Visual Display of Graphic Data in Single Case Design Studies:
Systematic Review and Expert Preference Analysis

Jennifer R. Ledford
Erin E. Barton
Katherine E. Severini
Kathleen N. Zimmerman
Elizabeth A. Pokorski

Final version of this paper has been accepted to *Education and Training in Autism and Developmental Disabilities* (2019). This manuscript (10/10/2018) is not the final version of record and may differ from the final, authoritative version.

Abstract

Two studies regarding the graphic display of single case data are presented. First, measurement and graphical display of data from studies in highly-ranked journals in special education were assessed. Measurement of desirable behaviors rather than undesirable behaviors was common and most studies used sessions and percentages as x and y-values, respectively. Data presentation (graph ratios and number of data points and sessions per cm) was highly variable; average ratios did not align with previously-published recommendations. In the second study, 50 editorial board members of special education journals were surveyed to determine preferences for graphing ratios. Preferences did not align with recommended graphing practices and varied based on the number of sessions depicted on the graph.

Visual Display of Graphic Data in Single Case Design Studies:
Systematic Review and Expert Preference Analysis

Introduction

Single case design (SCD) research plays a critical role in establishing evidence-based practices for individuals with developmental disabilities (Horner et al., 2005; Wong et al., 2015). Given its focus on the individual case and repeated measurement that is sensitive to change, SCD research aligns well with the goals of practice in special education and related fields (Ledford, 2018). Thus, assessing the nature and extent of SCD research available to support a specific practice is critical for improving outcomes for individuals with disabilities.

Visual analysis is the primary means by which SCD data are analyzed (Horner et al., 2005; Lane & Gast, 2013; What Works Clearinghouse [WWC], 2017). Visual analysis allows for formative decision-making (Ledford, Lane, & Severini, 2017) and can be used summatively to answer research questions. Specifically, SCD researchers should systematically assess for level, trend, and variability within and across conditions (WWC, 2017). Researchers should also examine consistency and immediacy of behavior change and degree of data overlap across conditions (Ledford et al., 2017; WWC, 2017). To facilitate accurate visual analysis, researchers should use appropriate visual display of their time series data (Morley & Adams, 1991; Tufte, 2001).

Several studies have shown that visual analysis of SCD data is dependent on a number of factors, including variability and magnitude of behavior change (Matyas & Greenwood, 1990), analyst expertise (Ledford & Wolery, 2013), and immediacy of behavior change (Lieberman, Yoder, Reichow, & Wolery, 2010). Factors that may increase the difficulty of SCD data analysis are the use of disparate graphing formats (Kubina, Kostewicz, Brennan, & King, 2017); number of data paths (Hagopian et al., 1997); direction of behavior change (Bartlett, Rapp, & Henrickson, 2011); and use of interval-based recording systems, which have known disadvantageous properties that could influence visual analysis (Ledford, Ayres, Lane, & Lam, 2015; Pustejovsky, 2016). Similarly, studies have shown that data display can considerably influence decision-making, including visual analysis of SCD data (Dart & Radley, 2017; Lauer & Post, 1989). Graphing guidelines include choosing an appropriate format (i.e., a line graph for time series data; Ledford & Gast, 2018) and using consistent height-to-length ratios (e.g., 2:3, 3:4 or 3:5; Cooper, Heron, & Heward, 2007; Johnston & Pennypacker, 1980). However, consistent guidelines regarding graphing of single case designs are not available (Ledford & Gast, 2018).

Previous studies regarding graph characteristics in journals related to behavioral analysis have concluded that graphs were often published absent crucial information (e.g., axis labels) and with height to length ratios that were not within previously suggested ranges and units of measurement (Kubina et al., 2017). Kubina and colleagues found that only 15% of graphs had ratios falling between their identified “optimal” range of 63 to 75%. They also found inconsistencies within and across studies regarding other graph characteristics (e.g., length of axes, labeling of axes, consistency of axes values). The extent to which inconsistencies reflect author or publisher errors, preferences, or lack of published standards is unclear. Graphing errors and inconsistencies may be less likely in research with individuals with disabilities given recently published SCD standards by agencies with ties to special education (e.g., Council for Exceptional Children, 2014; WWC, 2017), even though these standards do not explicitly address high-quality visual display of data.

Variability in the graphic display of data, in conjunction with the aforementioned factors, may result in inconsistent visual analysis—for example, a single graph presented at two different ratios might result in differential conclusions. Understanding data display and graphing practices in the field may lead to increased consistency of use of graphing conventions and subsequent improvements in the consistency of visual analysis conclusions; this may be especially critical given the increased emphasis on synthesis of data across studies.

The purpose of this manuscript is to characterize graphing practices and preferences for SCD data presentation in special education. The purpose of Study #1 was to identify the state of visual display of data in SCD studies that meet minimum standards for rigor (i.e., three potential demonstrations of effect; WWC, 2017) in journals related to special education. Specifically, research questions were: (1) What design types are widely used in special education research? (2) What dependent variable types are widely measured in special education research? (3) What height x length ratios are used? The purpose of Study #2 was to evaluate the extent to which expert SCD researchers preferred commonly-cited optimal graphing ratios versus commonly-published graphing ratios or other ratios with the assumption that raters prefer graphs that are easiest for them to analyze. The research questions were: (1) What ratio of y-axis to x-axis is reported as most preferred by SCD experts? (2) Do preferred ratios vary based on number of sessions or design type?

STUDY #1

Method

Published SCD studies were located by identifying highly-rated journals in special education, identifying which of those journals regularly published SCD studies, identifying all SCD studies in those journals, and then coding information for all SCD graphs in each study. Graphs were located and coded by six graduate students in special education, all of whom were working towards eligibility to sit for the certification exam for applied behavior analysts (Behavior Analysis Certification Board® [BCBA] exam) and their supervising professor, an expert in single case design with a doctoral degree in special education.

First, we identified the top 15 ranked journals in the “Special Education” category for Google Scholar and Web of Science. At the time of the search, eight journals were included on both lists (*Exceptional Children* [EC], *Focus on Autism and other Developmental Disabilities* [FADD], *Journal of Deaf Studies and Deaf Education*, *Journal of Emotional and Behavioral Disorders*, *Journal of Learning Disabilities* [JLD], *Journal of Positive Behavior Interventions* [JPBI], *Journal of Special Education* [JSE], *Remedial and Special Education*); 7 journals were included in the Google list only (*Education and Training in Autism and Developmental Disabilities* [ETADD], *Education and Treatment of Children* [ETC], *Gifted Child Quarterly*, *International Journal of Inclusive Education*, *Learning Disability Quarterly* [LDQ], *Teaching Exceptional Children*, *Topics in Early Childhood Special Education* [TECSE]); and 7 journals were on the Web of Science list only (*American Journal on Intellectual and Developmental Disabilities* [AJIDD], *Dyslexia*, *Journal of Fluency Disorders*, *Intellectual and Developmental Disabilities*, *Journal of Intellectual Disability Research*, *Research in Autism Spectrum Disorders* [RASD], *Research in Developmental Disabilities* [RDD])—for a total of 22 journals.

Following the initial selection of journals relevant to special education, we identified journals which regularly published SCD studies that met the minimum criterion of including at least three potential demonstrations of effect (cf. Council for Exceptional Children, 2014; Ledford & Gast, 2018; WWC, 2017). For example, a multiple baseline (MB) design with two tiers or A-B-A designs were not included, while 3-tiered MB designs (with concurrent baselines

and three separate intervention start points) and A-B-A-B designs were included. To determine which of the 22 identified journals regularly published SCD studies, we hand-searched all issues of the most recent complete volume for each journal. Any journal that published two or more eligible SCD studies was included. Twelve journals met this criterion (*AJIDD, EC, ETADD, ETC, FADD, JLD, JPBI, JSE, LDQ, RASD, RDD, TECSE*).

After identifying eligible journals, a hand-search of each was conducted to identify each eligible article that included a single case design study (using the same criterion listed above) published in the previous five years. Recent publications were identified because graphing practices may be considerably different over time given technological advances, software availability and capabilities, and changing standards in the field. For each study, we separately coded each graph (if applicable). For example, an article including A-B-A-B designs for two participants in two different graphs was coded in two separate rows.

For each identified graph, we coded a variety of elements. First, coders identified the graph design type, including multiple baseline and multiple probe (MB & MP; across contexts, participants, or behaviors); alternating treatments (ATDs); adapted alternating treatments (AATDs); and withdrawal designs (A-B-A-B). Additional variations were coded as “other”, with a descriptive label (e.g., changing criterion design). Coders also recorded the number of data paths and counted the number of data points; if multiple data paths were present, the data points in the longest data path were counted. Intermittent data collection (e.g., generalization sessions) outside of the primary data path were counted as a separate data path if the number of data points was greater than half of the total number of sessions (e.g., if the primary data path had 20 data points and five generalization sessions were conducted intermittently throughout the study, this was not coded as a data path); otherwise, these data points were disregarded.

The intended direction of behavior change was coded based on whether the intervention targeted *increasing* desirable behaviors (e.g., number of words read) or *decreasing* undesirable behaviors (e.g., percentage of intervals with aggression). If one graph included multiple data paths intended to change in opposite directions, the coder recorded *both*. Data collectors recorded the measurement units of the x- and y-axes (e.g., percent, number, rate, duration) based on the axes labels, along with description of measurement in text (if needed). The maximum values of the x- and y-axes were recorded based on the maximum value denoted on each axis. To measure the length of both axes, the graphs were printed at 100% scale on a letter sized sheet of paper and were measured with a ruler (to the closest tenth of a cm). Multiple tiers or panels of a single graph (e.g., MB across participants) were measured separately and averaged. We calculated additional variables using calculations in the Microsoft Excel spreadsheet:

1. Using axis lengths, we derived a ratio by dividing y-axis values by x-axis values.
2. Using the length of the x-axis and the maximum value of the x-axis, we derived a number of sessions per centimeter.

Agreement for coding was conducted for 20% of issues for each journal (i.e., all issues in a randomly selected year), using the previously identified graphs. Agreement was calculated using point-by-point dichotomous assessment for all codes except for measurement of axis lengths, which was calculated by dividing the smaller number by the larger number. Average agreement was 95% (range across journals = 90-99%). Following initial coding some consistent errors were identified for a single coder, despite high-average levels of agreement. Data from this coder were discarded and the third author re-coded those data. The first author then coded 20% of these articles for the purposes of establishing reliability of the re-coding (95%).

Results

Across journals, 1123 graphs were identified in 267 articles. The number of designs per article ranged from 1 to 54, with a mean of 4 and a median of 2 designs per article. As shown in Table 1, some journals published many single case graphs across 5 years (i.e., *ETADD* = 277, *RASD* = 247, *RDD* = 198). Others published relatively few (i.e., *JLD* = 5, *LDQ* = 12).

Research Designs

Table 2 shows graph frequency by design type. There were many ATD graphs ($n = 346$), MP design graphs ($n = 241$), and MB design graphs ($n = 220$). There were fewer A-B-A-B design graphs ($n = 133$) and AATD graphs ($n = 91$). Other designs (combined in Table 2) include multitreatment, combination, changing criterion, and simultaneous treatments designs. Because some articles included multiple graphs, the number of articles including each design was somewhat different, with 119 articles including MB designs, 112 articles including MP designs, 58 articles including ATDs, 42 articles including A-B-A-B designs, 32 articles including “other” designs, and 22 articles including AATDs. MP designs were somewhat balanced between MP designs across behaviors ($n = 119$) and across participants ($n = 108$), with few MP designs across contexts ($n = 14$). MB designs were mostly across participants ($n = 144$), with fewer designs across behaviors ($n = 64$), contexts ($n = 8$), or both participants and behaviors ($n = 4$). The majority of the MB designs included three tiers ($n = 143$) with 47 including four tiers and 30 including five or more tiers. Similarly, MP designs generally included three tiers ($n = 200$), with fewer including four ($n = 38$) or more ($n = 3$) tiers.

Dependent Variables

Most graphs (79%) included data for desirable behaviors, 17% included data for problematic behaviors, and 4% included both types of data. As shown in Table 3, nearly all graphs depicting undesirable behaviors (e.g., challenging behaviors) were ATDs. As expected given their use primarily with non-reversible behaviors, no MP designs or AATDs included measurement of undesirable behaviors, which are typically reversible in nature.

More than half of graphs included a y-value that represented a percentage ($n = 658$); other common metrics were count ($n = 291$) and rate ($n = 113$). Less commonly used were time ($n = 16$), rating scale or quality score ($n = 15$), cumulative count ($n = 6$), ratio ($n = 2$), and physiological measures ($n = 3$; heart rate, moderate-to-vigorous physical activity). Only 18 graphs had unlabeled y-axes (1.6%). Of the graphs including a dependent variable measured as a percentage, 471 (72%) reported a percentage of opportunities or trials (e.g., percentage correct for academic tasks, percentage of independent responses to social initiations). Most of the remainder (178; 27%) reported a percentage of intervals as an estimate of behavior occurrence. Six graphs reported a percentage derived from a duration measure (percentage of session). Three graphs included multiple data paths with different metrics.

Of the graphs that included use of interval-based systems (IBS; 16%), 132 reported the use of partial interval recording (PIR; 73%), 18 reported the use of momentary time sampling (MTS; 10%), and 21 reported the use of whole interval recording (WIR; 12%). Ten graphs included interval measurement but did not provide the name of the interval system or enough information to determine which was used. Authors used a wide range of interval durations (5-600 s), with an average duration of 22 s and a median duration of 10 s. Graphs using MTS used interval sizes of 10-120 s in duration (*mean* = 41 s; *median* = 5 s). Graphs using WIR included interval sizes from 5-60 s in duration (*mean* = 20 s, *median* = 15 s). Graphs using PIR included intervals 5 s to 10 min in duration (*mean* = 20 s, *median* = 10 s).

Data Presentation

Across research designs, the average number of data paths was 2 (range = 1-7). ATDs had an average of 2.91 and AATDs had an average of 2.28. The remaining designs had averages below 2, with the exception of “other” designs (see Table 2). The number of data points per design (or tier, for MB and MP designs) also varied; ATDs included an average of only 10.5 data points, while MB designs included an average of 31.36. Within design types, the number of data points varied widely—the smallest range was 58 data points (AATDs, minimum = 7, maximum = 65). As expected, some graphs included many more “sessions” than data points. This generally occurred in ATDs when each session represented a datum point on a different path and in MP designs, which have missing data by design; both of these design types had nearly twice the number of sessions as data points.

When viewed at 100% (fit to a typical page size), the average graph was 3.83 cm tall and 9.34 cm long, with a ratio of 0.45. The average ratios across journals were relatively similar (all fall within the 0.39-0.54 range) while the average ratios across graph types were slightly more variable. Specifically, A-B-A-B, MB, and MP designs included graphs with smaller ratios (i.e., had longer length given a similar height) and ATDs and AATDs had larger ratios (i.e., had shorter length given similar height). Although average ratios were similar, some journals had extremely wide-ranging ratios. Several journals included ratios that were different by a multiple of 10 (e.g., 0.1 to 1.0), including *JSE*, *LDQ*, and *RDD*—this list includes journals from the bottom, middle, and top tertile (third) in terms of total number of SCD graphs. Other journals had much smaller ranges, including *TECSE*, *JLD*, and *FADD*—these journals are in the bottom or middle tertile in terms of total number of SCD graphs. Only 103 graphs (9.2%) included ratios that were near previously reported ideal ratios, which we identified as those with ratios between 0.6 and 0.7 (i.e., approximating 3/5 to 2/3). As shown in Table 4, most graphs across design types had ratios between 0.25 to 0.5.

Across design types and journals, the average number of sessions per cm was 4.15, with a fairly restricted average across journals (from 2.45 in *AJIDD* to 5.26 in *RASD*), but with a very large range at the individual graph level (range: 0.4-28.07). The number of sessions per cm was fairly consistent across research designs, with the highest numbers for “other” designs (6.74 sessions per cm) and the lowest for A-B-A-B designs (3.74 sessions per cm).

Discussion

The measurement and presentation of data in SCD studies in special education is critical for understanding the state of the field. Line graphs are an important tool for evaluating progress, documenting behavior change, and identifying functional relations. Information about the extent to which various SCDs, measurement systems, and graphing conventions are used is relevant for understanding the state of SCD research and evidence-based practice.

Alignment with Similar Reviews and Recommendations

Research designs. Several previous studies have evaluated the extent to which different designs are used in behavioral and special education journals (Hammond & Gast, 2010) and in education and psychology research (Shadish & Sullivan, 2011; Smith, 2012). Although coded somewhat differently across studies, previous reviews have found that time-lagged designs are the most commonly used, followed by withdrawal (A-B-A-B) designs. Hammond and Gast reported the use of MP designs separately, and reported MB designs were used approximately four times as often as MP designs. Interestingly, the current review showed that MB and MP designs were used with approximately the same frequency and that ATDs were used most often when measured at the graph level (rather than at the published article level). This suggests that

the increasing trend regarding the use of MP designs reported by Hammond and Gast (through 2007) has continued over time.

Dependent variables. Several previous studies have evaluated the use of different measurement systems in early childhood special education (Lane & Ledford, 2014), specific areas of special education (functional behavior assessment; Lloyd, Weaver, & Staubitz, 2016), and behavior analytic research (Kubina et al., 2017; Mudford, Locke, & Jeffrey, 2011). Compared with data collected by Kubina and colleagues, a larger proportion of graphs included a dependent variable measuring percentages (59% versus 27%). Compared with data collected by Lane and Ledford and Lloyd and colleagues, a greater number of dependent variables in this broader review were measured via percentage of opportunities or trials and less often measured with interval-based recording systems (16% versus 45% in both previous reviews). Special education researchers might be more likely to conduct studies in typical educational environments (as compared to clinical and research contexts); some measures may be more difficult to conduct in these applied settings. The difference may also suggest that much of SCD research in special education is focused on trial-based behaviors (which tend to be non-reversible, although they are not necessarily so; Ledford & Gast, 2018) rather than free-operant behaviors (which tend to be reversible). Frequent measurement of non-reversible behaviors may be important because much work regarding single case standards and synthesis has focused on designs that are most appropriate for reversible behaviors (e.g., A-B-A-B, AATD, & MB designs; Shadish et al., 2014; WWC, 2017). This suggests that more methodological work is needed for designs associated with non-reversible behaviors. For example, specific visual analysis and internal validity guidelines for MP and AATD designs may be needed. Moreover, because measurement systems could impact functional relation conclusions and effect size estimations, understanding differences in measurement across studies is critical for interpretation.

Data presentation. The construction and display of line graphs can vary along critical dimensions, which might impact their usefulness, particularly when synthesizing results across studies. Although graphing conventions have been suggested, with a height to length ratio of 3:5 to 3:4 (i.e., between 0.6 and 0.75; Cooper et al., 2007; Johnston & Pennypacker, 1980), one review showed the percentage of authors approximating this ratio in published graphs in behavior analytic research was only 15% (0.63-0.75 ratio; Kubina et al., 2017). However, Kubina and colleagues did not further report the nature of “errors” from suggested ratios (e.g., range across graphs or journals). The current review suggests the average ratio in published single case research is 0.45, although the range across studies was large. Ratio discrepancies cannot be solely attributed to large numbers of data points given the average number of data points was only 22. Interestingly, many fewer studies in this review (2%) failed to label axes compared to Kubina and colleagues’ review of SCD studies in behavior analytic journals.

Implications and Future Research Needs

Outcomes of this review hold several implications for the field. First, measurement of non-reversible behaviors in the context of appropriate designs (i.e., MP, AATD) occurred frequently and as a greater proportion of total SCD studies when compared with other reviews. This is reasonable given that many behaviors expected of students with disabilities are non-reversible in nature (e.g., academic and self-help skills) and may reflect a difference between special education and behavior analytic research. Similarly, most studies measured dependent variables that were intended to increase, suggesting special education research largely focuses on improvement of desirable behaviors rather than reduction of problematic behaviors. In addition, data presentation is highly variable and published ratios (mean of 0.45) do not correspond with

previously-suggested standards (0.6 to 0.75). Updated standards may be needed to allow for more accurate cross-study comparison and ease of data interpretation.

Limitations

There are several limitations worth noting. First, we did not code several graphing features, including marker sizes, marker shapes, line weights, tick marks, or graph labels. These features might be important to consider when examining graph characteristics and when developing graphing standards. Second, we intentionally focused on special education journals but might have missed relevant studies. For example, we did not include the *Journal of Applied Behavior Analysis*, *Journal of Behavioral Education*, or *Remedial and Special Education*, which publish a large volume of SCD studies, some of which include individuals with disabilities. Nonetheless, this study provides important data regarding graphing practices in SCD studies.

STUDY #2

Given graphs in special education journals did not follow previously-suggested standards regarding graphing ratios, the purpose of Study #2 was to determine whether experts in SCD preferred ratios consistent with standards (e.g., 0.65, 0.75), publication practices (0.45), or neither. Procedures were in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki), and were approved via the University's Institutional Review Board. Informed consent was obtained from all participants.

Method

Participants

SCD experts were recruited who served on the editorial board of a journal publishing SCD studies and who had either (a) published 10 SCD studies over the course of their career or (b) published an SCD study in the previous 3 years. First, a graduate student in special education retrieved the names of individuals serving on editorial boards of the 12 journals included in Study #1 ($N = 684$). Then, the student retrieved publicly-available email addresses for editorial board members. Next, a different graduate student in special education performed a Google Scholar search for each identified email address ($n = 630$; 54 email addresses not identified) along with the terms “single case” or “single subject.” Individuals with at least three articles identified via this search were recruited ($n = 204$); in addition, the first two authors hand-searched the list of editorial board members and added 30 participants as potential SCD experts who were not identified via the Google Scholar search. In general, these individuals were not identified because previous SCD research had been conducted and associated with a different email address (e.g., at an institution where the individual was previously employed). A total of 234 recruitment emails were sent; of those, 12 self-identified as ineligible or unavailable. A total of 62 individuals responded and began the survey (28% of 222 potentially eligible recruits). Of these, 50 participants completed the survey in its entirety (81%); their responses were analyzed.

Procedures

Graph generation. Graphs identified during the course of Study #1 were divided into five design types (ATD, AATD, MB, MP, A-B-A-B; no “other” design types were included) and into quartiles based on the number of sessions on the x-axis. One graph was randomly selected from each group (20 total groups, 5 design types $\times$ 4 quartiles) using the *random* function in Microsoft Excel. Then, data from these 20 graphs were extracted using PlotDigitizer (2015), a free program designed for extracting the values of graphed data. Following data extraction, new graphs were generated with Microsoft Excel. All graphs were generated to be 5 cm tall (e.g., y-axis length); to generate graphs with varying ratios of length to height, x-axis lengths were varied

from 20 cm (ratio of 0.25) to 6.67 cm (ratio of 0.75). For each original graph, six variations were generated, each with a different y-axis to x-axis ratio (0.25, 0.35, 0.45, 0.55, 0.65, 0.75).

We generated the graphs using a number of rules designed to ensure differences between graphs were relegated to ratios only. For example, given the same ratio, authors might choose to label every session or every other session; this decision might impact analysis of the graph. Rules were: (1) All data provided by the author in the original graph were included in the generated graphs, including all data paths and any extraneous data points (e.g., generalization, maintenance). (2) All numbers and words were presented in Times New Roman, 12 pt font. (3) The number of unit labels on the y-axis was held constant at 5-8 (e.g., for a range of 0-100%, the values 0, 20, 40, 60, 80, and 100 were used). (4) Data paths were connected with 1.0 pt black lines, and lines were not continued across condition change lines. (5) Filled black circles were used for all data paths; any second data path was generated with unfilled triangles; any third data path was generated with gray-filled squares. (6) Axis tick marks were included for every value for the x-axis (e.g., 10 tick marks = 10 sessions), but were included only for labeled values for the y-axis (e.g., 5-8 tick marks per axis). (7) The number of value labels on the x-axis varied by length, with 1.0-1.6 labels per cm, to ensure similar readability and spacing of labels across ratio presentations. This rule was modified for studies with 33-39 sessions for the 0.25 ratio, because equal spacing and labels every 1.0-1.6 cm was not possible; in this case, a slightly lower rate was used (e.g., labels every 1.75 cm). Similarly, if there were not a sufficient number of sessions for maintaining a label at least every 1.6 cm, every session was labeled (e.g., on a 20 cm axis with 10 sessions, there were only 10 labels, one label every 2 cm).

Survey generation. Following graph development, graphs were saved as PDFs, preserving the original ratios and ensuring consistent quality and presentation across graphs. Each graph set was assigned a letter according to a random function that determined order presentation (e.g., Graph A was presented first). Then, graphs were uploaded using a common web-based survey program. For each graph, five forced-choice two-option items were generated (Graph A with ratio 0.25 versus 0.35, 0.35 versus 0.45, 0.45 versus 0.55, 0.55 versus 0.65, 0.65 versus 0.75). Graph pairs were presented with the question “Which graph presentation do you prefer?”. Pairs for each graph were presented in clusters (e.g., Graph A was presented in items 1-5). The first pair to be presented was determined via random selection, and subsequent questions were response-based. For example, if the first item was randomly determined to be 0.45 versus 0.55, and the participant chose 0.55 as their preferred ratio, the next presented pair would be 0.55 versus 0.65. If the participant chose 0.55 again, this was determined to be the most preferred ratio. If the participant chose 0.65, the next item presented would be 0.65 versus 0.75, with the answer to that item corresponding with the participants most preferred ratio. Errors were made in pair presentation or branching logic for two graphs; thus the final number of graph sets to be analyzed was 18. The minimum number of pairs presented per graph was 1 or 2 (based on whether the randomly determined start point was at the extreme of the scale), and the maximum number of pairs presented was 3-5 (again based on the location of the randomly determined start point). The total possible number of graph comparisons presented for all graphs was 32 to 87.

Recruitment. The first author sent a recruitment email to identified editorial board members and a reminder email 10 days later. The email described the purpose of the study and the inclusion criteria, and included a survey link. The survey was available for approximately four weeks after the initial email was sent. Respondents ranged from 28-71 years of age (mean: 48 years). They reported being employed as Assistant Professors ($n = 13$), Associate Professors ($n = 13$), or Professors ($n = 23$), with one participant reporting employment as a practitioner.

Most ($n = 43$) reported Caucasian (Non-Hispanic) ethnicity. Four reported Asian, two reported multiple, and one reported “other” race/ethnicities. Fewer than half ($n = 21$) reported behavior analysis certifications (BCBA or BCBA-D). Forty-nine reported publishing a single case study in the past three years, and one reported “no” to that question but “yes” regarding whether they had published 10 total studies. Thirty-one participants reported female gender, 17 reported male gender, and 2 did not report gender.

Analysis. Data were downloaded directly from the online program using Microsoft Excel. Descriptive data were calculated and analyzed regarding participant demographics and preferred ratio choices.

Results

Survey Completion

The median time spent on the survey (as reported by the website) was 8 min, 55 seconds. The fastest completion occurred in 5 min; two participants completed in between 45 and 60 min; one participant completed in 100 min; all remaining participants completed in 5 to 33 minutes.

Preferred Ratios

Across graphs and participants, the average preferred ratio was 0.33. Nearly two-thirds (64%) of total responses indicated a preferred ratio at the minimum value (0.25; see Table 5). Participants rarely chose 0.75 (3%), 0.65 (5%), or 0.55 (5%) as the most preferred ratio and sometimes chose 0.45 (10%) or 0.35 (13%). Individually, respondents’ average preferred ratio ranged from 0.28 to 0.50, with 58% of respondents having a preferred average between 0.25 and 0.35, 32% having an average between 0.35 and 0.45, and 10% having an average between 0.45 and 0.55. Each person, with one exception (98%), chose 0.25 as their most preferred ratio for at least one graph and 36% of respondents chose 0.75 as the most preferred ratio for at least one graph.

Variability in Preferred Ratios

Average preferred ratios varied based on starting pair; that is, respondents tended to choose smaller ratios when presented with those first. However, even when the largest ratio pairs were initially presented (e.g., 0.55 versus 0.65 and 0.65 versus 0.75), the average preferred ratio remained low (0.40 and 0.36, respectively). Likewise, ratios varied by design type, with lower preferred ratios for MB and MP designs (0.30 for both), followed by A-B-A-B designs (0.32), AATDs (0.36), and ATDs (0.42). As shown in Figure 1, much of the variability can be explained by the number of data points; a large portion of respondents consistently chose the minimum value (0.25) as preferred when the graph included 20-40 sessions. When the number of sessions was fewer than 20, the proportion of respondents choosing the minimum value was lower and the average preferred ratio was higher. Three outliers appear for graphs with the largest number of data points; these graphs were all randomly assigned to an initial pair of 0.65 versus 0.75, which may partially explain the discrepancy.

Discussion

Notable findings from Study #2 are that (1) self-reported expert preference is even more divergent from previous guidelines than publication practices; (2) regardless of design, experts prefer low ratios (i.e., longer lengths [x-axes] with shorter heights [y-axes]); and (3) experts generally preferred lower ratios (longer lengths) when there was a larger number of sessions. The three outliers, all of which had many data points, also happened to have been graphs with a randomly selected initial ratio comparison of 0.65 versus 0.75. Thus, additional research is needed to confirm the finding that preferred graphing ratios vary according to the number of sessions. In future research, initial graph presentation should be randomly selected separately for

each participant, rather than overall random selection to allow for more fine-grained analysis of this potential relation.

Limitations

The participate rate was somewhat low; the recruitment email included information regarding inclusion criteria (i.e., SCD expertise), so it may be likely that a number of the originally-recruited individuals were not eligible and did not participate due to ineligibility, artificially deflating participation rates. In addition, inaccurate pairs or branching logic was used for two graphs; thus data from those graphs were discarded. Finally, because many responses indicated a preference for the lowest ratio (0.25), it is likely that even lower ratios may be preferred by some individuals. We chose not to use lower ratios given that 0.25 was considerably lower than published recommendations, but additional research is needed including a wider range of ratios. Additional research is also needed regarding consistency in graphing details (e.g., marker size, line weight, and use of arrows to label data series); these may be important for comparing graphs across studies and authors.

Proposed Guidelines

Given results from these studies, other scientific data, and historical standards, we propose several guidelines for the field.

1. Given visual analysis is the recommended method of data analysis for SCD research by which functional relations are identified, quality features of the graphical display of data should be incorporated into SCD standards and visual analysis protocols. Consistent, clear, and accurate graphical data displays might improve the consistency with which visual analysis is conducted and reduce Type I errors.
2. In addition to development of evidence-based standards for graphical display, that the use of publication guidelines is needed to improve the consistency of publication practices across journals. All included journals had minimal guidelines for figures; most commonly, journals only had guidelines related to resolution, although *RASD* and *RDD* also included guidelines related to font type as well as uniformity of font and sizing.
3. Although specific ratios have been proposed, it may be important to consider graph size relative to the number of data points. We suggest that ratio sizes between 0.25 and 0.55 may be appropriate for most graphs, with smaller ratios for graphs with a larger number of data points.
4. Sessions are the most commonly used x-axis value. Others have indicated the x-axis should represent calendar time (i.e., calendar days; Kubina et al., 2017) rather than sessions, arguing that the term “sessions” is not time-based. However, sessions are time-based, in at least an ordinal sense (Session 1 occurred before Session 2 in time, and both occurred prior to Session 3). We recommend continued use of sessions at the current time; empirical data are needed that evaluate the use of sessions versus calendar days and whether either adversely impacts visual analysis. X-axis scales should be consistent across graphs within a study and should accurately represent the number of sessions. Authors should take care to accurately describe the unit(s) of time for each session, the time between sessions (e.g., twice per day with two hours in between, once per day, every other day), and the total duration of the study (e.g., study sessions occurred across three calendar months).

Conclusion

SCD research and the subsequent use of visual analysis play a critical role in establishing evidence-based practices in special education (Gast & Ledford, 2018; WWC, 2017). In SCD research the graph is a critical tool used to make both formative and summative decisions.

However, our results point to a divergence between graphing standards, published graphs, and expert preferences, which might impact reliability across visual analysts, decisions about functional relations, and analysis of results across multiple studies. As SCD research continues to advance to meet current research standards, standards related to the graphical display of data also should progress and evolve.

References

- Bartlett, S. M., Rapp, J. T., & Henrickson, M. L. (2011). Detecting false positives in multielement designs: Implications for brief assessment. *Behavior Modification*, 35, 531-552.
- Cooper, J. O., Heron, T. E., & Heward, W. L. (Eds.) (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson.
- Council for Exceptional Children (2014). Standards for evidence-based practices in special education. Arlington, VA: Author. Retrieved from: <https://www.cec.sped.org/Standards/Evidence-Based-Practice-Resources-Original>
- Dart, E. H., & Radley, K. C. (2017). The impact of ordinate scaling on the visual analysis of single-case data. *Journal of School Psychology*, 63, 105-118.
- 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.
- Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983-2007. *Education and Training in Autism and Developmental Disabilities*, 45, 187-202.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71, 165-179.
- Johnston, J. M., & Pennypacker, H. (1980). *Strategies and tactics of human behavior research*. Hillsdale, NJ: Lawrence Erlbaum.
- Kubina, R. M., Kostewicz, D. E., Brennan, K. M., & King, S. A. (2017). A critical review of line graphs in behavior analytic journals. *Educational Psychology Review*, 29, 583-598.
- Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, 24, 445-463.
- Lauer, T. W., & Post, G. V. (1989). Density in scatterplots and the estimation of correlation. *Behavior & Information Technology*, 8, 235-244.
- Ledford, J. R. (2018). No randomization? No problem. Experimental control and random assignment in single case research. *American Journal of Evaluation*, 39, 71-90.
- Ledford, J. R., Ayres, K. A., Lane, J. D., & Lam, M. F. (2015). Accuracy of interval-based measurement systems in single case research. *Journal of Special Education*, 49, 104-117.
- Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology. Applications in special education and behavioral sciences*. New York: Routledge.
- Ledford, J. R., Lane, J. D., & Severini, K. E. (2017). Systematic use of visual analysis for assessing outcomes. *Brain Impairment*, 19, 4-17
- Ledford, J. R., & Wolery, M. (2013). Effects of plotting a second observer's data on A-B-A-B graphs when disagreement is present. *Journal of Behavioral Education*, 22, 312-324.
- Lieberman, R., 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.
- Lloyd, B., Weaver, E., & Staubitz, J. (2016). A review of functional analysis methods conducted in public school classroom settings. *Journal of Behavioral Education*, 25, 324-356.

- Matyas, T. A., & Greenwood, K. M. (1990). 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.
- Morley, S., & Adams, M. (1991). Graphical analysis of single-case time series data. *British Journal of Clinical Psychology*, 30, 97-115.
- Mudford, O. C., Locke, J. M., & Jeffrey, K. (2011). Rates of responding measured by continuous recording in applied behavioral research. *Behavioral Interventions*, 26, 41-49.
- Pustejovsky, J. E. (2016). Procedural sensitivities of effect sizes for SCDs with behavioral outcome measures. Retrieved from: <https://osf.io/pxn24/>
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, 43, 971-980.
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, 17, 510-550.
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
- What Works Clearinghouse. (2017). Standards handbook. Version 4.0. Retrieved from: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf
- Wong, C., Odom, S. L., Hume, K. A., Cox, A., Fettig, A.,... Schultz, T. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders*, 45, 1951-1966.

Table 1.
Graph Characteristics by Journal

	<i>N</i> (1123)	Data paths	Data points	Sessions	Height	Length	Ratio	Sessions per cm
<i>AJIDD</i>	34	2.3 (1-5)	17.6 (2-95)	21.1 (3-94)	3.3 (2.3-4.6)	8.2 (4.8-13.8)	0.45 (0.25-0.81)	2.45 (0.51-8.50)
<i>ETADD</i>	277	1.7 (1-5)	19.2 (3-93)	34.5 (4-190)	4.1 (1.6-12.4)	10.9 (3.7-21.3)	0.39 (0.15-0.99)	3.21 (0.57-18.10)
<i>ETC</i>	120	2.0 (1-7)	21.8 (2-185)	32.5 (4-171)	3.7 (1.1-7.0)	8.2 (3.1-12.4)	0.46 (0.11-0.86)	4.36 (0.49-18.88)
<i>FADD</i>	68	2.5 (1-4)	14.3 (3-49)	24.9 (6-71)	3.9 (2.1-6.2)	8.6 (4.6-13.5)	0.46 (0.21-0.74)	3.30 (0.81-10.87)
<i>JLD</i>	5	1.6 (1-4)	20.4 (17-30)	33.8 (18-90)	3.5 (3.1-4.7)	9.0 (6.6-14.0)	0.42 (0.22-0.48)	3.23 (2.54-5.43)
<i>JPBI</i>	86	1.95 (1-4)	21.5 (4-65)	26.6 (4-65)	3.9 (1.7-7.7)	8.9 (4-15.8)	0.48 (0.20-0.78)	4.68 (0.86-14.37)
<i>JSE</i>	41	1.8 (1-4)	23.7 (8-56)	37.0 (8-94)	4.0 (1.1-6.2)	10.7 (6.7-14.6)	0.42 (0.09-0.93)	3.85 (1.38-13.94)
<i>LDQ</i>	12	1.7 (1-3)	21.3 (13-36)	33.7 (15-53)	4.06 (1.4-9.2)	8.20 (5.3-10.4)	0.54 (0.13-1.74)	4.43 (1.88-11.32)
<i>RASD</i>	247	1.9 (1-7)	24.6 (2-215)	33.9 (7-215)	3.7 (1.3-7.2)	8.4 (3.5-15)	0.49 (0.17-0.86)	5.26 (0.40-28.07)
<i>RDD</i>	198	2.2 (1-7)	26.0 (2-262)	39.7 (4-660)	3.61 (1.5-9.3)	9.17 (1.6-15.3)	0.46 (0.13-2.25)	4.51 (0.76-21.38)
<i>TECSE</i>	35	1.4 (1-2)	27.4 (11-46)	42.1 (12-125)	4.1 (2.3-9.4)	10.76 (4.5-15.2)	0.39 (0.17-0.66)	3.48 (1.43-6.83)
Average across graphs		2.0 (1-7)	22.1 (2-262)	33.8 (3-660)	3.83 (1.1-12.4)	9.34 (1.6-21.3)	0.45 (0.09-2.25)	4.15 (0.4-28.07)

Note: All cells include mean values with minimum and maximum values in parentheses.

Table 2.
Graph Characteristics by Design Type

	<i>N</i> (1123)	Data paths	Data points	Sessions	Height	Length	Ratio	Sessions per cm
ATD	346	2.91 (1-7)	10.50 (2-83)	20.32 (4-80)	3.83 (1.6-9.3)	7.62 (1.6-14.6)	0.54 (0.18-2.25)	4.00 (0.49-40.34)
AATD	91	2.28 (1-3)	19.36 (7-65)	24.86 (11-65)	4.25 (1.5-9.2)	8.84 (4-12.7)	0.54 (0.15-1.74)	5.10 (1.12-22.94)
ABAB	133	1.29 (1-3)	30.32 (10-69)	40.89 (10-660)	3.69 (2-7.4)	10.63 (5.1-15.3)	0.37 (0.16-0.71)	3.74 (0.76-11.55)
MB	220	1.36 (1-4)	31.36 (8-131)	38.30 (8-131)	3.42 (1.1-6.3)	9.96 (2.1-15.8)	0.36 (0.09-.95)	4.70 (0.75-28.07)
MP	241	1.30 (1-4)	23.48 (5-215)	42.28 (8-215)	3.80 (1.2-12.4)	10.50 (3.7-21.3)	0.38 (0.12-0.99)	3.58 (0.40,18.38)
Other	92	2.12 (1-7)	39.14 (6-262)	53.45 (11-260)	4.6 (1.6-9.4)	10.45 (4.6-14.2)	0.46 (0.13-1.00)	6.74 (1.63-21.38)
Average across graphs		2.0 (1-7)	22.1 (2-262)	33.8 (3-660)	3.83 (1.1-12.4)	9.34 (1.6-21.3)	0.45 (0.09-2.25)	4.15 (0.4-28.07)

Note: ATD=alternating treatments design; AATD=adapted alternating treatments design; ABAB=Withdrawal design; MB=multiple baseline design; MP=multiple probe design Other=multitreatment (28), combination (48), changing criterion (13), and simultaneous treatments (3).

Table 3.

Study 1: Dependent Variable by Design Type

	<i>N</i>	Desirable	Undesirable	Both
ATD	346	196	150	0
AATD	91	91	0	0
ABAB	133	106	16	11
MB	220	201	7	12
MP	241	234	0	7
Other	92	61	16	15
Totals	1123	889	189	45

Note: ATD=alternating treatments design; AATD=adapted alternating treatments design; ABAB=Withdrawal design; MB=multiple baseline design; MP=multiple probe design
 Other=multitreatment (28), combination (48), changing criterion (13), and simultaneous treatments (3).

Table 4.

Study 1: Graph ratios by design type

	<i>N</i>	Q1	Q2 (Median)	Q3
ATD	346	0.42	0.49	0.62
AATD	91	0.37	0.52	0.71
ABAB	133	0.24	0.30	0.52
MB	220	0.27	0.35	0.45
MP	241	0.31	0.35	0.43
Other	92	0.33	0.44	0.55
Total	1123			

Note: ATD=alternating treatments design; AATD=adapted alternating treatments design; ABAB=Withdrawal design; MB=multiple baseline design; MP=multiple probe design; Other=multitreatment (28), combination (48), changing criterion (13), and simultaneous treatments (3).

Table 5.

Study 2: Preferred ratio across graphs (ordered by number of data points)

Graph	Design	Initial Pair ¹	Data Paths	Data Points	Percentage of Participants Selecting Ratio as Most Preferred						Average
					0.25	0.35	0.45	0.55	0.65	0.75	
L	ATD	0.55, 0.65	3	6	36	8	12	6	36	2	0.454
R	ATD	0.55, 0.65	3	10	46	16	16	2	16	4	0.388
K	AATD	0.45, 0.55	2	11	42	24	18	4	8	4	0.374
H	AATD	0.55, 0.65	2	13	42	20	6	18	14	0	0.392
O	MB	0.35, 0.45	2	15	80	10	6	4	0	0	0.284
I	ABAB	0.35, 0.45	2	16	60	14	14	6	4	2	0.336
Q	MB	0.35, 0.45	1	18	60	16	20	4	0	0	0.318
G	AATD	0.65, 0.75	2	19	52	18	2	10	6	12	0.386
D	MP	0.55, 0.65	1	21	42	6	48	2	2	0	0.366
A	MP	0.25, 0.35	3	22	88	8	2	2	0	0	0.268
M	MP	0.25, 0.35	2	24	82	12	2	2	2	0	0.280
S	ABAB	0.25, 0.35	2	26	78	18	4	0	0	0	0.276
F	AATD	0.45, 0.55	2	32	70	18	8	4	0	0	0.296
B	MP	0.35, 0.45	1	34	88	8	4	0	0	0	0.266
P	MB	0.45, 0.55	2	36	82	8	6	4	0	0	0.282
N	MB	0.65, 0.75	1	44	68	16	6	6	2	2	0.314
C	ATD	0.65, 0.75	3	51	60	0	8	10	2	20	0.404
T	ABAB	0.65, 0.75	1	54	68	6	6	14	4	2	0.336

Note: 1=Initial pair was randomly selected; remaining pairs were presented based on responses.

Graphs are lettered according to the order in which they were presented. E and J were excluded from analysis due to errors in item administration.

Figure 1. The relation of the number of sessions to respondents selecting minimum value for graph ratio (0.25; top panel). The relation of the number of sessions to average ratio (bottom panel). Filled triangles=ATDS, unfilled triangles=MB designs, filled circles=A-B-A-B designs, unfilled circles=AATDs, filled squares=multiple probe designs