Reliability and Accuracy of Visually Analyzing Graphed Data From Single-Subject Designs

Kenneth J. Ottenbacher

Key Words: data collection • research design • research, method

As the number of single-subject research reports in the occupational therapy literature increases, the reliability of clinical decisions made from single-subject data must be determined. This study examined the ability of occupational therapists to reliably and accurately analyze data from single-subject designs using visual inspection of graphed data. Forty-six therapists provided a rating on whether or not a clinically significant change in performance had occurred across the baseline and treatment phases for five graphs of hypothetical data. The results revealed that considerable disagreement existed in making visual judgments based on the five graphs. The data sets were also analyzed using a recently developed statistical procedure for use with single-subject data. The level of agreement between the quantitative analysis and the visual analysis was also poor, suggesting inconsistency in the ability to make accurate visual judgments of single-subject data. The advantages and limitations of graphic presentation and visual inspection of single-subject data are discussed and the argument is made that some form of quantitative analysis should generally be used in conjunction with the visual inspection of graphed data.

Kenneth J. Ottenbacher, PhD, OTR, is Associate Professor, Occupational Therapy Program, University of Wisconsin-Madison, Madison, Wisconsin 53706.
employed in virtually all of the studies reviewed. The accuracy and reliability of judgments based on visual analysis have recently been the object of some concern (9). One of the major criticisms of graphic analysis and visual inspection of single-subject data has been the lack of any formal rules or guidelines for decision making associated with visual inspection. As a result, the process of visual analysis has received little empirical attention (10). Kazdin (7) has accurately served that "the process of visual inspection would seem to permit if not actively encourage, subjectivity and inconsistency in the evaluation of intervention effects" (p. 239). Some evidence exists to support the concern expressed by Kazdin. For example, Jones, Weinrott, and Vaught (11) selected 58 pairs of adjacent phases from graphs published in a professional journal and had 11 judges rate whether the intervention had produced a change in performance across the phases. The median intercorrelation indicating agreement among the 11 judges was only .39. Jones and associates also found poor agreement between the results of visual judgments and the result of statistical tests applied to the same set of single-subject data. They concluded that this latter finding suggests that statistically reliable experimental effects may be more often overlooked by visual appraisals of data than nonmeaningful effects" (p. 280).

Since clinical decisions may be based on graphic presentation and visual interpretation of single-subject data, it is important to determine the accuracy and reliability of visual inferences. The purpose of this study was to examine the reliability and accuracy of visual judgments and inferences made by therapists based on data from single-subject designs.

Methods

Participants

To examine the reliability and accuracy of visual inferences from graphed data, a series of five graphs depicting a typical A-B design was presented to 46 therapists practicing in the upper Midwest. The raters were all occupational therapists with 1 to 34 years of experience in various areas of clinical practice, administration, and education. Each rater was asked to decide whether or not a clinically significant change in patient performance was demonstrated across the A-B phases in each of the graphs (see Procedures section).

Stimulus Materials

In order to present those characteristics relevant to visual interpretation, sets of simulated data for five A-B designs were constructed. The five graphs presented to each of the raters appear in Figure 1. Each graph contains an A phase which represented a baseline or no treatment period and a B phase during which the treatment was implemented. All graphs contained a total of 16 data points.

Each graph varied across four factors that were believed to influence visual interpretation. The four factors were as follows: (a) degree of mean shift across phases, (b) degree of variation across phases, (c) the slope across phases, and (d) the degree of serial dependency.

Mean Shift. The degree of mean shift referred to the percentage of mean change from the A phase to the B phase. For example, for an A phase (baseline) with a mean of 10 and a B phase (treatment) with a mean of 12.5 the degree of mean shift was computed by dividing the difference between the two phases (12.5 - 10 = 2.5) by the mean of the A phase (2.5/10 = .25). In computing the mean shift, the A phase was subtracted from the B phase. A positive value indicated an increase in mean level across the two phases, while a negative value represented a decrease in mean level from phase A to phase B.

Variability. Variability referred to the amount of fluctuation that occurred within a phase, or across the phases. The measure of variability for each phase was obtained by computing the standard deviation for the data points contained in each phase of the design. An overall variability coefficient for all the data points combined across the phases was also computed. To

---

*Figure 1: Hypothetical Single-Subject Designs Rated by Therapists*
determine the degree of change in variability across phases the standard deviation for the A phase was subtracted from the standard deviation for the B phase. A positive number indicated that the degree of variability had increased between the two phases, while a negative value meant that the variability had decreased across the phases.

**Slope.** Slope referred to the angle or pith of any trend that existed within the data for each phase. Slope values were computed for the data in each phase using the method described by White and Haring (12). To determine the degree of change in slope from one phase to the next, the slope for the A phase was subtracted from the slope for the B phase. In the case of the slope change values, the largest absolute number indicates the greatest change in slope across the design phases, while the smallest absolute value is associated with the smallest change in slope.

**Serial Dependency.** Serial dependency is a characteristic of single-subject data related to the fact that data from single-subject designs are invariably of a repeated nature. That is, measurements on the outcome variable are collected repeatedly based on the performance of an individual over time (13). Serial dependency refers to the fact that sequential responses emitted by the same person may be correlated. The higher the correlation between responses (data points) the better one can predict performance over time. Jones, Weinrott, and Vaught (11) have suggested that the presence of serial dependence in a data series can influence the results of visual inferences made from that data series. A measure of serial dependency referred to as the autocorrelation coefficient was computed for the data points in each of the graphs using a modified procedure available in the SAS package (14).

Information related to the degree of mean shift, degree of variability, slope, and serial dependency for each of the five graphs is presented in Table 1.

**Procedures**

Each of the 46 raters was provided with a set of the five graphs shown in Figure 1 and one additional sample graph, which was used to orient the raters to the procedure. The purpose of single-subject designs was briefly discussed and the basic A-B design format was explained to the raters. Using the sample graph, the raters were instructed to respond yes or no to the following question: Is there a significant change in performance from one phase to the next? All raters were cautioned that significant referred to a clinically significant change based on their own professional experience. Any questions related to the procedure were also addressed at this time. Following the evaluation of the sample graph the remaining five graphs were presented to the raters. Each rater was provided with all five of the graphs on separate pieces of 8½ by 11 in. bond paper. The graphs were also presented on a screen via an overhead projector. The raters were allowed as much time as they desired to make a decision regarding each graph. The ratings were conducted over two sessions separated by approximately 3 weeks. Twenty-six therapists participated in the first rating session and 20 therapists in the second. The sequence of presentation of the graphs was reversed across the two sessions to reduce the likelihood of any order effects.

**Results**

The range of clinical experience for the 46 raters was from 1 to 34 years (M = 12.15, SD = 8.14). Each rater was asked to identify his or her primary area of practice. Thirty-three percent of the respondents (n = 15) identified administration/education as their primary area. Twenty-four percent (n = 11) identified physical dysfunction, 21% (n = 16) pediatrics, 15% (n = 7) mental health, and 7% (n = 3) the category of other as their area of practice. The highest academic degree held was the bachelor’s for 29 raters (63%). The remaining 17 raters (37%) possessed a master’s degree.

The ratings by the 46 therapists regarding whether a clinically significant change in patient performance occurred across the two phases are presented in Table 2. The results suggest a relatively low rate of agreement across most of the graphs. The responses for graphs 3 and 4 were particularly poor and approached the level of agreement that would be

<p>| Table 1: Descriptive Information for the Components of Visual Analysis for Each Graph |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|</p>
<table>
<thead>
<tr>
<th>Graph</th>
<th>Mean shift</th>
<th>Variability</th>
<th>Change</th>
<th>Slope</th>
<th>Serial Dependency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.56</td>
<td>.92</td>
<td>.99</td>
<td>.13</td>
<td>1.20</td>
</tr>
<tr>
<td>2</td>
<td>.80</td>
<td>2.14</td>
<td>1.59</td>
<td>1.00</td>
<td>1.17</td>
</tr>
<tr>
<td>3</td>
<td>.96</td>
<td>1.80</td>
<td>1.03</td>
<td>1.52</td>
<td>1.16</td>
</tr>
<tr>
<td>4</td>
<td>.16</td>
<td>-.34</td>
<td>-.56</td>
<td>.52</td>
<td>-.01</td>
</tr>
<tr>
<td>5</td>
<td>.17</td>
<td>.40</td>
<td>.50</td>
<td>.25</td>
<td>.37</td>
</tr>
<tr>
<td></td>
<td>.31</td>
<td>.18</td>
<td>.32</td>
<td>-.44</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>.14</td>
<td>-.22</td>
<td>-.18</td>
<td>.69</td>
<td>-.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Ratings by 46 Therapists Regarding Whether a Clinically Significant Change in Performance Occurred Across the Two Phases for Each Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

*See Figure 1 for a depiction of each of the five graphs.*
expected based on chance. To further examine which components of visual inspection were associated with rater agreement and disagreement, a ratio was computed by dividing the smaller percentage of agreement for each graph by the larger percentage. For example, for graph 1, 24% (the smaller percentage) was divided by 76% (the larger percentage) resulting in a ratio of 0.32. The larger the ratio, the more disagreement existed among the raters. As the ratio increases up to a maximum of 1.0, the level of agreement approaches what would be expected based on chance, that is, a 50:50 split. This ratio of disagreement was then correlated with the following variables for each graph: autocorrelation coefficient ($r = -0.11$); mean shift ($r = -0.58$); change in variability ($r = 0.19$); and change in slope ($r = 0.59$). This analysis revealed that the amount of rater agreement was weakly associated with the autocorrelation coefficient and the change in variability across phases for the five graphs. There was a moderate ($r = -0.58$) indirect relationship between the rate of disagreement and the mean shift, indicating that raters were more likely to agree if a large mean shift occurred from one phase to the next. There was also a moderate correlation ($r = 0.59$) between the ratio of disagreement and change in slope. This was a direct correlation and suggests that changes in slope across phases were associated with an increase in the amount of rater disagreement. These correlations are based on information from only five graphs and should be interpreted with caution.

To further clarify and quantify the reliability of the ratings across examiners and graphs, intraclass correlation coefficients were computed based on the generalizability theory approach presented by Cronbach and colleagues (15). Tinsley and Weiss (16) have recently argued that "at the present time the intraclass correlation is the most appropriate measure of interrater reliability available" (p. 363). Nontechnical descriptions of generalizability theory and illustrations of its application are available in works by Berk (17), Brennan and Kane (18), and Mitchell (19). Actually, there are several versions of the intraclass correlation coefficient. The particular method used in this study corresponds to a random effects analysis of variance (ANOVA) model that produced two reliability coefficients. The first coefficient of 0.53 estimated the average agreement between examiners and reflects the generalizability for a single rater. The second coefficient of 0.67 estimates the average agreement between the sample of raters employed and a hypothetical set of other raters from the same population.

The above analysis indicated that the rate of agreement based on visual inspection was relatively poor for the data sets included in this study. The second question of interest concerned the accuracy of the visual judgments made by the raters. To determine the "accuracy" of the judgments a point of comparison is required. To obtain such a reference point the data in each graph were analyzed using a quasi-statistical procedure referred to as the split-middle method of trend estimation (12). The split-middle method of trend estimation is a relatively easy-to-compute, quantitative procedure that can be used on small data sets and is not affected by the presence of directional serial dependency (20).

The split-middle technique has been proposed primarily to quantitatively describe the process of change across phases rather than to serve as an inferential statistical test (7). Statistical significance, however, can be evaluated using the procedure. White (21) has proposed a simple technique involving the use of the binomial test with the split-middle procedure to statistically evaluate change across phases of a single-subject design.

The split-middle method of trend estimation was computed for the data contained in each of the five graphs presented to the raters. The results of this analysis revealed statistically significant ($p < .05$) changes in performance (trends) across the two phases for graphs 4 and 5 (see Figure 2).
A majority of the raters (more than 50%) identified a significant change in performance across the A and B phases in graphs 2, 3, and 5. In contrast, the quantitative analysis revealed a statistically significant change in performance across the A and B phases for graphs 4 and 5. There is obviously disagreement between the ratings based on visual analysis and those produced by the quantitative procedure. The main advantage of employing the statistical technique as the point of reference is that the results are consistent. Assuming that the computations are performed correctly, the quantitative procedure will produce the same result for a given set of data regardless of who does the analysis. The same consistency in outcome was not obtained when only visual inference was employed. Consistent interpretation of data is obviously a requirement of valid clinical decision making.

Discussion

Visual analysis involves a professional judgment regarding the effect, meaningfulness, and generalizability of graphed data generated by single-subject studies (9). The results of this investigation reveal that the conclusions drawn from a visual inspection of data by one practitioner may not be consistent with the conclusions drawn by another practitioner. Nevertheless, visual analysis of single-subject data remains a clinically valuable procedure and should continue to be routinely employed in analyzing data from single-subject designs. However, the results of this analysis suggest that visual analysis should not be used in isolation.

Advantages of Visual Analysis

Visual analysis of graphed data has a number of advantages related to the assessment of patient performance in clinical environments. For instance, the procedure is appropriate for use with only a single individual or a small group of patients. In contrast, many of the quantitative data analysis procedures used in the behavioral sciences require a large number of subjects and, therefore, are not practical for use in clinical settings where the number of patients receiving treatment is often small. Visual analysis is also a dynamic process. The data are recorded continuously over time. The graphing of the data allows the therapist to systematically evaluate the patients’ performance over time. Thus, visual analysis of data encourages the investigation of process as well as outcome.

The argument is often made that visual analysis of single-subject data reveals only those intervention effects that are powerful enough to have produced clinically meaningful results. The “insensitivity” of visual analysis to weak treatment effects that are not immediately obvious by simply looking at the graphed data is considered a major advantage of visual analysis by some authorities (22). For example, Parsonson and Baer (23) contend that the potency of treatment effects revealed through visual inspection assures the therapist that the intervention effects are important because these effects appear in spite of the fact that a relatively insensitive analytic process was employed.

Finally, graphic presentation and visual analysis are easy to understand and intuitively appealing to therapists and patients. Graphically presenting the data allows the therapist, patient, or other interested party to evaluate the range and stability of client performance, understand the degree of control obtained by the design, and determine if the requirements of the design were met.

Limitations

The results of this investigation have obvious implications for the interpretation of single-subject data and clinical decisions based on such interpretations. Prior to a discussion of clinical implications the limitations of the present study must be considered. First, the data sets rated by the therapists were artificial and constructed for the purpose of the study. They were developed to illustrate and emphasize specific properties believed to be related to visual inspection and did not represent actual patient performance. A second limitation involves the raters. No attempt was made to train the therapists serving as raters regarding the process of visual analysis (i.e., identifying trend, level, slope, etc.). Their reliability and accuracy might have improved had they received such instruction. This is a question for future investigations to address. The final limitation concerns the quantitative procedure used to assess the accuracy of the therapists’ ratings. The split-middle method of trend estimation was computed for each data set using the eight data points available in the baseline phase to establish the trend or celeration line. Eight observations is the minimum recommended for use in computing the celeration line (7). If additional data points had been included in the baseline phase, a more accurate estimate of the baseline trend could have been obtained. It should be noted, however, that most clinical applications of single-subject designs result in phases with a relatively small number of data points (8).

A particular advantage of the split-middle method is that it cannot be computed without graphing the data. Plotting all of the data points in a series is an integral step in the analytic process for the split-middle method. Thus, the therapist is required to combine elements of both visual and quantitative analysis when employing the split-middle procedure.

Implications and Conclusions

Visual analysis, or what Sidman (24) refers to as “criterion by inspection,” is the standard method of data evaluation employed by practitioners using single-subject designs. However, as the results of this
investigation reveal, relying exclusively on visual analysis may lead to inconsistent interpretations of patient performance. The poor agreement associated with visual inspection indicates the need to develop adjunctive data analysis strategies. Christensen (25) has observed that “as single-subject designs have become more popular in the applied research areas, there has been increased emphasis on the need for statistical analysis of collected data” (p. 261). The use of various quantitative techniques with single-subject designs is a controversial topic. Many applied researchers using single-subject designs argue that one picture (graph) is worth a thousand p values. Barlow and Hersen (26) present an excellent overview of the controversy surrounding clinical versus statistical significance related to single-subject designs.

The low interrater agreement and inconsistent interpretation associated with visual inspection of graphed data are strong arguments for the use of a reliable supplemental strategy in conjunction with visual analysis. As noted previously, one of the advantages of quantitative procedures designed for use with single-subject data is that they will produce consistent results regardless of who does the computations. As Bloom and Fischer (20) state, “visual inspection of data should be considered a very useful beginning point. But unless the patterns are very clear, with sufficient numbers of observations and with stable baseline data, other methods of analysis should also be employed” (p. 439). Based on the results of the present investigation, this appears to be sound advice for therapists using single-subject designs to evaluate their clinical practice.

References