance, oral reading fluency has been shown to be sensitive to instructional changes (Good & Kaminski, 2003; Shinn, 1989), and it is frequently used as a measure...
to evaluate the effects of reading interventions. However, sequential assessments with a single individual typically show some random variability or “bounce” in addition to the actual changes in reading skill. This variability in oral reading rate can reduce the measure’s sensitivity to changes in reading skill, thereby hindering its effectiveness for monitoring progress in reading. In such cases, judgmental aids may be more helpful in describing the overall efficacy of the intervention.

Over the years, researchers have offered many suggestions for summarizing and synthesizing single-subject research in terms of trend, slope, and variability. Some of the many examples are the percentage of non-overlapping data (PND; Scruggs, Mastropieri, & Castro, 1987), the percentage of zero data (PzD; Scotti, Evans, Meyer, & Walker, 1991), the mean baseline reduction (MBLR; Kahng, Iwata, & Lewin, 2002), the C statistic (Nourbakhsh & Ottenbacher, 1994), the percentage of all non-overlapping data (PAND; Parker, Hagan-Burke & Vannest, 2007), Kruskal-Wallis W, and the improvement rate difference (IRD; Parker et al., 2009).

Campbell (2003; 2004) synthesized the literature for reducing problem behavior in persons with autism by quantifying 117 single-subject research articles and comparing the effect sizes for the PND, PzD, MBLR, and regression-based d metrics. Pearson’s product-moment correlations between all four were found to be statistically significant, except for PzD and d. This finding suggests that each effect size provides a similar interpretation of the data, so that multiple measures (i.e., both PND and PzD) are unnecessary. Campbell (2004) calls for future research to continue comparing and contrasting additional effect sizes so as to better understand their use in summarizing single-subject research.

One measure of single-subject research, which has long been used to measure change in frequency over time, is celeration (Graf & Lindsley, 2002; McGreevy, 1983; White & Haring, 1980). A celeration line is a trend line, drawn through multiple behavioral frequencies on a standard celeration chart (SCC), which quantifies the amount of learning over a given period of time. A frequent criticism of visual analysis in single-subject research is the lack of formal decision rules for analyzing data (Nourbakhsh & Ottenbacher, 1994). However, with standard displays such as the SCC, multiple practitioners interpret the same data in a more consistent manner: They bring the viewer’s reaction under control of the data, rather than the less pertinent features of the graph (e.g., scale; Johnson & Pennypacker, 1993).

Using the SCC, a specific value is computed for each celeration line, thereby providing a judgmental aid for comparing celerations. Celeration offers the rate of behavior over time as the measure of effect. Clearly, a reading intervention designed to increase words correct per minute with a celeration of x2.0 has a greater effect than a similar intervention with a celeration of x1.4. Even though celerations are frequently compared with one another to measure the effects of behavioral interventions, celeration has not yet been systematically compared with other types of single-subject effect sizes. The purpose of this study is therefore to examine the extent to which celeration and celeration change relate to PND, PzD, and MBLR. Specifically, this research sought to answer the following question: To what extent does celeration offer a unique effect size for single-subject research?

**METHOD**

**Selection of Studies**

Campbell (2003; 2004) identified the 117 articles used in this research. According to an a priori power analysis, this sample size was sufficient for computing a Pearson product-moment coefficient (r; Faul, Erdfelder, Lang, & Buchner, 2007) to examine the correlation between celeration and other measures of single-subject effect size. Individual data sets were selected, based on four criteria:

1. Only single-subject research was included to ensure that behavioral data for each participant were readily available.
2. Baseline and treatment phases in each single-subject design had to be presented as repeated measures.
3. Treatment targeted the reduction of problem behavior (e.g., self-injurious behavior, stereotypy, aggression, or property destruction).
4. At least one participant was diagnosed with autism.
If the article included multiple participants, only the behavers who fit these criteria were included in this analysis.

**Single-Subject Effect Sizes**

As noted, a variety of methods can be used to summarize single-subject data. Three of the more common methods found throughout single-subject literature are the percentage of non-overlapping data (PND), the percentage of zero data (PzD), and the mean baseline reduction (MBLR). The PND summarizes the effects of treatment by counting the number of data points in the intervention phase that do not overlap with the highest or lowest data points in the baseline phase, dividing by the total number of data points in the treatment phase, and multiplying by 100 (Scruggs et al, 1987; Scruggs, Mastropieri, Cook, & Escobar, 1986). Figure 1 shows hypothetical data on an intervention designed to reduce self-injurious behavior (SIB). The circled data point in the baseline phase represents the lowest level of SIB observed during baseline. A dashed line has been extended from this point into the intervention phase. The three data points circled in the intervention phase are those overlapping with the lowest data point in the baseline phase. The PND for this data set is 70%.

**Figure 1. Hypothetical data demonstrating the calculation of percentage of non-overlapping data (PND).**
The PZD measures behavior reduction by locating the first data point in an intervention based on a count of zero; for the remainder of the phase, the percentage of data points remaining at zero is calculated (Scotti et al., 1991). Figure 2 presents the same hypothetical data. In this figure, the three data points that reach zero are circled, and a dashed line is drawn at the first zero data point. The PZD is calculated from this point forward and equals 50%.

Figure 2. Hypothetical data demonstrating the calculation of percentage of zero data (PZD).

The MBLR is found by subtracting the mean treatment value from the mean baseline value, next dividing by the mean baseline value, and then multiplying the result by 100 (Kahng et al., 2005). Figure 3 shows the hypothetical data set once again. The average count of the 5 observations in the baseline is 7, whereas the average of the 10 observations in treatment is 2.3. These are calculated to give a MBLR of 67%.
This analysis also examined the celeration line of the first treatment phase and the celeration change between the baseline and the intervention. To calculate the celeration lines and the MBLR, the graphically presented data were converted to raw numbers. Using a drafting divider, the distance between the horizontal axis and each data point was measured in millimeters and rounded to the nearest half-millimeter (Huitema, 1985). An approximate value was then produced by measuring this distance against the vertical axis of the same graph. This data-conversion procedure has been used with a high degree of reliability (Allison, Faith, & Franklin, 1995; Kahng et al., 2005; Skiba, Casey, & Center, 1985-86).

Recharting on the Standard Celeration Chart

To compare celeration with the above-listed effect sizes, the data in each article were recharted on the SCC. The only graphs considered were those with a behavior or product of a behavior on the vertical axes and a unit of time on the horizontal axes. Using the guidelines Porter (1985) provided, each of the 117 articles was screened and recharted. A summary of these procedures follows.

The Dpmin-11EC SCC was used to replot data from each article. This chart consists of calendar days along the horizontal axis, allowing for a comparison of studies that use various observation schedules (e.g., daily versus twice weekly). Additionally, the SCC measures frequency on the vertical axis, so that studies using different

Figure 3. Hypothetical data demonstrating the calculation of mean baseline reduction (MBLR).
measures or interval lengths (e.g., number versus percent-interval) could be compared. Therefore, all the original details from the research are preserved on the SCC.

The frequencies were charted on the Microsoft Excel Standard Celeration Chart Template (Harder, 2008). A new chart was used for each data set from each study. In some cases, as with multi-element designs, the same baseline was used with multiple intervention phases — each replotted on its own chart. Record floors and ceilings were marked with dashes, and data points were placed between. Frequencies based on a count of zero were plotted $\div 2$ below the record floor (White & Neely, 2004).

Separate celeration lines were drawn for both the initial baseline and the first intervention phase. For the purposes of comparing effect sizes, both the celeration of the first intervention phase and the celeration change between the baseline and intervention phases were recorded for every chart. Celeration lines were automatically computed for each phase by the Excel Standard Chart Template, using the median slope method (White, 2005). The median slope is found by drawing lines passing through all possible pairs of data points, then selecting the line that falls in the middle of that array. If all the slopes in the distribution are arranged in numerical order and there is an odd number of scores, the median slope would be the score in the middle. With an even number of slopes, either the line representing the most conservative slope can be selected, or the two middle slopes can be averaged. White (2005) notes that the median slope is generally more useful in predicting future performance than other methods of calculating trend lines.

Celeration changes were determined by comparing the celeration of the baseline phase to the celeration of the intervention phase. Using the same hypothetical data as above, Figure 4 displays a celeration turn down from $x1.3$ to $+3.1$. This yields a celeration change of $+4.03$.

Figure 4. Hypothetical data demonstrating the calculation of celeration and celeration change.
Celeration lines were not calculated for any phase that had fewer than five daily frequencies. In cases where the intervention had fewer than five data points, the data set was excluded. If the baseline phase contained fewer than five data points but the intervention phase had at least five points, the intervention celeration was calculated, but the celeration change could not be determined.

Each article was closely examined to determine the frequency of observation. When this information was not provided, an assumption was made of once daily excluding weekends. When an article listed multiple sessions per day, only the initial daily data point was recharted. For example, if an article stated that two sessions were run each day, only the sequentially odd-numbered data points were replotted. Articles that listed a variety of sessions (i.e., between 3 and 5 sessions run daily) were excluded.

Additional information was required to rechart percent-interval data, including the total observation time and the interval length. Articles that did not include this information could not be recharted. Recharting percent-interval data requires converting each data point to an assumed frequency. However, three factors must be determined first: (a) the record floor, (b) the record ceiling, and (c) the total number of intervals observed in each session.

The minimum frequency that can be recorded during a session is called the record floor. In percent-interval graphs, this is the total observation time. For most articles, the observation time remained constant throughout the study. If observation time was given as a range (e.g., sessions ranging from 10 to 15 minutes), the shorter observation time was used as the record floor. When interrupted-interval recording procedures were used (e.g., a 5-second observe, a 5-second record cycle used for 10 minutes), only the actual observation time was used as the record floor.

The maximum frequency that can be recorded during a session is called the record ceiling. This is directly defined by the interval length used in each study. To find the record ceiling, divide 60 by the interval length (e.g., 60 divided by 6-second intervals yields a record ceiling of 10).

For converting a percentage of intervals to a frequency estimate, the total number of intervals observed in each session is needed. This can be found by multiplying the record floor by the record ceiling (e.g., a record floor of 10 multiplied by a record ceiling of 10 equals 100 intervals). A percentage of intervals can then be converted to the number of intervals by multiplying the percentage by the number of intervals observed (e.g., 75% of 100 intervals equals 75 intervals scored). Finally, dividing the number of intervals observed by the observation time yields a frequency estimate (Porter, 1985). This number can now be recharted on the SCC.

RESULTS

This study examined the extent to which celeration offers an independent effect size for single subject research. Of the original 117 articles Campbell (2003) identified, 75 fit the criteria for eligibility in this study. The data sets for two articles could not be located and were therefore not included in this analysis. The remaining articles examined 112 behavers, and a total of 176 behaviors that were recharted and included in this review. Interestingly, out of initial 117 articles, only two (Bierly & Billingsley, 1983; Sugai & White, 1986) originally plotted their data on standard celeration charts.

Correlation coefficients were computed among the five single-subject effect sizes by using the R statistical computing environment. The Bonferroni approach to control for Type I error was used across the 10 correlations, thereby requiring a p value of less than 0.005 to show statistical significance (0.05÷10 = 0.005; Green & Salkind, 2008). Table 1 shows that 4 out of the 10 correlations were statistically significant and were greater than or equal to 0.23. The largest correlation occurred between the celeration of the intervention phase and the celeration change of $r = 0.54$, $p < 0.001$. This is understandable since the intervention celeration is used to determine the celeration change.

A moderate correlation was found between the celeration of the intervention phase and the mean baseline reduction of $r = -0.33$, $p < 0.001$, and a small correlation was found between celeration change and MBLR, $r = -0.26$, $p = 0.002$. These negative coefficients can be explained by examining the manner in which each effect size was determined. For example, imagine the data set in which problem
behavior was high during baseline and immediately dropped to zero at the start of the intervention, where it remained. This would result in a high MBLR (e.g., 100%) and a low intervention celeration (e.g., x1.00). Conversely, a data set in which the baseline numbers were high, but gradually decreased over several intervention sessions, would result in a lower MBLR (e.g., 50%) and a greater celeration value (e.g., ÷4.00).

Another small correlation was found between MBLR and PzD, r = 0.23, p = 0.001. This is consistent with Campbell (2003, 2004), suggesting that these two effect sizes are measuring somewhat similar aspects of the effects of treatment. Conversely, no significant correlations were found between the intervention celeration or the celeration change and PND or PzD, indicating that these statistics measure different aspects of effectiveness.

**DISCUSSION**

Single-subject research has always relied on the graphical analysis of data to determine the effects of an intervention. This is primarily done by comparing level, trend, or variability across phases. Although several researchers have attempted to convert these effects into numbers that can be compared across studies, no single statistic appears to account for all methods of visual analysis. The data presented here suggest that celeration and celeration change are independent evaluations of single-subject research, which measure an effect that is entirely unrelated to PND and PzD. One reason for this may be because celeration measures slope, whereas the other statistics measure level or variability.

An interim step in determining effect size may be to select the appropriate statistic based on visual analysis. That is, multiple graphs demonstrating a change in level may then be compared using PND or PzD, whereas celeration or an improvement rate difference may be used to compare graphs showing a change in slope. What is important to note in the present study is that the mean baseline reduction did show some amount of correlation with both celeration and celeration change. Therefore, the effect sizes measuring level and slope are not mutually exclusive. To date, there has been no consensus on which effect sizes best represent raw data.

This research has other limitations that must be addressed. Most notably, recharting data does not result in a true frequency. Interval recording produces only an estimate of the actual frequency of behavior. Additionally, converting intervals to a percentage and back again results in some error (Porter, 1985). As a result, many of the charts included in this study were not precise.

Although the number of publications about individuals with autism continues to rise, there is an obvious dearth of data being presented on standard charts. Whether this is due to the multiply/divide scale on the SCC or the inability to manipulate axes is unclear. However, the ease with which it allows users to calculate a celeration line and compare data across charts makes a compelling argument for an increase in standard celeration charting. While the results of this study demonstrate that both celeration and celeration change are related to other single-subject effect sizes, future researchers are strongly

---

**Table 1**

<table>
<thead>
<tr>
<th></th>
<th>Celeration</th>
<th>Celeration Change</th>
<th>MBLR</th>
<th>PND</th>
<th>PzD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celeration</td>
<td>–</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Celeration Change</td>
<td>.54*</td>
<td>–</td>
<td>-.33*</td>
<td>-.26*</td>
<td>–</td>
</tr>
<tr>
<td>MBLR</td>
<td>-</td>
<td>-</td>
<td>.05</td>
<td>-.12</td>
<td>.23*</td>
</tr>
<tr>
<td>PND</td>
<td>.07</td>
<td>-.11</td>
<td>.08</td>
<td>–</td>
<td>.06</td>
</tr>
<tr>
<td>PzD</td>
<td>.07</td>
<td>-.11</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Note: MBLR = Mean Baseline Reduction; PND = Percentage of Non-overlapping Data; PzD = Percentage of Zero Data.

*p < .01
urged to continue examining and comparing additional methods for synthesizing single-subject designs.

Salzberg, Strain, and Baer (1987), as well as Michael (1974), note that the idiosyncrasies and familiarity accompanying prolonged and intense interaction with time-series data do not occur in a one-number summary. The experimenter is forced to rely on theory and other people’s research, and then attempt to draw conclusions about the relative merits of broad categories of intervention. Although Michael suggests that the use of these judgmental aids may produce a stimulus the teacher or behavior analyst can more easily react to, he cautions that these statistics may be worthwhile only when the time spent learning how to use such techniques and the effort in determining which one to use is relatively small compared with the simplifying effect achieved.

The term “effect size” has been used here to talk about comparing the effectiveness of interventions across single-subject research; however, other methods, such as metacharting, may also function to compare celerations. Lindsley, Calkin, and White (1993) emphasize the importance of analyzing chart collections, and Cooper, Kubina, and Malanga (1998) provide a variety of ways in which collections of standard celeration charts can be synchronized and displayed. Charting repeated measures not only helps users to stay connected with the data, but metacharting also allows them to make instructional or intervention decisions based on multiple sources of data (thereby also acting as a judgmental aid).

For celeration to truly function as a measure of the magnitude of effect for single-case interventions, future research should address the classification of large, medium, and small celeration effect sizes. Green and Salkind (2008) note that “as with all effect size indices, there is no good answer to the question ‘What value indicates a strong relationship between two variables?’” (p. 259). Effect size is dictated by the discipline within which the research is conducted. For celeration charting, each SCC includes a celeration fan ranging from x16 to ÷16 that may act as a guideline for talking about the magnitude of a celeration (e.g., 1.4, 2.0, and 4.0 – irrespective of sign – can be interpreted as small, medium, and large effect sizes, respectively).

For years, educators and researchers have been using data, or practice-based evidence, to make instructional decisions in their classrooms and clinics. These measures help to demonstrate that adequate progress is being made towards a specified goal. Recent educational policy may have just begun mandating the use of evidence in the classroom, but the practice is hardly new. Many practitioners have argued that the prescription of evidence-based practices results in the loss of autonomy. However, the specific educational gains of each student are more important than the generalization of practices across settings and participants. Cook, Tankersley, and Landrum (2010) conclude that evidence-based practices “will not and should not ever take the place of professional judgment but can be used to inform and enhance the decision making of special education teachers” (p. 380). Ultimately, effect size and other statistics are simply additional judgmental aids to help practitioners make data-based decisions.

REFERENCES

Note: Asterisks denote articles that were included in the synthesis.


*Bierly, C., & Billingsley, F. F. (1983). An investigation of the educative effects of...


AN ANALYSIS OF EFFECT SIZES FOR SINGLE SUBJECT RESEARCH


*Kern, L., Carberry, N., & Haidara, C. (1997). Analysis and intervention with two topographies of challenging behavior exhibited by a young...


Therapy and Experimental Psychiatry, 18, 285-288


