Article

Behavior Modification
2015, Vol. 39(4) 510–541
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0145445515581327
bmo.sagepub.com
SAGE

Interrater Agreement Between Visual Analysts of Single-Case Data: A Meta-Analysis

Jennifer Ninci[1], Kimberly J. Vannest[1], Victor Willson[1], and Nan Zhang[1]

Abstract

Visual analysis is the most widely applied method of data interpretation for single-case research as it encompasses multifaceted considerations relevant to evaluating behavior change. However, a previous research synthesis found low levels of interrater agreement between visually analyzed ratings of graphed data across all variables under analysis. The purpose of this meta-analysis was to evaluate the peer-reviewed literature to date for potential moderators affecting the proportion of interrater agreement between visual analysts. Nineteen articles with 32 effects were assembled. Potential moderators evaluated included (a) design families, (b) rater expertise, (c) the provision of contextual information for graphs, (d) the use of visual aids, (e) the provision of an operational definition of the construct being rated, and (f) rating scale ranges. Results yielded an overall weighted interrater agreement proportion of .76. Moderator variables identified produced low to adequate levels of interrater agreement. Practical recommendations for future research are discussed.

Keywords

visual analysis, agreement, reliability, single-case research

[1]Texas A&M University, College Station, USA

Corresponding Author:

Jennifer Ninci, Department of Educational Psychology, 4225 Texas A&M University, College Station, TX 77843-4225, USA.
Email: jninci@tamu.edu

In research and practice, visual analysis is the most widely used and recommended method for interpretation of single-case data (Gast, 2010; Kennedy, 2005; Kratochwill et al., 2010, 2013; Smith, 2012). Visual analysis consists of analyzing data according to six indices of change: level, variability, trend in the data path, overlap of data points between phases, immediacy of effect, and consistency across analogous phases (Franklin, Gorman, Beasley, & Allison, 1996; Kratochwill et al., 2010, 2013). These indices of change allow a holistic evaluation to best understand the idiosyncrasies present within the data (Kennedy, 2005; Kratochwill et al., 2010, 2013). Methods of statistical analysis run the danger of obscuring this inductive process and may be misinterpreted to demonstrate causality (Carter, 2013; Horner, Swaminathan, Sugai, & Smolkowski, 2012; Parker & Vannest, 2012; Perone, 1999). The reliability of visual analysis deserves further inspection as it is critical in the evaluation of single-case data.

Visually analyzing graphic representations of data can be subject to sources of disagreement or error. Extragraphic contextual variables such as the acceptability of the treatment (Spirrison & Mauney, 1994), the characteristics of the subject (Grigg, Snell, & Loyd, 1989), and the social significance of effects (DeProspero & Cohen, 1979) can contribute to the subjectivity involved in visual analysis. Charting and scaling techniques that transform the *y*-axis of graphs have influenced visually analyzed judgments (Furlong & Wampold, 1982; Knapp, 1983) as well as characteristics of the data itself such as the degree of autocorrelation (Fisch, 2001; Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1990a, 1990b) and specific indices of intervention effect (Furlong & Wampold, 1982; Gibson & Ottenbacher, 1988; Lieberman, Yoder, Reichow, & Wolery, 2010). As visual analysis takes multiple considerations into account, the variability of experience and expertise among visual analysts can produce disparate interpretations (Hojem & Ottenbacher, 1988; Wampold & Furlong, 1981).

Not only is agreement between visually analyzed ratings of single-case data subject to variability, but it is also historically low to moderate. A previous quantitative synthesis of the literature by Ottenbacher (1993) revealed a mean of .58 (range: .39-.84) interrater agreement (IRA) among 14 sources with 22 effects. Despite low levels of IRA overall and within moderator categories analyzed, Ottenbacher found agreement improved somewhat through the use of trend lines displayed over graphed data, particularly when raters were trained in the use of such methods. However, in analyzing rater expertise, Ottenbacher found that experts in single-case research produced lower levels of IRA among each other than did students or clinicians.

The meta-analysis by Ottenbacher (1993) has a number of limitations regarding variation in the methods among the included studies. First, the included studies did not use a common metric to calculate IRA. Studies included effect sizes based on both proportional (e.g., Hojem & Ottenbacher, 1988) and correlational metrics (e.g., Jones et al., 1978). This presents an issue given these metrics lie on different scales (0 to 1 vs. −1 to 1) and thus have different benchmarks of interpretation. Second, the number of options in rating scales varied across studies. Although this was not evaluated as a moderator, it is likely that effects had variation attributable to differences in the number of rating scale options for respondents, as the chance of agreement and the number of options in the scale would be inversely related. Third, the effects were not weighted to account for the widely varying numbers of graphs analyzed or raters used per study effect. Despite these limitations, findings are consistent with some of the extended literature base to date suggesting less than acceptable levels of reliability (e.g., Bobrovitz & Ottenbacher, 1998; Danov & Symons, 2008; James, Smith, & Milne, 1996; Lieberman et al., 2010; Normand & Bailey, 2006).

Variable and low IRA between visual analysts can affect our confidence in data-based decisions (Franklin et al., 1996). Furthermore, the issue of reliable methods in data interpretation may compromise the credibility of advancements in single-case research to contribute maximally in the establishment of evidence-based practices (Burns, 2012; Wampold & Furlong, 1981). For decades, there lacked a clear and systematic method of visual analysis to facilitate the reliability of judgments (Wampold & Furlong, 1981). Over the years, actions have been made to refine and comprehensively describe the process of visual analysis for single-case research. Structured criteria available from the What Works Clearinghouse systematize quality indicators of experimental control and evidence of effectiveness for single-case research designs (Kratochwill et al., 2010, 2013). These criteria have been adopted in a protocol to structure decision making for which acceptable levels of IRA have been found between two independent raters (Maggin, Briesch, & Chafouleas, 2013; Ninci et al., 2015), supporting the idea that elucidating the specific constructs that raters are evaluating could improve IRA between visual analysts.

Another influence to the visual analysis process that could potentially affect IRA is the inclusion of contextually relevant information about the graphic displays. Although we have a limited understanding of how contextual variables influence judgments, factors related to the context such as the abilities of the recipient of behavior change or the relative malleability of a particular dependent variable to change are important matters when making data-based decisions. DeProspero and Cohen (1979) reported a number of respondents to decline participation due to the absence of contextual information with graphic displays. Both experts of single-case research and teachers with relatively less expertise in single-case research report to use contextual information to inform their judgments (DeProspero & Cohen, 1979; Grigg et al., 1989).

The purpose of the current study was to statistically examine potential moderators affecting the variability in the proportion of IRA between visual analysts in the interest of identifying factors that support the scientific principles of visual analysis. No studies since Ottenbacher (1993) have systematically reviewed or meta-analyzed the visual analysis literature to determine the overall IRA or moderators of IRA between visual analysts. Therefore, we sought to extend the previous review by (a) including new studies advancing the literature base, (b) addressing the aforementioned methodological limitations threatening the internal validity of findings, and (c) evaluating new variables as potential moderators. Based on the existing literature and methodological considerations, the following variables were selected as possible moderators of influence to evaluate: (a) design families, (b) rater expertise (i.e., training in single-case research), (c) provision of contextual information related to cases displayed in graphs, (d) use of a visual aid, (e) use of guidelines or criteria defining the construct to be rated, and (f) differences in rating scales. An ancillary purpose was to summarize descriptive statistics of the included studies to supplement findings.

Method

A comprehensive literature review of peer-reviewed journal articles was conducted among PsycINFO, Education Resource Information Center (ERIC), and MEDLINE (PubMed) electronic databases in May 2013. No date limits were used. In each individual database, a keyword search was conducted by combining all variations of the following three sets of terms: (a) *visual analysis* or *visual inspection*, (b) *reliability* or *agreement*, and (c) *single-subject, single-case, time-series*, or *intrasubject*. A total of 37 articles were retrieved with duplicates removed. To retrieve articles focused on visual analysis that were not gathered from the previous search, a second keyword search was conducted in PsycINFO and ERIC electronic databases for articles including either the terms "*visual analysis*" or "*visual inspection*" restricted to be included within the title; no additional terms were combined. The restricted title keyword search produced an additional 108 novel articles with duplicates removed for a total of 145 articles to be screened for inclusion.

Gate 1 Criteria: Initial Screening

Article titles and abstracts were screened to initially determine potential exclusion from further analysis in a full-text screening. Articles were excluded in this initial evaluation if it was determined they were not in English (*n* = 4), did not refer to "visual analysis" or "visual inspection" as used in single-case research (*n* = 92), descriptive in their purpose (*n* = 5), single-case research studies (*n* = 4), or qualitative studies (*n* = 1; Grigg et al., 1989). From the keyword search, 26 articles met these criteria, and from the restricted title word search, an additional 13 articles met these criteria (*n* = 39).

Gate 2 Criteria: Full-Text Screening

In the full-text screening, studies were excluded for which the purpose was to evaluate the accuracy of visual analysis ratings through agreement with various statistical approaches (*n* = 8). Literature reviews and meta-analyses were excluded (*n* = 3). Studies were excluded that did not report visual analysis agreement coefficients or sufficient data to yield this information (*n* = 6). Articles were included if numerical data were reported on agreement between visually analyzed ratings of graphed time-series data. Reanalysis of previously published data was not replicated for use (*n* = 1; Matyas & Greenwood, 1990a). Study effects were excluded that used fewer than three graphic displays for visual analysis or that used ratings from fewer than three respondents (*n* = 1; Bonner & Barnett, 2004). These minimum requirements were set to control for quality and focus of included studies by ensuring representativeness of the potential variability existing among distributions of graphed data sets and visual analysts. Through Gate 2, 15 articles fit the inclusion criteria from the keyword search and 5 articles fit the inclusion criteria from the restricted title keyword search. An ancestral search was then conducted consisting of screening the reference lists among the 20 articles. An additional 7 articles were determined to fit the Gate 2 inclusion criteria in the iterative ancestral search process. The final 27 articles were then screened for final inclusion based on the Gate 3 criteria described below.

Interrater reliability on the full-text screening. Two independent reviewers (the first and second authors) read each qualifying article from the initial screening (*n* = 39; 100%) in full text, using the Gate 2 criteria with a dichotomous (i.e., yes or no) rating over inclusion of the article. Interrater reliability was calculated by dividing the number of articles agreed upon for inclusion or exclusion by the total number of articles and multiplying by 100. There were four disagreements resulting in 90% interrater reliability. The raters then discussed each disagreement until a consensus was reached between them for a final determination of inclusion or exclusion.

Gate 3 Criteria: Computation Screening

Studies included through the final gate had to report effects specifically on the *interrater* agreement computed between visual analysts (i.e., different subjects). Studies were excluded that evaluated *intrarater* agreement (i.e., within subjects; Knapp, 1983; Rojahn & Schulze, 1985; Ximenes, Manolov, Solanas, & Quera, 2009). Although Knapp (1983) provided a supplemental report of proportions of IRA between raters for one graphic display plotted in two different ways, this was deemed to be an insufficient number of graphic displays for inclusion per Gate 2 criteria. Studies reporting *intergroup* agreement (i.e., composited and averaged ratings within a group of visual analysts) were excluded if proportions of IRA could not be derived from the provided information (i.e., Rojahn & Schulze, 1985; Spirrison & Mauney, 1994; Ximenes et al., 2009).

Finally, studies that did not use a proportion of IRA computation (i.e., agreements divided by agreements and disagreements) or report sufficient information to calculate the proportion of IRA were excluded for the purpose of this meta-analysis yielding a homogenous sample of studies. In other words, studies reporting IRA via correlational measures or evaluating the variability around the mean were excluded (i.e., Brossart, Parker, Olson, & Mahadevan, 2006; DeProspero & Cohen, 1979; Jones et al., 1978; Mercer & Sterling, 2012). Proportion of IRA calculations was included because the majority of studies used this computation or provided sufficient information to compute it. At Gate 3, 19 peer-reviewed journal articles with 32 total effects were found to report or provide information to calculate proportions of IRA between visual analysts.

Derived Effects

Each effect was calculated as the mean proportion of IRA between raters among the graphic displays under analysis. IRA between raters was either calculated as the number of raters in the largest rating group divided by the total number of raters for each graphic display (e.g., Gibson & Ottenbacher, 1988) or as the number of agreements between pairs of ratings divided by all possible pairs for each graphic display (e.g., Park, Marascuilo, & Gaylord-Ross, 1990). Effects were included from two studies that analyzed pairs of agreements between a sample of visual analysts and a priori determined visually analyzed judgments (Normand & Bailey, 2006; Roane, Fisher, Kelley, Mevers, & Bouxsein, 2013). A priori judgments in these studies were made using a visual analysis procedure through which an established criterion was obtained prior to collecting ratings from a sample of visual analysts.

Hagopian et al. (1997) reported proportions of IRA between three raters pre- and post-training in the use of structured criteria and also reported the IRA between those ratings and a priori determined ratings. Because IRA proportions were reported between the three raters, IRA proportions against the a priori determined ratings were excluded from this study as not to report additional effects derived from reanalysis of the raters' responses. Per initial inclusion criteria, effects provided between only two visual analysts were excluded. That is, although Roane et al. (2013) reported pre- and post-training effects between two raters, only the IRA proportions that represented three sets of judgments (two raters compared with a priori determined ratings) were included.

Mean proportional effects were derived directly from study reports or were derived by calculations made to data reported in studies under the following circumstances. First, the mean proportion of IRA across graphic displays was calculated if studies only reported IRA proportions by individual graphs (Hojem & Ottenbacher, 1988; Johnson & Ottenbacher, 1991; Kahng et al., 2010; Ottenbacher, 1986; Ottenbacher, 1990a, 1990b) or only reported the number of raters per category by graph (Bobrovitz & Ottenbacher, 1998; James et al., 1996). Thus, data across raters with varied levels of expertise had to be aggregated from Johnson and Ottenbacher (1991) to calculate the mean proportion of IRA across graphs. Second, reports of a ratio of disagreement effect size had to be converted to a proportion of IRA effect size in one study (Ottenbacher & Cusick, 1991). Third, all studies besides those using multielement designs that did not already report data on a dichotomous scale or a scale of 3 happened to provide enough raw data to convert proportions of IRA to ratings on a scale of 2 to ensure a more homogenous sample of effects. For instance, a continuous scale of 6 (0-5) ranging from strongly agree to strongly disagree that there is significant change across phases was converted to a scale of 2 by combining the ratings of 0 to 2 and that of 3 to 5 (Gibson & Ottenbacher, 1988; Harbst, Ottenbacher, & Harris, 1991). All studies using multielement designs used graphs of functional analyses and derived ratings on a categorical scale of 12 based on identification of behavioral function (Danov & Symons, 2008, Hagopian et al., 1997, Roane et al., 2013). Thus, multielement scales were not converted.

Multiple items rated from the same presentation of a graphic display were aggregated. That is, in Bailey (1984), respondents rated each data set presentation on their interpretations of both level and trend, so the effects of level and trend were averaged per derived effect. Multiple measures of effects were included when studies reused a sample of graphs or a sample of raters, as long as different sets of ratings were taken from independent presentations of data to evaluate an independent variable. For instance, several studies used the same graphs among the same raters and used pre- and post-measures to evaluate the effects of a visual aid or training (Bailey, 1984; James et al., 1996; Normand & Bailey, 2006; Roane et al., 2013; Skiba, Deno, Marston, & Casey, 1989). Also, Bailey (1984) evaluated effects of semi-logarithmic charts versus equal interval charts, with and without a visual aid to produce four total effects. Although raters and graphic displays were sometimes resampled and thus raters were exposed to the same sample of data sets, separate effects were included because independent sets of ratings from separate presentations of graphic displays were obtained.

Moderator Analysis and Descriptive Variable Coding

The 32 effects were coded for potential moderator variables and variables for descriptive review. Coded variables included respondent characteristics, characteristics of the graphic displays evaluated, methodological characteristics, and the derived proportion of IRA for each effect. Respondent characteristics coded included the number of raters and their levels of expertise. In analyzing the number of raters, a priori judgments were counted as one rater per study effect. Moderator categories of respondent expertise are defined in Table 1. Effects that did not include a homogenous group of raters based on the categories of expertise were classified as "mixed" and were not evaluated in moderator analyses due to variability within the category and a limited number of study effects.

Characteristics of the data sets included the number of graphic displays examined per rater, the design family used (i.e., multielement, AB or reversal variations, and multiple baseline designs), and whether contextual information was provided with the graphic displays. Methodological characteristics evaluated included whether a visual aid was provided to supplement respondent decisions, whether the construct being rated was defined, the type of judgment made (e.g., judgment of "clinically significant change"), the type of scale used (i.e., dichotomous, categorical, continuous), and the number of rating scale options. Descriptions of data set and methodological characteristics that were evaluated as dichotomous moderators are presented in Table 2.

Among the aforementioned variables, the following were analyzed for potential moderators: (a) design families, (b) categories of rater expertise, (c) whether contextual information was provided, (d) whether a visual aid was provided, (e) whether the construct being rated was defined, and (f) the number of rating scale options (2 vs. 3). As a result of substantial differences in study characteristics between design families, effects from multielement designs were only evaluated within the overall effects and as a design family moderator in the interest of maintaining a homogenous sample of studies in other moderator analyses. For the moderator analysis of design family, AB

Table 1. Moderator Categories of Rater Expertise.

Variable codes	Description
Novices	Experience in SCR was limited to brief overview prior to ratings or experience was not explicitly stated but suggested to be relatively low.
Beginners	Received basic training (in-service or formal lesson related to SCR data analysis), had experience conducting SCR, or were receiving training in SCR (e.g., individuals completing ABA coursework).
Experienced	Attaining or attained doctorates with extensive SCR/ABA experience or attained BCBA/master's degree with SCR training and experience.

Note. SCR = single-case research; ABA = applied behavior analysis; BCBA = Board Certified Behavior Analyst.

Table 2. Characteristics Evaluated as Dichotomous Moderators Among AB/Reversal Design Variations.

Variable codes	Description
Context–No Context	Provided description of the context of graphs such as dependent variables, independent variables, and/or participant information, or did not.
Visual Aid–No Visual Aid	Provided lines of progress or criterion lines (training in the use of the aid was not necessitated), or did not.
Defined–Not Defined	Stated objective operational definition of what was to be visually analyzed (e.g., "functional relation") by raters or taught them a method to use a criterion, or did not.
Scale of 2–Scale of 3	All relevant studies used measures on a scale of 2, scale of 3, or provided data for the scale to be reduced to 2 in calculating the proportion of IRA between visual analysts.

Note. IRA = interrater agreement.

and reversal design variations were compared against multielement designs. Because only one study with one effect used multiple baseline designs (Lieberman et al., 2010), this study was only included in the analysis of overall effects. Most studies used AB and reversal design variations, and thus only these study effects were included in the remainder of moderator analyses.

Interrater reliability on coded variables. Articles were coded on the six aforementioned moderator analyses using a spreadsheet organized by the first author. In addition, the number of raters, the number of graphs used, the type of judgment made, the type of scale used, and the derived proportion of IRA were coded per effect. Notes were made along the side of the table in the spreadsheet explaining how effects were derived. The second author read each article in full and determined the accuracy of each code (100%) by marking whether agreement or disagreement was obtained. Interrater reliability was computed by dividing the number of codes agreed upon by the total of possible agreements and multiplying that figure by 100. There were two instances of disagreement across the 352 cells (32 effects by 11 codes) resulting in .006% disagreement or 99% agreement. The disagreements were resolved by both authors reexamining the article and discussing the discrepancy until consensus was reached.

Effect Size Computations

Most calculations were conducted using standard procedures articulated by Lipsey and Wilson (2001). The logit proportion of IRA was converted from each proportion effect size derived. The logit proportion was used for all moderator analyses as an effect size because variability around the mean was of interest. While the proportion statistic ranges in values from 0 to 1, logit proportions can be of any numerical value. The logit proportion has an approximately normal distribution with a mean of zero and standard deviation of 1.83 (Lipsey & Wilson, 2001). Logit proportion effect sizes were converted back to weighted mean proportions to supplement reports. Median proportions of each set of study effects were also calculated as some categories of moderator analyses had few total effects, potentiating a non-normal distribution of the data. The number of graphs evaluated per set of raters was deemed as appropriate for the weights of effects. Although the number of raters may have been suitable as the application of weights, Ottenbacher (1993) did not find that sample size correlated with IRA while the number of graphs examined had an inverse correlation with IRA. The number of graphs was chosen to be the weight per effect because large samples of graphs under analysis would best represent the variability that occurs in patterns of time-series data.

The Q and I^2 statistics were used to analyze homogeneity among samples of effect sizes. The Q statistic represents the sum of the weighted square distances of each individual effect around the mean effect (Lipsey & Wilson, 2001). Statistical significance in the Q analysis indicates that the variation around the mean could be accounted for by one or more moderators. The I^2

statistic complements the Q, informing the percentage of variance of the total variance among effect sizes that is a result of the true between-study variation (Higgins & Thompson, 2002). Higgins, Thompson, Deeks, and Altman (2003) suggested that an I^2 of 25%, 50%, and 75% would respectively represent low, moderate, and high heterogeneity. Potential moderators were analyzed by running the Q_{between} and Q_{within} statistics for each variable. The Q_{between} represents the variance between categories of a potential moderator variable, and the Q_{within} represents the pooled variance within categories of a potential moderator variable. Significance tests were calculated for the Q_{between} and Q_{within}. A statistically significant Q_{between} rejects that sampling error alone explains the variance between moderator categories. A statistically significant Q_{within} indicates that there could be other moderators to explain the variance.

Results

Table 3 displays a matrix of information over each effect ordered by date of publication. Columns display variables including (a) the number of raters, (b) the level of expertise in single-case research among raters, (c) the number of graphs or phase changes rated, (d) the design used, (e) whether context was provided, (f) whether visual aids were provided, (g) whether questions were defined to raters, (h) the number of scale options, (i) the derived proportion of IRA, and (j) the logit proportion effect size with confidence intervals set at 95%.

Descriptive Statistics

Overall, the number of raters included in studies ranged from 3 to 61. Effects had relatively equal numbers of raters across experienced (25%; K = 8), beginner (28%; K = 9), and novice (22%; K = 7) levels while the remainder (25%; K = 8) included mixed levels of expertise in single-case research. Effects including mixed levels of expertise consisted of five effects with a mix of beginner and experienced raters and three effects with primarily novice raters. One study evaluated respondent expertise (between novices and beginners) as an independent variable and did not find differences between the IRA proportions of each group (James et al., 1996).

The number of graphic displays rated among the 32 effects ranged from 4 to 141. The majority of effects (72%; K = 23) used AB and reversal design variations, most of which displayed only one replication of effect (i.e., AB). Eight effects (25%) used multielement designs and one effect (3%) used multiple baseline designs. Eighteen effects (56%) supplemented graphic displays

Table 3. Descriptive Information of Study Effects and Results.

	Raters		Data sets			Methods			Results	
Study effects	*n*	Level of expertise	No. rated	Design	Context provided	Visual aid or criteria provided	Specific question(s) defined	No. of scale options	Proportion agreement	Logit proportion (95% CI)
Bailey (1984; *equal interval*)	13	Beginners	19	AB and reversal	Context	*No visual aid*	Defined	2	.70	0.85 [0.24, 2.04]
Bailey (1984; *equal interval*)	13	Beginners	19	AB and reversal	Context	*Visual aid*	Defined	2	.85	1.73 [−0.23, 2.52]
Bailey (1984; *semilog*)	13	Beginners	19	AB and reversal	Context	*No visual aid*	Defined	2	.74	1.05 [0.17, 2.12]
Bailey (1984; *semilog*)	13	Beginners	19	AB and reversal	Context	*Visual aid*	Defined	2	.87	1.90 [−0.37, 2.66]
Ottenbacher (1986)	46	Novices	5	AB	No context	No visual aid	Not defined	2	.74	1.05 [−2.89, 5.18]
Gibson and Ottenbacher (1988)	20	Novices	24	AB	No context	No visual aid	Not defined	2	.76	1.15 [0.30, 1.99]
Hojem and Ottenbacher (1988)	20	Beginners	5	AB	No context	*No visual aid*	*Not defined*	2	.77	1.21 [−3.27, 5.55]
Hojem and Ottenbacher (1988)	19	Beginners	5	AB	No context	*Visual aid*	*Defined*	2	.84	1.66 [−4.66, 6.94]
Skiba, Deno, Marston, and Casey (1989)	27	Beginners	4	ABC	Context	*No visual aid*	Not defined	3	.56	0.24 [−2.79, 5.08]
Skiba et al. (1989)	27	Beginners	4	ABC	Context	*Visual aid*	Not defined	3	.78	1.27 [−5.45, 7.73]
Ottenbacher (1990a)	61	Novices	6	AB	No context	No visual aid	Defined	3	.69	0.80 [−1.63, 3.92]
Ottenbacher (1990b)	30	Novices	24	AB	No context	No visual aid	Not defined	2	.72	0.94 [0.37, 1.92]
Park, Marascuilo, and Gaylord-Ross (1990)	5	Experienced	44	AB	Context	No visual aid	Not defined	3	.60	0.41 [0.74, 1.55]
Harbst, Ottenbacher, and Harris (1991)	30	Mixed (primarily novices)	24	AB	No context	Visual aid	Not defined	2	.73	0.99 [0.35, 1.93]

(continued)

Table 3. (continued)

	Raters		Data sets			Methods			Results	
Study effects	*n*	Level of expertise	No. rated	Design	Context provided	Visual aid or criteria provided	Specific question(s) defined	No. of scale options	Proportion agreement	Logit proportion (95% CI)
Johnson and Ottenbacher (1991)	20	Mixed (primarily novices)	24	AB	No context	Visual aid	Not defined	2	.76	1.15 [0.30, 1.99]
Ottenbacher and Cusick (1991)	39	Novices	21	AB	No context	*No visual aid*	Not defined	2	.76	1.15 [0.20, 2.08]
Ottenbacher and Cusick (1991)	40	Novices	21	AB	No context	*Visual aid*	Not defined	2	.71	0.90 [0.30, 1.99]
James, Smith, and Milne (1996)	21	*Novices*	14	AB	No Context	No visual aid	Not defined	3	.72	0.94 [−0.08, 2.36]
James et al. (1996)	21	*Beginners*	14	AB	No context	No visual aid	Not defined	3	.74	1.05 [−0.14, 2.42]
Hagopian et al. (1997)	3	Experienced	26	Multielement	Context	*No visual aid*	*Not defined*	12	.46	−0.16 [0.66, 1.63]
Hagopian et al. (1997)	3	Experienced	26	Multielement	Context	*Visual aid*	*Defined*	12	.81	1.45 [0.24, 2.05]
Bobrovitz and Ottenbacher (1998)	32	Mixed (primarily novices)	42	ABAB	No context	No visual aid	Not defined	3	.68	0.75 [0.67, 1.61]
Normand and Bailey (2006)	5 to a priori determined set	Experienced	12	AB and ABA	Context	*No visual aid*	Not defined	3	.78	1.27 [−0.50, 2.79]
Normand and Bailey (2006)	5 to a priori determined set	Experienced	12	AB and ABA	Context	*Visual aid*	Not defined	3	.67	0.71 [−0.12, 2.41]
Danov and Symons (2008)	43	Mixed (beginners and experienced)	26	Multielement	Context	No visual aid	Not defined	12	.63	0.53 [0.52, 1.76]
Kahng et al. (2010)	45	Experienced	36	ABAB	No context	No visual aid	Defined	2	.96	3.18 [−0.87, 3.15]
Lieberman, Yoder, Reichow, and Wolery (2010)	36	Experienced	16	MBL across participants	context	No visual aid	Not defined	3	.40	−0.41 [0.49, 1.80]

(continued)

Table 3. (continued)

	Raters		Data sets			Methods			Results	
Study effects	*n*	Level of expertise	No. rated	Design	Context provided	Visual aid or criteria provided	Specific question(s) defined	No. of scale options	Proportion agreement	Logit proportion (95% CI)
Roane, Fisher, Kelley, Mevers, and Bouxsein (2013; Study 1)	2 to a priori determined set	Experienced	141	Multielement	Context	Visual aid	Defined	12	.92	2.44 [0.76, 1.53]
Roane et al. (2013; Study 2)	2 to a priori determined set	Mixed (beginners and experienced)	141	Multielement	Context	Visual aid (*no training*)	Defined	12	.73	0.99 [0.92, 1.37]
Roane et al. (2013; Study 2)	2 to a priori determined set	Mixed (beginners and experienced)	141	Multielement	Context	Visual aid (*training*)	Defined	12	.98	3.89 [0.15, 2.14]
Roane et al. (2013; Study 2)	3 to a priori determined set	Mixed (beginners and experienced)	141	Multielement	Context	Visual aid (*no training*)	Defined	12	.80	1.39 [0.89, 1.39]
Roane et al. (2013; Study 2)	3 to a priori determined set	Mixed (beginners and experienced)	141	Multielement	Context	Visual aid (*training*)	Defined	12	.95	2.94 [0.64, 1.65]

Note. Italics represent variables under analysis within studies. CI = confidence interval; MBL = multiple baseline design.

with contextual information while 14 effects (44%) provided no context. Those effects using no context tended to use generated data in graphic displays, whereas those effects providing context tended to use real data such as a sample of published graphs. None of the included studies evaluated the provision of contextual information as an independent variable to potentially affect IRA.

Eighteen effects (56%) did not make use of a visual aid, and 14 effects (44%) used a visual aid. Studies using AB or reversal design variations with a visual aid used trend estimation lines. The majority of these studies used the split-middle trend estimation approach (e.g., Bailey, 1984) while one study used the quarter-intersect method of trend estimation (Skiba et al., 1989). Studies with multielement designs exclusively used criterion lines as a visual aid (Hagopian et al., 1997; Roane et al., 2013). Several studies did not appear to report if raters were trained to use the lines of progress (e.g., Bailey, 1984; Johnson & Ottenbacher, 1991; Normand & Bailey, 2006; Ottenbacher & Cusick, 1991). It is possible that raters, especially those with more expertise in single-case research, were familiar with using visual aids. Among studies evaluating the effects of a visual aid, visual aids appeared to produce higher IRA with the exception of two studies finding the opposite (Normand & Bailey, 2006; Ottenbacher & Cusick, 1991).

The majority of effects (59%; $K = 19$) did not appear to objectively define what raters were evaluating. For example, for Ottenbacher (1990a), raters were instructed that “clinically significant change” was to be based on their own professional experiences. Thirteen effects (41%) operationally defined what was being visually analyzed. For example, Bailey (1984) defined "significant change" as concluding that experimental control was obtained. Various types of judgments were made including determinations of a “significant change” ($K = 13$; 41%), “behavioral function” ($K = 8$; 25%), “clinically significant change” ($K = 4$; 13%), “improvement” ($K = 2$; 6%), “systematic change in level and/or trend” ($K = 2$; 6%), “experimental control” ($K = 1$; 3%), “functional relation” ($K = 1$; 3%), and “significant treatment effect in Phase B” ($K = 1$; 3%). No studies evaluated the exclusive effects of defining constructs on proportions of IRA, but two studies evaluated this variable in addition to training in the use of a visual aid, each finding that a visual aid and clear criteria increased IRA (Hagopian et al., 1997; Hojem & Ottenbacher, 1988).

Fourteen effects (44%) used a dichotomous scale or provided information for a scale to be converted from 6 to 2. Ten effects (31%) used a scale of 3, and 8 effects (25%; those using multielement designs) used a scale of 12. Several effects used continuous scales such as an original range of 6 options from strongly agree to strongly disagree ($K = 8$; 25%) or a range of three options such as “yes, no, or uncertain” and “increase, decrease, or no change” (K = 8; 25%). Six effects (19%) used dichotomous (e.g., yes/no) scales. Categorical scales included categories on functions of behavior (K = 8; 25%) and categories on improvement per phase (K = 2; 6%).

Summaries and Homogeneity Tests

Table 4 includes overviews of findings overall and from each sample of effects analyzed for moderators, as well as results from corresponding statistical tests of homogeneity. In the analysis of all 32 effects, the median was 0.74 and the weighted mean logit proportion effect size was 1.14, or .76 when converted to proportion. For proportions of IRA, necessary agreement is suggested to be .70 or above, adequate or minimally acceptable agreement is generally considered to be .80 or above, and above .90 is deemed as good (Hartmann, Barrios, & Wood, 2004; House, House, & Campbell, 1981). Therefore, the mean and median proportions of IRA reached necessary but not acceptable levels overall and within each sample of study effects analyzed for moderators. The overall effects had a high level of heterogeneity (I^2 = 89%), with a statistically significant Q. The sample of effects comparing design families also had a high level of heterogeneity (I^2 = 88%) with a statistically significant Q. The remaining samples of effects under analysis had low levels of heterogeneity (rater expertise I^2 = 15%; graphic and methodological characteristics I^2 = 7%) and statistical significance was not found; however, moderators were still analyzed as results may be affected by low power due to inclusion of few effects (Higgins & Thompson, 2002).

Moderator Analyses

Table 5 displays results from the moderator analysis of AB and reversal design variations versus multielement designs. Multielement designs produced a minimally adequate median proportion (.81) and weighted mean proportion of IRA (.80). The AB and reversal design variations produced necessary levels of median (.74) and mean proportions of IRA (.71). There was no overlap in confidence interval ranges between the two design families. While finding statistical significance for the Q_{between} revealed that these findings can be attributed to the design family moderator, a statistically significant Q_{within} indicated that significance could be accounted for by other variables, because the pooled heterogeneity indicates too great variability within categories.

Table 6 shows results from the moderator analysis of respondent expertise among effects using AB and reversal design variations. The mean proportions of IRA across categories of expertise ranged from low to necessary

Table 4. Descriptive Summaries and Tests of Homogeneity.

No. of studies	K	No. of raters	No. of data sets	Median proportion	Mean proportion	Mean logit proportion (95% CI)	Q	I^2
Overall								
19	32	669	1,216	.74	.76	1.14 [1.05, 1.23]	285.82***	89.15
AB/reversal design variations and multielement designs								
18	31	633	1,200	.74	.77	1.19 [1.10, 1.29]	249.82***	87.99
AB/reversal design variations: Rater expertise								
12	20	485	327	.74	.71	0.87 [0.68, 1.06]	22.45	15.37
AB/reversal design variations: Graphic and methodological characteristics								
15	23	567	417	.74	.71	0.88 [0.72, 1.04]	23.57	6.65

Note. K = number of effects; CI = confidence interval; I^2 values are percentages.
***p = .001.

Table 5. Moderator Analysis: AB/Reversal Design Variations—Multielement Designs.

Moderator variables	K	No. of data sets	Median proportion	Mean proportion	Mean logit proportion (95% CI)	Q_{between}	Q_{within}
Reversal/AB variations	23	417	.74	.71	0.88 [0.72, 1.04]	22.44***	227.38***
Multielement	8	783	.81	.80	1.36 [1.25, 1.47]		

Note. K = number of effects; CI = confidence interval.
***p = .001.

Table 6. Moderator Analysis: Level of Rater Expertise Among AB/Reversal Design Variations.

Moderator	*K*	No. of data sets	Median proportion	Mean proportion	Mean logit proportion (95% CI)	Q_{between}	Q_{within}
Experienced	4	104	.73	.64	0.58 [0.27, 0.89]	5.00*	15.80
Beginner	9	108	.77	.76	1.14 [0.75, 1.53]		
Experienced	4	104	.73	.64	0.58 [0.27, 0.89]	3.71	11.95
Novice	7	115	.72	.73	1.01 [0.70, 1.32]		
Beginner	9	108	.77	.76	1.14 [0.75, 1.53]	0.27	4.90
Novice	7	115	.72	.73	1.01 [0.70, 1.32]		

Note. *K* = number of effects; CI = confidence interval.
*p < .05.

levels among studies using AB and reversal designs. Beginners produced the highest median (.77) and mean (.76) proportion of IRA. Novice raters produced a higher mean proportion (.73) relative to that of experienced raters (.64), while median proportions were comparable between novice (.72) and experienced (.73) raters. Confidence interval ranges around effect sizes were overlapping for each of the comparisons. Still, experienced raters produced statistically significant lower effects in comparison with raters with a beginner level of expertise. Between novice and experienced raters, effects bordered on statistical significance (p = .054). The Q_{within} scores did not indicate statistically significant variances within the moderator levels of expertise, indicating little variability within moderator levels that would potentially be attributable to other variables.

Table 7 reports results from the dichotomous moderator analyses on data set and methodological characteristics among studies using AB and reversal design variations. The mean proportions of IRA from the context variable resulted in low (.67 for context) and necessary (.73 for no context) levels. However, the median proportions for the context variables were equal (.74). The confidence interval ranges yielded some overlap. The provision of context failed to produce statistically significant findings attributable to the variance associated with IRA between categories. It is noteworthy that the p value of the variance between context categories was somewhat low (p = .10), favoring effects that provided no context. The Q_{within} did not indicate statistically significant differences within the moderator categories for this variable.

The mean proportions of IRA were low (.69) for study effects using no visual aid and reached a necessary level of IRA (.75) for effects using visual aids to supplement graphic displays. Median proportions reached necessary levels for effects using no visual aid (.74) and for effects using visual aids (.77). There was some overlap between the confidence interval ranges. Effects of a visual aid were not found to be statistically significant between moderator categories. However, the p value was still relatively low (p = .10), indicating that visual aids potentially contributed to a somewhat higher mean of IRA, although not to a statistically significant degree. Findings from the Q_{within} did not indicate statistically significant variability within the moderator categories for this variable.

The variable of whether studies defined what raters were analyzing resulted in a necessary mean proportion of IRA (.79) for those that provided operational definitions and low mean proportion of IRA (.69) for those that did not. Median proportions reached an adequate level for effects that provided operational definitions (.84) and a necessary level (.74) for those that did not. The confidence interval ranges around the effect sizes had some

Table 7. Moderator Analyses: Graphic and Methodological Characteristics Among AB/Reversal Design Variations.

Moderator	*K*	No. of data sets	Median proportion	Mean proportion	Mean logit proportion (95% CI)	Q_{between}	Q_{within}
Context	9	152	.74	.67	0.72 [0.47, 0.97]	2.64	20.92
No context	14	265	.74	.73	0.99 [0.78, 1.20]		
Visual aid	8	128	.77	.75	1.11 [0.78, 1.44]	2.69	20.88
No visual aid	15	289	.74	.69	0.80 [0.61, 0.99]		
Defined	7	123	.84	.79	1.30 [0.89, 1.71]	4.77*	18.80
Not defined	16	294	.74	.69	0.80 [0.63, 0.97]		
Scale of 2	14	265	.76	.76	1.13 [0.90, 1.36]	8.89**	14.68
Scale of 3	9	152	.69	.65	0.64 [0.41, 0.87]		

Note. *K* = number of effects; CI = confidence interval.
*p < .05. **p < .01.

overlap between the categories. The Q_{between} for this variable was statistically significant, indicating that operationally defining what was to be rated produced statistically greater effects than not defining. The Q_{within} was not found to be statistically significant, demonstrating that the findings between effects were likely not confounded by variability within moderator categories for this variable.

Study effects that included a scale of 2 resulted in a necessary mean proportion of IRA (.76) while effects using a scale of 3 resulted in a low mean proportion of IRA (.65). Median proportions of IRA also reached a necessary level among effects using a scale of 2 (.76) and a low level among effects using a scale of 3 (.69). There was no overlap in the ranges of confidence intervals around effects between categories. The Q_{between} analysis reached statistical significance for the moderator analysis of scale, demonstrating that a scale of 2 produced a statistically greater proportion of IRA relative to that of a scale of 3. The Q_{within} did not reach a statistically significant level, validating the statistically significant Q_{between} in that there was low variance within moderator categories.

Discussion

This purpose of this meta-analytic review was to determine the factors contributing to variable proportions of IRA between visual analysts of single-case data. We also descriptively analyzed the peer-reviewed literature in this area. The overall weighted mean proportion of IRA reached .76 for all 32 effects, an improvement relative to the findings of Ottenbacher (1993) in which a mean of .58 was found. Effects did not appear to improve linearly over the years since Ottenbacher's review, but there are potential reasons why the current meta-analysis produced larger effects. First, the current review only included peer-reviewed literature while Ottenbacher's review included effects from resources that were not peer-reviewed. Thus, the current review may include a degree of publication bias. Second, the scales of several effects were reduced in the current review to compute effects while Ottenbacher's review appeared to use the originally reported effects from studies that included a range of scale options. The number of rating scale options was found to be inversely related to IRA in the current review, so the scale reduction may have produced larger effects relative to Ottenbacher's findings. Finally, the current review only included proportional effect sizes, whereas Ottenbacher's review included both correlational and proportional metrics. Our inclusion of only proportional effect sizes may have produced disparate findings relative to the sample of effects used by Ottenbacher.

The first moderator analysis conducted in this meta-analysis was over categories of design families. Multielement designs versus AB and reversal design variations yielded large differences in effects supporting multielement designs with statistical significance between and within categories. Therefore, statistically significant findings between categories for this moderator were likely not only due to the design alone but also due to the characteristics of the studies within categories. Studies using multielement designs included relatively similar purposes, that is, to evaluate graphs of functional analyses on challenging behavior and identify the function(s) of those behaviors either with or without the use of or training in the use of criterion lines. Most effects using multielement designs included visual aids and defined the constructs to be rated ($K = 6$; 75%). The two effects without visual aids and defined constructs likely produced variance found within the multielement design variable, but this was not analyzed as a moderator due to the limited sample of effects. Also within studies using multielement designs, all raters had training in single-case research. These variables may have attributed to the increased IRA found among effects using multielement designs relative to findings from AB and reversal design variations.

Another finding was that the level of expertise among visual analysts may be influential to the variance of IRA proportions obtained across the included studies using AB and reversal designs. Interestingly, the lowest weighted mean proportion IRA was found between visual analysts who had the most expertise in single-case research. These finding are not unlike the outcomes of Ottenbacher (1993). It is unclear why raters with the most expertise in single-case research produced the lowest levels of IRA. One possible reason is the variety of backgrounds and training those experienced individuals received. Differing foci of training can produce incongruent effects (Hojem & Ottenbacher, 1988). It appears that raters classified as beginners (i.e., Bailey, 1984; Hojem & Ottenbacher, 1988; James et al., 1996; Skiba et al., 1989) received a common training in single-case research among one another, potentially affecting their IRA in a positive manner. However, novices did not receive a common training within studies and produced a somewhat higher weighted mean proportion of IRA relative to experienced raters. Still, the median proportions of novice raters and experienced raters were nearly equal. Also, there was overlap in confidence interval ranges around effects between the expertise categories. Because there were only four effects within the experienced category, this sample of effects may not adequately represent the normal distribution and should be interpreted with some caution.

The provision of context (e.g., dependent variable) to supplement graphic displays was another variable evaluated as a moderator. Although most study effects overall provided contextual information, those using AB and reversal design variations primarily did not. Although contextual information is an important concern according to visual analysts (DeProspero & Cohen, 1979; Grigg et al., 1989), this information does not appear to assist in increasing the consistency of ratings between respondents. In fact, it may contribute to some disagreement. However, the effects between categories of this variable were not largely different, confidence intervals overlapped, and statistical significance was not found. Therefore, contextual information may be more of a value to visual analysts than a moderator of IRA.

This meta-analysis also revealed that visual aids may account for some improvement in IRA between ratings of visual analysts, although overlap in confidence intervals was noted and statistical significance was not found. Although not evaluated as a moderator, explicit training in the use of evaluating data with a visual aid clearly produced large effects that were discrepant from no training in the use of such aids (Hagopian et al., 1997; Hojem & Ottenbacher, 1988; Roane et al., 2013; Skiba et al., 1989). Roane et al. (2013) compared the provision of written instruction over the use of criterion lines with written plus verbal instruction, finding that verbal instruction led to improved proportions of IRA. The two studies that did not find a visual aid to improve IRA also did not appear to explicitly train raters in the use of the visual aids (Normand & Bailey, 2006; Ottenbacher & Cusick, 1991). This indicates that professional development in the use of visual aids is crucial to benefit from their use, coinciding with an analysis conducted by Ottenbacher (1993).

Also based from our findings, specifically defining what visual analysts were to rate increased IRA to meet near acceptable standards according to the weighted mean proportion and acceptable standards according to the median proportion. Rating a concept such as “experimental control,” for instance, would require raters to understand the necessity of demonstrating control through the use of an experimental design to make valid judgments. To illustrate how important this may be, Kahng et al. (2010) reportedly used methods similar to DeProspero and Cohen (1979), with an exception being that they defined “experimental control” to participating raters. In comparison, Kahng et al. (2010) found considerably higher IRA. Heeding that the majority of studies in this review did not appear to define the constructs rated, along with the fact that that this moderated proportions of IRA, it may be interpreted that many studies did not demonstrate an important indicator of validity to support the certainty of overall IRA.

Finally, the last variable analyzed as a potential moderator in this meta-analysis was the use of a rating scale of 2 versus 3. As probable with respect to chance levels of agreement, effects obtained via a scale of 2 produced greater IRA than that obtained between raters using a scale of 3. Although a dichotomous scale produced larger proportions of IRA, the appropriate scale for use depends on the purpose of a study. However, this issue is somewhat complex. For instance, the construct of "experimental control" could lie on a continuous scale (Horner et al., 2012), but could also be appropriately represented on a dichotomous scale (Kahng et al., 2010).

Limitations

There are a number of biases and limitations to this meta-analysis regarding inclusionary criteria and methods. We only included studies yielding proportion of IRA effect sizes. Furthermore, only peer-reviewed journal articles were sought. It may have been beneficial to take more of an inclusionary approach by including descriptive results from studies using correlational metrics as well as effects reported in book chapters and dissertations. Accordingly, the results of this meta-analysis are not fully comprehensive of the existing literature. Also, the quality of research designs used among included studies varied. There is an evident need for more quality research in this area. It should be noted that some studies used methods to control for the potential confounds of resampling by regenerating aspects of the data (Hagopian et al., 1997) or by using different respondent samples (Hojem & Ottenbacher, 1988; Ottenbacher & Cusick, 1991).

Regarding methods, although dichotomizing data allowed for systematization and the evaluation of scale as a moderator, a dichotomous scale may not represent the range of potential interpretations for a given data set within research and practice (Horner et al., 2012). Relatedly, although the proportion or percentage of agreement calculation is commonly used, it presents with the limitation that chance agreement is not corrected (House et al., 1981). However, using proportion allowed us to determine the degree to which chance affected IRA. Finally, several study effects compared rater judgments with a priori determined ratings based on criteria used with visual analysis (Normand & Bailey, 2006; Roane et al., 2013). This could technically be regarded as more of a measure of accuracy than agreement (cf. Kazdin, 1977). However, these effects were included in this meta-analysis because a priori judgments were made by using visual analysis, and the use of such criteria would be evaluated as a moderator.

Future Research

Concerns over the literature base point to the need for research that represents the true elements of the visual analysis process for single-case experimental research (Parsonson & Baer, 1992). More research is needed to determine

IRA between visual analysts using experimental designs. Although most studies in this review used AB and reversal design variations, few used reversal designs demonstrating experimental control via replications of effect. In the current review, only one study used multiple baseline designs (across participants), and no studies used alternating treatments designs. It is recommended to evaluate IRA of visual analysis for alternating treatment (i.e., two alternating conditions) and multielement designs with data representing other uses of these designs beyond functional analysis, such as intervention comparisons. Continued study is warranted that evaluates IRA between visual analysts of multiple baseline designs, as this design is commonly used in research (Smith, 2012). Contextual descriptions of graphs, although not positively influential to IRA in this meta-analysis, is relevant in the decision-making process of visual analysis (DeProspero & Cohen, 1979; Grigg et al., 1989) and thus should be utilized when possible in future studies. Several studies included in this review evaluated IRA between clinicians with little to no experience in the use of single-case research. Because experienced visual analysts represent professionals in the field of applied behavior analysis, peer-reviewers, and educators of single-case research, this population would be useful to include in continued research. Sidman (1960) provided the first seminal discussion of visual analysis in which he describes "criterion-by-inspection" as a procedure an experienced decision maker uses to determine whether a predictable stability criterion is reached. Few prior studies evaluated the use of criterion lines as a visual aid (Hagopian et al., 1997, Roane et al., 2013) and none have appeared to evaluate a criterion-referenced behavioral objective to influence IRA between raters. It is likely that this fundamental practice influences IRA positively, as criteria provide a clear and objective outcome to be inspected. Finally, future studies on IRA between visual analysts should evaluate categorical scales that reflect possible options for data-based decisions (e.g., continue treatment, cease treatment, or add a component to the treatment).

Future research on the IRA of visual analysis should also face contemporary issues and evaluate practical solutions. It would be beneficial for studies to continue determining the effects of visual analysis training. For instance, studies should assess the IRA between visual analysts using structured criteria such as that established by the What Works Clearinghouse (Kratochwill et al., 2010, 2013) and adopted by Maggin et al. (2013). This systematic methodology for interpretation of single-case data could potentially increase consistency between visual analysts and provide a means of valid and reliable data aggregation of single-case research. Another area for research is the synthesized use of visual and statistical analyses, as these methods are often recommended to be used in conjunction (Carter, 2013; Fisch, 2001; Horner et al., 2012; Kratochwill & Levin, 2014; Manolov, Sierra, Solanas, & Botella, 2014; Parker & Vannest, 2012). Because effect sizes and visual analysis indices evaluate different constructs (Ninci et al., 2015), it is unclear whether supplementing visual analysis with effect sizes would result in increased proportions of IRA between raters.

Conclusion

Single-case research requires visual interpretation of across- and within-phase data patterns to reveal comprehensive information. Through peer review, that involves agreement in contextualized results interpretation, published single-case data sets may be defended against chance and probability of error (Baer, 1977). However, the levels of reliability in regard to visually analyzed ratings are often considered unacceptable. Including effect sizes provides a means of interpreting the reliability and generalizability of results; this can be useful for the acceptance of single-case research methods to influence the movement in identifying evidence-based practices (Gage & Lewis, 2013; Vannest & Davis, 2013). Yet parametric approaches applied to single-case data often violate critical assumptions that the data are independent (Parker & Vannest, 2012) and nonparametric statistical options available do not encompass all considerations salient to single-case data analysis (Carter, 2013). These issues have driven the refining of both visual analysis and statistical models in single-case research. Although single-case researchers have not demonstrated agreement in how best to refine analytic methods, one idea that many scholars in this area appear to agree upon is that visual analysis cannot be abandoned (e.g., Carter, 2013; Horner et al., 2012; Kratochwill & Levin, 2014; Parker & Vannest, 2012). Baer (1977) noted a clear yet clearly attainable need to develop techniques of visual analysis. Over more recent years, this challenge has been met with a new luster. Still, it is evident that systematized visual analytic methods must continue to be evaluated, replicated, and ultimately agreed upon to improve the reliability between visual analysts.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

*Indicates study included in this review.

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis*, *10*, 167-172. doi:10.1901/jaba.1977.10-167

*Bailey, D. B. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. *Journal of Applied Behavior Analysis*, *17*, 359-365. doi:10.1901/jaba.1984.17-359

*Bobrovitz, C. D., & Ottenbacher, K. J. (1998). Comparison of visual inspection and statistical analysis of single-subject data in rehabilitation research. *American Journal of Physical Medicine & Rehabilitation*, *77*, 94-102. doi:10.1097/00002060-199803000-00002

Bonner, M., & Barnett, D. W. (2004). Intervention-based school psychology services: Training for child-level accountability; preparing for program-level accountability. *Journal of School Psychology*, *42*, 23-43. doi:10.1016/j.jsp.2003.10.001

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, *30*, 531-563. doi:10.1177/0145445503261167

Burns, M. K. (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education*, *21*, 175-184. doi:10.1007/s10864-012-9158-9

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data: What they tell us and what they don't. *Behavior Modification*, *37*, 378-390. doi:10.1177/0145445513476609

*Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification*, *32*, 828-839. doi:10.1177/0145445508318606

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *12*, 573-579. doi:10.1901/jaba.1979.12-573

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes*, *54*, 137-154. doi:10.1016/S0376-6357(01)00155-3

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, NJ: Lawrence Erlbaum.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis*, *15*, 415-421. doi:10.1901/jaba.1982.15-415

Gage, N. A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology*, *25*, 46-60. doi:10.1080/10413200.2012.660673

Gast, D. L. (2010). *Single subject research methodology in behavioral sciences*. New York, NY: Routledge.

*Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *The Journal of Applied Behavioral Science*, *24*, 298-314. doi:10.1177/0021886388243007

Grigg, N. C., Snell, M. E., & Loyd, B. (1989). Visual analysis of student evaluation data: A qualitative analysis of teacher decision making. *Journal of the Association for Persons With Severe Handicaps*, *14*, 23-32.

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis*, *30*, 313-326. doi:10.1901/jaba.1997.30-313

*Harbst, K. B., Ottenbacher, K. J., & Harris, S. R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy*, *71*, 107-115.

Hartmann, D. P., Barrios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In S. N. Haynes & E. M. Hieby (Eds.), *Comprehensive handbook of psychological assessment, behavioral assessment* (Vol. 3, pp. 108-127). New York, NY: John Wiley.

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

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557-560. doi:10.1136/bmj.327.7414.557

*Hojem, M. A., & Ottenbacher, K. J. (1988). Empirical investigation of visual-inspection versus trend-line analysis of single-subject data. *Physical Therapy*, *68*, 983-988.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education & Treatment of Children*, *35*, 269-290. doi:10.1353/etc.2012.0011

House, A. E., House, B. J., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Journal of Behavioral Assessment*, *3*, 37-57. doi:10.1007/BF01321350

*James, I. A., Smith, P. S., & Milne, D. (1996). Teaching visual analysis of time series data. *Behavioural and Cognitive Psychotherapy*, *24*, 247-262. doi:10.1017/S1352465800015101

*Johnson, M. B., & Ottenbacher, K. J. (1991). Trend line influence on visual analysis of single-subject data in rehabilitation research. *Disability and Rehabilitation*, *13*, 55-59. doi:10.3109/03790799109166685

*Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, *11*, 277-283. doi:10.1901/jaba.1978.11-277

*Kahng, S. W., Chung, K. M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *43*, 35-45. doi:10.1901/jaba.2010.43-35

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, *10*, 141-150. doi:10.1901/jaba.1977.10-141

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

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment*, *5*, 155-164.

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

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26-38. doi:10.1177/0741932512452794

Kratochwill, T. R., & Levin, J. R. (2014). Meta- and statistical analysis of single-case intervention research data: Quantitative gifts and a wish list. *Journal of School Psychology*, *52*, 231-235. doi:10.1016/j.jsp.2014.01.003

*Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly*, *25*, 28-44. doi:10.1037/a0018600

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

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the What Works Clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education*, *34*, 44-58. doi:10.1177/0741932511435176

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*, 878-913. doi:10.1177/0145445514545679

Matyas, T. A., & Greenwood, K. M. (1990a). The effect of serial dependence on visual judgment of single-case charts: An addendum. *Occupational Therapy Journal of Research*, *10*, 208-220.

Matyas, T. A., & Greenwood, K. M. (1990b). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341-351. doi:10.1901/jaba.1990.23-341

Mercer, S. H., & Sterling, H. E. (2012). The impact of baseline trend control on visual analysis of single-case data. *Journal of School Psychology*, *50*, 403-419. doi:10.1016/j.jsp.2011.11.004

Ninci, J., Neely, L. C., Hong, E. R., Boles, M. B., Gilliland, W. D., Ganz, J. B., . . .Vannest, K. J. (2015). Meta-analysis of single-case research on teaching functional living skills to individuals with ASD. *Review Journal of Autism and Developmental Disorders*. Advance online publication. doi:10.1007/s40489-014-0046-1

*Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification*, *30*, 295-314. doi:10.1177/0145445503262406

*Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject designs. *American Journal of Occupational Therapy*, *40*, 464-469. doi:10.5014/ajot.40.7.464

*Ottenbacher, K. J. (1990a). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation*, *28*, 283-290.

*Ottenbacher, K. J. (1990b). When is a picture worth a thousand *p* values? A comparison of visual and quantitative methods to analyze single subject data. *Journal of Special Education*, *23*, 436-449. doi:10.1177/002246699002300407

Ottenbacher, K. J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal on Mental Retardation*, *98*, 135-142.

*Ottenbacher, K. J., & Cusick, A. (1991). An empirical investigation of interrater agreement for single-subject data using graphs with and without trend lines. *Journal of the Association for Persons With Severe Handicaps*, *16*, 48-55. doi:10.1177/154079699101600106

*Park, H. S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *The Journal of Experimental Education*, *58*, 311-320. doi:10.1080/00220973.1990.10806545

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education*, *21*, 254-265. doi:10.1007/s10864-012-9153-1

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15-40). Hillsdale, NJ: Lawrence Erlbaum.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, *22*, 109-116.

*Roane, H. S., Fisher, W. W., Kelley, M. E., Mevers, J. L., & Bouxsein, K. J. (2013). Using modified visual-inspection criteria to interpret functional analysis outcomes. *Journal of Applied Behavior Analysis*, *46*, 130-146. doi:10.1002/jaba.13

Rojahn, J., & Schulze, H. H. (1985). The linear regression line as a judgmental aid in visual analysis of serially dependent A-B time-series data. *Journal of Psychopathology and Behavioral Assessment*, *7*, 191-205. doi:10.1007/BF00960752

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York, NY: Basic Books.

*Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subject familiarity on practitioners' judgments of intervention effectiveness. *The Journal of Special Education*, *22*, 433-446. doi:10.1177/002246698902200405

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*, 510-550. doi:10.1037/a0029312

Spirrison, C. L., & Mauney, L. T. (1994). Acceptability bias: The effects of treatment acceptability on visual analysis of graphed data. *Journal of Psychopathology and Behavioral Assessment*, *16*, 85-94. doi:10.1007/BF02229067

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.), *Advances in Learning and Behavioral Disabilities: Vol. 26. Evidence-based practices* (pp. 93-119). West Yorkshire. England: Emerald Group.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. *Behavioral Assessment*, *3*, 79-82.

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual analysis in single-case designs. *Spanish Journal of Psychology*, *12*, 823-832.

Author Biographies

Jennifer Ninci, MEd, BCBA, is a doctoral student in the Special Education program at Texas A&M University. Her research interests include educational interventions for individuals with developmental and intellectual disabilities and single-case research methodology.

Kimberly J. Vannest is a professor in the Special Education program in the Department of Educational Psychology at Texas A&M University. Her research interests include interventions for students with emotional and behavioral disorders, the measurement of effects, and single-case research methodology.

Victor Willson is head and professor of educational psychology at Texas A&M University. He conducts research in multilevel longitudinal modeling and structural equation modeling. He has conducted a number of meta-analytic studies over the last 30 years.

Nan Zhang, MA, is a doctoral student in the Special Education program at Texas A&M University. Her research interests include interventions for individuals with internalizing disorders and meta-analytic methodology.