# Validity and Judgment Bias in Visual Analysis of Single-Case Data

*Jürgen Wilbert[1], Jannis Bosch[1], and Timo Lüke[2]*
*[1]University of Potsdam*
*[2]University of Graz*

## Abstract

Analysis of data from single-case intervention studies commonly involves visual analysis. Previous research indicates that visual analysis may suffer from low reliability and unpromising error rates. We investigated the reliability and validity of visual analysis and explored to what extent data trends affect judgments. We administered a within-subject experiment in which 186 teacher-education students visually analyzed specifically constructed single-case graphs that included either an intervention effect, a trend effect, both effects, or no effect. Participants identified intervention effects in 75% of the graphs, regardless of the existence of a trend. Type I error rates were low (5%) in graphs without a trend but increased fivefold (25%) for graphs with a trend. Inter- and intra-rater reliability was low, particularly when a trend was present in the data.

*Keywords:* Single-case research, visual analysis, judgment, trend, curriculum-based measurement, reliability, validity

Single-case research has become an important and broadly accepted way to gain insight into educational processes (Gast & Ledford, 2018; Horner et al., 2005). Particularly in the field of special education, single-case research has been adopted as an appropriate method of evaluating the effectiveness of an intervention or the developmental processes underlying difficulties in acquiring academic skills (Kratochwill et al., 2010, 2012). Furthermore, single-case methods can be used by teachers and educators who are interested in evaluating the effects of their interventions or the learning progress of their students (e.g., in combination with curriculum-based measurements). The information resulting from single-case research designs is helpful for decision-making regarding future teaching processes for an individual student but also helps to decide whether or how to implement certain teaching methods in the classroom.

One of the major concerns with single-case studies is the validity of conclusions drawn from the data, with respect to both internal and external validity. Internal validity addresses the question whether a correct causal relation can be inferred from an intervention applied during a single-case study, whereas external validity refers to the generalizability of results across persons, settings, and measurements found in the study (Shadish et al., 2002).

Several strategies have been developed to counter these two methodological issues. These strategies focus either on aspects of the design or on methods for analyzing single-case data. In the present paper, we take a closer look at visual inspection, one of the major methods for analyzing single-case data. We specifically focus on aspects of the internal and external validity of conclusions derived from visual inspections.

## Visual Inspection

Visual inspection (or visual analysis; Barton et al., 2018) is one of the most common strategies for analyzing single-case data (Davis et al., 2013; Lane & Gast, 2014). However, it is also one of the most controversial. In visual analysis, a person, usually the

Wilbert, J., Bosch, J., & Lüke, T. (2021). Validity and Judgment Bias in Visual Analysis of Single-Case Data. *International Journal for Research in Learning Disabilities, 5*(1), 13-24. https://doi.org/10.28987/ijrld.5.1.13

investigator, draws a conclusion about the effectiveness of an intervention solely based on inspection of a diagram comprising the measurement times on the x-axis and the measured values on the y-axis with a vertical line indicating the beginning of the intervention (Spriggs et al., 2018).

Proponents of visual inspection argue that this procedure is practitioner friendly, no further in-depth statistical knowledge is required, and the results are directly and easily understandable (Parsonson & Baer, 2015). They assume that any effect large enough to be practically significant will be detected with visual analysis and that advanced statistical procedures sensitive enough to detect smaller effects do not provide additional clinically significant information (Kazdin, 2011; Parsonson & Baer, 2015).

Critics, on the other hand, argue that visual analyses yield low interrater reliabilities (Danov & Symons, 2008; Ottenbacher, 1990; Park et al., 1990; van den Bosch et al., 2017). Furthermore, the presence of a data trend beginning prior to the intervention (i.e., a positive lag 1 autocorrelation) substantially increases the rate of type I judgment errors (i.e., an unsuccessful intervention is erroneously judged as being successful; Allison et al., 1992; Greenwald, 1976; Jones et al., 1978; Matyas & Greenwood, 1990). However, the proclaimed conservative nature of visual analysis, which should lead to an increase in type II error rates (i.e., a successful intervention is erroneously judged as being unsuccessful; Kazdin, 2011; Parsonson & Baer, 2015), has not been corroborated empirically (Matyas & Greenwood, 1990).

Previous research exploring the strategies used by inservice (Espin et al., 2017) and preservice teachers (Wagner et al., 2017) as well as board-certified behavior analysts (Normand & Bailey, 2006) when interpreting single-case data has demonstrated that the same errors occurred independent of previous experience with visual analysis of single-case data (Espin et al., 2017). This cannot be explained by poor general graph-reading skills alone, as Zeuch et al. (2017) found a correlation of $r = .45$ between general graph-reading skills (extracting information from pie, bar, line and other charts) and the accuracy of visual judgments of learning-progress charts (interpreting the first and last data point of the graph and judging the development throughout all data points).

A major challenge in determining the validity of visual analysis involves selecting a standard with which to compare raters' judgments. Most studies compare visual judgments to the results of statistical procedures (e.g., Brossart et al., 2006; Brossart et al., 2014). This approach, estimating the correctness of visual judgment by comparing it to the results of a statistical analysis, implies that the respective statistical procedures are the best possible ways to analyze the data and that raters cannot be more efficient than such statistical analysis. Both assumptions are highly problematic. Hence, statistical and visual analyses cannot be compared properly. It gets even more complicated when different statistical procedures are applied, with some corroborating an effect while others reject it (Parsonson & Baer, 2015).

## Model-Based Data Generation as a Standard of Comparison

One way to overcome these problems is to simulate single-case data with highly controllable and known statistical properties. These properties are systematically varied, and single-case graphs are provided to raters with which to judge the presence of an effect or other criteria of interest to the researcher (e.g., Matyas & Greenwood, 1990; Ximenes et al., 2009).

The challenge here is determining which model to base the data generation on. An inappropriate model might threaten the ecological validity of a study (i.e., the extent to which the material approaches the conditions of real-world data). While all models are reductions of the complexity of real-world events, an oversimplified model impairs the generalizability of the conclusions and diminishes the external validity. Moreover, the data-generation process must be deduced from a model that is based on a theory of the factors influencing the measurements across time. Without a sound theoretical foundation, inference from the results of a particular study to the higher-order construct it presents (Shadish et al., 2002) is not possible. That is, a study lacks construct validity.

Huitema and McKean (2000) suggested a general model for single-case data: two-phase single-case designs with a pre-intervention phase comprising measurements before the start of the intervention (Phase A) and an intervention phase containing measurements beginning at the intervention's start and lasting throughout the intervention (Phase B). In this model, four factors predict the outcome at a specific measurement point: the performance at the beginning of the study (intercept), a developmental effect leading to a continuous increase throughout all measurements of both phases (trend effect), an immediate intervention effect leading to an immediate and enduring increase in the level of performance (level effect), and a continuous intervention effect that leads to a continuous increase in the slope of the learning curve (slope effect).

Most investigations that have used artificially constructed single-case graphs to study the quality of visual analyses have focused on participants' accuracy in detecting a level effect of an intervention under varying circumstances (e.g., Brossart et al., 2006; Matyas & Greenwood, 1990; Ximenes et al., 2009). For example, Normand and Bailey (2006) systematically varied level and slope effects in a study on visual aids in visual analysis; however, they did not present completely controlled data but manipulated data of two real-world single case graphs.

Espin et al. (2018) explored the evaluation accuracy and difficulty (i.e., response time) of preservice teachers comparing different "graph patterns;" that is, combinations of level and slope effects and display of goal lines (slope-to-goal and slope-to-slope comparison). However, the stimulus material appeared to be "error-free" as it used straight lines within phases. While their findings are new to the field and relevant as they show that even comparing straight lines was not easy for the participants, the material was rather artificial. In practice, virtually no single-case experiment results in a graph with straight lines for Phases A and B – ideally with a level *and* an additional slope effect recognizable.

## Research Questions

In an attempt to fill the above research gaps, the present study addressed the following research questions:

1. How accurate are judgments on single-case graphs? To what extent do baseline trends influence the accuracy of judgments on single-case graphs?
2. How reliable are judgments on single-case graphs? To what extent do baseline trends influence the reliability of judgments on single-case graphs?

For the study, graphs were created using naturalistic – though simulated – data, based on a model including several factors (Huitema & McKean, 2000). Hence, we examined the reliability (intra- and inter-rater) and judgment correctness (power and type I error probability) of visual inspections. Additionally, we examined the impact of a baseline trend on judgment accuracy and reliability regarding the effectiveness of an intervention. The current research focused on an intervention effect that exerts its influence as a continuous increase in performance, starting with the beginning of the intervention (a slope effect). Therefore, we wanted to determine the influence of a trend effect (a positive lag 1 autocorrelation) on judgment correctness and reliability.

## Hypotheses

We expected that judgments on the effectiveness of an intervention based on visual analyses would yield high power and low type I error rates when no trend effect is present (Hypothesis 1a). In contrast, when a trend effect is present, we expected increased type I error rates. (Hypothesis 1b). Moreover, we expected a high consistency of judgments between raters and low uncertainty within each rater when no trend effect is present. Hence, both inter- and intra-rater judgment reliabilities should be high (Hypothesis 2a). However, judgments should become unstable and inconsistent between raters resulting in decreased inter- and intra-rater reliabilities when a trend effect is present (Hypothesis 2b).

Furthermore, we wanted to differentiate between a technical judgment on the existence of an intervention effect (intervention effectiveness) and a pedagogical judgment on the efficacy of the intervention (intervention efficacy). The terms *intervention effectiveness* and *intervention efficacy* are used throughout this manuscript in order to distinguish between these two judgment processes. This distinction has not been made before and might provide further insights into the topic.

# Method

To answer the research questions and test the hypotheses, we implemented a computer-based within-subject experiment, in which teacher-education students conducted visual analyses of graphs from a fictitious single-case research intervention study on reading. Their judgments are then compared to the graphs' underlying statistical properties.

## Participants

A sample of 186 first-year teacher-education majors (89% female) from a research university in Germany participated in our study, ranging in age from 18 to 37 ($M$ = 22.3, $SD$ = 4.6). In self-report ratings 147 (79%), participants reported having no prior knowledge of assessing learning development; 36 (19%) reported having basic and only three (2%) reported having substantial knowledge in this field. Similarly, participants reported having little previous knowledge about single-case data analysis. Specifically, a majority, 143 (77%), had no prior knowledge, 39 (21%) had basic knowledge, and 4 (2%) reported having substantial knowledge. All participants attended the lecture Introduction to Inclusive Education and received partial course credit as compensation. Nevertheless, participation in the study was voluntary as participants had the option of completing an assignment instead of participating. Only two students picked the assignment.

## Procedure

After giving written informed consent, participants were instructed and tested in groups of (up to) four in a lab located on campus, seated in front of computers, separated by panel screens. All instruction and the test were computer-administered. To make sure participants understood the principles underlying visual analysis, participants learned about the difference between baseline and intervention phase and the difference between trend and intervention effect in single-case research designs. Further, the instruction included a cover story about single-case research on reading speed.

Afterwards, participants evaluated 80 single-case graphs. The graphs were presented in a randomized order. With the current graph visible, they answered three questions:

1. Does the reading speed of this child change throughout the data? Response options: "it declined,""no change," and "it increased."
2. Does the intervention have an effect? Response options: *Negative effect, No effect*, and *Positive effect* (technical judgment or *intervention effectiveness*).
3. Do you think it is useful to apply this intervention to a child with similar skills? Rated on a 5-point Likert scale with *Certainly not* (0) and *Certainly* (4) as semantic anchors (pedagogical judgment or *intervention efficacy*).

No time limit was set for answering the questions. Participants responded to Question 1 in $M = 4.0$ seconds ($SD = 5.2$), to Question 2 in $M = 3.0$ seconds ($SD = 3.0$), and to Question 3 in $M = 7.2$ seconds ($SD = 6.4$).

Three weeks later, 87 participants, randomly drawn from the first sample, were again presented with a random sample of 40 single-case graphs (10 per condition; details on the four conditions follow) drawn from the original item pool to determine test-retest reliability. The procedure was identical to the first measurement.

## Design and Materials

We generated AB single-case graphs using a regression-based method. We adopted a common method to visualize single-case data (Spriggs et al., 2018; see Figure 1 for an example). To distinguish between trend effect (i.e., lag 1 autocorrelation throughout all data points) and intervention effect (i.e., an additional slope effect in Phase B), we implemented a 2 x 2 within-subject design. Both, trend and intervention effect, could be either present or non-present, resulting in the following conditions: trend effect ($T^+I^0$), trend and intervention effect ($T^+I^+$), intervention effect ($T^0I^+$), and no effect ($T^0I^0$).

A linear model applying the following formula created each of the 80 single-case graphs (20 per condition):

$y_i = \beta_{0i} + \beta_{trend|condition} \times MT + \beta_{intervention|condition} \times (MT\text{-}9) \times D + \varepsilon$, where $\beta_{0i}$ is the intercept (i.e., the starting value) of case *i*, $\beta_{trend|condition}$ is the trend effect size, $\beta_{intervention|condition}$ is the intervention effect size, *MT* is the measurement time, *D* is a dummy-vector showing whether or not an intervention was present, and ε a measurement error.

Although each case was randomly created, simulation parameters were set according to empirical values reported by Klicpera and Schabmann (1993), who investigated the reading speed (words per minute) of German primary school students. The starting value ($\beta_{0i}$) was randomly chosen from a normal distribution with $M = 130$ and $SD_{between} = 20$ for each case. Trend effect size was set to one standard deviation across all 30 measurement points of a single case. Therefore, changes per measurement ($\beta_{trend}$) for conditions with a trend ($T^+I^0$ & $T^+I^+$) was  and zero for conditions without trend ($T^0I^+$ & $T^0I^0$). The intervention effect size was set to three standard deviations across the 20 Phase B measurements (representing a shift from a very weak to an average reader based on the values reported by Klicpera and Schabmann, 1993). Accordingly, for conditions with an intervention effect ($T^0I^+$ & $T^+I^+$) $\beta_{intervention}$ was  and for conditions without an intervention effect ($T^+I^0$ & $T^0I^0$) $\beta_{intervention}$ was zero.

Variability was introduced as a measurement error affecting each single measurement. The measurement error $\varepsilon_{ij}$ for each data point was randomly drawn from a normal distribution with $M = 0$ and $SD = \sqrt{\frac{(1-r_{tt})}{r_{tt}} \times SD_{between}{}^2}$ with the measurement reliability $r_{tt}$ set to .80 and $SD_{between} = 20$ (the standard deviation of the intercept $\beta_{0i}$ between cases). Please compare Figure 1 for single-case graphs for each condition prior to and after the addition of measurement errors.

All graphs were created using the package *scan* (Wilbert & Lüke, 2019) for *R* (R Core Team, 2018).

As a data check, we reanalyzed the resulting 80 single-case data sets. For each data set (and phase) we calculated a regression with the criteria (words per second) regressed on measurement time, providing the slope for

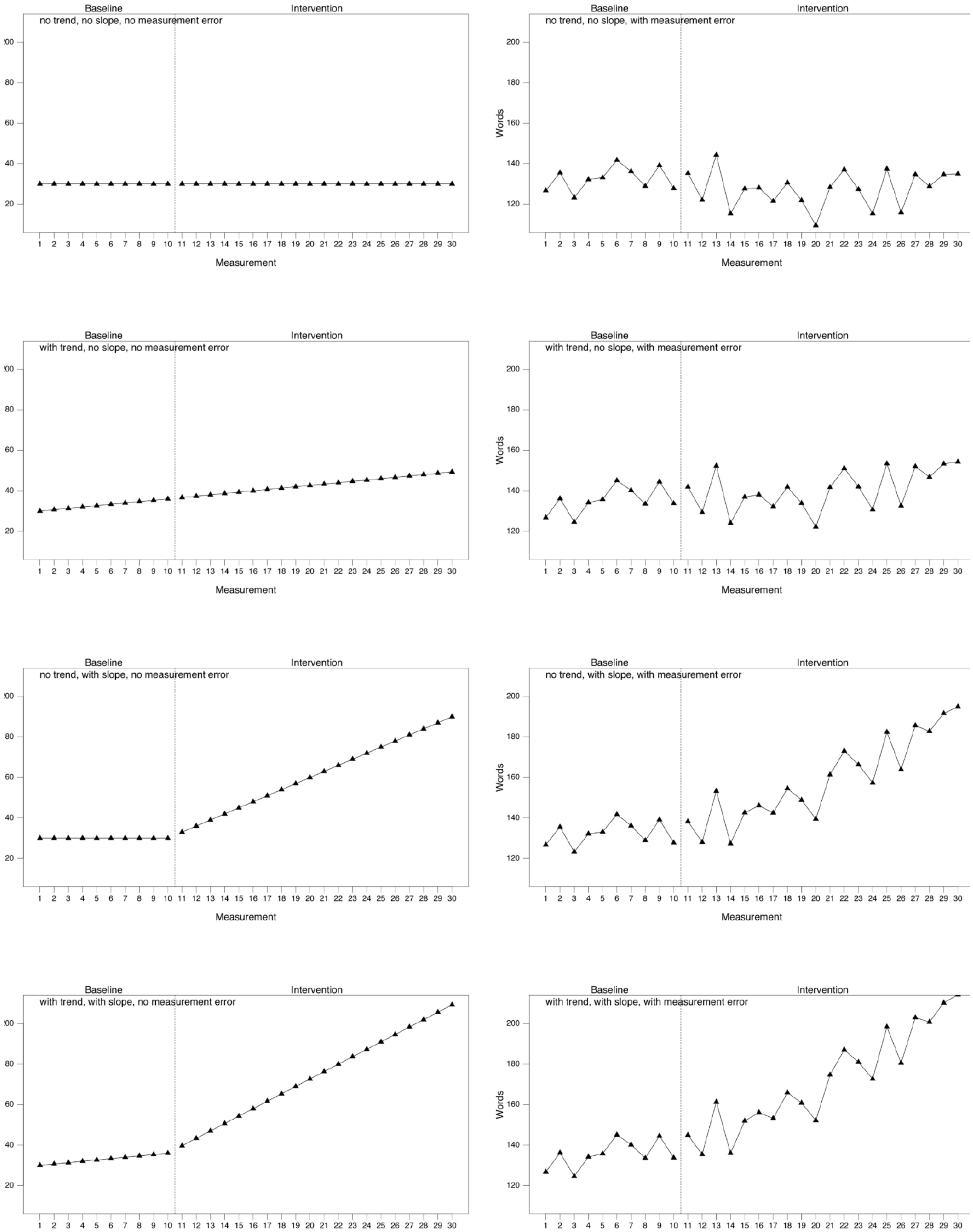

Figure 1.
*Sample Items for Each of the Four Conditions Prior to the Addition of Measurement Errors and After the Addition of a Measurement Error*

each phase. Additionally, we calculated the difference between the Phase B and Phase A slope for each graph. The values for Phase A indicate the trend effect, and the values for Phase B – Phase A indicate the intervention effect. (See Table 1 for further information on mean regression weights of the 20 items [single-case graphs] per condition.) Overall, the intended effects are represented by the 80 items: The two conditions with intervention effect showed an intervention effect (B-A) of 3.28 and 3.06 words per measurement ($d = 3.3$ and $d = 3.1$ for the complete intervention phase), while there was a trend effect of 0.43 and 0.77 words per measurement ($d = 0.65$ and $d = 1.15$ for all phases) for the condition with trend effect and nearly no trend effect (-0.14 and 0.11 words per measurement; that is, $d = -0.21$ and $d = 0.17$) for the conditions without trend effect.

Table 1
*Mean Regression Weights for Each Condition and Phase (Dependent Variable Regressed on Measurement Time)*

| Condition[a] | PHASE | | | |
|---|---|---|---|---|
| | A | B | ALL[b] | B-A[c] |
| $T^+I^+$ | 0.43 | 3.72 | 2.97 | 3.28 |
| $T^0I^+$ | -0.14 | 2.92 | 2.26 | 3.06 |
| $T^+I^0$ | 0.77 | 0.63 | 0.67 | -0.15 |
| $T^0I^0$ | 0.11 | -0.08 | -0.04 | -0.19 |

*Note.* $T^+I^+$ = trend and intervention effect; T+I0 = trend effect only; $T^0I^+$ = intervention effect only; $T^0I^0$ = no effect.

[a] N = 20 items per condition.

[b] All values, ignoring phase separation.

[c] Difference between regression weights of Phases B and A.

Item presentation order was randomized and a second list was created with an inverse order. Each participant was randomly assigned to one of the two orders. Participants' judgments for an item were not influenced by the presentation order; therefore, we ruled out an influence of the serial position or participants fatigue on the results and did not include the presentation order in further analyses.

## Data Analyses

Because judgments of both intervention effectiveness and intervention efficacy are ordinal data, cumulative link models were applied. As we were not only interested in the overall impact of the trend effect and the intervention effect manipulations, but also the variability of this impact between subjects, we implemented multilevel regression models. Trials on level-1 were nested within subjects on level-2. All predictors were modeled as fixed and random effects. Each regression model included two dummy variables representing presence of trend and intervention effects, and the interaction term, as predictor variables and judgment of intervention effectiveness and rating of intervention efficacy, respectively, as the criterion variable.

We calculated likelihood ratio chi-square tests (Winship & Mare, 1984) to determine the significance of the random slope effects. Thus, the complete model was compared to a model without the target random slope (e.g., in order to calculate the significance of the random slope of the trend effect, the full model was compared to a model with all predictors except the random slope trend effect).

We used intraclass correlation coefficients (ICC) to determine the degree of inter-rater agreement (i.e., the extent of agreement between raters). The ICC conceptualizes inter-rater agreement as the proportion of variance determined by the object of observation (Shrout & Fleiss, 1979). Because we were interested in the degree of absolute agreement rather than consistency of ratings, we used case 2 ICC (2, 1), based on the formalization of McGraw and Wong (1996). To determine whether a trend effect impacts inter-rater agreement, we used separate ICCs for trials with and without trend effect and an $F$-test based on the procedure suggested by Donner (1986). Additionally, we calculated Fleiss' Kappa, which only assumes categorical data to check for the stability of the results.

Because judgments of both intervention effectiveness and intervention efficacy are ordinal data, we calculated intra-rater reliability (i.e., the stability of ratings within a person) by means of non-parametric correlation coefficients. Average correlations across participants were Fisher's $z$-transformed in order to account for their skewed distribution.

# Results

First, we checked if participants perceived the trend and intervention effect manipulation. As shown in Table 2, on average, more than 90% of the graphs with an intervention effect ($T^+I^+$ & $T^{0+}I^+$) were correctly identified as displaying an overall increase in reading performance (Question 1); about 6% were rated as showing no change. In the condition without intervention effect but with a positive trend effect ($T^+I^0$), the average ratings identified no change (40%) or an increase (50%). Likewise, 10% attested a decrease in performance.

Table 2

*Average Percentage of Judgments on Overall Development of Change in Reading Performance by Condition*

| Condition | Judgment | | |
|---|---|---|---|
| | Decline | No Change | Increase |
| $T^+I^+$ | 0.3 | 5.7 | 94.0 |
| $T^0I^+$ | 0.6 | 6.2 | 93.2 |
| $T^+I^0$ | 10.3 | 39.4 | 50.3 |
| $T^0I^0$ | 42.3 | 48.0 | 9.7 |

*Note.* $T^+I^+$ = *trend and intervention effect; T+I0 = trend effect only;* $T^0I^+$ = *intervention effect only;* $T^0I^0$ = *no effect.*

Table 3

*Average Percentage of Judgment on Intervention Effectiveness by Condition*

| Condition | Judgment | | |
|---|---|---|---|
| | Negative effect | No effect | Positive effect |
| $T^+I^+$ | 0.7 | 23.6 | 75.5 |
| $T^0I^+$ | 1.4 | 23.8 | 74.8 |
| $T^+I^0$ | 15.8 | 57.6 | 26.6 |
| $T^0I^0$ | 44.8 | 49.9 | 5.3 |

*Note.* $T^+I^+$ = *trend and intervention effect; T+I0 = trend effect only;* $T^0I^+$ = *intervention effect only;* $T^0I^0$ = *no effect.*

Table 4

*Average Percentage of Judgments on Intervention Efficacy by Condition*

| Condition | Judgment | | | | |
|---|---|---|---|---|---|
| | Certainly not (efficacious) | Rather not (efficacious) | Uncertain | Rather (efficacious) | Certainly (efficacious) |
| $T^+I^+$ | 0.5 | 9.6 | 12.1 | 42.8 | 35.0 |
| $T^0I^+$ | 0.3 | 1.3 | 14.2 | 46.6 | 28.5 |
| $T^+I^0$ | 12.9 | 38.5 | 27.5 | 17.9 | 3.2 |
| $T^0I^0$ | 35.8 | 45.0 | 14.7 | 3.8 | 0.7 |

*Note.* $T^+I^+$ = *trend and intervention effect; T+I0 = trend effect only;* $T^0I^+$ = *intervention effect only;* $T^0I^0$ = *no effect.*

Graphs in the condition without intervention or trend effect ($T^0I^0$) were considered as showing no change (48%) or even a decline in performance (42%). Ten percent were rated as increasing reading performance. Hence, participants were able to correctly identify the total increase in reading fluency in the vast majority of graphs with an intervention effect. Judgments on graphs without an intervention effect were also correct in a majority of cases. However, they were a little less accurate than those on graphs with an intervention effect.

## Descriptive Analyses

Participants' ratings of the intervention effectiveness (Question 2) are depicted in Table 3. As illustrated, when an intervention effect was present ($T^+I^+$ and $T^0I^+$), it was detected on about three out of four occasions with about one fourth of the average ratings identifying *no effect*. Similar to the results on Question 1, there were only marginal differences between judgments in the $T^+I^+$ and $T^0I^+$ conditions. When only a trend effect was present ($T^+I^0$), graphs were mainly regarded as representing no intervention effect (58%), but a substantial proportion was judged as showing a negative (16%) or even positive (27%) intervention effect. For the condition with no effects ($T^0I^0$), the majority of participants responded *no effect* (50%) or a *negative effect* (45%). However, 5% of graphs were interpreted as showing a *positive effect*.

Participants' ratings on the intervention efficacy (Question 3) are depicted in Table 4. Once again, ratings were similar for the conditions with ($T^+I^+$) and without ($T^0I^+$) trend effect if an intervention effect was present: Support for further implementation of this intervention was on the same level as the identification of a positive intervention effect. As expected, support for the intervention was lower for the conditions without intervention effect ($T^+I^0$ & $T^0I^0$) and corresponded with the portion of ratings indicating positive intervention effects for these graphs.

Taken together, these results show that the addition of a smaller trend effect had almost no effect on participants' judgments when an intervention effect was already present in the data. When no intervention effect was present, however, a trend effect had a stronger influence on participants' judgment. This pattern was similar for participants' ratings on both intervention efficacy and intervention effectiveness.

We then applied ordinal regression models to further investigate potential interferences of a trend effect on judgment accuracy.

## Hypotheses 1a and 1b: Intervention Effectiveness and Intervention Efficacy

Results of the multilevel ordinal regression models are presented in Table 5 (intervention effectiveness) and Table 6 (intervention efficacy). With respect to intervention effectiveness, the odds ratios suggested that the presence of a trend effect led to a 5.8 times higher chance for the choice of a higher category (i.e., from *negative* to *no* intervention effect or from *no* to *positive* intervention effect) in trials without an intervention effect. However, in trials with an intervention effect, the presence of a trend effect only very slightly increased the chance of a higher category answer than the intervention effect itself, as shown by the odds ratio of .2 for the trend x intervention effect interaction. The intervention effect itself led to a 81.5 times higher chance for the choice of a higher category. Random slope effects documented that all presented effects showed considerable and significant variations, suggesting differentiated influences of trend and intervention effects on the effectiveness judgment between persons.

Table 5

*Multilevel Ordinal Regression Model (Logit) for Participants' Judgment on Intervention Effectiveness (Negative Effect, No Effect, Positive Effect)*

| | β | *SE* | OR[a] | *p* |
|---|---|---|---|---|
| Fixed | | | | |
| Trend effect | 1.76 | 0.06 | 5.8 | <.001 |
| Intervention effect | 4.40 | 0.11 | 81.5 | <.001 |
| Trend x intervention effect | -1.61 | 0.10 | 0.20 | <.001 |
| Thresholds | | | | |
| 0 \| 1[b] | -0.22 | 0.05 | 0.8 | <.001 |
| 1 \| 2[c] | 2.97 | 0.06 | 19.5 | <.001 |
| Random | *SD* ß | LR | *df* | *p* |
| Intercept | 0.49 | | | |
| Trend effect | 0.50 | 75.2 | 4 | <.001 |
| Intervention effect | 1.18 | 306.1 | 4 | <.001 |
| Trend x intervention effect | 0.39 | 3.1 | 4 | <.001 |
| Model fit | | | | |
| LogLik | -10541 | | | |
| AIC | 21113 | | | |

*Note.* Analyses were conducted with the R package ordinal (Christensen, 2019). Trend and intervention effect were dummy-coded (0 and 1).

[a] Odds ratio.

[b] Intercept for judgment negative effect to no effect.

[c] Intercept for judgment no effect to positive effect.

Regression models for intervention efficacy showed a similar pattern (see Table 6). Odds ratios indicated that the trend effect influenced intervention efficacy ratings in trials without an intervention effect (odds ratio of 5 for the trend effect), but not in trials with intervention effect (odds ratio of .3 for the trend x intervention effect interaction). Once again, the intervention effect increased the chance of the choice of a higher category by a factor of 83.7. All effects showed significant random slopes, suggesting that trend and intervention effects also influenced efficacy judgments differentially from person to person.

Table 6

*Multilevel Ordinal Regression Model (Logit) for Rating Intervention Efficacy on a 5-Point Likert Scale (0 – Certainly not to 4 – Certainly). Judgments Nested in Individuals*

| | β | *SE* | OR[a] | *p* |
|---|---|---|---|---|
| Fixed | | | | |
| Trend effect | 1.62 | 0.06 | 5.1 | <.001 |
| Intervention effect | 4.43 | 0.13 | 83.7 | <.001 |
| Trend x intervention effect | -1.30 | 0.04 | 0.3 | <.001 |
| Thresholds | | | | |
| 0 \| 1 | -0.67 | 0.08 | 0.5 | <.001 |
| 1 \| 2 | 1.76 | 0.09 | 5.8 | <.001 |
| 2 \| 3 | 3.10 | 0.07 | 22.1 | <.001 |
| 3 \| 4 | 5.53 | 0.10 | 253.1 | <.001 |
| Random | *SD* ß | LR | *df* | *p* |
| Intercept | 1.04 | | | |
| Trend effect | 0.58 | 107.5 | 4 | <.001 |
| Intervention effect | 1.52 | 88.2 | 4 | <.001 |
| Trend x intervention effect | 0.65 | 24.2 | 4 | <.001 |
| Model fit | | | | |
| LogLik | -17364 | | | |
| AIC | 34762 | | | |

*Note.* Analyses were conducted with the R package ordinal (Christensen, 2019). Trend and intervention effect were dummy-coded (0 and 1).

[a] Odds ratio.

In summary, results of the regression models showed that in trials without an intervention effect the addition of a trend effect led to a roughly five times higher chance for the choice of a higher category answer for both the intervention effectiveness and the intervention efficacy ratings. In contrast, in trials with an intervention effect, the presence of a trend had only minor effects on participants' answers.

## Hypotheses 2a and 2b: Reliability of Visual Judgments

### *Inter-Rater Reliability (Consistency of Judgments Between Raters)*

The overall intraclass correlation between participants was ICC = .50 (95% CI: .43 – .58, $F$[79, 14378] = 203, $p$ < .001) for intervention effectiveness and ICC = .54 (95% CI: .47 – .62, $F$[79, 14378] = 251, $p$ < .001) for intervention efficacy, indicating a relatively low inter-rater reliability (see Table 7). For intervention effectiveness, trials without a trend effect ($T^0$) showed an ICC of .59 (95% CI: .49 – .70, $F$[39, 7098] = 283, $p$ < .001), while the ICC was .34 (95% CI: .25 – .46, $F$[39, 7098] = 107, $p$ < .001) for trials with a trend effect ($T^+$). For intervention efficacy, trials without a trend effect showed an ICC of .62 (95% CI: .52 – .73, $F$[39, 7098] = 340, $p$ < .001) while the ICC was .42 (95% CI: .33 – .55, $F$[39, 7098] = 162, $p$ < .001) for trials with a trend effect. Hence, trials with a trend effect had lower inter-rater reliability than trials without a trend for both intervention effectiveness and intervention efficacy.

### *Intra-Rater Reliability (Stability of Judgments Within Raters)*

Median intra-rater reliability coefficients (Kendall's Tau) for the 40 items administered at Measurement Times 1 and 2 are presented in Table 8 for both intervention effectiveness and intervention efficacy. Intra-rater reliability was relatively low overall, with an average Kendall's Tau of .56 for intervention effectiveness and .57 for intervention efficacy. In line with the results for inter-rater reliability, trials with a trend effect showed lower intra-rater reliabilities for both questions. For intervention effectiveness, trials with a trend showed an average Kendall's Tau of .43 compared to .66 for trials without a trend. For intervention efficacy, trials with a trend showed an average Kendall's Tau of .53, while trials without a trend showed an average Kendall's Tau of .60. Hence, similar to the pattern observed for inter-rater reliability, trials with a trend effect had lower intra-rater reliability than trials without a trend for both intervention effectiveness and intervention efficacy.

Table 7
*Inter-Rater Reliability and Agreement*

| | Intervention effectiveness | | | | Intervention efficacy | | | |
|---|---|---|---|---|---|---|---|---|
| | ICC2.1 | ICC2.K | $r_{wg}$ | Fleiss' K[a] | ICC2.1 | ICC2.K | $r_{wg}$ | Fleiss' K[a] |
| $T^+I^+$ | .045 | .898 | .980 | .042 | .068 | .932 | .965 | .020 |
| $T^0I^+$ | .067 | .930 | .978 | .064 | .107 | .957 | .969 | .033 |
| $T^+I^0$ | .145 | .969 | .946 | .086 | .170 | .974 | .962 | .044 |
| $T^0I^0$ | .151 | .970 | .962 | .128 | .088 | .947 | .976 | .041 |
| $T^1$ | .335 | .989 | .983 | .223 | .423 | .993 | .981 | .121 |
| $T^0$ | .587 | .996 | .985 | .356 | .616 | .997 | .986 | .193 |
| ALL | .499 | .995 | .992 | .304 | .538 | .995 | .992 | .163 |

*Note.* $r_{wg}$ = Within-group correlation comparing the variance of all raters to random variance.
[a]Fleiss' Kappa for nominal and ordinal data (Bliese, 2000).

Table 8
*Mean Correlation (Kendall's Tau) Between Measurements 1 and 2*

| Condition | Intervention effectiveness | | | | Intervention efficacy | | | |
|---|---|---|---|---|---|---|---|---|
| | *N* | *MD* | 1st quartile | 3rd quartile | *N* | *MD* | 1st quartile | 3rd quartile |
| No trend ($T^+I^+$ & $T^+I^0$) | 76 | .66 | .55 | .75 | 76 | .60 | .50 | .68 |
| Trend ($T^0I^+$ & $T^0I^0$) | 76 | .43 | .31 | .59 | 76 | .53 | .39 | .64 |
| All | 76 | .56 | .47 | .67 | 76 | .56 | .48 | .64 |

*Note.* MD is the median correlation for all participants; 1st quartile and 3rd quartile for all participants.

## Discussion

In this study we addressed two central questions: First, how reliable are students' evaluations of single-case graphs? Second, to what extent do baseline trends impact judgment of an intervention's efficacy and effectiveness? We conducted a computer-based within-subject experiment, in which students judged 80 AB single-case graphs. As suggested by Ximenes et al. (2009), artificial data sets were created to enable us to vary intervention and trend effects independently.

In line with Matyas and Greenwood (1990) and corroborating Hypotheses 1a and 1b, judgments were found to be quite accurate when no baseline trend was present, with accuracy dropping considerably (type I errors rates increased fivefold) in the presence of a baseline trend. Unfortunately, the most common areas of application for single-case research – and accordingly visual analysis of the resulting graphs – are interventions targeting learning processes where baseline trends are common (e.g., reading fluency, basic arithmetic). Indeed, our findings support the argument that the presence of a data trend is indeed a major reason for type I errors in the visual analysis of single-case graphs.

However, the baseline trend did not reduce type II error rates (i.e., the power remained about 80%). This might be because the intervention effects used in this study were much larger than the trend effects, therefore, possibly overshadowing them in trials where an intervention effect was present. Future research should investigate whether the increase in type I error rates is replicable even when trend and intervention effect sizes are comparable.

In line with Hypotheses 2a and 2b and previous work by other researchers (Jones et al., 1978; Park et al., 1990), inter- and intra-rater reliabilities dropped for items including a baseline trend compared to those without a trend. However, contrary to our expectations in Hsypothesis 2a, reliabilities were low even for items where no trend was present. Judgments appeared both inconsistent across raters and unstable over time. This pattern was similar for judgments on the effectiveness as well as the efficacy of the intervention.

It is often argued that any intervention effect large enough to be relevant in practice is detectable by visual inspection (e.g., Kazdin, 2011; Parsonson & Baer, 2015). However, in line with other studies (e.g., Danov & Symons, 2008; Ottenbacher, 1990; Park et al., 1990), our results indicate that even under relatively clear conditions, with a large intervention effect size and the intervention effect exceeding the Phase A trend, visual judgments were not reliable both within and between raters.

Generally, the distinction between intervention effectiveness and intervention efficacy did not yield insight into the decision process. Trend effects on judgments were slightly more pronounced for intervention effectiveness, and reliabilities were a bit higher for intervention efficacy but overall, the results were very similar for all analyses. Thus, either the experimental variations exerted a consistent effect on both dependent variables or participants did not differentiate between effectiveness and efficacy and both measured practically the same.

## Conclusion

In summary, the results of our study suggest that first-year teacher-education majors' visual judgments are unreliable and highly prone to type I errors in the presence of a baseline trend in the data. However, this conclusion must be balanced by several limitations: First, participants were university students with limited experience of visual analysis of single-case data. Note, however, that previous studies resulted in comparable reliabilities and error rates – even in experienced raters (Espin et al., 2017; Normand & Bailey, 2006).

Second, although we did our best to explain the difference between a trend and an intervention effect to the students, no empirical evidence verifies that they truly understood the distinction.

Third, the results were based on an arbitrary decision with regard to the relation between the effect size of the intervention, the trend effect, and the measurement error. Arguably, a change in the proportion between these effects sizes might have led to different results.

Finally, we assumed a linear trend and a linear intervention effect. While we think this is appropriate for a reading intervention, other kinds of interventions (e.g., behavioral modifications or medical treatments) might be better represented with non-linear developments or even performance shifts. Our results, therefore, are not directly applicable to these contexts.

Despite these limitations, the present study provides a novel approach to investigating these effects and reveals that, given certain defined conditions, visual inspections are only of limited value. Two ways to overcome such a limitation have been proposed: First, enriching visual graphs with lines (Kratochwill et al., 2010). A widely applied method consists of drawing a mean line of the A phase extrapolated across the B phase to improve visual inspection accuracy. Similarly, Kadzin (2011) recommended inserting a split-middle line, a type of regression line, for the A phase extrapolating across the B phase. However, Fisher, Kelley, and Lomas (2003) found an increased type I error rate when inserting a split-middle line. Instead, they proposed a combination of a mean and a split-middle line (the dual criterion, DC) and showed that this procedure leads to an improved visual inspection accuracy. Evaluation studies on variations of this procedure (the conservative dual criterion with 0.25 standard deviations raised lines) have corroborated these findings (Stewart et al., 2007; Young & Daly, 2016).

A second way to improve the accuracy of visual inspections is to validate the interpretations with the results from statistical analyses (Harrington & Velicer, 2015, Park et al., 1990). However, this approach raises the question if a statistical analysis is the most reliable and valid approach to analyzing single-case data in the first place. From this perspective, visual graphs have an important but mere illustrative function.

## References

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and type I error rate in single-case designs. *The Journal of Experimental Education*, *61*, 45–51. https://doi.org/10.1080/00220973.1992.9943848

Barton, E. E., Lloyd, B. P., Spriggs, A. D., & Gast, D. L. (2018). Visual analysis of graphic data. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 179–214). Taylor and Francis.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Frontiers of industrial and organizational psychology. Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

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

Brossart, D. F., Vannest, K. J., Davis, J. L., & Patience, M. A. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation*, *24*, 464–491. https://doi.org/10.1080/09602011.2013.868361

Christensen, R.H.B. (2019). *Ordinal: Regression models for ordinal data* (Version 2019.3-9). R Package.

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

Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., & Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behavior Modification*, *37*, 62–89. https://doi.org/10.1177/0145445512453734

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, *54*, 67–82. https://doi.org/10.2307/1403259

Espin, C. A., Saab, N., Pat-El, R., Boender, P.D.M., & van der Veen, J. (2018). Curriculum-based measurement progress data: Effects of graph pattern on ease of interpretation. *Zeitschrift Für Erziehungswissenschaft*, *21*, 767–792. https://doi.org/10.1007/s11618-018-0836-9

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & Rooij, M. de (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice*, *32*, 8–21. https://doi.org/10.1111/ldrp.12123

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. https://doi.org/10.1901/jaba.2003.36-387

Gast, D. L., & Ledford, J. R. (2018). Research approaches in applied settings. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 1–26). Taylor and Francis.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*, 314–320. https://doi.org/10.1037/0033-2909.83.2.314

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research*, *50*(2), 162. https://doi.org/10.1080/00273171.2014.973989

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.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, *60*, 38–58. https://doi.org/10.1177/00131640021970358

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

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

Klicpera, C., & Schabmann, A. (1993). Do German-speaking children have a chance to overcome reading and spelling difficulties? A longitudinal survey from the second until the eighth grade. *European Journal of Psychology of Education*, *8*, 307–323. https://doi.org/10.1007/BF03174084

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case design technical documentation*. What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2012). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26–38. https://doi.org/10.1177/0741932512452794

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: brief review and guidelines. *Neuropsychological Rehabilitation*, *24*, 445–463. https://doi.org/10.1080/09602011.2013.815636

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. https://doi.org/10.1901/jaba.1990.23-341

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037//1082-989X.1.1.30

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

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

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

Parsonson, B. S., & Baer, D. M. (2015). The visual analysis of data and current researh into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Routledge.

R Core Team. (2018). *R: A language and environment for statistical computing*. Author. http://www.R-project.org/

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Spriggs, A. D., Lane, J. D., & Gast, D. L. (2018). Visual representation of data. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 157–178). Taylor and Francis.

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

van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice*, *32*, 46–60. https://doi.org/10.1111/ldrp.12122

Wagner, D. L., Hammerschmidt-Snidarich, S. M., Espin, C. A., Seifert, K., & McMaster, K. L. (2017). Pre-service teachers' interpretation of CBM progress monitoring data. *Learning Disabilities Research & Practice*, *32*, 22–31. https://doi.org/10.1111/ldrp.12125

Wilbert, J., & Lüke, T. (2019). *scan: Single-case data analyses for single and multiple designs* (Version 0.40). https://cran.r-project.org/package=scan

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, *49*, 512–525. https://doi.org/10.2307/2095465

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology*, *12*, 823–832. https://doi.org/10.1017/S1138741600002195

Young, N. D., & Daly, E. J. (2016). An evaluation of prompting and reinforcement for training visual analysis skills. *Journal of Behavioral Education*, *25*(1), 95–119. https://doi.org/10.1007/s10864-015-9234-z

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice*, *32*, 61–70. https://doi.org/10.1111/ldrp.12126