

Exploring the Potential of Generative AI as a Support Tool for Single-Case Data Analysis

Li-Ting Chen
University of Nevada, Reno

Isaac Akuoko-Mensah
University of Nevada, Reno

Visual analysis is the primary method used to analyze data in single-case experimental designs (SCEDs). Yet, whether valid conclusions can be drawn from visual analysis remains a concern. There is a need to explore new methods that not only are able to make accurate conclusions from SCED data graphs but are also easy to use. In the present study, we extended the findings of Akuoko-Mensah and Chen (2025) and examined the capabilities of ChatGPT and Claude for analyzing SCED graphs. SCED graphs available at www.singlecase.org were used to train and test ChatGPT and Claude. The findings revealed that Claude's ratings closely aligned with expert ratings on both ABAB and multiple-baseline graphs. In contrast, ChatGPT's ratings were closely aligned with expert ratings on ABAB graphs only. In addition, the graph descriptions and rating justifications generated by Claude included more quantitative detail and terminology related to specific design forms. Based on these findings, we discuss study implications and future research directions.

Keywords: visual inspection, generative AI, ChatGPT, single-subject design, intervention effect, functional relationship

INTRODUCTION

Single-case experimental designs (SCEDs) are a family of experimental designs that aim to determine whether changes in an outcome are in fact a function of an intervention, using repeated assessment of the outcome over time. A causal or functional relation in SCED research is defined as establishing a consistent effect of an outcome by systematically manipulating an independent variable (e.g., intervention implementation). Compared to randomized controlled trials, which are often regarded as the gold standard for determining causal effects using a large number of cases being assigned randomly to one of the control and experimental conditions (Kazdin, 2021; Shadish et al., 2002; Steeger et al., 2021), SCEDs establish causal relationships using a small number of cases that serve as their own controls (Kazdin, 2021; Vlaeyen et al., 2020).

According to Hersen (1990), the roots of SCEDs can be traced back to Johannes Müller and Claude Bernard's work in physiology in the 1830s. Since then, SCEDs have been used in various fields, including counseling (Brown et al., 2024), clinical psychology (Vlaeyen et al., 2020), school psychology (Radley et al., 2020), sports psychology (Barker et al., 2013), education (Dayo et al., 2024; Enders et al., 2024; Shepley

et al., 2024), speech pathology (Binns et al., 2021), and neuroscience (Soto, 2020). In educational technology, researchers may employ SCEDs to examine the effectiveness of technology-enhanced learning or technology integration on improving learning outcomes. For instance, Guinness and Feil (2025) employed an SCED to evaluate the effects of a digital training package on rapport-building skills among four caregivers of children with autism, using 10 to 16 repeated assessments.

SCEDs represent a family of experimental designs, and the primary five design forms are phase designs (e.g., AB, ABA, ABAB design), multiple baseline designs, alternating treatment designs, changing criterion designs, and combinations of two or more of the previous four designs (Dayo et al., 2024; Shadish & Sullivan, 2011; Smith, 2012; Tanious & Onghena, 2022; What Works Clearinghouse, 2022). Among the five design forms, phase designs and multiple baseline designs are the most popular designs (Dayo et al., 2024; Shadish & Sullivan, 2011; Smith, 2012; Tanious & Onghena, 2021). In their review of 136 SCED studies on evaluating technology-enhanced learning, Dayo et al. (2024) reported that 33% used a phase design and 46% used a multiple baseline design.

Visual analysis of data plotted on a line graph is the primary method used to evaluate whether the data pattern reflects a systematic intervention effect in SCED research (Horner et al., 2005; Kazdin, 2019; Ledford et al., 2018). Visual analysis is performed through examining the characteristics of outcome change within and across the same phases, and between different phases of graphed data, based on the features of level, trend, variability, overlap, immediacy, and consistency across similar phases (Ledford et al. 2018; What Works Clearinghouse, 2022). Although visual analysis is a common strategy to determine an intervention effect, the extent to which reliable data-based decisions can be made using visual analysis remains a concern (Dart & Radley, 2024; Fisher et al., 2003; Lanovaz & Hranchuk, 2021; Tarlow et al., 2021; Wolfe & Seaman, 2023; Wolfe et al., 2024). Lanovaz and Hranchuk (2021) found that when visual analysts were less likely to claim an intervention with no true effect to be effective (low Type I error rate), they were also less likely to correctly determine an intervention with a true effect to be effective (low statistical power). These findings implied that visual analysts may struggle to balance Type I error with statistical power. Ninci (2023) suggested that analysts' backgrounds and expertise, and their familiarity with terminology used to describe the effect of SCED data, may affect interrater reliability. Data characteristics such as trend (Wolfe & Seaman, 2023) and effect size (Pierce et al., 2024; Wolfe & Seaman, 2023) have also been shown to affect interrater reliability.

Rapid advancements in generative artificial intelligence (GenAI) have the potential to assist visual analysis of SCED data and enhance valid inferences to be drawn about an intervention effect. A recent study conducted by Akuoko-Mensah and Chen (2025) reported high agreement between Chat Generative Pre-Trained Transformer (ChatGPT) and expert ratings of ABAB graphs. In another evaluation on six GenAIs for their capabilities of graph analysis, Sartori et al. (2025) found that Claude-3.5-Sonnet, Gemini-Pro-1.5, and ChatGPT-4o outperformed Llama3.2, Qwen-2-VL, and Pixtral (Sartori et al., 2025). Among the three best rated GenAIs in Sartori et al. (2025), ChatGPT is especially popular. As of June 2025, ChatGPT has approximately 800 million weekly active users and about 122.58 million daily users (Singh, 2025). Yet Claude 3 can remember and use more information when creating relevant answers than ChatGPT-4o (Belcic & Stryker, 2024). The new version of Claude, released on May 22, 2025, may further improve its ability to conduct graph analysis.

To better understand GenAIs' capabilities to support visual analysis of SCED data, the present study extended findings from Akuoko-Mensah and Chen (2025) by investigating (1) the degree of ChatGPT ratings align with expert ratings and the degree of Claude ratings align with expert ratings on ABAB and multiple baseline graphs, and if the degree of agreement depends on (2) the magnitude of effect, and (3) SCED design forms. The present study aimed to address three research questions.

- RQ1. When GenAI (i.e., ChatGPT and Claude) is tasked to determine the functional relation between independent and dependent variables using SCED data graphs, to what extent do GenAI ratings align with expert ratings as measured by weighted Kappa coefficients?
- RQ2. Does the degree of agreement depend on the magnitude of effect?
- RQ3. Does the degree of agreement depend on SCED design forms?

LITERATURE REVIEW

A single case is the unit of intervention and data analysis in SCEDs, and a case may be a person, a classroom or a community (Horner et al., 2005). The conditions or phases in SCED research are often designated by capital letters. Traditionally, A refers to a baseline condition, and B indicates an intervention condition. Each subsequent condition introduced in a study is denoted in sequential alphabetical order, such as C, D (Ledford et al., 2018).

In the fundamental AB design, a comparison of the data pattern in the baseline condition (A phase) with the data pattern from the same case in the intervention condition (B phase) provides the basis for evaluation of the effectiveness of an intervention. The focus on within-case comparisons makes SCED useful for applied studies (Kazdin, 2021; Ledford et al., 2019). SCEDs are common in the context where interventions are tailored to meet the unique needs of cases (Krasny-Pacini & Evans, 2018). SCEDs also fit well when interventions target persons with low-prevalence conditions (Odom et al., 2005); therefore, it is difficult to obtain a sufficient number of participants for a group design. Furthermore, SCEDs allow for the gradual or small-scale implementation of the intervention (Kazdin, 2021). Before applying an intervention to a large-scale setting (e.g., a state or nation), the intervention can be implemented with a few cases to examine its effectiveness. Kazdin (2021) noted that “first starting out on a modest scale ... helps the investigator preview the impact of the intervention as well as major implementation and some of the practical issues that may relate to its effectiveness” (p. 69).

In the following, we first introduce the two most used SCED forms (Dayo et al., 2024; Shadish & Sullivan, 2011; Smith, 2012; Tanious & Onghena, 2021), namely phase designs and multiple baseline designs. We then provide an overview of visual analysis of SCED data, along with discussions of issues, tools, criteria, and reporting practices. Next, we provide a summary of recent advancements in AI and its potential to support the visual analysis of SCED data. Finally, the focus of the study is outlined, detailing the key research questions guiding the investigation.

PHASE DESIGNS

There are variations in phase designs (e.g., AB, ABA, ABAB, ABCCAB). The most basic phase design is the AB phase design, which is characterized by analyzing two adjacent conditions, namely baseline and intervention conditions, for evaluating an intervention (Kennedy 2005; Lane & Gast, 2014; Michiels & Onghena, 2019). Although changes in data patterns in an AB phase design may be attributed to intervention implementation, it is possible that extraneous variables are influencing the outcome, such as spontaneous recovery or developmental effect. To control threats to internal validity, a minimum of two phases per condition is recommended for a phase design, namely an ABAB design (Kazdin, 2019; What Works Clearinghouse, 2022).

The ABAB design allows a potential demonstration of effect to be replicated across identical baseline and intervention conditions at three different points in time (Ledford et al., 2018). The first potential demonstration of effect is conducted using data in the first baseline (A1) and intervention (B1) phases, the second demonstration is performed in the first intervention phase (B1) and the second baseline phase (A2), and the third demonstration uses data in the second baseline (A2) and intervention phases (B2). If the outcome improves when the intervention (B1) is introduced, reverts to or near baseline levels after the intervention is withdrawn (A2), and again improves when the intervention is reintroduced (B2), then the data pattern suggests strongly that the change was attributed to the intervention (Kazdin, 2021). Figure 1 illustrates hypothetical data collected from a single participant using an ABAB design. The y-axis shows engagement time in minutes, and the x-axis shows the session number. Assuming that blended learning, an intervention of face-to-face instruction combined with computer-based instruction, is used to increase academic engagement time of a student with an emotional behavior disorder, engagement time can be measured to examine the effectiveness of the intervention. In baseline conditions (A1 and A2), only face-to-face instruction is used; in intervention conditions (B1 and B2), a blended learning approach is used.

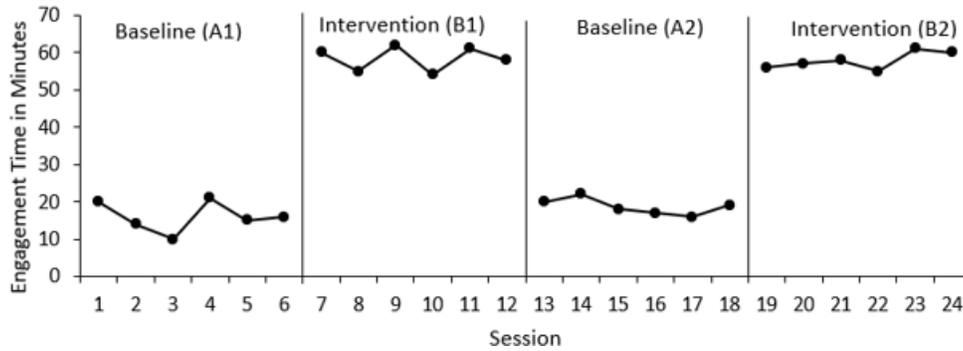


Figure 1. *A Hypothetical Example of an ABAB Design with a Single Participant and Higher Scores Showing a Desired Outcome*

MULTIPLE BASELINE DESIGNS

In multiple baseline designs, multiple baselines are established concurrently, and the intervention is sequentially introduced across the baselines. A major practical advantage of multiple baseline designs is that they do not require withdrawing an intervention (Kazdin, 2019; Kennedy 2005; Ledford et al., 2019). Multiple baseline designs can be used to demonstrate the effectiveness of an intervention across multiple people, behaviors or settings, by comparing data patterns in A and B conditions across people, behaviors or settings. Each A-B comparison is graphed separately and referred to as a tier (Ledford et al., 2019). To control threats to internal validity, a minimum of six phases (What Works Clearinghouse, 2022) or at least three people (or behaviors, settings) is recommended for a multiple baseline design (Ledford et al., 2019).

The multiple baseline design with three tiers allows for an opportunity for a potential demonstration of effect to be replicated across baseline and intervention conditions at three different points in time. The effect of an intervention is demonstrated by showing that outcome changes when and only when the intervention is implemented. For example, assuming that a multiple-baseline-across-people design is used to evaluate whether game-based learning increases academic engagement time, and three college students (Victoria, Max, Luke) with low engagement participate in the study. At the beginning, baseline data are collected for all three students. After baseline data on engagement time are stable for each student, game-based learning is introduced to Victoria. The first demonstration of effect is achieved if Victoria's engagement time increases when game-based learning is used, but both Max's and Luke's engagement time remains the same as the baseline. After engagement time stabilizes across three students, game-based learning is then introduced to Max. The second demonstration of effect is achieved if Max's engagement time changes when game-based learning is introduced to him, but Luke's engagement time remains the same as the baseline. After engagement time again stabilizes across three students, game-based learning is introduced to Luke. The third demonstration of an effect is achieved if Luke's engagement time changes when game-based learning is used.

Figure 2 presents the data from the hypothetical example in the previous paragraph to illustrate a multiple baseline design. The y-axis shows engagement time in minutes, and the x-axis shows the session number. As shown in Figure 2, Victoria's engagement time increases when game-based learning is implemented, while the engagement time of the two other students remains the same. Max's engagement time also increases when the intervention of game-based learning is implemented, while Luke's engagement time remains the same. When the intervention is introduced to Luke, his engagement time then increases.

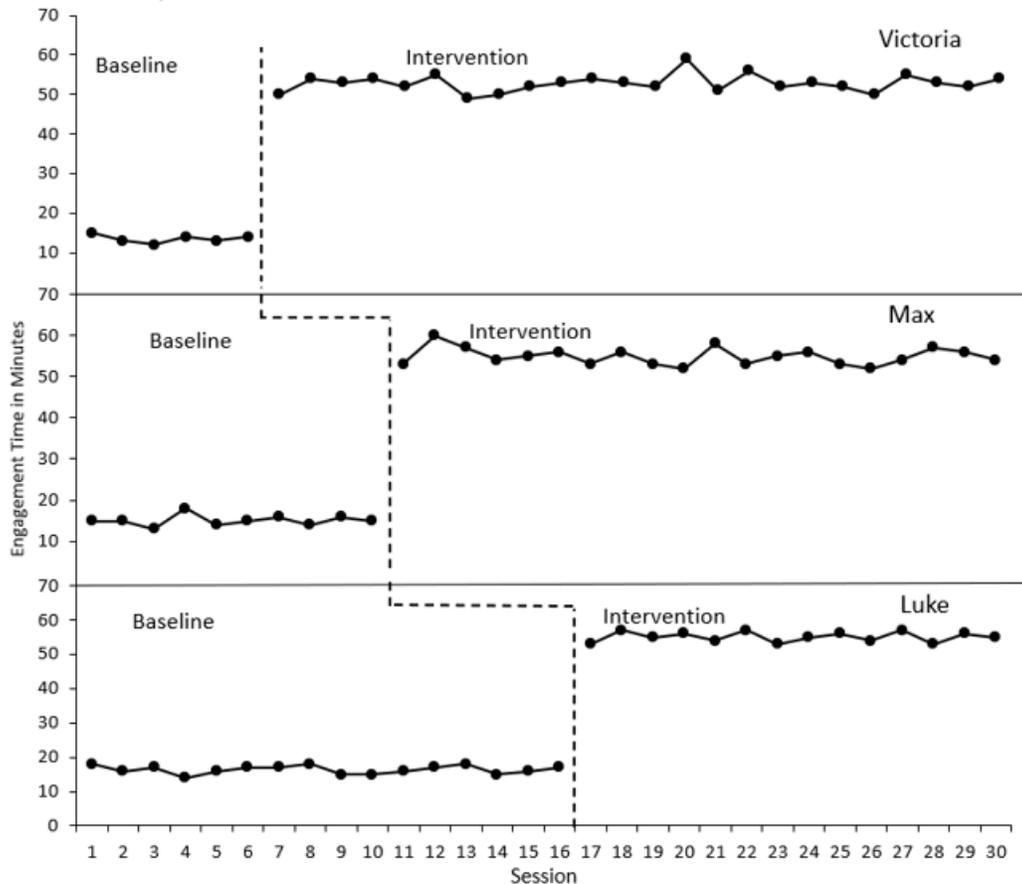


Figure 2. A Hypothetical Example of a Multiple Baseline Design

VISUAL ANALYSIS OF SCED DATA

In 2002, the What Works Clearinghouse was established as a part of an initiative of the U.S. Department of Education's Institute of Education Sciences to evaluate scientific research evidence on education interventions (Institute of Education Sciences, n.d.). According to the What Works Clearinghouse (2017), visual analysis has been used for SCEDs to determine (a) the existence of a functional relation between an independent variable and an outcome variable and (b) the strength of that relation. An inferred functional (or causal) relation requires that changes in the outcome result from the manipulation of the independent variable and at least three demonstrations of an effect at a minimum of three different points in time (WWC, 2017).

Level, trend, variability, immediacy, overlap, and consistency of data in similar phases are the six features to be examined within- and between-phase data patterns for visual analysis in SCEDs (Horner et al., 2005; Ledford et al., 2018; WWC, 2017). These six features are used collectively to “compare the observed and projected patterns for each phase with the actual pattern observed after manipulation of the independent variable” (WWC, 2017, A-7). Specifically, *level* refers to the outcome mean score within a phase (WWC, 2017) and researchers are most often interested in changes in level between adjacent phases (Ledford et al., 2018). *Trend* is “the slope of the best fitting straight line for the data within a phase” (WWC, 2017, A-7). When comparing adjacent phases, an effect is observed when directionality shifts across phases in the desired direction (Ledford et al., 2018). However, when a desired trend occurs in a baseline phase, it presents a potential threat to internal validity because the outcome change may result from a factor (e.g.,

spontaneous recovery or developmental effect) other than the independent variable. *Variability* refers to “the extent of fluctuation in the data within a phase” (Kazdin, 2021, p. 62). Ledford et al. (2018) further explained that when baseline data are considerably variable, a change in level may be due to the variability of the data but not the manipulation of the independent variable. *Immediacy* refers to “the extent to which data change simultaneously with a condition change” (Ledford et al., 2018, p. 10), and an immediate change is preferable. *Overlap* describes the “proportion of data from one phase that overlaps with data from the previous phase” (WWC, 2017, A-7). The fewer overlapping data points, the more compelling the demonstration of an effect. *Consistency of data in similar phases* involves examining data from phases within the same condition, such as all baseline phases or all intervention phases. Great consistency indicates a convincing causal relation between the independent variable and the outcome.

ISSUES, TOOLS, CRITERIA, & REPORTING PRACTICES OF VISUAL ANALYSIS

Even though visual analysis is the primary method used to determine an intervention effect in SCEDs, whether reliable data-based decisions can be made using visual analysis remains a concern (Dart & Radley, 2024; Fisher et al., 2003; Lanovaz & Hranchuk, 2021; Tarlow et al., 2021; Wolfe & Seaman, 2023; Wolfe et al., 2024). Using AB phase design data graphs generated by manipulating six data characteristics, Wolfe and Seaman (2023) found that trend and effect size had a greater impact on interrater agreement than number of outcome measures in Phase A, number of outcome measures in Phase B, autocorrelation and variability. Specifically, higher interrater agreement occurred when there was no trend. Interrater agreement was also higher when there was no effect or large effect than when the effect was moderate (Wolfe & Seaman, 2023). Likewise, using multiple baseline graphs from peer-reviewed articles and dissertations, Pierce et al. (2024) reported that agreement among experts was higher for graphs displaying no or large effect sizes than for graphs displaying moderate effect sizes.

To facilitate consistent and accurate visual analysis, researchers have developed tools and criteria to interpret SCED data systematically (e.g., Cox & Friedel, 2022; Fisher et al., 2003; Hagopian et al., 1997; Manolov & Vannest, 2023; Roane et al., 2013). The dual-criteria and conservative dual-criteria methods developed by Fisher et al. (2003) for an AB contrast can be used for both phase designs and multiple baseline designs. The dual-criteria method requires a trend line based on the split-middle method (Kazdin, 1982) and a horizontal line based on the baseline mean to be drawn and overlaid in the intervention phase. The number of data points that fall above or below the two lines in the desired direction is used to conclude whether an intervention effect has occurred. The conservative dual-criteria method shifts the height of the two lines based on the dual-criteria method by 0.25 standard deviations to control for the Type I error rate (Fisher et al., 2003). The structured criteria developed by Hagopian et al. (1997) for alternating treatment designs require an upper criterion line (UCL) and a lower criterion line (LCL) to be drawn based on 1 standard deviation above and below the mean of the control condition, respectively. The number of outcomes that fall above the UCL and below the LCL, along with other expert-developed guidelines, is used to determine the interpretation of a target behavior (e.g., undifferentiated, maintained by tangible reinforcement). Hagopian et al. (1997) reported that using the structured criteria for visual analysis increased the interrater agreement from .54 to .90. The criteria developed by Hagopian et al. (1997) is limited to 10 outcomes per condition. Roane et al. (2013) modified the criteria developed by Hagopian et al. (1997) to allow the criteria to interpret SCED outcomes of varying lengths. Cox and Friedel (2022) used the criteria established by Roane et al. (2013) to create an R script for assisting visual analysis, and they reported 81% agreement between experts and the R script outcomes. The updated R script can further achieve 89% agreement with experts' rating (Friedel et al., 2024). Web-based applications have also been developed to facilitate visual analysis. For instance, VAIOR (visual aid implying an objective rule) developed by Manolov and Vannest (2023) allows decision making regarding the presence of an immediate effect, a delayed effect, or an overall effect for an AB contrast in phase designs, multiple baseline designs, and alternating treatment designs.

Structured visual analysis can be defined as “a tool or criteria used to supplement and structure response-guided or post-hoc visual analysis, aimed at reducing interpretation heterogeneity that includes traditional SCED properties” (Dowdy et al., 2022, p. 454). Based on this definition, the dual-criteria and

conservative dual-criteria methods developed by Fisher et al. (2003), the structured criteria developed by Hagopian et al. (1997), and the modified criteria based on Fiser et al. (2003) or Hagopian et al. (1997) are all considered as structured visual analysis. Although these methods have shown better results on visual analysis, they were infrequently used. In their review of articles published by the *Journal of Applied Behavior Analysis* between 2015 and 2020, Dowdy et al. (2022) reported that only eight (1.47%) articles included a form of structured visual analysis. Moreover, seven of the eight articles were published in 2020 (Dowdy et al., 2022).

In a review article of visual analysis reporting procedures in the functional communication training literature, Wolfe et al. (2024) found that while many authors used the key term “level” explicitly in their reports (88% used specific term and 12% described it), trend (45% used specific term and 25% described it) and variability/stability (66% used specific term and 6% described it) were less likely to be referred to using specific terms. Immediacy or delay to change between adjacent phases was described in 65% of the studies, and the degree of overlap was described in 11% of the studies. Consistency of data patterns was reported in 40% of ABAB design studies, while vertical analysis was reported in 14% of multiple baseline design studies. Based on these findings, Wolfe et al. (2024) recommended that authors be transparent about their conclusions by describing features that justify their conclusions.

ARTIFICIAL INTELLIGENCE TO ANALYZE SCED DATA

According to the U.S. Government Accountability Office (2024), artificial intelligence (AI) refers to “computer systems that can solve problems and perform tasks that have traditionally required human intelligence.” According to this definition, machine learning (ML) and GenAI are two subsets of AI. ML refers to “AI that involves training models on data to make predictions or decisions without being explicitly programmed” (Rashidi et al., 2025, Table 1). GenAI refers to computational techniques that can create content, including text, images, audio, or video, when prompted by a user, and GenAI systems create responses using algorithms that are often trained on open-source information, such as text and images from the internet (U.S. Government Accountability Office, 2023). In the fields of pathology and medicine, Rashidi et al. (2025) indicated that ML methods such as convolution neural networks are often employed for classifying images, GenAIs such as ChatGPT are frequently used to analyze and generate text responses that mimic human language, and ML methods such as logistic regression, support vector machine, neural network-based method are often used to conduct predictive modeling for tasks including predicting disease states and patient outcomes.

Recent publications have demonstrated the use of machine learning in analyzing SCED data (Lanovaz et al., 2020; Lanovaz & Bailey, 2024; Lanovaz & Hranchuk, 2021; Taylor & Lanovaz, 2022). For instance, when experts, conservative dual-criteria method, and machine learning were tasked to analyze A-B comparison data graphs, Lanovaz and Hranchuk (2021) found that the two machine learning methods (i.e., stochastic gradient descent and support vector classifier) outperformed experts and the conservative dual-criteria method in accurately identifying a true effect and showed the best balance between Type I error rate and power. A clinical case application of using machine learning to conduct a pediatric feeding treatment evaluation was reported by Taylor and Lanovaz (2022). Researchers can refer to the tutorials written by Lanovaz and Bailey (2024) and Turgeon and Lanovaz (2020) to apply different machine learning algorithms to analyze SCED data.

Since OpenAI released ChatGPT to the public on November 30, 2022, GenAI has become more and more popular (Maslej et al., 2025; Singh, 2025). Rapid improvements in GenAI’s capabilities have the potential to facilitate problem solving and decision making, but they also require researchers to examine their strengths and weaknesses carefully. The first publicly available ChatGPT was text-only. The new ChatGPT-4o can use multiple modalities, including text, audio, images, and video to answer elaborate questions (OpenAI, 2024). Since November 2023, custom versions of ChatGPT have been available for users to create for a specific purpose (OpenAI, 2023). Because of their custom instructions and ability to keep a knowledge base in addition to what ChatGPT has already been trained on, custom versions of ChatGPT allow users to address a specific need (e.g., analyzing data in SCEDs) that might be hard for ChatGPT to achieve on its own (MIT Sloan Teaching & Learning Technologies, 2025). A report by

Akuoko-Mensah and Chen (2025) provided initial findings on using ChatGPT to analyze SCED graphs. Akuoko-Mensah and Chen (2025) developed a custom version of ChatGPT. They used ABAB and multiple baseline graphs with expert ratings available from the Assessing Visual Analysis of Single Case Research Design website. Their findings revealed that higher agreement between ChatGPT and experts occurred when the knowledge file included the graphs for training, expert ratings, and graph descriptions generated by ChatGPT itself than when it included the graphs for training and expert ratings only (Akuoko-Mensah & Chen, 2025). In addition, agreement between ChatGPT and expert ratings was higher for ABAB graphs (88.8% within ± 1 point) than for multiple baseline graphs (66.7% within ± 1 point).

The emergence of ChatGPT has significantly fueled excitement and investment in the AI sector (Maslej et al., 2025). Alternatives to ChatGPT may perform similarly or even outperform ChatGPT on graph analysis. Sartori et al. (2025) used 990 graph images that fall into seven task categories (detection, classification, segmentation, pattern recognition, link prediction, reasoning, and matching) to evaluate the capabilities of six GenAIs in tackling abstract visual tasks. Their findings showed that Claude-3.5-Sonnet, Gemini-Pro-1.5, and ChatGPT-4o outperformed Llama3.2, Qwen-2-VL, and Pixtral (Sartori et al., 2025). The overall accuracy rate of Claude-3.5-Sonnet, Gemini-Pro-1.5, and ChatGPT-4o ranged between 70% to 80%. Among the three, ChatGPT is particularly popular (Singh, 2025), and the newly released Claude 4 may further improve its earlier graph analysis capabilities.

THE PRESENT STUDY

Although visual analysis is a primary method of evaluating SCED data, to what extent accurate and consistent data-based decisions can be made using this method remains a concern (Dart & Radley, 2024; Fisher et al., 2003; Lanovaz & Hranchuk, 2021; Tarlow et al., 2021; Wolfe & Seaman, 2023; Wolfe et al., 2024). The rapid advancement of multimodal GenAI has the potential to support visual analysis of SCED data. A recent study conducted by Akuoko-Mensah and Chen (2025) revealed that ChatGPT provided ratings more similar to expert ratings on ABAB graphs than multiple baseline graphs. Additional research is needed to further explore the capabilities of other GenAIs on visual analysis of SCED data. In addition, literature has suggested that agreement among visual analysts was higher for graphs displaying no or large effect sizes than for those displaying moderate effect sizes (Pierce et al., 2024; Wolfe & Seaman, 2023). It brings up the question, “Will agreement between expert and GenAI’s ratings also depend on effect size?”

To better understand the potential of GenAIs to analyze SCED data, the present study aimed to extend findings from Akuoko-Mensah and Chen (2025) by investigating (1) the degree of ChatGPT ratings align with expert ratings as well as the degree of Claude ratings align with expert ratings, and if the degree of alignment depends on (2) effect size, and (3) design forms. Three research questions guided the study.

- RQ1. When ChatGPT and Claude are asked to determine the functional relation between the independent and dependent variables using ABAB and multiple baseline graphs, to what extent do their ratings align with expert ratings, as measured by weighted Kappa coefficients?
- RQ2. Does the degree of alignment between GenAI’s ratings and expert ratings depend on the magnitude of effect?
- RQ3. Does the degree of alignment between GenAI’s ratings and expert ratings depend on SCED design forms (i.e., ABAB and multiple baseline design)?

Findings from the present study provide insights into the opportunities and challenges of using GenAI to facilitate the analysis of SCED data, contributing to broader research on AI for graph analysis.

METHOD

Part of the data generated by ChatGPT in Akuoko-Mensah and Chen (2025) were used in the present study. To compare the performance of ChatGPT and Claude on SCED data analysis, the same graphs that were used for training and testing ChatGPT in Akuoko-Mensah and Chen (2025) were used to train and test Claude. The approach of training ChatGPT with graph descriptions was shown to provide ratings more like expert ratings (Akuoko-Mensah & Chen, 2025). Therefore, this approach was adopted to train Claude to perform graph analysis. Training and testing Claude on analyzing SCED graphs were conducted in May

2025, right after the release of Claude Opus 4. Details of the study materials, procedures, and data analysis are provided below.

MATERIALS

Data graphs were obtained from the Assessing Visual Analysis of Single Case Research Designs website (www.singlecase.org). The graphs on the website include expert ratings as part of the Visual Analysis Training Protocol developed by Swoboda et al. (2012). There is a total of 53 ABAB, 47 multiple baseline, and 36 alternating treatment data graphs. Because phase designs and multiple baseline designs were used more frequently than alternating treatment designs (Dayo et al., 2024; Shadish & Sullivan, 2011; Smith, 2012; Tanious & Onghena, 2021) and alternating treatment graphs were not used in Akuoko-Mensah and Chen (2025), we used ABAB and multiple baseline graphs only. All the ABAB graphs include 40 outcome measures, 10 for each phase. All the multiple baseline graphs include 39 outcome measures, and each includes four tiers. Specifically, six baseline and 33 intervention measures for Individual A, 11 baseline and 28 intervention measures for Individual B, 16 baseline and 23 intervention measures for Individual C, 21 baseline and 18 intervention measures for Individual D.

Instructions on the website ask trainees to consider features of level, trend, variability, immediacy of effect and score overlap to determine the degree to which they believe the intervention was “successful.” Trainees are asked to respond to the question “Is there a functional relationship between the treatment and the outcome?”, using a 7-point rating scale. A rating of 1 indicates no functional relation between the independent and dependent variables, a rating of 5 represents a moderate but publishable functional relation, and a rating of 7 indicates a strong functional relation.

The graphs with the same expert ratings were divided into two groups, one for training and the other for testing. Specifically, among the 53 ABAB graphs, six received an expert rating of 1 (3 for training, 3 for testing), eight received an expert rating of 2 (4 for training, 4 for testing), seven received an expert rating of 3 (3 for training, 4 for testing), seven received an expert rating of 4 (3 for training, 4 for testing), 11 received an expert rating of 5 (6 for training, 5 for testing), 10 received an expert rating of 6 (5 for training, 5 for testing), and four received an expert rating of 7 (2 for training, 2 for testing). Hence, a total of 26 ABAB data graphs were used for training, and 27 were used for testing. Among the 47 multiple baseline data graphs, 13 received an expert rating of 1 (6 for training, 7 for testing), three received an expert rating of 2 (1 for training, 2 for testing), four received an expert rating of 4 (2 for training, 2 for testing), four received an expert rating of 4 (2 for training, 2 for testing), 10 received an expert rating of 5 (5 for training, 5 for testing), nine received an expert rating of 6 (5 for training, 4 for testing), and four received an expert rating of 7 (2 for training, 2 for testing). Hence, a total of 23 multiple baseline data graphs were used for training, and 24 were used for testing.

PROCEDURES

The following three steps were taken to train and test Claude. The same steps were used in Akuoko-Mensah and Chen (2025).

Step 1: Generate descriptions for the training graphs. As we mentioned earlier, 26 ABAB graphs and 23 multiple baseline graphs were used for training. Akuoko-Mensah and Chen (2025) generated descriptions for training graphs using ChatGPT-4o. The same prompts were used to generate graph descriptions from Claude Opus 4, except that ChatGPT was replaced with Claude. The prompts used to guide Claude are presented in Figure 3. After the third prompt, the training data graphs and their corresponding expert ratings were provided to Claude one by one for obtaining graph descriptions.

Step 2: Design the custom version of Claude. A Custom Claude project was created to facilitate SCED graph analysis. This custom version of Claude was specifically designed to address the need for systematic analysis of SCED graphs. During this step, both the instruction prompt and knowledge file were developed and uploaded to the project. The instruction prompt guided Claude to play the role of an expert in SCED graph analysis. The instruction prompt included a description of the six data features as criteria for evaluation, explanations of the ratings from 1 to 7, two reference examples (with corresponding expert ratings and graph descriptions from Step 1), and the two tasks assigned (i.e., assign a rating to the graph

and justify rating based on the features of level changes, trend, variability, immediacy of effect, overlap, and consistency across phases). The knowledge file included a brief introduction, expert ratings of training graphs, images of the training graphs, and corresponding graph descriptions generated from Step 1.

Step 3: Test Claude on data analysis. As we mentioned earlier, 27 ABAB graphs and 24 multiple baseline graphs were used for testing. In this step, testing graphs were provided to Claude one by one in the order of experts' ratings from 1, 2, 3, 4, 5, 6, to 7. The prompt "Rate this ABAB graph" or "Rate this multiple baseline graph" was used to request a rating.

First Prompt

You are going to help me to generate descriptions of SCD graphs based on the six criteria for visual analysis (level changes, trend, variability, immediacy of effect, overlap, and consistency across phases) for rating SCD graphs. These descriptions are going to help me to train a Custom Claude to accurately rate any [ABAB or multiple baseline] graph image that is fed to the Claude.

Second Prompt

I will be feeding some few examples to the instruction prompt and the rest with their corresponding graphs and rating will be uploaded as a knowledge file to the Custom Claude. I want the best descriptions to train the Claude model that will yield the best results (ratings) when any [ABAB or multiple baseline] graph is uploaded to it.

Third Prompt

I will be uploading the graphs with their expert ratings one after the other for you to generate the descriptions.

Figure 3. Prompts for Generating Graph Descriptions. In the brackets, either ABAB or multiple baseline was used, depending on the type of graphs in the training phase.

DATA ANALYSIS

Kappa coefficient can be used to measure the level of agreement between experts and GenAI based on the difference between the observed agreement and the expected agreement that would be present by chance alone. Kappa coefficient ranges from -1 to 1, where 1 is a perfect agreement, 0 is an agreement by chance, and a value less than 0 indicates an agreement less than expected by chance. McHugh (2012) classified level of agreement using Kappa coefficient as none (0 – .20), minimal (.21 – .39), weak (.40 – .59), moderate (.60 – .79), strong (.80 – .90), and almost perfect (> .90). For behavioral observation, Hartmann et al. (2004) suggested at least .60 if measured by Kappa coefficient for agreement. While Kappa coefficient can be used for two raters with two categories or for nominal variables with three or more categories, weighted Kappa coefficient is more suitable for ordered categorical variables that consist of three or more categories (Li et al., 2023). Linear weighted Kappa coefficient assigns weights to categories of agreement and disagreement based on the linear distance between the categories on the rating scale. In contrast, quadratic weighted Kappa coefficient assigns weights based on the quadratic distance (Li et al., 2023). To answer RQ1, we computed linear and quadratic weighted Kappa coefficients to measure the level of agreement between (1) expert and ChatGPT ratings using ABAB graphs, (2) expert and Claude ratings using ABAB graphs, (3) expert and ChatGPT ratings using multiple baseline graphs, and (4) expert and Claude ratings using multiple baseline graphs. A value of .60 for both linear and quadratic weighted Kappa coefficients was considered as a good level of agreement between experts and GenAI.

To answer RQ2, the graphs for testing were categorized into 0 = moderate effect (expert ratings of 3, 4, or 5) and 1 = minimum or strong effect (expert ratings of 1, 2, 6, or 7). The level of agreement between expert and GenAI ratings was categorized as 0 = same rating, and 1 = a difference in rating (e.g., -3, -2, -1, 1, 2, 3). Because we examined the performance of ChatGPT and Claude using two design forms, four 2

(same rating) \times 2 (effect magnitude) chi-square tests were performed to compare the frequency of no difference in expert and GenAI ratings with those showing a difference in ratings between moderate and minimum or strong effect data graphs. The tests were performed separately for ChatGPT and Claude and for ABAB and multiple baseline graphs.

Same as RQ2, to answer RQ3, the level of agreement between expert and GenAI ratings was categorized as 0 = same rating, and 1 = a difference in rating (e.g., -3, -2, -1, 1, 2, 3). Two 2 (same rating) \times 2 (design form) chi-square tests were performed to compare the frequency of no difference in expert and GenAI ratings with those having a difference in ratings between ABAB and multiple baseline data graphs. The tests were performed separately for ChatGPT and Claude. A priori power analysis was conducted using G*Power 3.1.9.7 to estimate the minimum sample size required for the chi-square test to achieve a statistical power of .8, with effect size $w = .3$, $\alpha = .05$ and $df = 1$. The result showed that a minimum of 88 was required for the total sample size.

RESULTS

This section presents results for RQ1, RQ2, and RQ3. Table 1 presents a cross-tabulation of ChatGPT and expert ratings on ABAB graphs (upper panel) and multiple baseline graphs (lower panel), and Table 2 presents a cross-tabulation of Claude and expert ratings. At the end, we explain similarities and differences between ChatGPT and Claude on graph descriptions and rating justifications.

EXPERT-GENAI RATING AGREEMENT

According to Table 1, among the 27 ABAB graphs used for testing, 11 or 40.7% (the total of shaded numbers) received the same rating from experts and ChatGPT, 14 or 51.9% received discrepancies of ± 1 rating, and two or 7.4% received discrepancies of ± 2 rating. There was a good level of agreement between expert and ChatGPT ratings on ABAB graphs (linear weighted Kappa = .66, 95% CI [.51, .81]; quadratic weighted Kappa = .86, 95% CI [.77, .95]). Among the 24 multiple baseline graphs used for testing, seven or 29.2% received the same rating from experts and ChatGPT, nine or 37.5% received discrepancies of ± 1 rating, five or 20.8% received discrepancies of ± 2 rating, and three or 12.5% received discrepancies of ± 3 rating. The level of agreement between expert and ChatGPT ratings on multiple baseline data graphs was lower than their agreement on ABAB graphs (linear weighted Kappa = .50, 95% CI [.30, .70]; quadratic weighted Kappa = .72, 95% CI [.54, .89]).

Table 2 presents the contingency table of expert and Claude ratings on ABAB and multiple baseline data graphs. The shaded numbers show the frequency of the same ratings from experts and Claude. Among the 27 ABAB graphs used for testing, 21 or 77.8% received the same rating from experts and Claude, five or 18.5% received discrepancies of ± 1 rating, and one or 3.7% received discrepancies of ± 2 rating. The level of agreement between expert and Claude ratings was substantial on ABAB data graphs (linear weighted Kappa = .87, 95% CI [.77, .97]; quadratic weighted Kappa = .95, 95% CI [.90, 1.00]). Among the 24 multiple baseline graphs used for testing, 14 or 58.3% received the same rating from experts and Claude, seven or 29.2% received discrepancies of ± 1 rating, and three or 12.5% received discrepancies of ± 2 rating. Although agreement between expert and Claude ratings on multiple baseline data graphs was lower than their agreement on ABAB graphs, weighted Kappa coefficients still showed a good level of agreement (linear weighted Kappa = .77, 95% CI [.63, .90]; quadratic weighted Kappa = .90, 95% CI [.83, .98]).

MAGNITUDE OF EFFECT & EXPERT-GENAI RATING AGREEMENT

For ABAB designs, moderate and no and strong effect graphs did not differ significantly with respect to having the same expert and GenAI ratings (chi-square (1) = 1.03, $p = .310$, Cramer's $V = .20$ for ChatGPT; chi-square (1) = 0.678, $p = .410$, Cramer's $V = .16$ for Claude). Likewise, for multiple baseline designs, moderate and no and strong effect data graphs did not differ significantly with respect to having the same expert and GenAI ratings (chi-square (1) = 2.27, $p = .132$, Cramer's $V = .31$ for ChatGPT; chi-square (1) = 1.143, $p = .285$, Cramer's $V = .22$ for Claude).

Table 1. Contingency Table of Expert and ChatGPT Ratings

ABAB		ChatGPT rating							Total
		1	2	3	4	5	6	7	
Expert rating	1	1	2	0	0	0	0	0	3
	2	1	2	1	0	0	0	0	4
	3	0	0	2	1	1	0	0	4
	4	0	0	2	1	1	0	0	4
	5	0	0	0	4	1	0	0	5
	6	0	0	0	1	1	3	0	5
	7	0	0	0	0	0	1	1	2
Total		2	4	5	7	4	4	1	27

MB		ChatGPT rating							Total
		1	2	3	4	5	6	7	
Expert rating	1	3	3	1	0	0	0	0	7
	2	0	0	1	1	0	0	0	2
	3	0	0	0	1	0	1	0	2
	4	0	0	1	0	0	0	1	2
	5	0	0	1	2	1	0	1	5
	6	0	0	1	1	0	2	0	4
	7	0	0	0	0	0	1	1	2
Total		3	3	5	5	1	4	3	24

Note. The upper panel shows the ratings for ABAB data graphs, and the lower panel shows the ratings for multiple baseline (MB) data graphs. The shaded numbers show the frequencies of the same expert and ChatGPT ratings.

Table 2. Contingency Table of Expert and Claude Ratings

ABAB		Claude rating							Total
		1	2	3	4	5	6	7	
Expert rating	1	3	0	0	0	0	0	0	3
	2	0	3	0	1	0	0	0	4
	3	0	0	4	0	0	0	0	4
	4	0	0	0	3	1	0	0	4
	5	0	0	0	1	4	0	0	5
	6	0	0	0	0	1	3	1	5
	7	0	0	0	0	0	1	1	2
Total		3	3	4	5	6	4	2	27

MB		Claude rating							Total
		1	2	3	4	5	6	7	
Expert rating	1	5	2	0	0	0	0	0	7
	2	0	1	0	1	0	0	0	2
	3	0	0	0	1	1	0	0	2
	4	0	1	0	1	0	0	0	2
	5	0	0	0	2	3	0	0	5
	6	0	0	0	0	1	3	0	4
	7	0	0	0	0	0	1	1	2
Total		5	4	5	5	5	4	1	24

Note. The upper panel shows the ratings for ABAB data graphs, and the lower panel shows the ratings for multiple baseline (MB) data graphs. The shaded numbers show the frequency of the same ratings from experts and Claude.

SCED DESIGN FORM & EXPERT-GENAI RATING AGREEMENT

When ChatGPT was used to provide ratings, ABAB and multiple baseline data graphs did not differ significantly with respect to having the same expert and GenAI ratings (chi-square (1) = 0.745, $p = .388$, Cramer’s $V = .12$). Likewise, when Claude was used to provide ratings, ABAB and multiple baseline data graphs did not differ significantly with respect to having the same expert and GenAI ratings (chi-square (1) = 2.231, $p = .135$, Cramer’s $V = .21$).

SIMILARITIES & DIFFERENCES IN GRAPH DESCRIPTIONS & RATING JUSTIFICATIONS

To better understand the similarities and differences between ChatGPT's and Claude's capabilities to analyze SCED graphs, their graph descriptions and justifications for ratings were compared. Because of the prompts we used, ChatGPT and Claude consistently described the training graphs and justified their ratings on testing graphs based on the six data features. The most striking differences between the two GenAIs laid in (1) their level of descriptive precision and quantitative detail, (2) terminology used, and (3) interpretive confidence and causal attribution. Analyses revealed similar patterns for graph descriptions and justifications for ratings. Hence, we focused on graph descriptions below. To delineate the differences, Table 3 presents an example of ABAB training graph descriptions generated by the two GenAIs, and the expert rating of the graph is 1. Table 4 presents an example of multiple baseline training graph descriptions generated by the two GenAIs, and the expert rating of the graph is 6.

Descriptive precision and quantitative detail. As shown in the third column of Tables 3 and 4, Claude provided specific numerical information across the two design forms, offering precise measurements such as "A1 averages ~45, B1 averages ~47, A2 averages ~52, and B2 averages ~40." and "Individual A drops from ~35 to ~22 (37% reduction), B from ~34 to ~21 (38% reduction), C from ~35 to ~22 (37% reduction), and D from ~33 to ~21 (36% reduction)." Claude also consistently specified data ranges and overlap percentages with precision, stating values such as "data points ranging from ~25 to ~75 throughout. Notable spikes (session 9: ~73, session 24: ~72) occur randomly across phases" and "Minimal to no overlap between phases across all tiers (0-5%)". In contrast, ChatGPT employed more general descriptive language across both design forms, using broader characterizations such as "There is no meaningful or sustained difference in level between baseline and treatment phases" or describing "minimal overlap between baseline and intervention phases" without providing specific numerical details.

Claude frequently pointed out specific session numbers, such as "notable spikes (session 9: ~73, session 24: ~72)." By comparison, ChatGPT infrequently referred to sessions. When sessions were mentioned, ChatGPT used "middle sessions," "later sessions," "early sessions." In one of the multiple baseline training graphs, outcome values dropped at the sixth measures after intervention introduction across all four tiers. ChatGPT incorrectly described the feature of immediacy of effect as "Each transition point marks an abrupt and immediate level shift, with no delay or gradual onset. The data sharply descend at the exact session the intervention begins." Although Claude did not correctly identify the session for which a drastically change occurred, Claude was able to identify a delay effect. Claude explained the feature of immediacy of the same graph as "Individual A drops from ~40 to ~15 within 2 sessions, B from ~38 to ~12 within 2 sessions, C shows slight delay then drops from ~36 to ~10 by session 20, and D plummets from ~32 to ~8 within 3 sessions. These rapid, large-magnitude changes strongly support causal attribution."

Terminology used. When analyzing ABAB graphs, Claude explained data features within and between phases using specific terms of A1, A2, B1, and B2 to refer to the four phases, whereas ChatGPT explained data features using terms of the first or second A-B sequence or A-B-A-B sequence. When Claude was tasked to describe multiple baseline graphs, it used terms specifically related to multiple baseline designs, such as "staggered introduction," "temporal control," and "no evidence of co-variation (untreated tiers remain stable while others receive intervention)". In contrast, ChatGPT described multiple baseline graphs with general terms used for SCEDs (e.g., functional relation, experimental control).

Because possible ratings for the graphs range from 1 to 7, it required the two GenAIs to determine the magnitude of effect or effect size. When examining the graph descriptions, it was found that "effect size" was not mentioned in any of the graph descriptions generated by ChatGPT. For Claude, "Effect size" was mentioned in two ABAB graphs and six multiple baseline graphs.

Interpretive confidence and causal attribution. It was a notable difference in how confidently ChatGPT and Claude make interpretive statements and causal attributions. For instance, the expression "flaw," "flawless," or "strongly support" was never mentioned by ChatGPT in describing training graphs. Yet among the ABAB training graphs, Claude Opus 4 used the term "flawless" in describing five graphs. Expert ratings of these graphs were 5, 5, 6, 6, and 7. Specifically, "This flawless replication across both AB cycles, combined with zero overlap and large effect sizes, represents the gold standard for ABAB designs and

justifies the highest rating” was commented on the graph with an expert rating of 7. Among the multiple baseline training graphs, Claude Opus 4 used the term “flawless” in describing six graphs with an expert rating of 5, 6, or 7 and “flaw” in describing two graphs with an expert rating of 1. Specifically, “While all tiers show eventual decreases, the presence of baseline trends and lack of clear phase changes violate multiple baseline logic. The decreasing patterns begin before intervention in most tiers, suggesting maturation, practice effects, or other confounds rather than treatment effects. The failure to demonstrate stable baselines before intervention introduction represents a fundamental design flaw” was commented by Claude on one of the graphs with an expert rating of 1. The corresponding comment provided by ChatGPT read “There is no consistent or replicated pattern of improvement across the four individuals. The weak and inconsistent effects across tiers result in no compelling evidence of a functional relation.”

In short, ChatGPT's descriptions relied more on qualitative assessments rather than precise quantitative measurements, a pattern that remained consistent when analyzing ABAB and multiple baseline graphs. Claude generated more numeric values to describe the graphs in detail and used terminology related to the specific assigned tasks. Claude also showed greater confidence in making definitive interpretations than ChatGPT.

Table 3. *ABAB Graph Descriptions (Expert Rating = 1) Generated by ChatGPT-4o and Claude Opus 4*

Features	ChatGPT-4o	Claude Opus 4
Level	The average level of data varies substantially across and within phases. There is no meaningful or sustained difference in level between baseline and treatment phases.	Mean levels across phases show no meaningful differentiation. A1 averages ~45, B1 averages ~47, A2 averages ~52, and B2 averages ~40. While means vary slightly, the ranges overlap extensively and differences are obscured by high variability, preventing any clear phase distinction.
Variability	Variability is high throughout the entire graph. Data fluctuate widely within phases, with swings exceeding $\pm 25\text{--}30\%$ from the mean. This level of instability obscures any potential treatment effect.	Extremely high variability characterizes all phases, with data points ranging from ~25 to ~75 throughout. Notable spikes (session 9: ~73, session 24: ~72) occur randomly across phases. This excessive variability masks any potential treatment effects and indicates lack of behavioral control.
Trends	There is no clear or consistent trend in any phase. Trends frequently reverse direction, and no systematic shift is observed following treatment introductions.	No consistent trends emerge within or between phases. Each phase shows erratic up-and-down patterns without systematic direction. The chaotic nature of the data prevents identification of any treatment-related trend changes.
Overlap	There is extensive overlap of data across all phases. Data points in treatment phases fall within the full range of baseline values, with no visual separation between conditions	Complete overlap (100%) exists between all phases. Every phase contains data points spanning from approximately 25 to 75, making visual discrimination between baseline and treatment conditions impossible.
Immediacy	No immediate change in level or slope occurs after the onset of treatment. Phase transitions are followed by continued irregular fluctuation rather than abrupt shifts.	No immediate effects occur at phase transitions. Phase changes are marked only by vertical lines, not by any observable behavioral shifts. Data points continue their erratic pattern regardless of phase changes.
Consistency	No consistent pattern is replicated across the A-B-A-B sequence. The data appear random and highly unstable, with treatment phases showing no clear improvement or effect.	Neither baseline nor treatment phases replicate their patterns. A1 and A2 show different variability patterns and levels, as do B1 and B2. The inconsistent and unpredictable nature of the data across all phases demonstrates absence of experimental control.

Note. Words mentioned in the main text that explain the differences in graph descriptions between ChatGPT-4o and Claude Opus 4 are bold.

Table 4. Multiple Baseline Graph Descriptions (Expert Rating = 6) Generated by ChatGPT-4o and Claude Opus 4

Features	ChatGPT-4o	Claude Opus 4
Level	There is a clear and substantial reduction in level across all individuals, with average post-intervention values approximately 10–15 percentage points lower than baseline values.	Consistent moderate to strong level changes across all tiers. Individual A drops from ~35 to ~22 (37% reduction), B from ~34 to ~21 (38% reduction), C from ~35 to ~22 (37% reduction), and D from ~33 to ~21 (36% reduction). The remarkably uniform effect sizes demonstrate reliable treatment impact.
Variability	Baseline phases display low to moderate variability, while intervention phases exhibit consistently low variability across all four individuals, suggesting stable responding post-treatment.	Very low variability throughout enhances interpretation. Baselines show tight clustering (ranges 30-40), and intervention phases maintain equally low variability (ranges 18-26). This exceptional stability creates clear visual separation and maximizes confidence in effects.
Trends	Baseline trends are relatively flat or slightly increasing, whereas intervention trends are flat or slightly decreasing, showing no regression toward baseline levels.	All phases display remarkably stable, flat patterns. Baselines show no directional trends, and intervention phases maintain stable levels at consistently lower values. The absence of any trends eliminates confounds and highlights pure level-based effects.
Overlap	There is minimal overlap between baseline and intervention phases. Most intervention data points fall well below the baseline range, indicating a non-random change.	Minimal to no overlap between phases across all tiers (0-5%). Clear separation exists with 8-15 point gaps between the lowest baseline values and highest intervention values. This minimal overlap provides strong visual evidence despite moderate effect sizes.
Immediacy	Each participant shows a noticeable and immediate drop in level within the first 1–2 sessions following the introduction of the intervention.	Perfect immediate effects within 1 data point across all tiers. Individual A drops from ~33 to ~21, B from ~32 to ~22, C from ~32 to ~24, and D from ~34 to ~21. These instantaneous, uniform changes strongly support causal attribution.
Consistency	The pattern of change is consistently replicated across all four individuals. Each tier demonstrates a systematic decrease in behavior following intervention, providing strong evidence for a functional relation.	Exceptional consistency in both effect magnitude and pattern. All four tiers demonstrate nearly identical 36-38% reductions with identical response patterns - immediate drop to stable lower level. The staggered introduction (sessions 6, 11, 17, and 22) is perfectly executed with flawless temporal control. This remarkable uniformity of moderate effects, combined with minimal variability, perfect immediacy, and zero overlap, demonstrates exceptional experimental precision. The highest rating is justified by the flawless execution and replication rather than effect magnitude alone.

Note. Words mentioned in the main text that explain the differences in graph descriptions between ChatGPT-4o and Claude Opus 4 are bold.

DISCUSSION

Single-case experimental designs (SCEDs) are increasingly used in technology-enhanced learning and other education-related fields (Dayo et al., 2024). A core feature of SCED research is graphic display of data (Wolfe et al., 2024), and visual analysis of graphs is the primary method used for data analysis (Horner et al., 2005; Kazdin, 2019; Ledford et al., 2018). However, literature has shown mixed findings regarding agreement among visual analysts and inaccurate conclusions drawn from visual analysis (Dart & Radley, 2024; Fisher et al., 2003; Lanovaz & Hranchuk, 2021; Tarlow et al., 2021; Wolfe & Seaman, 2023; Wolfe et al., 2024). Depending on factors such as trend (Wolfe & Seaman, 2023) and magnitude of effect (Pierce et al., 2024; Wolfe & Seaman, 2023), the level of agreement among visual analysts varied. Moreover, although tools and criteria have been developed to assist visual analysis, they were used infrequently by SCED researchers (Dowdy et al., 2022). It is thus critical to explore new methods and strategies that not

only are able to make accurate conclusions but also are easy to use to facilitate reliable and accurate visual analysis.

The rapid evolution of artificial intelligence (AI) has generated growing excitement and discussion about its potential to improve various aspects of our lives, including applying AI in SCED research (Neely et al., 2023). Machine learning (ML) and generative AI (GenAI) are two subsets of AI. Prior studies have shown promising results on using ML to analyze SCED data (Lanovaz et al., 2020; Lanovaz & Hrančuk, 2021). A recent report by Akuoko-Mensah and Chen (2025) demonstrated the potential of using GenAI to support data analysis in SCED research, especially when graph descriptions of the training graphs were provided in the knowledge file to analyze new ABAB graphs. However, only ChatGPT was examined in Akuoko-Mensah and Chen (2025). To better understand GenAI's capability to interpret and analyze SCED graphs, the present study extended findings from Akuoko-Mensah and Chen (2025) and it aimed to investigate (1) alignment of ChatGPT ratings with expert ratings and alignment of Claude ratings with expert ratings, and if degree of alignment would depend on (2) magnitude of effect and (3) design forms (ABAB design versus multiple baseline design).

To achieve these aims, three research questions were constructed. RQ1 asked "To what extent do GenAI ratings align with expert ratings as measured by weighted Kappa coefficients?" Findings and implications related to RQ1 are discussed under the heading *ALIGNMENT OF GENAI RATINGS WITH EXPERT RATINGS*. RQ2 asked "Does degree of alignment depend on the magnitude of effect?" and RQ3 asked "Does degree of alignment depend on design forms?" Results and implications related to RQ2 and RQ3 are summarized under the section titled *IMPACT OF EFFECT SIZE AND DESIGN FORM ON RATING ALIGNMENT*. The comparisons between ChatGPT and Claude on graph descriptions and rating justifications and their implications are explained under the heading *GRAPH DESCRIPTIONS, RATING JUSTIFICATIONS AND THEIR IMPLICATIONS*. The section titled *PRACTICAL CONSIDERATIONS* presents possible challenges when using GenAI to analyze SCED data. Limitations and future research directions are presented in the section *STUDY LIMITATIONS AND FUTURE DIRECTIONS*. In the last section, the conclusion is discussed.

ALIGNMENT OF GENAI RATINGS WITH EXPERT RATINGS

Using at least .60 for both linear and quadratic weighted Kappa as a cutoff for acceptable agreement, agreement between ChatGPT and expert ratings on ABAB graphs was acceptable, but not on multiple baseline graphs. Agreement between Claude and expert ratings was acceptable for both ABAB and multiple baseline graphs. Moreover, the two GenAI's ratings on ABAB graphs aligned more closely with expert ratings than on multiple baseline graphs. In fact, analysis of ABAB graphs requires evaluation of data patterns in the same and different conditions horizontally. In contrast, analysis of multiple baseline graphs requires similar horizontal analysis of the baseline and intervention phases, plus the vertical analysis of comparisons across tiers. Wolfe et al. (2024) defined vertical analysis in multiple baseline designs as "whether levels of the dependent variable remained stable when the independent variable was introduced to a prior tier" (p. 551). Findings from the present study suggested that guidelines for vertical analysis of multiple baseline data can be provided to GenAI to improve analysis results. In addition, it may be more challenging to conduct vertical analysis for ChatGPT than for Claude.

To confirm difficulties in performing vertical analysis of graphs in multiple baseline designs, the graph descriptions generated by ChatGPT-4o and Claude Opus 4 for one multiple baseline training graph (expert rating = 1) were examined. On this graph, immediacy effect was evident for Individual A. For Individual A, the introduction of the intervention at Session 7 resulted in a drop of about 20 points in the outcome value. Individual B maintained the same level throughout the baseline and intervention phases. Individuals C and D both had their outcome values dropped more than 20 points at Session 7, several sessions before their intervention start points. When tasked to describe the graph, Claude Opus 4 explained, "Individual C shows a precipitous drop during baseline (from 38 to 12) before intervention. Individual D similarly drops during baseline (from 40 to 10). These pre-intervention changes violate the stable baseline requirement and suggest extraneous variables rather than treatment effects" for the data feature of trend. These descriptions implied a good understanding of vertical analysis. In contrast, ChatGPT interpreted the data pattern

incorrectly as “An immediate drop is visible in Individuals C and D at the point of intervention, but not in A or B, weakening the evidence for a causal link across the dataset.” ChatGPT incorrectly explained the introduction of intervention for Individual A as the start point of intervention for Individuals B and C.

In one of the multiple baseline testing graphs (expert rating = 1), all four individuals had outcomes drop from above 30 to below 20 at the introduction of intervention for Individual A at Session 7. ChatGPT gave a 2, and Claude gave a 1 on rating the functional relation between the independent and dependent variables. Examining justifications provided by ChatGPT and Claude revealed that they both detected changes that occurred before the implementation of intervention for Individuals B, C and D. However, Claude provided a more comprehensive explanation than ChatGPT on the data feature of immediacy. Justifications for the feature of immediacy by Claude read “immediacy cannot be properly assessed due to timing violations” and Claude commented that the changes occurred to all four tiers at Session 6-8, before intervention introducing to Individuals B (Session 11), C (Session 17) and D (Session 22). For changes that occurred across four tiers at the start point of intervention for Individual A, Claude further explained, “Only Individual A’s change coincides with intervention introduction, while the other three tiers change prematurely.” In contrast, ChatGPT described the justification for immediacy as “the transitions appear abrupt, but because of early decreasing trends, the immediacy may be misleading or coincidental rather than a clear intervention effect.”

The analyses above demonstrated that Claude outperformed ChatGPT in conducting and interpreting vertical analysis of multiple baseline data. The analysis results also explained why ChatGPT’s ratings on multiple baseline data were not closely aligned with expert ratings, while Claude’s ratings on multiple baseline data remained at a good level of agreement with expert ratings.

Although both ChatGPT and Claude ratings on ABAB graphs reached a good level of alignment with expert ratings, Claude ratings were more aligned with expert ratings across ABAB and multiple baseline graphs than ChatGPT. These findings were consistent with findings from Sartori et al. (2025). Sartori et al. (2025) reported that Claude-3.0-Sonnet and GPT-4o performed much better than the other three GenAIs on analyzing seven types of visual tasks. However, Claude-3.0-Sonnet showed higher average performance than GPT-4o.

IMPACT OF EFFECT SIZE AND DESIGN FORM ON RATING ALIGNMENT

Although the level of agreement among visual analysts has been shown to depend on the magnitude of effect (Pierce et al., 2024; Wolfe & Seaman, 2023), findings from the present study did not reveal a significant difference in having the same ratings from GenAI (i.e., ChatGPT and Claude) and experts between a moderate effect and a minimum or strong effect. While Akuoko-Mensah and Chen (2025) suggested that alignment of ChatGPT ratings with expert ratings may depend on design forms, our findings did not show a significant difference in having the same GenAI and expert ratings between ABAB and multiple baseline graphs.

It should be noted that these non-significant results may be due to low statistical power. The priori power analysis revealed that a minimum total sample size of 88 is required for a statistical power of .8. The numbers of testing ABAB and multiple baseline graphs were 27 and 24, respectively. Even for the chi-square test to examine if having the same rating is independent of design forms, the number of 51 was far less than 88. A larger number of training SCED data graphs would have allowed for greater statistical power.

GRAPH DESCRIPTIONS, RATING JUSTIFICATIONS AND THEIR IMPLICATIONS.

ChatGPT and Claude were prompted to generate descriptions for training graphs based on features of level, trend, variability, immediacy, overlap and consistency of data in similar phases. ChatGPT and Claude were also prompted to evaluate testing graphs according to the same six features. This strategy was used to facilitate a comprehensive and systematic evaluation of SCED graphs based on these six features. Although descriptions of data features may be inaccurate, the two GenAIs consistently provided interpretations of the six features for both training and testing graphs. When reporting findings in SCED research, most researchers explained level explicitly in their report, but the other five data features were less likely to be

referred to using specific terms (Wolfe et al., 2024). GenAI can be used to encourage researchers to examine and report the six features collectively for determining an intervention effect.

Compared to the graph descriptions and rating justifications generated by ChatGPT to those generated by Claude, we concluded that Claude outperformed ChatGPT in terms of (1) descriptive precision and quantitative detail, (2) frequent use of terminology related to specific SCED design forms, and (3) confidence in graph interpretation and causal attribution. For both ABAB designs and multiple baseline designs, Claude provided more specific numeric data to interpret the graphs, whereas ChatGPT employed more general descriptive language. Although details of the logic behind ABAB designs and multiple baseline designs were not provided to the two GenAIs in either the training or testing phase, Claude was able to interpret the graphs using technical language, such as A1, A2, B1, and B2 phases in ABAB designs and staggered introduction in multiple baseline designs. These findings suggest that Claude is preferred over ChatGPT for training and assisting researchers and practitioners to explicitly report their identification of effects by describing the data features that led to their conclusions.

PRACTICAL CONSIDERATIONS

ChatGPT Plus and Claude Pro subscriptions were required for this study, each costing \$20 monthly. This subscription requirement may limit accessibility for researchers with constrained budgets or those seeking to conduct large-scale analyses. Significant differences emerged between the platforms' usage constraints during data analysis. Claude Opus 4 imposed stricter message limits, allowing a maximum of eight graph descriptions or rating justifications per session before displaying a "maximum limit reached" notification. Once this limit was reached, a mandatory 4-hour waiting period was required before resuming work. In contrast, ChatGPT 4o demonstrated more generous usage limits during the study period, with no comparable restrictions encountered. These limitations substantially impacted workflow efficiency. The frequent interruptions with Claude Opus 4 extended the time required to complete analytical tasks, making ChatGPT 4o considerably faster for processing large volumes of SCED data. Researchers planning similar analyses should factor these constraints into their timeline and methodology, particularly when working with extensive datasets that require multiple AI interactions.

GenAI systems like ChatGPT and Claude operate based on probabilistic language generation rather than deterministic rules, which may result in inconsistent interpretations, especially in cases requiring nuanced vertical analysis of SCED data. The same prompt applied to identical graphs could yield variations in descriptions or quality ratings across different sessions. This non-deterministic nature poses challenges for research reproducibility and requires careful documentation of model versions, timestamps, and multiple validation approaches.

STUDY LIMITATIONS AND FUTURE DIRECTIONS

The present study investigated the capabilities of ChatGPT and Claude to assist with data analysis in SCED research, using data graphs from the Assessing Visual Analysis of Single Case Research Designs website (www.singlecase.org). The expert rating of each graph available on the website served as a reference for assessing the performance of ChatGPT and Claude in graph analysis. Although findings from the present study contribute to a better understanding of the application of GenAI in conducting SCED data analysis, additional research is needed before GenAI can be fully applied to support analysis. Below, we suggest several directions for future research in light of the limitations of the present study.

First, rating options used in the present study ranged from 1 (no functional relation) to 7 (strong functional relation). These options align with the purposes for visual analysis based on What Works Clearinghouse (2017), namely determining the existence of a functional relation between an independent variable and an outcome variable and the strength of that relation. Nevertheless, some existing research evaluating the performance of a new method or strategy for analyzing SCED data used a binary decision (e.g., Fisher et al., 2003; Lanovaz & Hranchuk, 2021; Manolov & Vannest, 2023). To compare GenAI's performance with prior findings, future research can examine its ability to make accurate binary decisions.

Moreover, future research can investigate the capabilities of GenAI using simulated data. Simulated data allow researchers to systematically examine factors (e.g., number of measurements, trend) that may affect ratings and interpretations generated from GenAI. Second, the results from ChatGPT were obtained in January 2025, and the results from Claude were obtained in May 2025. ChatGPT's and Claude's capabilities to generate accurate graph descriptions and analyze SCED graphs may improve as they evolve. Future research can continue examining the use of ChatGPT and Claude to facilitate the analysis of SCED data. Future research may also examine the capabilities of other GenAIs in assisting visual analysis, such as DeepSeek and Gemini. Third, the findings of the present study were influenced by the information provided to GenAI and the degree of human involvement. The logic behind ABAB and multiple baseline designs was not provided to the two GenAIs. No interaction was made between the researchers and GenAI to communicate the quality of graph descriptions generated by the GenAIs either. Future research may include interaction between the research team and GenAI to optimize graph descriptions provided in the knowledge file before testing GenAI's capabilities to analyze SCED graphs. Lastly, a small number of ABAB and multiple baseline graphs were used to train and test ChatGPT and Claude for visual analysis. A large number of training and testing graphs with systematic manipulation of potential factors that may affect GenAI's ratings will provide a more comprehensive understanding of GenAI's capabilities for conducting SCED graph analysis.

CONCLUSION

In sum, the present study found that Claude outperformed ChatGPT in analyzing SCED graphs. Yet the stricter message limits with Claude required extended time allocations to complete tasks. Due to greater complexity in analyzing multiple baseline data graphs, ratings on multiple baseline data graphs were less aligned to expert ratings than ratings on ABAB data graphs for both Claude and ChatGPT. The findings provide insights into the opportunities and challenges of using generative AI as a support tool for analyzing SCED data, contributing to broader research on AI for graph analysis. These preliminary findings should be further explored in future research, particularly with systematic manipulation of data characteristics using a large number of SCED graphs.

REFERENCES

- Akuoko-Mensah, I. & Chen, L.-T. (2025). ChatGPT as a support tool for visual analysis in single-case design studies. In R. Jake Cohen (Ed.), *Proceedings of the Society for Information Technology & Teacher Education International Conference* (pp. 3403–3412). Association for the Advancement of Computing in Education (AACE).
- Barker, J. B., Mellalieu, S. D., McCarthy, P. J., Jones, M. V., & Moran, A. (2012). A review of single-case research in sport psychology 1997–2012: Research trends and future directions. *Journal of Applied Sport Psychology*, 25(1), 4–32. <https://doi.org/10.1080/10413200.2012.709579>
- Belcic, I., & Stryker, C. (September 24, 2024). What is Claude AI? <https://www.ibm.com/think/topics/claude-ai>
- Binns, A. V., Smyth, R., Andres, A., Lam, J., & Oram Cardy, J. (2021). Looking back and moving forward: A scoping review of research on preschool autism interventions in the field of speech-language pathology. *Autism & Developmental Language Impairments*, 6. <https://doi.org/10.1177/23969415211033171>
- Brown, C. L., Peltier, C., Smarinsky, E. C., Lewis, O. J., Lee, D. Y. (2024). A review of counseling research using single-case research design. *Journal of Counseling & Development*, 102(3), 334–349. <https://doi.org/10.1002/jcad.12517>
- Cox, A., & Friedel, J. E. (2022). Toward an automation of functional analysis interpretation: A proof of concept. *Behavior Modification*, 46(1), 147–177. <https://doi.org/10.1177/0145445520969188>

- Dart, E. H., & Radley, K. C. (2024). Single-case design in school psychology: Issues of design, analysis, and training. *School Psychology, 39*(6), 545–547. <https://doi.org/10.1037/spq0000682>
- Dayo, N., Metwaly, S. S., & Van Den Noortgate, W. (2024). Evaluating technology enhanced learning by using single-case experimental design: A systematic review. *British Journal of Educational Technology, 55*(6), 2457–2477. <https://doi.org/10.1111/bjet.13468>
- Dowdy, A., Jessel, J., Saini, V., & Peltier, C. (2022). Structured visual analysis of single-case experimental design data: Developments and technological advancements. *Journal of Applied Behavior Analysis, 55*(2), 451–462. <https://doi.org/10.1002/jaba.899>
- Enders, O., King, S. A., Nysten, B., Wang, L., & Opeoluwa, O. (2024). A review of measurement characteristics within single-case designs in special education. *Behavioral Disorders, 50*(1), 46–57. <https://doi.org/10.1177/01987429241249141>
- 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>
- Friedel, J. E., Cox, A. D., & Davis, S. (2023). Further progress toward automating functional analysis interpretation. *Behavior Modification, 48*(1), 75–106. <https://doi.org/10.1177/01454455231195825>
- Guinness, K. E., & Feil, E. G. (2025). A digital intervention package to teach rapport-building skills to caregivers of children with autism. *Journal of Applied Behavior Analysis, 58*(2), 303–318. <https://doi.org/10.1002/jaba.70004>
- Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis, 30*(2), 313–326. <https://doi.org/10.1901/jaba.1997.30-313>
- Hartmann, D. P., Barrios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In S. N. Haynes & E. M. Hieby (Eds.), *Comprehensive handbook of psychological assessment. Vol. 3: Behavioral assessment* (pp. 108–127). John Wiley & Sons.
- Hersen, M. (1990). Single-case experimental designs. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.) *International handbook of behavior modification and therapy* (pp. 175–210). Springer. https://doi.org/10.1007/978-1-4613-0523-1_9
- 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*(2), 165–179. <https://doi.org/10.1177/001440290507100203>
- Institute of Education Sciences (n.d.). *What is the What Works Clearinghouse (WWC)?* <https://ies.ed.gov/ncee/wwc/FAQ>
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press.
- Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy, 117*, 3–17. <https://doi.org/10.1016/j.brat.2018.11.015>
- Kazdin, A. E. (2021). Single-case experimental designs: Characteristics, changes, and challenges. *Journal of the Experimental Analysis of Behavior, 115*(1), 56–85. <https://doi.org/10.1002/jeab.638>
- Krasny-Pacini, A., & Evans, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicine, 61*(3), 164–179. <https://doi.org/10.1016/j.rehab.2017.12.002>
- Kennedy, C. H. (2005). *Single-case designs for educational research*. Person Education, Inc.
- Lane, J. D., & Gast, D. L. (2013). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*(3–4), 445–463. <https://doi.org/10.1080/09602011.2013.815636>
- Lanovaz, M. J., & Bailey, J. D. (2024). Tutorial: Artificial neural networks to analyze single-case experimental designs. *Psychological Methods, 29*(1), 202–218. <https://doi.org/10.1037/met0000487>
- Lanovaz, M. J., Giannakakos, A. R., & Destras, O. (2020). Machine learning to analyze single-case data: A proof of concept. *Perspectives on Behavior Science, 43*, 21–38. <https://doi.org/10.1007/s40614-020-00244-0>

- Lanovaz, M. J., & Hranchuk, K. (2021). Machine learning to analyze single-case graphs: A comparison to visual inspection. *Journal of Applied Behavior Analysis*, 54(4), 1541–1552. <https://doi.org/10.1002/jaba.863>
- Ledford, J. R., Barton, E. E., Severini, K. E., & Zimmerman, K. N. (2019). A primer on single-case research designs: Contemporary use and analysis. *American Journal on Intellectual and Developmental Disabilities*, 124(1), 35–56. <https://doi.org/10.1352/1944-7558-124.1.35>
- Ledford, J. R., Lane, J. D., Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment*, 19(1), 4–17. <https://doi.org/10.1017/BrImp.2017.16>
- Li, M., Gao, Q. & Yu, T. (2023). Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer*, 23, Article 799. <https://doi.org/10.1186/s12885-023-11325-z>
- Manolov, R., & Vannest, K. J. (2023). A visual aid and objective rule encompassing the data features of visual analysis. *Behavior Modification*, 47(6), 1345–1376. <https://doi.org/10.1177/0145445519854323>
- Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., . . . Oak, S. (2025). *The AI index 2025 annual report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf
- McHugh, M. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276–282. <https://doi.org/10.11613/BM.2012.031>
- Michiels, B., & Onghena, P. (2019). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods*, 51, 2454–2476. <https://doi.org/10.3758/s13428-018-1084-x>
- MIT Sloan Teaching & Learning Technologies. (2025). Custom GPTs at MIT Sloan: A comprehensive guide. <https://mitsloanedtech.mit.edu/ai/tools/writing/custom-gpts-at-mit-sloan-a-comprehensive-guide>
- Neely, L., Carnett, A., Quarles, J., MacNaul, H., Park, S.-W., Oyama, S., Chen, G., Desai, K., & Najafirad, P. (2023). The case for integrated advanced technology in applied behavior analysis. *Advances in Neurodevelopmental Disorders*, 7, 415–425. <https://doi.org/10.1007/s41252-022-00309-y>
- Ninci, J. (2023). Single-case data analysis: A practitioner guide for accurate and reliable decisions. *Behavior Modification*, 47(6), 1455–1481. <https://doi.org/10.1177/0145445519867054>
- Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, 71(2), 137–148. <https://doi.org/10.1177/001440290507100201>
- OpenAI. (2023, November 6). Introducing GPTs. <https://openai.com/index/introducing-gpts/>
- OpenAI. (2024, May 13). Introducing GPT-4o and more tools to ChatGPT free users. <https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/>
- Pierce, A. M., Sanetti, L. M. H., Collier-Meek, M. A., & Johnson, A. H. (2024). Evaluating the correspondence between expert visual analysis and quantitative methods. *School Psychology*, 39(6), 557–571. <https://doi.org/10.1037/spq0000644>
- Radley, K. C., Dart, E. H., Fischer, A. J., Collins, T. A. (2020). Publication trends for single-case methodology in school psychology: A systematic review. *Psychology in the Schools*, 57(5), 683–698. <https://doi.org/10.1002/pits.22359>
- Rashidi, H. H., Pantanowitz, J., Hanna, M. G., Tafti, A. P., Sanghani, P., Buchinsky, A., Fennell, B., Deebajah, M., Wheeler, S., Pearce, T., Abukhiran, I., Robertson, S., Palmer, O., Gur, M., Tran, N. K., Pantanowitz, L. (2025). Introduction to artificial intelligence and machine learning in pathology and medicine: Generative and nongenerative artificial intelligence basics. *Modern Pathology*, 38(4), Article 100688. <https://doi.org/10.1016/j.modpat.2024.100688>
- Roane, H. S., Fisher, W. W., Kelley, M. E., Mevers, J. L., & Bouxsein, K. J. (2013). Using modified visual-inspection criteria to interpret functional analysis outcomes. *Journal of Applied Behavior Analysis*, 46(1), 130–146. <https://doi.org/10.1002/jaba.13>

- Sartori, C. C., Blum, C., & Bistaffa, F. (2025). VisGraphVar: A benchmark generator for assessing variability in graph analysis using large vision-language models. *IEEE Access*, *13*, 21788–21810. <https://doi.org/10.1109/access.2025.3535837>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Cengage.
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980. <http://doi.org/10.3758/s13428-011-0111-y>
- Shepley, C., Shepley, S. B. & Spriggs, A. D. (2024). On the history of single-case methodology: A data-based analysis. *Journal of Behavioral Education*, *33*, 163–183. <https://doi.org/10.1007/s10864-022-09477-2>
- Singh, S. (2025, June 5). *ChatGPT statistics 2025 – DAU&MAU data (Worldwide)*. <https://www.demandsage.com/chatgpt-statistics/>
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–550. <http://doi.org/10.1037/a0029312>
- Soto, P. L. (2020). Single-case experimental designs for behavioral neuroscience. *Journal of the Experimental Analysis of Behavior*, *114*(3), 447–467. <https://doi.org/10.1002/jeab.633>
- Steeger, C. M., Buckley, P. R., Pampel, F. C., Gust, C. J., & Hill, K. G. (2021). Common methodological problems in randomized controlled trials of preventive interventions. *Prevention Science*, *22*(8), 1159–1172. <https://doi.org/10.1007/s11121-021-01263-2>
- Swoboda, C., Kratochwill, T., Horner, R., Levin, J., & Albin, R., (2012). Visual Analysis Training Protocol: Applications with the Alternating Treatment, Multiple Baseline, and ABAB Designs.
- Tanious, R., & Onghena, P. (2021). A systematic review of applied single-case research published between 2016 and 2018: Study designs, randomization, data aspects, and data analysis. *Behavior Research Methods*, *53*(4), 1371–1384. <https://doi.org/10.3758/s13428-020-01502-4>
- Tanious, R., & Onghena, P. (2022). Applied hybrid single-case experiments published between 2016 and 2020: A systematic review. *Methodological Innovations*, *15*(1), 73–85. <https://doi.org/10.1177/20597991221077910>
- Tarlow, K. R., Brossart, D. F., McCammon, A. M., Giovanetti, A. J., Belle, M. C., & Philip, J. (2021). Reliable visual analysis of single-case data: A comparison of rating, ranking, and pairwise methods. *Cogent Psychology*, *8*(1), Article 1911076. <https://doi.org/10.1080/23311908.2021.1911076>
- Taylor, T., & Lanovaz, M. J. (2021). Machine learning to support visual inspection of data: A clinical application. *Behavior Modification*, *46*(5), 1109–1136. <https://doi.org/10.1177/01454455211038208>
- Turgeon, S., & Lanovaz, M. J. (2020). Tutorial: Applying machine learning in behavioral research. *Perspectives on Behavior Science*, *43*(4), 697–723. <https://doi.org/10.1007/s40614-020-00270-y>
- U.S. Government Accountability Office. (2023). Science & Tech Spotlight: Generative AI. (GAO Publication No. 23-106782). <https://www.gao.gov/assets/gao-23-106782.pdf>
- U.S. Government Accountability Office. (2024). Artificial Intelligence: GAO's Work to Leverage Technology and Ensure Responsible Use (GAO Publication No. 24-107237). <https://www.gao.gov/products/gao-24-107237>
- Vlaeyen, J. W. S., Wicksell, R. K., Simons, L. E., Gentili, C., De, T. K., Tate, R. L., Vohra, S., Punja, S., Linton, S. J., Sniehotta, F. F., & Onghena, P. (2020). From boulder to Stockholm in 70 years: Single case experimental designs in clinical research. *The Psychological Record*, *70*(4), 659–670. <https://doi.org/10.1007/s40732-020-00402-5>
- What Works Clearinghouse. (2017). *What Works Clearinghouse standards handbook, Version 4.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf
- What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, Version 5.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

- Wolfe, K., McCammon, M. N., LeJeune, L. M., Check, A. R., & Slocum, T. A. (2024). A review of visual analysis reporting procedures in the functional communication training literature. *School Psychology, 39*(6), 548–556. <https://doi.org/10.1037/spq0000660>
- Wolfe, K., & Seaman, M. A. (2023). The influence of data characteristics on interrater agreement among visual analysts. *Journal of Applied Behavior Analysis, 56*(2), 365–376. <https://doi.org/10.1002/jaba.980>