




Systematic Protocols for the Visual Analysis of Single-Case Research Data

Katie Wolfe¹  · Erin E. Barton² · Hedda Meadan³

Published online: 28 January 2019

© Association for Behavior Analysis International 2019

Abstract

Researchers in applied behavior analysis and related fields such as special education and school psychology use single-case designs to evaluate causal relations between variables and to evaluate the effectiveness of interventions. Visual analysis is the primary method by which single-case research data are analyzed; however, research suggests that visual analysis may be unreliable. In the absence of specific guidelines to operationalize the process of visual analysis, it is likely to be influenced by idiosyncratic factors and individual variability. To address this gap, we developed systematic, responsive protocols for the visual analysis of A-B-A-B and multiple-baseline designs. The protocols guide the analyst through the process of visual analysis and synthesize responses into a numeric score. In this paper, we describe the content of the protocols, illustrate their application to 2 graphs, and describe a small-scale evaluation study. We also describe considerations and future directions for the development and evaluation of the protocols.

Keywords Visual analysis · Single-case research · Visual inspection · Data analysis

Single-case research (SCR) is the predominant methodology used to evaluate causal relations between interventions and target behaviors in applied behavior analysis and related fields such as special education and psychology (Horner et al., 2005; Kazdin, 2011). This methodology focuses on the individual case as the unit of analysis and is well suited to examining the effectiveness of interventions. SCR facilitates a fine-grained analysis of data patterns across experimental phases, allowing researchers to identify the conditions under which a given intervention is effective for particular participants (Horner et al., 2005; Ledford & Gast, 2018). In addition, the dynamic nature of SCR allows the researcher to make adaptations to phases and to conduct component analyses of intervention packages with nonresponders to empirically identify optimal treatment components (Barton et al., 2016; Horner et al., 2005).

Visual analysis is the primary method by which researchers analyze SCR data to determine whether a causal relation (i.e., functional relation, experimental control) is documented (Horner et al., 2005; Kratochwill et al., 2013). Visual analysis involves examining graphed data within and across experimental phases. Specifically, researchers look for changes in the level, trend, or variability of the data across phases that would not be predicted to occur without the active manipulation of the independent variable. *Level* is the amount of behavior that occurs in a phase relative to the *y*-axis (Barton, Lloyd, Spriggs, & Gast, 2018). *Trend* is the direction of the data over time, which may be increasing, decreasing, or flat (Barton et al., 2018). *Variability* is the spread or fluctuation of the data around the trend line (Barton et al., 2018). A change in the level, trend, or variability of the data between adjacent phases is a basic effect; to determine whether there is a causal relation, the researcher looks for multiple replications of the effect at different and temporally related time points (Kratochwill et al., 2013).

Despite this reliance on visual analysis, there have been long-standing concerns about *interrater agreement*, or the extent to which two visual analysts evaluating the same graph make the same determination about functional relations and the magnitude of change. In general, these concerns have been borne out by empirical research (e.g., Brossart, Parker, Olson, & Mahadevan, 2006; DeProspero & Cohen, 1979; Wolfe, Seaman, & Drasgow, 2016). In one study, Wolfe et al.

✉ Katie Wolfe
kmwolfe@mailbox.sc.edu

¹ Department of Educational Studies, University of South Carolina, 820 Main St, Columbia, SC 29208, USA

² Department of Special Education, Vanderbilt University, Box 228 GPC, Nashville, TN 37203, USA

³ Department of Special Education, University of Illinois at Urbana-Champaign, 1310 South Sixth Street, Champaign, IL 61820, USA

(2016) asked 52 experts to report whether each of 31 published multiple-baseline design graphs depicted (a) a change in the dependent variable from baseline to intervention for each tier of the graph and (b) an overall functional relation for the entire multiple-baseline design graph. Interrater agreement was just at or just below minimally acceptable standards for both types of decisions (intraclass correlation coefficient [ICC] = .601 and .58, respectively). The results of this study are generally representative of the body of literature on interrater agreement among visual analysts (cf. Kahng et al., 2010). Given that visual analysis is integral to the evaluation of SCR data (Horner & Spaulding, 2010; Kazdin, 2011), research indicating that it is unreliable under many circumstances presents a significant challenge for the field—particularly the acceptance of SCR as a credible and rigorous research methodology.

Many researchers have argued that poor agreement among visual analysts may be due to the absence of formal guidelines to operationalize the process (Furlong & Wampold, 1982), which leaves the analysis vulnerable to idiosyncratic factors and individual variability related to “history, training, experience, and vigilance” (Fisch, 1998, p. 112). Perhaps due to the lack of formal guidelines, single-case researchers rarely identify, let alone describe, the methods by which they analyze their data. Smith (2012) reported that authors in fewer than half of the SCR studies published between 2000 and 2010 ($n = 409$) identified the analytic method they used; only 28.1% explicitly stated that they used visual analysis. Even less frequently do authors describe the specific procedure by which visual analysis was conducted. In a review of SCR articles published in 2008 ($n = 113$), Shadish and Sullivan (2011) found that only one study reported using a systematic procedure for visual analysis (Shadish, 2014). Barton, Meadan, and Fetting (2019) found similar results in a review of parent-implemented functional assessment interventions; study authors rarely and inconsistently used visual analysis terms and procedures across SCR studies and were most likely to discuss results using only mean, median, and average rather than level, trend, or variability. Overall, it is difficult to identify specifically how single-case researchers are conducting visual analysis of their data, which might lead to high rates of disagreement and adversely impact interpretations of results and syntheses across SCR. In other words, unreliable data analysis may impede the use of SCR to identify evidence-based practices, which has important and potentially adverse practical and policy implications.

There have been a few recent efforts to produce and disseminate standards that may promote greater consistency in visual analysis. The What Works Clearinghouse (WWC) Single-Case Design Standards (Kratochwill et al., 2013; WWC, 2017) describe four steps for conducting visual analysis that consider six data characteristics (i.e., level, trend, variability, immediacy, overlap, and consistency). However,

the WWC standards were not designed to provide a systematic, step-by-step protocol to guide the visual analysis process (Hitchcock et al., 2014) and do not assist researchers in synthesizing information about the data characteristics and across experimental phases. For example, the four steps do not explain the relative importance of the data characteristics in making determinations about basic effects and experimental control. This ambiguity could introduce subjectivity into the analysis and result in two visual analysts reaching different conclusions about the same graph despite using the same procedures.

To increase agreement among visual analysts working on reviews of SCR literature, Maggin, Briesch, and Chafouleas (2013) developed a visual analysis protocol based on the WWC SCR standards (Kratochwill et al., 2013). Using this protocol, the analyst answers a series of questions about the graph and then uses these responses to determine the number of basic effects and the level of experimental control demonstrated by the graph (Maggin et al., 2013). Maggin et al. (2013) reported high agreement between the three authors following training on the protocol (e.g., 86% agreement), which suggests that structured, step-by-step protocols could be an effective way to increase consistency among visual analysts. Their protocol guides researchers through visual analysis procedures; however, it does not assist the researcher in synthesizing the six data characteristics within and across phases to make determinations about basic effects, experimental control, or weighing conflicting data patterns for making a judgment about functional relations. This introduces potential variability that could produce inconsistencies across different individuals and studies. The study by Wolfe et al. (2016) provides empirical evidence of this variability. They found that experts vary in the minimum number of effects they require to identify a functional relation. Some experts identified functional relations when there were three basic effects, but other experts identified a functional relation with only two basic effects. In other words, two experts may come to the same conclusions about the presence of basic effects in a particular graph, but they may translate that information into different decisions about the presence of a functional relation. Structured criteria that systematize the process of translating the within- and across-phase analysis into a decision about the overall functional relation may reduce this variability and improve agreement.

Researchers have developed structured criteria for the analysis of a specific type of SCR design used for a specific purpose. Hagopian et al. (1997) developed criteria for evaluating multielement graphs depicting the results of a functional analysis. The criteria consist of a step-by-step process that leads to a conclusion about the function of the behavior depicted in the graph. Hagopian et al. (1997) evaluated the effects of the criteria with three predoctoral interns in a multiple-baseline design and showed that participants' agreement with the first

author increased from around 50% in baseline to an average of 90% following training in the use of the structured criteria. The work of Hagopian et al. (1997) demonstrates that structured criteria can be developed for SCR that synthesize the user's responses and lead directly to a conclusion about the data. Further, the use of the criteria improved agreement between raters and experts. However, the Hagopian et al. (1997) criteria apply only to multielement graphs used for a specific purpose and cannot be applied to other SCR designs.

To address the shortcomings of current practice and standards in visual analysis, we developed systematic, web-based protocols for the visual analysis of A-B-A-B and multiple-baseline design SCR data that consist of a series of questions for the analyst to answer that synthesizes the analyst's responses to produce a numerical rating of experimental control for the graph. We designed our protocols to emphasize the six data characteristics outlined in the WWC (2017) SCR standards (i.e., level, trend, variability, immediacy, overlap, and consistency) and to support single-case researchers in making decisions about data patterns based on these characteristics. Further, our protocols guide the researchers in systematically making decisions about data patterns within and across phases and tiers to make judgments about functional relations. In this paper we describe the protocols, illustrate their application to two SCR graphs, and discuss findings from an initial evaluation study.

Content and Structure of the Protocols

We developed two step-by-step protocols, one for A-B-A-B designs, and one for multiple-baseline designs, to guide the analyst through the process of evaluating SCR data. The protocols are accessible as web-based surveys and as Google Sheets; both formats can be accessed from <https://sites.google.com/site/scrvaprotocols/>. Each protocol consists of a series of questions with dichotomous response options (i.e., yes or no) about each phase and phase contrast in the design. The questions in each protocol are based on current published standards for SCR (Kratochwill et al., 2013), as well as guidelines for visual analysis published in textbooks on SCR (e.g., Cooper, Heron, & Heward, 2007; Kazdin, 2011; Ledford & Gast, 2018). Table 1 lists the relevant sources that support the inclusion of the questions in each protocol and also provides evidence of the protocols' content validity. Each question in the protocols includes instructions and graphic examples illustrating potential "yes" and "no" responses. In the web-based survey, these instructions appear when the user hovers over a question. In Google Sheets, the instructions are accessed by clicking on a link in the spreadsheet.

The basic process for assessing each phase using the protocols includes examining both within- and between-phase data patterns (Kratochwill et al., 2013). First, the protocol

prompts the visual analyst to evaluate the stability of the data within a given phase. Second, if there is a predictable pattern, the visual analyst projects the trend of the data into the subsequent phase and determines whether the level, trend, or variability of the data in this subsequent phase differs from the pattern predicted from the previous phase. If there was a change in the data between the two phases, then the analyst identifies if that change was immediate and measures the data overlap between the two phases. If there is not a change between the two phases, the analyst is directed to proceed to the next phase contrast. If multiple data paths are depicted on an A-B-A-B or multiple-baseline design graph, the data paths typically represent different dependent variables. In these cases, each data path should be analyzed with a separate application of the protocol to determine the presence of a functional relation between the independent variable and each dependent variable.

The protocols are response guided (i.e., responsive to the analyst's input) and route the analyst through the process based on responses to previous questions. For example, if there are not sufficient data in the baseline phase to predict the future pattern of behavior, then the analyst cannot project the trend of the baseline data into the intervention phase to evaluate whether the data changed from the predicted pattern. In this case, the protocol skips ahead to questions about the next phase. Likewise, if the analyst responds that there is not a change in the dependent variable from one phase to the next, the protocol skips questions about immediacy and overlap, which are not relevant if the data did not change. The protocols are dynamic—some questions act as gatekeepers, making other questions available or unavailable based on the user's response.

Unlike other systematic guidelines for visual analysis (e.g., Maggin et al., 2013), the protocols generate an experimental control score for the graph based on the analyst's responses to the questions. Specific questions in the protocols have weighted values based on their importance to demonstrating a functional relation, and the sum of these values produces the experimental control score for the graph. Scores generated by the protocols range from 0 (no functional relation) to 5 (functional relation with large behavioral change), with 3 being the minimum score for evidence of a functional relation. Published guidelines for the analysis of SCR suggest that three basic effects, or changes in the level, trend, or variability of the dependent variable from one phase to the next, are required to demonstrate a functional relation (Barton et al., 2018; Kratochwill et al., 2013). Therefore, the questions pertaining to changes between adjacent phases (i.e., phase contrast questions) have a value of 1 in the protocols. As a result, a study depicting three basic effects would earn a minimum score of 3, which is the minimum criterion for demonstrating a functional relation based on our proposed interpretation guidelines.

Table 1 Alignment of protocol content with published recommendations for visual analysis

Protocol Content	Cooper et al. (2007)	Ledford and Gast (2018)	Kazdin (2011)	Kratochwill et al. (2013)
A-B-A-B and Multiple-Baseline Design Protocols				
Documentation of a predictable within-phase data pattern	X	X	X	X
Comparison of projected pattern to actual pattern in adjacent phases	X		X	X
Level, trend, or variability change between adjacent phases	X	X	X	X
Immediacy of change between adjacent phases	X	X	X	X
Overlap between adjacent phases	X	X	X	X
Consistency between similar phases	X	X		X
Multiple-Baseline Design Protocol Only				
Staggering of introduction of treatment across tiers	X	X	X	X
Vertical analysis	X	X	X	X

X = item is referenced in source

Other questions may not be critical to the demonstration of a functional relation but strengthen the evidence of a functional relation if one is present. For example, depending on the nature of the dependent variable, it may not be essential that the data change immediately after the introduction of the intervention (i.e., within 3–5 data points) to demonstrate a functional relation (Kazdin, 2011). However, an immediate change increases the analyst's confidence that the intervention caused the change in the dependent variable. Therefore, questions about the immediacy of the effect have a smaller weight (e.g., 0.25; A-B-A-B protocol) compared to questions about identifying basic effects.

Similarly, minimal overlap between the data paths in adjacent phases is generally considered desirable but not necessary nor always meaningful (e.g., data might have substantial overlap but contrasting trends) for demonstrating functional relations (Barton et al., 2018). Therefore, the overlap item also has a smaller weight (e.g., 0.25; A-B-A-B protocol). Phase contrasts must have 30% or fewer overlapping data points to receive points for this item in the protocol. This criterion is based on the interpretive guidelines proposed for the percentage of nonoverlapping data (Scruggs & Mastropieri, 1998), which suggest that 70% of nonoverlapping data between phases indicates an effective intervention (note that the protocol asks the analyst to calculate the inverse, or the amount of overlapping data, and thus the criterion is set at 30%).

In the multiple-baseline design protocol, we assigned the questions pertaining to vertical analysis a negative value. Vertical analysis refers to the examination of the data in tiers that remain in baseline when the intervention is introduced to a previous tier (Horner, Swaminathan, Sugai, & Smolkowski, 2012). Other sources refer to this same feature as *verification of the change* in the previous tier (Cooper et al., 2007). If the baseline data for any tiers still in baseline change markedly when the intervention is introduced to another tier, this indicates a potential alternative explanation for any observed

change (e.g., behavioral covariation, history, maturation) and decreases confidence that the intervention was causally related to the change in the dependent variable. This question has a negative value because if the analyst answers “yes,” it detracts from the overall experimental control score for the graph.

Although we have proposed interpretation guidelines for the scores generated by the protocols, the score should be interpreted within the context of the study's overall methodological quality and rigor; if the study has strong internal validity, minimizing plausible alternative explanations, then the score produced by the protocol can indicate the presence and strength of a functional relation. However, if the study is poorly designed or executed or is missing key features (e.g., interobserver agreement [IOA], procedural fidelity), or if key features are insufficient to rule out threats to internal validity (e.g., IOA is less than 80%, missing data), then the score produced by the protocol may be misleading because the methodological rigor limits interpretations of the data.

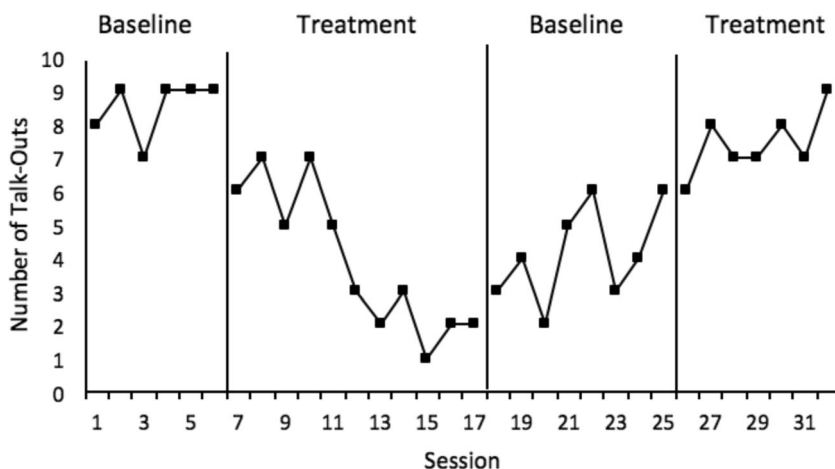
Application of the Protocols

Although we cannot demonstrate the dynamic and responsive nature of the protocols in this article, we will walk through two examples to illustrate how they are applied to SCR data. Both of the graphs used to illustrate the application of the protocols were used in our reliability and validity evaluations of the protocols. We encourage the reader to access the protocols in one or both formats to explore the content, structure, routing, and scoring that will be illustrated in the next sections.

A-B-A-B Design Protocol

Figure 1 depicts a hypothetical A-B-A-B graph showing the number of talk-outs within a session, and Fig. 2 shows the

Fig. 1 Sample A-B-A-B graph



completed protocol for this graph. Use of the protocol involves comparing the first baseline phase to the first treatment phase (A1 to B1), the first treatment phase to the second baseline phase (B1 to A2), and the second baseline phase to the

second treatment phase (A2 to B2). We also compare the data patterns in similar phases (i.e., A1 to A2 and B1 to B2).

The protocol starts by prompting the visual analyst to examine the first baseline phase. There are three data points, and

Fig. 2 Completed protocol for sample A-B-A-B graph

ABAB Protocol (Wolfe, Barton, & Meadan)			
Question	Select Response	Point Value	Protocol Routing
1 FIRST BASELINE: Are there at least 3 data points in the first baseline phase and can you predict the future pattern of the behavior?	Y	1	If #1 & #2 are yes, continues to #3
2 FIRST TREATMENT: Are there at least 3 data points in the first treatment phase and can you predict the future pattern of the behavior?	Y	1	If #1 or #2 is no, skips to #6
3 FIRST PHASE CONTRAST: Project the trend of the first baseline phase into the first treatment phase. Is the level, trend, or variability of the data in the treatment phase different than what you would predict based on the baseline data?	Y	1	If #3 is yes, continues to #4 If #3 is no, skips to #6
4 IMMEDIACY: Is there an immediate change from the last 3–5 data points in the first baseline phase to the first 3–5 data points in the first treatment phase?	Y	0.25	Continues to #5
5 OVERLAP: Do less than 30% of the data points in the first treatment phase overlap with the data points in the first baseline phase?	Y	0.25	Continues to #6
6 SECOND BASELINE: Are there at least 3 data points in the second baseline phase and can you predict the future pattern of the behavior?	Y	1	If #6 is yes, continues to #7 If #6 is no, skips to #11
7 SECOND PHASE CONTRAST: Project the trend of the first treatment phase into the second baseline phase. Is the level, trend, or variability of the data in the baseline phase different than what you would predict based on the preceding treatment data?	Y	1	If #7 is yes, continues to #8 If #7 is no, skips to #11
8 IMMEDIACY: Is there an immediate change from the last 3–5 data points in the first treatment phase to the first 3–5 data points in the second baseline phase?	Y	0.25	Continues to #9
9 OVERLAP: Do less than 30% of the data points in the second baseline phase overlap with the data points in the first treatment phase?	N	0	Continues to #10
10 CONSISTENCY: Are the data patterns of the first and second baseline phases similar in level, trend, or variability?	N	0	Continues to #11
11 SECOND TREATMENT: Are there at least 3 data points in the second treatment phase and can you predict the future pattern of the behavior?	Y	1	If #11 is yes, continues to #12 If #11 is no, skips to end
12 THIRD PHASE CONTRAST: Project the trend of the second baseline phase into the second treatment phase. Is the level, trend, or variability of the data in the treatment phase different than what you would predict based on the preceding baseline data?	N	0	If #12 is yes, continues to #13 If #12 is no, skips to #15
			Continues to #14
			Continues to #15
15 CONSISTENCY: Are the data patterns of the first and second treatment phases similar in level, trend, or variability?	N	0	End of protocol
	SCORE	2.75	

those data are stable—we predicted that if baseline continued, the data would continue to decrease—so we answered “yes” to the first question. The second question asks us to evaluate the first treatment phase in the same manner, and given the number of data points and the overall decreasing trend, we answered “yes” to this question as well. Next, we are directed to project the trend of the first baseline phase into the first treatment phase and evaluate whether the level, trend, or variability of the treatment data is different from our prediction. The level is different from our prediction, so we answered “yes,” identifying a basic effect between these phases. The identification of a basic effect for this phase contrast makes the next two questions available.

Regarding immediacy, the level of the data did change from the last three data points in baseline to the first three data points in treatment, so we selected “yes.” To identify the amount of overlap between the two phases, we drew a horizontal line extending from the highest baseline datum point into the first treatment phase because the goal of the intervention was to increase the behavior. Next, we counted the number of data points in the first treatment phase that are the same or “worse” than this line. Whether “worse” data are higher or lower than the line will depend on the targeted direction of behavior change. In this case, the goal was to increase the behavior, so treatment data points that are the same as or below the line would be considered worse. There are no treatment data points below the line, so there is no overlapping data between these two phases. If there were data points below the line, we would divide the number of data points below the line by the total number of data points in the treatment phase to get the percentage of overlapping data. We answered “yes” because less than 30% of the data overlaps between the two phases.

The majority of the remaining A-B-A-B protocol involves answering this same series of questions about the remaining phases and phase contrasts; however, it is important to note that in the second phase contrast (i.e., the comparison from the first treatment phase to the second baseline phase), a basic effect would be demonstrated by a decrease in the number of talk-outs relative to our prediction from the treatment phase. Because the expected direction of behavior change is different for this particular phase contrast, the procedure for calculating overlapping data differs slightly as well (see instructions for this question in the protocol). The A-B-A-B protocol also includes two questions about the consistency of the data patterns across like phases. These questions involve examining the similarity of the level, trend, or variability of the data across (a) both baseline phases and (b) both treatment phases to evaluate if any of these characteristics are similar. For this graph, the data in the first baseline phase have a low level, little variability, and a decreasing trend. The data in the second baseline phase have a medium level, medium variability, and no clear trend. Therefore, we answered “no” to the question

about consistency between the baseline phases. Based on our dichotomous responses to the questions in the protocol, the overall score for experimental control for this graph is 2.75, which does not provide evidence of a functional relation. To see answers and scoring for the complete protocol for this graph, as well as details about how the protocol routes the user through relevant questions based on responses, we encourage the reader to examine Fig. 2 in detail.

Multiple-Baseline Design Protocol

Similar to the A-B-A-B protocol, the multiple-baseline design protocol requires that the analyst examine each phase and phase contrast in the design. However, consistent with the logic of a multiple-baseline design, use of this protocol involves both comparing baseline to treatment for each tier (i.e., A to B) and determining if the introduction of the intervention was staggered in time across tiers and whether the dependent variable changed when and only when the intervention was applied (i.e., vertical analysis).

Figure 3 shows a hypothetical multiple-baseline design depicting the percentage of steps of a hygiene routine completed independently, and Fig. 4 is the completed protocol for this graph. The first question in the protocol involves the stability of the baseline data in the first tier. The phase does have three data points, but the variability of the data makes it difficult to project the overall pattern of the behavior, and as a result, we answered “no” to this question. This made the next four questions unavailable; if we cannot predict the future pattern of the baseline data, then we cannot project the trend into the treatment phase and make a confident determination about the presence of a basic effect. The next available question is about the stability of the baseline data in the second tier. This phase has more than three data points, and they are fairly stable around 10–20%, so we answered “yes.” Next, we looked at whether the baseline data in Tier 2 changed when the intervention began with Tier 1, which was after Session 3. The data in Tier 2 remain stable during and immediately after that session, so we answered “no” for this question. The next question asks if the treatment was introduced to Tier 2 after it was introduced to Tier 1; it was, so we answered “yes.” Had this question been answered “no,” the remaining questions for Tier 2 would become unavailable.

We continue by examining the stability of the Tier 2 treatment phase, and we have more than three data points and a clear upward trend, so we answered “yes.” Projecting the trend of the baseline phase into the treatment phase for Tier 2, we see there is a change in both the level and trend of the treatment data compared to our prediction from baseline, so we answered “yes.” That change was immediate (i.e., within the first 3–5 data points of treatment), so we answered “yes” to the next question about immediacy. Calculating overlap as

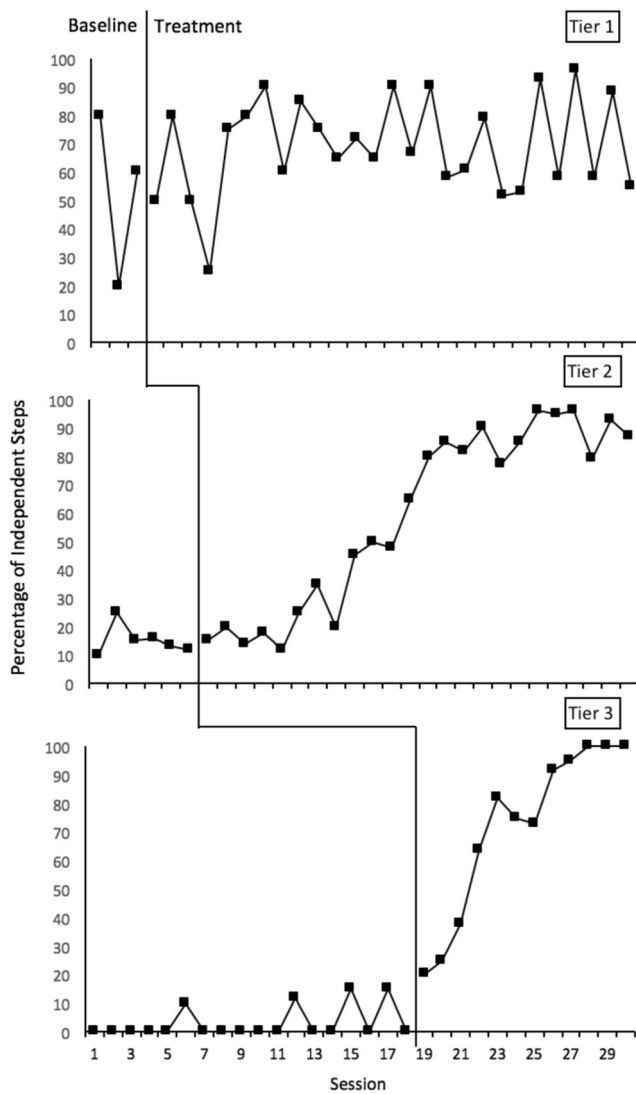


Fig. 3 Sample multiple-baseline design graph

previously described, we calculated 13% overlap between the two phases (1 overlapping datum point out of 8 total treatment data points), which is less than 30%, so we answered “yes.” The last question about this tier asks us to examine the similarity of data patterns between the treatment phases for Tier 1 and Tier 2. The tiers have similar levels, trends, and variability, so our response was “yes.”

The remainder of the multiple-baseline design protocol includes these same questions about the third tier in the design. Notably, the Tier 3 baseline data did change after Session 3, when the treatment was introduced to Tier 1, so we answered “yes” to the question about vertical analysis for Tier 3. Based on our dichotomous responses to the questions in the protocol, our overall score for experimental control for this graph was 2.32. To see answers and scoring for the complete protocol for this graph, as well as details about how the protocol routes the user through relevant questions based on responses, examine Fig. 4 in detail.

Evaluation of the Protocols

We conducted an initial evaluation of the reliability and validity of the protocols. We evaluated the reliability of the protocols by comparing the interrater agreement produced by the protocols to interrater agreement produced by a visual analysis rating scale. We evaluated the validity of the protocols by comparing scores produced by the protocols to scores assigned to the graphs by expert visual analysts using a rating scale.

Reliability Evaluation

To evaluate the reliability of the protocols, we recruited 16 attendees at an international early childhood special education conference held in a large city in the Southeastern United States. Attendees had to have taken a graduate-level course in SCR to participate in the evaluation. Nine participants reported that their terminal degree was a doctorate and designated their primary roles as university faculty or researchers, and seven reported that their terminal degree was a master’s and indicated that they were students. Participants were randomly assigned to the rating scale group ($n = 8$) or the protocol group ($n = 8$) and were split fairly evenly between the two groups based on highest degree earned (e.g., the protocol group consisted of three participants with doctorates and five with master’s degrees).

Each of the three authors independently used the protocols with 48 randomly selected published SCR graphs (24 A-B-A-B; 24 multiple-baseline design) during the iterative development process. From this set, we identified four A-B-A-B graphs and four multiple-baseline graphs with (a) ratings across the range of the protocol (i.e., 0–5) and (b) differences of 0.5 to 1.5 in our expert ratings based on our independent applications of the protocol. These criteria were used to ensure that we included diverse graphs in terms of both (a) the presence and absence of basic effects and functional relations and (b) graph difficulty (e.g., graphs with data with more variability or smaller changes might be difficult to visually analyze). We quantified difficulty using the range of scores produced by our independent applications of the protocol, such that graphs with more disparate scores between the authors were considered more difficult.

All study materials (i.e., graphs, rating scale, protocol) were uploaded into an online survey platform, and participants accessed the survey from the web browser on their personal laptop or tablet. All participants took a pretest on which they scored the eight graphs using a rating scale from 0 to 5. All points on the rating scale were defined as illustrated in Table 2, and the terms *basic effect* and *functional relation* were defined on each page of the pretest. Then, based on their random group assignments, participants rated the same eight graphs using either the rating scale or the systematic protocols.

Multiple-Baseline Design Protocol (Wolfe, Barton, & Meadan)			
Question	Response	Point Value	Protocol Routing
1 TIER 1 BASELINE: Are there at least 3 data points in baseline for Tier 1 and can you predict the future pattern of the behavior?	N	0	If #1 is yes, continues to #2 If #1 is no, skips to #6
2			If #2 is yes, continues to #3 If #2 is no, skips to #6
3			If #3 is yes, continues to #4 If #3 is no, skips to #6
4			Continues to #5
5			Continues to #6
6 TIER 2 BASELINE: Are there at least 3 data points in baseline for Tier 2 and can you predict the future pattern of the behavior?	Y	0	If #6 is yes, continues to #7 If #6 is no, skips to #13
7 TIER 2 VERTICAL ANALYSIS: Did the baseline data path of Tier 2 change when the treatment began with Tier 1?	N	0	Continues to #8
8 TIER 2 STAGGERED TREATMENT: Was the treatment introduced to Tier 2 AFTER it was introduced to Tier 1?	Y	0	If #8 is yes, continues to #9 If #8 is no, skips to #13
9 TIER 2 TREATMENT: Are there at least 3 data points in treatment for Tier 2 and can you predict the future pattern of the behavior?	Y	0	If #9 is yes, continues to #10 If #9 is no, skips to #13
10 TIER 2 PHASE CONTRAST: Project the trend of Tier 2 baseline into the treatment phase. Is the level, trend, or variability of the data in the Tier 2 treatment phase different than what you would predict based on the baseline data for Tier 2?	Y	1	If #10 is yes, continues to #11 If #10 is no, skips to #13
11 TIER 2 IMMEDIACY: Is there an immediate change from the last 3–5 data points in baseline to the first 3–5 data points in treatment for Tier 2?	N	0	Continues to #12
12 TIER 2 OVERLAP: Do less than 30% of the treatment data points overlap with the baseline datapoints for Tier 2?	Y	0.22	Continues to #13
13 CONSISTENCY: Are the data patterns of treatment for Tier 1 and Tier 2 similar in level, trend, or variability?	N	0	Continues to #14
14 TIER 3 BASELINE: Are there at least 3 data points in baseline for Tier 3 and can you predict the future pattern of the behavior?	Y	0	If #14 is yes, continues to #15 If #14 is no, skips to #21
15 TIER 3 VERTICAL ANALYSIS: Did the baseline data path of Tier 3 change when the treatment began with Tier 1 or Tier 2?	N	0	Continues to #16
16 TIER 3 STAGGERED TREATMENT: Was the treatment introduced to Tier 3 AFTER it was introduced to Tier 2?	Y	0	If #16 is yes, continues to #17 If #16 is no, skips to #21
17 TIER 3 TREATMENT: Are there at least 3 data points in treatment for Tier 3 and can you predict the future pattern of the behavior?	Y	0	If #17 is yes, continues to #18 If #17 is no, skips to #21
18 TIER 3 PHASE CONTRAST: Project the trend of Tier 3 baseline into the Tier 3 treatment phase. Is the level, trend, or variability of the data in the Tier 3 treatment phase different than what you would predict based on the baseline data for Tier 3?	Y	1	If #18 is yes, continues to #19 If #18 is no, skips to #21
19 TIER 3 IMMEDIACY: Is there an immediate change from the last 3–5 data points in baseline to the first 3–5 data points in treatment for Tier 3?	Y	0.22	Continues to #20
20 TIER 3 OVERLAP: Do less than 30% of the treatment data points overlap with the baseline datapoints for Tier 3?	Y	0.22	Continues to #21
21 CONSISTENCY: Are the data patterns of treatment for Tier 1 and Tier 3 similar in level, trend, or variability?	N	0	Continues to #22
22 CONSISTENCY: Are the data patterns of treatment for Tier 2 and Tier 3 similar in level, trend, or variability?	Y	0.22	End of Protocol
	SCORE	2.66	

Fig. 4 Completed protocol for sample multiple-baseline design graph

To evaluate interrater agreement, we calculated the ICC (Shrout & Fleiss, 1979) on the scores produced by the rating scale and the protocols (i.e., 0–5). The ICC is an index of agreement across multiple judges making multiple decisions that takes into account the magnitude of difference between judges’ decisions, unlike other agreement indices that are calculated based on exact agreement (Hallgren, 2012). Suggested interpretation guidelines for ICCs are as follows: Values below .40 are considered poor, values between .41 and .59 are considered fair, values between .60 and .74 are considered good, and values at .75 and above are considered excellent (Cicchetti, 1994). We calculated the ICC for each group at each time point, which enabled us to evaluate (a) if the use of the protocols improved agreement compared to the

use of the rating scale and (b) if we could attribute improvements in agreement to the protocols rather than to the evaluation of the same graphs a second time. We collected social validity data from the participants regarding the utility of each method for understanding the data and the extent to which each reflected how the analyst would typically analyze SCR data. We also asked the protocol group which method (i.e., rating scale or protocol) they would be more likely to use to conduct visual analysis and to teach others to conduct visual analysis.

Figure 5 shows the pretest and posttest ICCs for each group. Both groups had similar interrater agreement at pretest when using the rating scale (rating scale group ICC = .60; protocol group ICC = .58). However, the agreement of the protocol group improved at posttest

Table 2 Visual analysis rating scale

Score	Anchor
0	No basic effects; does NOT demonstrate a functional relation
1	One basic effect; does NOT demonstrate a functional relation
2	Two basic effects; does NOT demonstrate a functional relation
3	Three basic effects; DOES demonstrate a functional relation with small behavioral change
4	Three basic effects; DOES demonstrate a functional relation with medium behavioral change
5	Three basic effects; DOES demonstrate a functional relation with large behavioral change

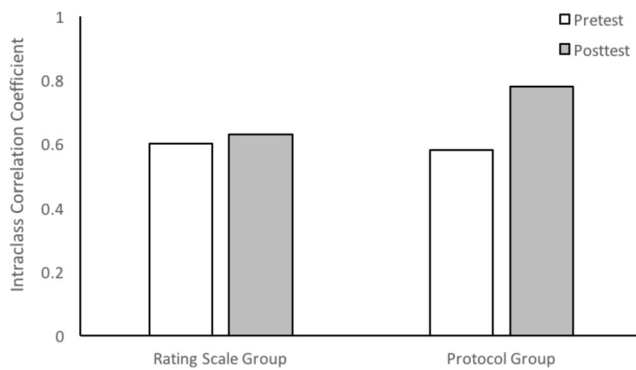


Fig. 5 Intraclass correlation coefficients for the rating scale group ($n = 8$) and the protocol group ($n = 8$) at pretest and posttest

($ICC = .78$), whereas the agreement of the rating scale group remained relatively stable ($ICC = .63$). Based on the proposed guidelines for interpreting ICCs (Cicchetti, 1994), the agreement of the protocol group improved from fair at pretest when using the rating scale to excellent at posttest when using the protocol.

We also examined percentage agreement across protocol questions, displayed in Table 3, to identify the types of questions that produced the most disagreement among participants. Participants disagreed most often about questions pertaining to phase stability, followed by questions about the presence of basic effects. Questions about immediacy, overlap, consistency, and staggered treatment introduction (multiple-baseline designs) produced the highest agreement. Most participants in the protocol group rated the protocol as easy or very easy to understand ($n = 6$), whereas half as many participants in the rating scale group reported the same about the rating scales ($n = 3$). Similarly, most participants who used the protocol rated it as either mostly or very reflective of how they would typically conduct visual analysis, whereas one participant in the rating scale group reported the same about the rating scale. Finally, almost all

Table 3 Percentage agreement on protocols by question type across graphs

Question type	A-B-A-B			Multiple baseline		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Number of data points (stability)	4	62%	0.21	6	67%	0.20
Basic effect	3	71%	0	3	71%	0.28
Immediacy of effect	3	73%	0.14	3	73%	0.28
Overlap	3	74%	0.23	3	73%	0.28
Consistency	3	80%	0.22	3	91%	0.22
Vertical analysis				2	77%	0.25
Staggered introduction				2	100%	0

n refers to the number of questions per graph

participants in the protocol group reported that they would choose the protocol over the rating scale to conduct visual analysis ($n = 6$) and to teach others to conduct visual analysis ($n = 7$).

Validity Evaluation

We also evaluated the validity of the protocols by comparing decisions produced by it to decisions made by expert visual analysts. We recruited eight researchers with expertise in SCR, which we defined as having a doctorate and being an author on at least five SCR publications (Wolfe et al., 2016), to participate. All experts identified their current position as faculty member or researcher and reported that they were an author on an average of 21 SCR publications (range = 5–65; median = 10).

Using the graphs from the reliability evaluation, we asked the experts (a) to make a dichotomous judgment about whether there was a functional relation and (b) to use the rating scale in Table 2 for each graph. Experts accessed the materials from a link sent via e-mail, and we allowed 10 days for experts to participate in the validity evaluation. We told the experts that we were evaluating the validity of systematic protocols for visual analysis, but they did not have knowledge of or access to the protocols.

To evaluate the validity of the protocols, we calculated the percentage of experts who said there was a functional relation and the percentage of participants whose protocol score converted to a functional relation (i.e., ≥ 3) for each graph. Although we asked the experts to answer “yes” or “no” about the presence of a functional relation and then use the rating scale for each graph, the experts’ dichotomous decisions always aligned with their score on the rating scale. There was some disagreement among the experts on their ratings and dichotomous decisions, so we calculated the mean score of the experts using the rating scale and compared it to the mean score of the participants using the protocols.

The ICC for the experts using the rating scale was .73, which is considered good according to interpretive guidelines for the statistic. Table 4 displays the percentage of experts who said there was a functional relation for each graph and the percentage of participants whose protocol score indicated a functional relation for each graph, as well as the mean scores for each graph for each group. These results indicate similar levels of agreement among experts using the rating scale and among participants using the protocol.

Figure 6 shows the mean scores for each graph for both groups of raters. Graphs 1–4 were multiple-baseline designs, and Graphs 5–8 were A-B-A-B designs. Across all graphs, the correlation between the mean scores produced by the experts using the rating scale and by the participants using the protocol was strong ($r = 0.83$). The mean difference between the expert rating scale score and the

Table 4 Percentage agreement and mean ratings for experts and protocol group

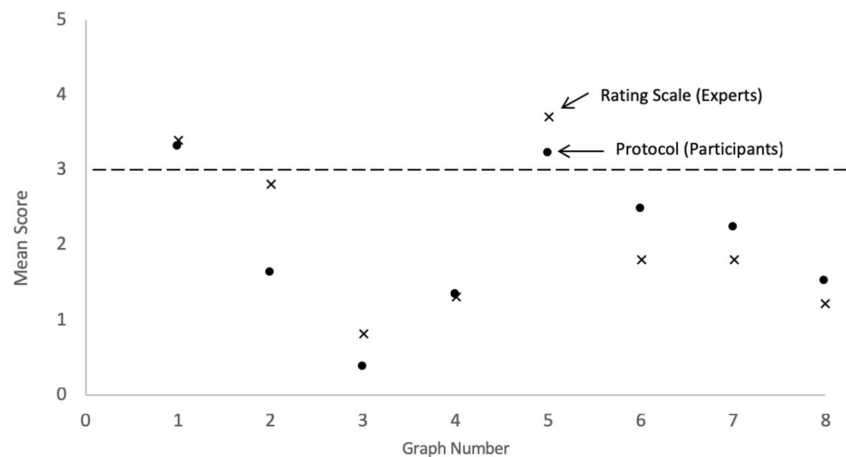
Graph	Percentage indicating functional relation		Mean rating	
	Experts	Protocol group	Experts	Protocol group
1	75	67	3.4	3.3
2	63	22	2.8	1.6
3	0	0	0.8	0.4
4	0	0	1.3	1.3
5	50	44	3.7	3.2
6	13	22	1.8	2.5
7	38	22	1.8	2.2
8	0	0	1.2	1.5

participant protocol score was 0.5, with a range of 0–1.2. For most of the graphs (63%), the difference between the scores was less than 0.5. Although the average difference score was 0.5 for both multiple-baseline designs and A-B-A-B designs, there was a larger range of difference scores for the multiple-baseline designs (0–1.2) than for the A-B-A-B designs (0.3–0.7). We dichotomized the mean scores for each group for each graph to obtain one “decision” for each group with respect to the presence or absence of a functional relation for the graph. The mean decision produced by the experts using the rating scale agreed with the mean decision produced by the participants using the protocol for all eight graphs. As shown in Fig. 6, the mean participant protocol score tended to be below the mean expert rating scale score for multiple-baseline designs, but the reverse was true for A-B-A-B designs. The lower score for the use of the protocol for multiple-baseline designs may be due to the question on vertical analysis, which subtracts a point if the participant indicated that the data in a tier that was still in baseline changed when the intervention was introduced to a previous tier.

Further Development and Evaluation of the Protocols

Visual analysis of SCR data is the primary evaluative method to identify functional relations between experimental variables (Horner et al., 2005; Kazdin, 2011). However, visual analysis procedures are not standardized, subjective judgments about behavior change and magnitude of effects can be idiosyncratic, and interpretations often result in low agreement across analysts, all of which has led to criticism of the method (Kazdin, 2011; Lieberman, Yoder, Reichow, & Wolery, 2010). We developed our protocols to address these issues and provide standardized and systematic procedures to guide visual analysts through the comprehensive processes involved in making judgments about two common SCR designs: A-B-A-B and multiple baseline. Our initial evaluation of the protocols indicates that they improved reliability among visual analysts from fair to excellent, and the correspondence with expert visual analysis provides evidence of criterion validity. In addition, participants reported that they found the protocols easy to understand and navigate, supporting the social validity

Fig. 6 Mean scores for each graph on the rating scale (expert visual analysis) and on the protocol (participant visual analysis). The dotted line indicates the criterion for demonstrating a functional relation



of the tools. These preliminary results are promising and highlight several areas for future research.

First, we plan to continue to examine the protocols' reliability in a number of ways. Our results support the use of transparent and consistent visual analysis procedures for improving reliability. However, we did include a small sample of participants, which impacts the interpretation of our results. Specifically, the limited number of participants in each group may influence the accuracy of the ICCs, and we were unable to statistically compare the ICCs between the two groups to identify whether the differences were likely due to chance. Evaluating the protocols across a larger pool of raters will increase the precision of our reliability estimates and provide important information about the utility of the protocols.

In addition, we only included eight graphs in this investigation, and only two of these received mean scores at or above 3, which is the cutoff for demonstrating a functional relation using either method. Although we did not purposefully select graphs that did not depict a functional relation, we did attempt to include graphs with a range of difficulty and may have eliminated graphs with large, obvious effects as a result. Thus, this evaluation provides more compelling evidence of the reliability and validity of the tool for graphs that do not demonstrate a functional relation than for those that do. Additional investigations of the protocols with graphs that demonstrate functional relations are warranted. The application of the protocols to a larger sample of graphs will allow us to (a) examine the validity of the scoring procedures for additional and varied data patterns and (b) evaluate the appropriateness of individual item weights and the proposed interpretation guidelines for the overall experimental control score. The scores produced by the protocols could also be compared to other analytical approaches, such as statistics, to expand on the evaluation of the protocols' validity.

In future investigations, we plan to compare the protocols to other methods of visual analysis with similar sensitivity. In the current study, we compared the protocols, which can produce scores with decimals (i.e., 2.5), to a rating scale, which could only produce integer-level scores (i.e., 2). It is possible that this differential sensitivity may have impacted our reliability estimates. There is some evidence that correlation coefficients increase but percentage agreement decreases when comparing reliability of a more sensitive rubric to a less sensitive version of the same rubric (Penny, Johnson, & Gordon, 2000a, 2000b). However, because these studies compared different versions of the same measure, it is not clear that their findings apply to the current results given the distinct structures of the protocols and the rating scale. Nonetheless, we could mitigate this factor in future studies by allowing raters using the rating scale to select a score on a continuum from 0 to 5 (i.e., including decimals).

Second, we developed the protocols to be comprehensive, transparent, and ubiquitous. We intend for visual analysts at

any level of training to be able to use the protocols to make reliable and sound decisions about data patterns and functional relations. Thus, we plan to continue to test agreement across different groups, including single-case researchers with expertise in visual analysis, practitioners, and students in SCR coursework who are learning to conduct visual analysis.

Third, the usability of the protocols is critical. The results of the social validity survey suggest that participants found the protocols to be user-friendly; however, all participants in the evaluation had already completed a course on SCR. Although even expert visual analysts are continually improving their visual analysis skills, we designed the protocols to support novice visual analysts who are acquiring their visual analysis knowledge and skills. Future research should involve testing the use of the protocols as an instructional tool for individuals who are learning how to visually analyze SCR data.

Fourth, we plan to continue the iterative development of the protocols. This pilot investigation identified questions that were likely to produce discrepant responses among users; future versions of the protocols could address this by providing more explicit instructions for how to examine the data to answer those questions. Additional examples embedded in the instructions for these questions could also improve agreement. We plan to update the protocols as additional information is published on the process of visual analysis and on the variables that influence agreement among visual analysts. For example, Barton et al. (2018) recommend that visual analysts examine the scaling of the y -axis to determine whether it is appropriate for the dependent variable and, in multiple-baseline designs, whether it is consistent across tiers. This initial step of the visual analysis process could be included in the next version of the protocol to ensure that it remains up-to-date with current recommended practices.

In conclusion, there is a clear need for standardized visual analysis procedures that improve consistency and agreement across visual analysts with a range of professional roles (e.g., researchers, practitioners). We developed and evaluated protocols for two common SCR designs and plan to use an iterative process to continue to test and refine our protocols to improve their reliability, validity, and usability. Improved consistency of visual analysis also might improve SCR syntheses, which is important for ensuring aggregate findings from SCR can be used to identify evidence-based practices.

Compliance with Ethical Standards

Conflict of Interest Katie Wolfe declares that she has no conflict of interest. Erin E. Barton declares that she has no conflict of interest. Hedda Meadan declares that she has no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

The University of South Carolina Institutional Review Board approved the procedures in this study.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Barton, E. E., Ledford, J. R., Lane, J. D., Decker, J., Germansky, S. E., Hemmeter, M. L., & Kaiser, A. (2016). The iterative use of single case research designs to advance the science of EI/ECSE. *Topics in Early Childhood Special Education, 36*(1), 4–14. <https://doi.org/10.1177/0271121416630011>.
- Barton, E. E., Lloyd, B. P., Spriggs, A. D., & Gast, D. L. (2018). Visual analysis of graphic data. In J. R. Ledford & D. L. Gast (Eds.), *Single-case research methodology: Applications in special education and behavioral sciences* (pp. 179–213). New York, NY: Routledge.
- Barton, E. E., Meadan, H., & Fettig, A. (2019). Comparison of visual analysis, non-overlap methods, and effect sizes in the evaluation of parent implemented functional assessment based interventions. *Research in Developmental Disabilities, 85*, 31–41. <https://doi.org/10.1016/j.ridd.2018.11.001>.
- 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. <https://doi.org/10.1177/0145445503261167>.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>.
- Cooper, C. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis*. St. Louis: Pearson Education.
- DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*(4), 573–579. <https://doi.org/10.1901/jaba.1979.12-573>.
- Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21*(1), 111–123. <https://doi.org/10.1007/BF03392786>.
- 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*(3), 415–421. <https://doi.org/10.1901/jaba.1982.15-415>.
- 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*(2), 313–326. <https://doi.org/10.1901/jaba.1997.30-313>.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorial in Quantitative Methods for Psychology, 8*(1), 23–34.
- Hitchcock, J. H., Horner, R. H., Kratochwill, T. R., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. M. (2014). The what works Clearinghouse single-case design pilot standards: Who will guard the guards? *Remedial and Special Education, 35*(3), 145–152. <https://doi.org/10.1177/0741932513518979>.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practices in special education. *Exceptional Children, 71*, 165–179. <https://doi.org/10.1177/001440290507100203>.
- Horner, R. H., & Spaulding, S. A. (2010). Single-subject designs. In N. E. Salkind (Ed.), *The encyclopedia of research design* (Vol. 3, pp. 1386–1394). Thousand Oaks: Sage Publications.
- Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children, 35*(2), 269–290. <https://doi.org/10.1353/etc.2012.0011>.
- 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*(1), 35–45. <https://doi.org/10.1901/jaba.2010.43-35>.
- Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.
- 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. <https://doi.org/10.1177/0741932512452794>.
- Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences*. New York: Routledge.
- 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*(1), 28–44. <https://doi.org/10.1037/a0018600>.
- 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*(1), 44–58. <https://doi.org/10.1177/0741932511435176>.
- Penny, J., Johnson, R. L., & Gordon, B. (2000a). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7*(2), 143–164. [https://doi.org/10.1016/S1075-2935\(00\)00012-X](https://doi.org/10.1016/S1075-2935(00)00012-X).
- Penny, J., Johnson, R. L., & Gordon, B. (2000b). Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education, 68*(3), 269–287. <https://doi.org/10.1080/00220970009600096>.
- Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*(3), 221–242. <https://doi.org/10.1177/01454455980223001>.
- Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science, 23*(2), 139–146. <https://doi.org/10.1177/0963721414524773>.
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971–980. <https://doi.org/10.3758/s13428-011-0111-y>.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>.
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510–550. <https://doi.org/10.1037/a0029312>.
- What Works Clearinghouse. (2017). *Procedures and standards handbook* (Version 4.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf. Accessed 9 Jan 2018.
- Wolfe, K., Seaman, M. A., & Drasgow, E. (2016). Interrater agreement on the visual analysis of individual tiers and functional relations in multiple baseline designs. *Behavior Modification, 40*(6), 852–873. <https://doi.org/10.1177/0145445516644699>.