Research Study

Hammill Institute on Disabilities

Remedial and Special Education
2023, Vol. 44(4) 283–293

DOI: 10.1177/07419325221128907
rase.sagepub.com
SAGE

Progress Monitoring Data for Learners With Disabilities: Professional Perceptions and Visual Analysis of Effects

Collin Shepley, PhD, BCBA-D[1], Justin D. Lane, PhD, BCBA-D[1], and Devin Graley, MA, BCBA[1]

Abstract

This study serves as an initial attempt to establish content validity for graphs likely to be included in trainings targeting progress monitoring for professionals serving learners with or at risk for disabilities. We created a survey containing 32 graphic displays of hypothetical learner data. These surveys were administered to a sample of special education teachers, behavior analysts, higher education faculty, and related service providers. Survey respondents rated each graph on its likelihood of being encountered in practice and whether a graph depicted a therapeutic effect. Results indicated that graphs displaying a therapeutic effect were most likely to be encountered and that graphs with variable data in either a baseline or intervention condition were associated with incorrect visual analysis determinations. Implications of our findings are discussed with respect to personnel who develop and provide trainings on analyzing learners' progress monitoring data.

Keywords

progress monitoring, visual analysis, data analysis, professional development

Progress monitoring is frequently conceptualized as a multi-component practice consisting of the (a) continuous assessment of a learner's performance over time, (b) analysis of the assessment data, and (c) data-based decision-making about the effectiveness of the instruction received by the learner (Akers et al., 2014). For children and school-age students with disabilities receiving special education services, legal mandates require that progress monitoring occur across goals and objectives outlined in individualized education plans or individualized family service plans (Etscheidt, 2006). In the recent U.S. Supreme Court decision of *Endrew F. v. Douglas County School District* (2017), legal mandates surrounding progress monitoring were strengthened by the implications of the Court's ruling, which stressed that student progress be monitored in a "systematic manner" to support instructional modifications when data indicate that a student is "not progressing toward his or her goals" (Yell & Bateman, 2017, p. 14). In response to the rise of tiered support frameworks, the significance of progress monitoring has extended to learners who may be at risk for a disability, given that collected data may be used to inform eligibility decisions for special education services as well as determine the effectiveness of the learner's received instruction (Ardoin et al., 2016). Both the legal requirements surrounding progress monitoring and the emphasis on progress monitoring within tiered support systems are supported by more than 30 years of empirical research (Fuchs & Fuchs, 1986; Lee et al., 2020).

Regarding the types of progress monitoring that a professional may employ, this will typically vary based on the needs of a learner. For learners struggling with grade-level academic content, a teacher may use a curriculum-based measure (CBM) to monitor performance across "various skills required for year-end performance" (Fuchs, 2004, p. 188). For learners with needs unrelated to grade-level academic content that generally comprise CBMs, teachers may use a mastery measurement approach and develop "single skill progress monitoring measures" that offer "strong criterion validity" (Fuchs, 2004, p. 191). Regardless of the type of progress monitoring approach used, one of the critical aspects of progress monitoring is the ability of a professional to correctly analyze a learner's data. Following the collection of valid and reliable data, correct analysis of these data permits a professional to make an accurate decision about whether to modify instruction or continue as planned. If analysis reveals little to no meaningful improvements in a leaner's performance on an objective, then the professional may identify an appropriate modification to their instruction

[1]University of Kentucky, Lexington, USA

Corresponding Author:
Collin Shepley, Department of Early Childhood, Special Education, & Counselor Education, University of Kentucky, 229 Taylor Education Building, Lexington, KY 40506, USA.
Email: collinshepley@uky.edu

Associate Editor: Eric Common

to better meet the needs of the learner. If a professional incorrectly analyzes data, then (a) a learner may continue receiving ineffective instruction or (b) the professional may devote resources to developing and implementing instructional changes that are not needed (Grigg et al., 1989; Sandall et al., 2004).

Various methods are available to support a professional's analysis of progress monitoring data. When collecting data using a CBM, professionals often use a continuous treatments design in which a learner's performance is compared to an aim line "which depicts a desired rate of growth" (Ardoin et al., 2013, p. 2). When using a mastery measurement approach, data are commonly compared across adjacent conditions within an A-B design, for which the first condition (e.g., baseline condition) represents the learner's performance prior to the introduction of an intervention or supports, and the second condition (e.g., intervention condition) represents the learner's performance following the introduction of an intervention or supports. Recommendations for using mastery measurement approaches typically stress the use of an A-B design (Wolery, 2004), as it can capture idiosyncratic responding (e.g., changes solely in level rather than trend), account for systematic instruction procedures that may initially suppress responding (e.g., constant time-delay, most-to-lest prompting), and provide an understanding about a learner's initial performance when no other data are available (e.g., target skill is not depicted in a school-wide curriculum-based assessment). Given (a) differences in the progress monitoring approaches previously discussed, (b) the applicability and appropriateness of each approach relative to a learner's needs, and (c) the focus of the study described in this manuscript, all future mentions and discussions of progress monitoring are specific to mastery measurement approaches that utilize A-B designs.

To facilitate correct analysis of progress monitoring data, general recommendations focus on plotting a learner's data within a time-series format across adjacent condition in an A-B graph and conducting a visual analysis (Gischlar et al., 2009). Visual analysis is a well-established and long-standing approach in special education and related professions for evaluating therapeutic improvements with instructional goals and objectives over time (Ninci et al., 2015; Wolery, 2004). For professionals who serve learners with individualized needs, such as those with or at risk for disabilities, trainings on how to visually analyze progress monitoring data are essential. Lane et al. (2019) found that repeated exposures to progress monitoring data alone are insufficient to improve visual analysis abilities with A-B graphs. Given the plethora of service providers who work with learners with and at risk for disabilities, visual analysis trainings targeting A-B graphs have been evaluated across a variety of professionals, including pre-service educators (Lane et al., 2019), in-service special education teachers (Jimenez et al., 2012, 2016), behavioral therapists (Fisher et al., 2003; Kipfmiller et al., 2019), psychology students (Stewart et al., 2007; Wolfe & Slocum, 2015), and attendees of a conference workshop (Fisher et al., 2003). Visual analysis trainings have and continue to receive attention in the literature to further professionals in practice.

As evidence-based practices are identified for training professionals in visually analyzing progress monitoring data within A-B graphs, an important area that needs to be considered pertains to the content validity of the measures used in these training studies. That is, to what extent are the data paths presented in the A-B graphs on which trainees are assessed, similar to the data paths that will be encountered in practice by the trainees? The development and validation of content included in trainings on analyzing data will better ensure trainees can generalize their data analysis skills to their in-service experiences working with individuals with or at risk for disabilities. Many of the previously discussed training studies noted that data patterns displayed in graphs were informed solely by those depicted in research studies (Fisher et al., 2003), whereas others did not indicate a rationale for the selection or construction of data patterns (Stewart et al., 2007). One study noted that data patterns were constructed and included based on the professional judgments of the study's authors with regard to a graph's perceived likelihood of being encountered in practice (Lane et al., 2019). To assess the validity of the authors' judgments, professionals unaffiliated with the study rated the graphs using a Likert-type scale to yield an estimate of the perceived likelihood of encountering each graph in practice. Lane et al. (2019) found that the graphs which were most frequently analyzed incorrectly were the ones rated as least likely to be encountered in practice. This finding has significant implications—if professionals without training in visual analysis are incorrectly analyzing data paths that are only ever *rarely* encountered, then this lack of proper training may have a negligible impact on learner outcomes. Thus, professional development and teacher preparation may be better tailored if focused on areas other than visual analysis.

To our knowledge, the findings from Lane et al. (2019) have not been explored in another study. Furthermore, Lane et al. had only eight professionals provide ratings of a graph's likelihood to be encountered; therefore, attempts to extrapolate their ratings to all populations of professionals who serve learners with or at risk for disabilities may not be appropriate. In considering Lane et al.'s findings alongside (a) the criticality of progress monitoring for learners with or at risk for disabilities regarding compliance with special education laws and the implementation of multi-tiered systems of support and (b) the need to establish valid content for inclusion in trainings on analyzing learners' progress monitoring data, we sought to systematically replicate Lane et al.'s study. In doing this, we hoped to provide content validity data to support the development of future trainings on visual analysis and to better understand the relationship

between a graph's likelihood of being encountered and the ease with which an A-B graph can be correctly analyzed. Our study's research questions are as follows:

Research Question 1 (RQ1): What progress monitoring graphs do professionals serving individuals with or at risk for disabilities report being most likely to encounter?
Research Question 2 (RQ2): What features of progress monitoring graphs predict how professionals report the likelihood of encountering such graphs?
Research Question 3 (RQ3): What progress monitoring graphs are most frequently visually analyzed correctly by professionals?
Research Question 4 (RQ4): What features of progress monitoring graphs are associated with correct visual analysis by professionals?
Research Question 5 (RQ5): Are graphs that are more likely to be encountered by professionals also more likely to be analyzed correctly?

It should be noted that additional research questions related to subgroups of professionals were initially included in the conceptualization of this study; however, due to issues in securing an adequate sample size, answers to these questions were not pursued within the presented study. Supplemental materials, tables, figures, syntax, and output tables referenced throughout the manuscript may be obtained at https://osf.io/dyq7h/.

Method

Sampling

The research team intended for the study's sample of participants to represent four professions that served individuals with disabilities: (a) special education teachers, (b) related service providers, (c) behavioral therapists, and (d) higher education faculty. Researchers selected to use virtual "snowball" sampling to recruit participants. Although this method is generally used to study hidden populations (e.g., drug-users, homeless persons) for which other sampling methods would not be possible, we selected this method given its efficiency and resource sensitivity in relation to the research team's effort and allocated time for the study. To account for community bias effects inherent in the "snowball" sampling design, we selected a set of variables we perceived to correlate with a participant's membership in a particular community or social network. These included (a) the institution at which a participant most recently attended and (b) board certification as a behavior analyst.

Initially, the research team created a list of known individuals who fell into one of the four professions, totaling 108 individuals, with 35 of these individuals having an affiliation with the institution at which all research team members were employed. The listed individuals were divided among the research team based on assumed levels of familiarity. Using a templated document, research team members contacted each of their assigned individuals through email within a 2-week time frame in mid-February 2020 (see supplemental materials). All templated emails provided a brief description of the study, a link to the survey, information about the estimated time commitment required to complete the survey, and a request that the individual also send the survey link to other relevant individuals. Of the 108 individuals initially contacted, 33 were asked to send the survey to teachers, three were asked to send to related service providers, 12 were asked to send to behavior analysts, and 60 were provided with directions that did not specify a particular group to send the survey. Forty-four individuals completed the survey from mid-February to early March in the year 2020, prior to state-of-emergency declarations and state-ordered lockdowns in response to the COVID-19 pandemic. From the remainder of March through mid-June, four additional individuals completed the survey. In mid-June, the research team sent a follow-up email to all individuals who received the initial email about the study. The follow-up email thanked the individuals for their tolerance of research-related emails during the pandemic and kindly requested that the individuals resend the survey link to relevant individuals. Nine additional individuals completed the survey in the second half of June, and no individuals completed the survey throughout the month of July. In August 2020, the research team decided to end recruitment with a total of 57 individuals completing the survey (henceforth referred to as *respondents*). Due to the final number of respondents being significantly smaller than anticipated, research questions related to sub-groups of professionals were dropped and the variables selected to control for community bias effects were not included in analyses.

Respondent Demographics

Supplemental Table S1 provides information on the survey respondents by profession, degree, average years of experience, and institutional affiliations. Due to a low number of respondents working as related service providers, we used the category of *other* to better reflect the professions of related service providers and those that regularly served individuals with disabilities but were not included in our initially developed categories (e.g., school psychologists). Of the 57 respondents, 16 were teachers (18%), 27 were behavioral therapists (47%), 10 were higher education faculty (17%), and four were categorized as other (7%). Seventeen institutions were represented across the 57 respondents (i.e., the higher education institution at which a

respondent most recently attended; see Supplemental Table S2). Slightly less than half of the respondents were board-certified behavior analysts ($n = 27$). The respondents' average years of experience was 10.01 ($SD = 8.55$), with a range of 1 to 42 years.

Instrumentation

The research team developed and provided a survey through the online Qualtrics platform. The survey contained 32 A-B graphs depicting hypothetical progress monitoring data from individuals with or at risk for disabilities. All graphs contained a baseline and intervention condition, with three data points included in the baseline condition and 10 data points included in the intervention condition. The size of each graph was presented at a ratio of 6 (height) by 13 (width). Data were graphed as either a frequency or a percentage with 10 intervals ranging from 0 to 10 (frequency) or 0 to 100 (percentage) along a graph's ordinate. The presentation for each of the 32 graphs was randomized for each respondent using Qualtrics' randomization feature.

Graph variation. The research team varied three variables across the 32 graphs: (a) type of behavior represented in the graph, (b) variability in the graphed data, and (c) presence of a therapeutic effect. Four types of behaviors were reflected across all graphs, with each behavior type included in eight graphs (8 graphs × 4 behavior types = 32 total graphs). The included behavior types were (a) prosocial, (b) academic, (c), daily living, and (d) challenging behavior. Within each set of eight graphs grouped by behavior type, two graphs displayed variable baseline data, two graphs displayed variable intervention data, two graphs displayed variable baseline and intervention data, and two graphs did not display variable data in either condition. Of the 32 graphs, 16 demonstrated a therapeutic effect from the baseline to intervention condition and 16 graphs did not demonstrate a therapeutic effect. Refer to Supplemental Figures S1a-S1d for each graph included in the survey. Supplemental Table S3 overviews the features of each graph included in the final analytical sample (discussed later in the manuscript).

Responding. For each respondent, the survey started by providing general information about (a) the graphs, (b) what each respondent would be asked to do throughout the survey, and (c) the general purpose of the survey. Each graph was then presented in isolation along with two directions. The first direction stated, "Please rate the graph on its likelihood of being encountered when working with individuals with or at risk for disabilities." A drop-down menu was provided in which respondents selected a likelihood rating. Ratings resembled a Likert-type scale with a rating of 1 indicating *highly unlikely to be encountered*, 2 indicating *unlikely to be encountered*, 3 indicating *no more likely than unlikely to be encountered*, 4 indicating *likely to be encountered*, and 5 indicating *highly likely to be encountered*. The second direction stated, "Please indicate, by selecting an option below, if there was a basic demonstration of effect (i.e., a significant change in responding from baseline to intervention)." Respondents were provided with a dichotomous choice of *Yes* or *No* to indicate whether a therapeutic effect was demonstrated within a graph.

At the end of the survey, a series of questions about each respondent's profession, work experience, and educational experience were presented. It should be noted that three additional graphs depicting data collected within a single-case experimental design (e.g., multiple baseline design) were presented throughout the survey. Given the low sample size, research questions related to the experimental design graphs were dropped as they pertained to subgroup differences in visual analysis abilities.

Technical adequacy. The number of data points displayed on the created graphs was informed by a combination of (a) minimum guidelines for conducting visual analysis (Ledford & Gast, 2018), (b) the Lane et al. (2019) study which we were replicating, and (c) contemporary standards for single-case research (What Works Clearinghouse, 2017). The total number of data points on each graph within the Lane et al. study ranged from 9 to 15, with the number included in baseline and intervention conditions differing across graphs. Given this, the research team discussed that were no generally accepted standards to guide practitioners in collecting a specified number of baseline sessions prior to introducing intervention. However, standards developed to guide researchers in collecting experimental single-case data specify that the minimum acceptable number of data points to collect within a condition range from 3 to 5 data points. In addition, three data points is the minimum required to detect a trend with visual analysis procedures. Therefore, we chose to adopt the lower end of the research standard and included three data points in the baseline condition of our graphs, as this was also the minimum number of data points required to permit visual analysis of trend. Given that we chose to adopt three data points in baseline, we selected to include 10 data points in intervention and have 13 data points total in each graph, as this provided an approximate middle-ground to Lane et al.'s range of 9 to 15 data points in each graph.

As previously discussed in the Introduction section of this article, Lane et al. (2019) had professionals unaffiliated with the study rate the likelihood of graphs being encountered in practice. Some of the professionals noted that the inclusion of multiple baseline data points was not something they were trained to do, nor was it something they observed in their field. Given this observation, we sought to better understand how our alignment with a standard developed to guide researchers in collecting an adequate number

of baseline data points (i.e., three baseline data points) was acceptable for practice. Thus, we included a question at the end of our survey in which respondents were asked to indicate the number of baseline data points they typically collected when working with individuals with or at risk for disabilities. Forty-three respondents (75%) indicated that they collected three to five baseline data points, five (9%) indicated they collected more than five baseline data points, two (4%) indicated they collected one to two baseline data points, and two (4%) indicated they collected 0 baseline data points. Five respondents provided a note indicating that the number was often conditioned on some other variable (e.g., "depends on the situation").

Regarding the behavior types selected for our graphs, we chose to use broad conceptualizations (e.g., daily living skills) rather than specific skills (e.g., tying shoes) given concerns from Lane et al. (2019) that participants in their study may have misinterpreted the ordinate label on some graphs due to the study authors varying the specific skill depicted on each graph. A detrimental consequence of misinterpreting an ordinate label may be that a participant would perceive an accelerating trend as being therapeutic, when, in fact, the accelerating trend is contratherapeutic (as is the case when graphing challenging behavior). The size of each graph in our survey was based on the visual preferences of the research team and established through consensus agreement.

The research team did not establish a priori quantitative parameters by which to determine what constituted *variable data* or a *therapeutic effect* within each graph. Rather, a consensus approach was used among members of the research team. First, one member of the research team created each of the 32 graphs based on the previously described features. Then, all members of the research team collaboratively reviewed each graph's alignment with its intended features (e.g., prosocial behavior, variable data in baseline and intervention conditions, no therapeutic effect demonstrated). Research team members suggested changes to each graph until all members achieved consensus that each graph accurately reflected its intended features.

Post hoc analysis. After all graphs were created, we conducted post hoc analyses to assess the extent to which the features of our created graphs aligned with quantitative measures of variability and effectiveness. The variability of the data in each graph was modeled using ordinary least squares (OLS) regression analyses; a separate model was created for the data in the baseline condition and the intervention condition of each graph, with robust standard errors applied to each model. The standard error of standardized coefficient estimates for the linear trend in each model served as the primary measure of interest with regard to variability. Supplemental Figure S2 displays the standard error reported in graphs that were determined to contain variable and non-variable data for the baseline or intervention condition. On average, graphs determined by the research team to contain variable data in the baseline condition yielded a standard error of 0.69, compared with graphs determined to contain non-variable data in the baseline condition as yielding a standard error of 0.05. Regarding data in the intervention conditions, on average, graphs determined by the research team to contain variable data yielded a standard error of 0.30, compared with graphs determined to contain non-variable data yielding a standard error of 0.16.

The presence of a therapeutic effect in each graph was modeled using an interrupted time-series framework within an OLS regression analysis. The outcome in each model was the percentage or frequency of the behavior depicted in a graph (e.g., frequency of challenging behavior). Three covariates were included in each model: *Sessions, Level_Change*, and *Intervention_Trend. Sessions* was a continuous variable that ranged from 1 to 13 and represented the number of each session in a graph. *Level_Change* was a dichotomous variable in which baseline sessions were coded as 0 and intervention sessions were coded as 1. *Intervention_Trend* was a continuous variable that ranged from 0 to 9 and represented each intervention session in chronological order starting with 0; all baseline sessions were coded as 0 for the *Intervention_Trend* variable. Within each model, *Sessions* estimated the linear trend in baseline data, *Level_Change* estimated the absolute level change from a baseline to intervention condition, and *Intervention_Trend* estimated the linear trend in the intervention data. The primary measures of interest with regard to a graph demonstrating a therapeutic effect were the *p* values for the *Level_Change* parameter and *Intervention_Trend* parameter. Supplemental Figure S3 displays a scatterplot of the *p* values from each graph's modeled data. Graphs determined by the research team to indicate a therapeutic effect yielded an average *p* value of 0.34 for the *Level_Change* and *Intervention_Trend* variables; and graphs determined by the research team to not indicate a therapeutic effect yielded an average *p* value of 0.58 for the variables of interest. Models including an estimate of autocorrelation were also created for comparison. Output tables of these results are available as supplemental materials. Estimates from these models do not meaningfully alter the findings.

As a secondary method to determine the extent to which our created graphs accurately demonstrated a therapeutic effect, we evaluated each graph in our analytical sample against Fisher et al.'s (2003) conservative dual criterion method. Graph 9 was the only graph for which the research team's determination about the presence of a therapeutic effect was in disagreement with determinations based on the conservative dual criterion method. Specifically, the research team determined that Graph 9 demonstrated a

therapeutic effect; however, when applying the conservative dual criterion method, sessions 7, 9, and 10 overlapped with the linear regression line of best fit calculated using baseline data point values, and thus the graph did not indicate a therapeutic effect according to the conservative dual criterion method.

Research team. Give that consensus agreement was used among members of the research team to select the final versions of the graphs included in the survey, we provide a description of the team. Three individuals with affiliation to the same institution of higher education comprised the research team: two tenure-track faculty and one doctoral student within the programs of early childhood education, special education, and applied behavior analysis. All members held board certification as behavior analysts, with differing experiences as former practitioners (i.e., preschool special education teacher, home- and school-based applied behavior analysis provider, supervisor within an out-patient serve behavior clinic). Collectively, the research team had over 50 peer-reviewed publications primarily pertaining to single-case methodology and the preparation of practitioners who serve individuals with or at risk for disabilities.

Analysis and Measurement

To answer the first research question, we employed Lawshe's (1975) method for quantifying content validity. We recoded respondents' likelihood ratings from a 1 to 5 scale to a dichotomous determination of *1* or *0*. Respondents' initial ratings of 1, 2, or 3 were recoded to *0*; and initial ratings of 4 or 5 were recoded to *1*. Recoded ratings of *1* indicated that a graph was *likely to be encountered* and recoded ratings of *0* indicated a graph was *not likely to be encountered*. Using the new codes, for each graph we calculated the percentage of respondents providing a rating of *likely to be encountered* (i.e., code of *1*). We then examined descriptive statistics and rank ordered the graphs starting with the graph with the highest percentage of respondents indicating the graph was likely to be encountered. Given that we used an ordinal ranking system to analyze the descriptive data, we did not apply Lawshe's critical values.

To answer the second research, we conducted a series of OLS regression models, with the graphs serving as the unit of analysis and the outcome in each model being the percentage of respondents rating a graph as *likely to be encountered* (*Percentage_Likelihood*). Variables were entered into each model based on a priori assumptions and were influenced by the number of graphs included in the final analytical sample ($n = 27$, discussed further in the "Analytical sample" section). In the first model, we regressed the outcome on the type of behavior displayed by a graph (i.e., academic, daily living, challenging, or prosocial). This variable was entered as a fixed effect (*Behavior_Type*), with prosocial behaviors as the reference group. In the second model, we examined the association between the percentage of respondents rating a graph as *likely to be encountered* and the variability present in a graph. We included three binary variables which indicated whether there were variable data in the (a) baseline condition (*Variable_Baseline*), (b) intervention condition (*Variable_Intervention*), and (c) baseline and intervention conditions (*Variable_Baseline* × *Variable_Intervention*). In the third model, we included only one variable, which was an indicator of whether a graph displayed the presence of a therapeutic effect (*Effect_Present*). In the fourth model, all variables were included. This model was used to serve as a test of robustness for any variables that previously achieved statistical significance ($\alpha = .05$). Robust standard errors were included in all models; subscript *i* indicated each graph; and ε served as the error term. The formula for Model 4 is as follows:

$$(4)\quad Percentage\_Likelihood_i = \beta_0 + \beta_1 Behavior\_Type_i + \beta_2 Variable\_Baseline_i + \beta_3 Variable\_Intervention_i + \beta_4 Variable\_Baseline \times Variable\_Intervention_i + \beta_5 Effect\_Present_i + \varepsilon_i$$

To answer the third research question, we calculated the percentage of individuals providing a correct determination about the presence of an effect for each graph. We then rank ordered these graphs, starting with the highest percentage of respondents providing a correct determination about the presence of a therapeutic effect.

To answer the fourth research question, we conducted another series of OLS regression models with the graphs serving as the unit of analysis. Different from the previous series of models described (i.e., Models 1–4), the outcome variable was the percentage of respondents providing a correct determination about the presence of a therapeutic effect in a graph (*Correct_Analysis*). The same variables used in Models 1 to 4 were included in the new series of models (i.e., Models 5–8). Robust standard errors were included in each model.

To answer the fifth research question, we regressed the percentage of respondents providing a correct determination about the presence of a therapeutic effect in a graph (*Correct_Analysis*) on the percentage of respondents rating a graph as *likely to be encountered* (*Percentage_Likelihood*). Robust standard errors were included in the model (Model 9). All model formulas are included in Supplemental Table S4, along with the research question to which they were aligned.

Given the number of hypothesis tests conducted across all research questions, we used Benjamini–Hochberg-corrected *p* values to determine the extent to which statistically significant variables ($\alpha = .05$) retained their

significance when correcting for the number of significance tests conducted. Variables that retained significance or approached significance with Benjamini–Hochberg-corrected p values are noted within tabled data and in the narrative of the "Results" section.

Analytical Sample

Upon initial analysis of our results, we identified five graphs as outliers with regard to respondents' correct visual analysis determinations as the graphs were 1.5 times the interquartile range. In examining four of these graphs, we observed that the data in the intervention conditions demonstrated contratherapeutic trends. Given the framing of the survey question asking respondents to visually analyze each graph (i.e., "Please indicate, by selecting an option below, if there was a basic demonstration of effect"), we recognized that we did not specify that respondents indicate whether a *therapeutic* effect was observed. Rather, we asked whether a *basic* demonstration of effect was observed, for which graphs reflecting a contratherapeutic effect may have met criteria for a basic demonstration of effect. In visually analyzing these four graphs, the research team observed that a contratherapeutic effect was present in each graph. We attribute these four graphs as outliers, due to the research team's misalignment between the purpose of the study specific to professionals' identification of *therapeutic* effects and the questions asked in the survey pertaining to *basic* demonstrations of effect. The research team determined to exclude these four graphs from analyses (i.e., Graphs 13, 14, 16, and 32), due to the likely confusion arising on the part of the respondents as a result of the misalignment.

Graph 27 was also identified as an outlier. In examining the graph, the research team perceived that respondents may have had concerns with data instability, given that data point values in both the baseline and intervention conditions ranged between the minimum and maximum values reported on the graph (i.e., 0–10). We excluded Graph 27 from our preliminary analyses but included it within sensitivity analyses (discussed in the "Results" section). We did not conduct sensitivity analyses with Graphs 13, 14, 16, and 23, due to their exclusion pertaining to a misalignment with the graphs' construction and the study's intended purpose; whereas, the features and construction of Graph 27 aligned with the purpose of the study despite meeting a generally accepted rule for classifying outliers.

Results

Please refer to the supplemental materials for syntax and related output tables for all analyses discussed in the "Results" section (https://osf.io/dyq7h/).

RQ1: What Progress Monitoring Graphs Are Most Likely to be Encountered by Professionals Serving Individuals With or at Risk Disabilities?

Of the 27 graphs included in the analytical sample, Graph 17 was most frequently rated by respondents as likely to be encountered (86% of respondents, $n = 49$) and Graph 2 was least frequently rated as likely to be encountered (23% of respondents, $n = 15$). The average percentage of respondents rating a graph as likely to be encountered was 61% ($n = 35$). Nineteen of the 27 graphs were each rated by more than 50% of the respondents as likely to be encountered. Supplemental Table S5 provides the rank order of all graphs.

RQ2: What Are the Features of Progress Monitoring Graphs That Are Associated With a Graph's Likelihood of Being Encountered by Professionals?

Results of regression Models 1 to 4 are detailed in Table 1. In Model 1, which examined the relationship between behavior types presented on a graph and their impact on a graph's likelihood of being encountered, average coefficient estimates ranged from 6.77 percentage points (PP) to 13.17 PP, and standard errors ranged from 9.20 PP to 10.72 PP. None of the behavior types achieved or approached statistical significance, suggesting that we do not have evidence that academic, daily living, or challenging behavior have a different likelihood than prosocial behaviors (reference group in the model) as being encountered in our sample of progress monitoring graphs. Relatively large standard errors were also observed in Model 2, which precluded any indicator of variability from achieving or approaching statistical significance. Therefore, we do not have evidence that the presence of variability affects a graph's likelihood of being encountered. Model 3 examined the presence of a therapeutic effect on a graph's likelihood of being encountered, for which the likelihood of a graph demonstrating a therapeutic effect was on average 26.79 PP higher than the likelihood of a graph not demonstrating a therapeutic effect. The standard error of this estimate was relatively small ($SE = 4.52$), and the estimate achieved statistical significance ($p < .001$) and maintained significance when employing the Benjamini–Hochberg correction ($p < .001$). To test the robustness of this finding, all previously tested variables were entered into Model 4. A graph's presence of a therapeutic effect retained statistical significance with the Benjamini–Hochberg correction ($p < .001$) in Model 4, and though the magnitude and precision of the estimate increased ($B = 26.92$, $SE = 5.20$). In addition, the standard error of regression increased from Model 3 to Model 4, suggesting that the inclusion of all variables led to worse model performance.

Table 1. Results of Regression Models Examining Features of Graphs Likely to be Encountered by Respondents.

	Model			
Variable	1	2	3	4
Behavior type				
Academic	13.17 (9.79)	—	—	Yes
Daily living	6.77 (10.72)	—	—	Yes
Challenging	8.27 (9.20)	—	—	Yes
Prosocial	RG	—	—	Yes
Variable BL	—	8.16 (10.64)	—	Yes
Variable INT	—	12.92 (11.15)	—	Yes
Variable BL and INT	—	−17.97 (14.24)	—	Yes
Effect present	—	—	26.79[a] (4.52)	26.92[a] (5.20)
r^2	.07	.08	.55	.66
SER	18.70	18.59	12.39	12.48

Note. $n = 27$ for all models; coefficient estimates reported as percentage points with standard errors in parentheses. BL = baseline; INT = intervention; RG = reference group; SER = standard error of regression; Yes = indicates variable was included in the model solely to test the robustness of other variables.
[a]Benjamini–Hochberg-corrected p value $<.001$.

RQ3: What Progress Monitoring Graphs Are Most Frequently Visually Analyzed Correctly by Professionals?

Graphs 6, 10, 18, and 31 were analyzed correctly for a therapeutic effect by all respondents ($n = 57$). Graph 19 was analyzed correctly by the fewest respondents ($n = 43$, 75%). The average percentage of respondents correctly analyzing a graph was 92% ($n = 53$). For 19 of the 27 graphs, at least 90% of respondents ($n = 52$) provided correct visual analysis determinations. Refer to Supplemental Table S5 for a rank order of all graphs.

RQ4: What Are the Features of Progress Monitoring Graphs That Are Associated With Correct Visual Analysis Determinations by Professionals?

Results of regression Models 5 to 8 are detailed in Table 2. Model 5 examined the relationship between behavior types reported on a graph and the percentage of respondents correctly analyzing the graphed data. None of the behavior type variables achieved or approached statistical significance, with coefficient estimates ranging from −2.32 PP to 2.12 PP and standard errors ranging from 3.40 PP to 4.09 PP. The results from Model 5 do not support that academic, daily living, or challenging behaviors reported on a graph meaningfully affected the percentage of respondents that correctly analyzed the graphed data when compared with graphs reporting on prosocial behavior (reference group in the model). Model 6 included indicators for the presence of variable data in a baseline condition, intervention condition, and both conditions. Estimates for variable baseline data ($B = -6.29$ PP, $SE = 2.29$, $p = .010$, Benjamini–Hochberg $p = .044$) and variable intervention data ($B = -6.77$ PP, $SE = 2.42$, $p = .012$, Benjamini–Hochberg $p = .044$) achieved statistical significance and retained significance with Benjamini–Hochberg corrections. Interpretations of these estimates indicate that when a graph contained variable baseline data, on average, the percentage of respondents correctly analyzing the graph was 6.29 PP less than the percentage for a graph not containing variable baseline data when controlling for the variability of data in the intervention condition and any additive effect of a graph having variable data across both baseline and intervention conditions. When a graph contained variable intervention data, on average, the percentage of respondents correctly analyzing the graph was 6.77 PP less than for a graph not containing variable intervention data when controlling for all other variables. Estimates for variable data across both baseline and intervention conditions did not achieve statistical significance ($B = 6.04$ PP, $SE = 4.97$, $p = .236$), indicating that there was no additive effect when variable data were present across both conditions and when controlling for the variable data in either single condition. The presence of a therapeutic effect in Model 7 did not meaningfully affect the percentage of respondents that correctly analyzed a graph's data ($B = -2.14$ PP, $SE = 2.62$, $p = .420$). To test the robustness of significant estimates for variable data, we included all tested variables in Model 8. The indicator for variable baseline data ($B = -6.35$ PP, $SE = 2.59$, $p = .024$, Benjamini–Hochberg $p = .066$) and the indicator for variable intervention data ($B = -6.77$ PP, $SE = 2.59$, $p = .017$, Benjamini–Hochberg $p = .053$) retained statistical significance in Model 8 and approached statistical significance with Benjamini–Hochberg corrections. In addition, the standard error of regression increased from Model 7 to Model 8, suggesting that the inclusion of all variables again led to decreased model performance.

Given that Graph 27 was excluded from the analytical sample, we conducted a sensitivity analysis to determine how the graph's inclusion changed our estimates of previously identified significant variables. We included Graph 27 in Model 7, given that this model achieved the lowest standard error of regression among the tested models for the outcome variable in question, and thus, we determined this model to be the best performing model. When including Graph 27 in Model 7, coefficient estimates and standard errors were relatively unchanged with all differences being

Table 2. Results of Regression Models Estimating the Effect of Graphic Features on the Percentage of Respondents That Correctly Identified the Presence of a Therapeutic Effect.

	Model			
Variable	5	6	7	8
Behavior type				
Academic	2.12 (3.40)	—	—	Yes
Daily living	−1.08 (4.09)	—	—	Yes
Challenging	−2.32 (3.64)	—	—	Yes
Prosocial	RG	—	—	Yes
Variable BL	—	−6.29[a] (2.29)	—	−6.35[b] (2.59)
Variable INT	—	−6.77[a] (2.42)	—	−6.77[b] (2.59)
Variable BL and INT	—	6.04 (4.97)	—	Yes
Effect present	—	—	−2.14 (2.62)	Yes
r^2	.05	.19	.03	.32
SER	7.11	6.55	6.90	6.61

Note. $n = 27$ for all models; coefficient estimates reported as percentage points with standard errors in parentheses; BL = baseline; INT = intervention; RG = reference group; SER = standard error of regression; Yes = indicates variable was included in the model solely to test the robustness of other variables in the model.
[a]Benjamini–Hochberg-corrected p value $< .05$. [b]Benjamini–Hochberg-corrected p value $< .10$.

less than 0.02 PP. Both the indicator for variable baseline data ($B = -6.28, SE = 2.29, p = .011$, Benjamini–Hochberg $p = .044$) and the indicator for variable intervention data ($B = -6.77$ PP, $SE = 2.41$, $p = .010$, Benjamini–Hochberg $p = .044$) achieved statistical significance and retained significance with Benjamini–Hochberg corrections.

RQ5: Are Graphs That Are More Likely to be Encountered by Professionals Also More Likely to be Analyzed Correctly?

Supplemental Figure S4 displays a scatterplot and line-of-best fit for the relationship between a graph's percentage of respondents providing a rating of likely to be encountered and the percentage of respondents correctly analyzing the graph. Results of our regression analysis indicate that, on average, a 1.00 PP increase in the percentage of respondents rating a graph as likely to be encountered was associated with a 0.06 PP decrease in the percentage of respondents correctly analyzing the graph. The precision of this estimate was relatively imprecise ($SE = 0.07$), precluding statistical significance ($p = .423$). This analysis does not support that a graph's likelihood of being encountered affects a professional's ability to correctly analyze a graph.

When including Graph 27 in the regression analysis to for sensitivity purposes, the magnitude of the effect decreased slightly ($B = -0.02$ PP) along with the precision of the estimate ($SE = 0.89$ PP); the estimate did not achieve statistical significance ($p = .828$).

Discussion

Our findings provide insight into the type of progress monitoring graphs that professionals perceive as likely to encounter when working with individuals with or at risk for disabilities, as well as how these graphs are analyzed by professionals. These findings build upon prior studies exploring professionals' abilities to accurately analyze progress monitoring data. Most directly, our findings are in contrast with those reported by Lane et al. (2019), given that we did not detect a relationship between graphs that were likely to be encountered in practice and graphs that respondents most frequently analyzed correctly. The practical implication of this finding is that practitioners will encounter "easy-to-analyze" progress monitoring graphs at a similar rate as "hard-to-analyze" progress monitoring graphs. Therefore, the extent to which pre- and in-service professionals are properly trained to analyze various data patterns of learner progress is of critical significance to ensuring that valid instructional decision-making is continually occurring. Reasons for the differing findings between studies may pertain to (a) the types of individuals who analyzed graphs for the presence of an effect in each study and (b) the number of professionals that rated graphs as likely to be encountered in each study. Within the present study, 57 in-service professionals analyzed graphs for the presence of an effect and rated graphs on their likelihood of being encountered, whereas pre-service educators (i.e., college students) visually analyzed graphs in Lane et al.'s study and only eight in-service professionals rated graphs on their likelihood to be encountered. We perceive that the experience of in-service professionals contributed to more accurate visual analysis determinations in our study and that the increased sample size in our study contributed needed precision to the ratings of perceived likelihood.

Regarding the implications of our study as they pertain to training professionals, the findings provide initial data to support the development of targeted content for trainings on analyzing progress monitoring data. Although there are numerous research studies in which trainings specific to analyzing progress monitoring data for learners with or at risk for disabilities have been evaluated, it is unclear the extent to which the data patterns included in those trainings aligned with those that are likely to be encountered in practice. In contrast, the findings from our

study provide initial support for the development of graphs depicting data patterns likely to be encountered in practice, specifically graphs demonstrating therapeutic effects. Our findings also extend the evidence about the types of graphs that are most likely to be analyzed incorrectly by professionals, with our findings suggesting that graphs with variable data in either the baseline or intervention condition (but not both) as the most difficult to analyze. We hope that professional development providers and content creators will refer to our data to support the development of materials included within future trainings on analyzing progress monitoring data.

Limitations

In our view, the primary limitation to our findings is the use of a virtual "snowball" sampling method. As a brief illustration on the need for these controls, the average percentage of respondents, who rated a graph in the analytical sample as *likely to be encountered*, is quite different when broadly grouping respondents by institutional affiliation. Of the respondents affiliated with the research team's home institution ($n = 15$), on average 57% ($SD = 29$ PP) of respondents rated a graph as *likely to be encountered*; whereas, of respondents who were not affiliated with the research team's home institution ($n = 42$), an average of 70% ($SD = 60$ PP) of the respondent rated a graph as *likely to be encountered*. These differences in magnitude and precision suggest that there may be meaningful differences between the groups of respondents who are and are not affiliated with the research team's home institution.

Another limitation that warrants discussion is our failure to develop an a priori plan for identifying and removing outliers from analyses. We attempted to correct for this by conducting sensitivity analyses; however, pre-registering our study and methods would have been ideal for handling this situation. In addition, our use of likelihood percentages represents professionals' perceptions, which may not align with a file-drawer review of progress monitoring data as collected by in-service professionals.

Conclusion

This study provides data on the validity of the content included in graphs that professionals serving learners with or at risk for disabilities are likely to encounter when engaging in progress monitoring. Our findings indicate that graphs demonstrating a therapeutic effect are most likely to be encountered, while graphs with variable baseline or intervention are most likely to be incorrectly analyzed by professionals. For trainers and higher education faculty, we hope that they will use our findings as they prepare professionals to analyze learner data.

Declaration of Conflicting Interests

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

Funding

Funding was provided by the Institute of Education Sciences, Grant No. R324B210002.

ORCID iD

Collin Shepley https://orcid.org/0000-0003-1967-9714

Supplemental material

Supplemental material for this article is available on the *Remedial and Special Education* website with the online version of this article and at https://osf.io/dyq7h/.

References

Akers, L., Del Grosso, P., Atkins-Burnett, S., Boller, K., Carta, J., & Wasik, B. A. (2014). *Tailored teaching: Teachers' use of ongoing child assessment to individualize instruction (Volume II)*. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding curriculum-based measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology*, *51*(1), 1–18. https://doi.org/10.1016/j.jsp.2012.09.004

Ardoin, S. P., Wagner, L., & Bangs, K. E. (2016). Applied behavior analysis: A foundation for response to intervention. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 29–42). Springer.

Endrew F. v. Douglas County School District, 580 U.S. ____ (2017).

Etscheidt, S. K. (2006). Progress monitoring: Legal issues and recommendations for IEP teams. *Teaching Exceptional Children*, *38*(3), 56–60. https://doi.org/10.1177/004005990603800308

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, *36*(3), 387–406. https://doi.org/10.1901/jaba.2003.36-387

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188–192. https://doi.org/10.1080/02796015.2004.12086241

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*(3), 199–208. https://doi.org/10.1177/001440298605300301

Gischlar, K. L., Hojnoski, R. L., & Missall, K. N. (2009). Improving child outcomes with data-based decision making: Interpreting and using data. *Young Exceptional Children*, *13*(1), 2–18. https://doi.org/10.1177/1096250609346249

Grigg, N. C., Snell, M. E., & Loyd, B. (1989). Visual analysis of student evaluation data: A qualitative analysis of teacher decision making. *Journal of the Association for Persons With Severe Handicaps*, *14*(1), 23–32. https://doi.org/10.1177/154079698901400104

Jimenez, B. A., Mims, P. J., & Baker, J. (2016). The effects of an online data-based decisions professional development for in-service teachers of students with significant disability. *Rural Special Education Quarterly*, *35*(3), 30–40. https://doi.org/10.1177/875687051603500305

Jimenez, B. A., Mims, P. J., & Browder, D. M. (2012). Data-based decisions guidelines for teachers of students with severe intellectual and developmental disabilities. *Education and Training in Autism and Developmental Disabilities*, *47*(4), 407–413. https://www.jstor.org/stable/23879634

Kipfmiller, K. J., Brodhead, M. T., Wolfe, K., LaLonde, K., Sipila, E. S., Bak, M. S., & Fisher, M. H. (2019). Training front-line employees to conduct visual analysis using a clinical decision-making model. *Journal of Behavioral Education*, *28*(3), 301–322. https://doi.org/10.1007/s10864-018-09318-1

Lane, J. D., Shepley, C., & Spriggs, A. D. (2019). Issues and improvements in the visual analysis of AB single-case graphs by pre-service professionals. *Remedial and Special Education*, *42*(4), 235–247. https://doi.org/10.1177/0741932519873120

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Ledford, J. R., & Gast, D. L. (Eds.). (2018). *Single-case research methodology: Applications in special education and behavioral sciences*. Routledge.

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: A systematic review. *Applied Measurement in Education*, *33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification*, *39*(4), 510–541. https://doi.org/10.1177/0145445515581327

Sandall, S. R., Schwartz, I. S., & Lacroix, B. (2004). Interventionists' perspectives about data collection in integrated early childhood classrooms. *Journal of Early Intervention*, *26*(3), 161–174. https://doi.org/10.1177/105381510402600301

Stewart, K. K., Carr, J. E., Brandt, C. W., & McHenry, M. M. (2007). An evaluation of the conservative dual-criterion method for teaching university students to visually inspect AB-design graphs. *Journal of Applied Behavior Analysis*, *40*(4), 713–718. https://doi.org/10.1901/jaba.2007.713-718

What Works Clearinghouse. (2017). *What Works Clearinghouse Standards Handbook Version 4.0*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Wolery, M. (2004). Monitoring children's progress and intervention implementation. In M. McLean, D. B. Bailery, Jr, & M. Wolery (Eds.), *Assessing infants and preschoolers with special needs* (pp. 545–584). Merrill.

Wolfe, K., & Slocum, T. A. (2015). A comparison of two approaches to training visual analysis of AB graphs. *Journal of Applied Behavior Analysis*, *48*(2), 472–477. https://doi.org/10.1002/jaba.212

Yell, M. L., & Bateman, D. F. (2017). Endrew F. v. Douglas county school district (2017) FAPE and the US supreme court. *Teaching Exceptional Children*, *50*(1), 7–15. https://doi.org/10.1177/0040059917721116