

# Violin Plots as Visual Tools in the Meta-Analysis of Single-Case Experimental Designs

René Tanious<sup>1</sup> , Rumen Manolov<sup>2</sup> 

[1] Faculty of Psychology and Educational Sciences, Methodology of Educational Sciences Research Group, KU Leuven, Leuven, Belgium. [2] Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Barcelona, Spain.

---

Methodology, 2022, Vol. 18(3), 221–238, <https://doi.org/10.5964/meth.9209>

Received: 2022-03-25 • Accepted: 2022-09-12 • Published (VoR): 2022-09-30

Handling Editor: Katrijn van Deun, Tilburg University, Tilburg, The Netherlands

Corresponding Author: René Tanious, Faculty of Psychology and Educational Sciences, KU Leuven, Andreas Vesaliusstraat 2, Box 3762, 3000 Leuven, Belgium. Phone: +32 16 32 82 19. E-mail: [rene.tanious@kuleuven.be](mailto:rene.tanious@kuleuven.be)

---

## Abstract

Despite the existence of sophisticated statistical methods, systematic reviews regularly indicate that single-case experimental designs (SCEDs) are predominantly analyzed through visual tools. For the quantitative aggregation of results, different meta-analytical techniques are available, but specific visual tools for the meta-analysis of SCEDs are lacking. The present article therefore describes the use of violin plots as visual tools to represent the raw data. We first describe the underlying rationale of violin plots and their main characteristics. We then show how the violin plots can complement the statistics obtained in a quantitative meta-analysis. The main advantages of violin plots as visual tools in meta-analysis are (a) that they preserve information about the raw data from each study, (b) that they have the ability to visually represent data from different designs in one graph, and (c) that they enable the comparison of score distributions from different experimental phases from different studies.

## Keywords

single-case experimental designs, meta-analysis, violin plots, visual analysis

In single-case experimental designs (SCEDs) a single entity (e.g., a classroom) is measured repeatedly under different conditions of one or several independent variables (e.g., token economy) (Barlow et al., 2009). Parametric significance tests such as *F* or *t* are usually not applicable for the analysis of SCED data (Toothaker et al., 1983), because data obtained via SCEDs do not meet the assumptions of these tests (e.g., normally



distributed data). Systematic reviews regularly indicate that visual analysis is by far the most preferred data analysis technique among applied SCED researchers (Smith, 2012; Tanious & Onghena, 2021, 2022) despite the availability of sophisticated statistical techniques (Busse et al., 2015; Manolov & Moeyaert, 2017). When aggregating results within or across SCEDs, a meta-analysis can be performed. Meta-analyses are relevant for identifying evidence-based practices (Schlosser & Sigafos, 2008), and are also useful in the context of the current emphasis on replication (Hedges, 2019). The application of meta-analyses for SCEDs has been rising steadily over the past 30 years (Jamshidi et al., 2018), with an exponential increase over the past decade (Becraft et al., 2020). However, in spite of the strong preference of applied researchers for visual analysis at the individual study level, visual tools for meta-analytical purposes in the SCED context have not yet received sufficient attention. Consequently, the present article presents a visual analytical tool for this purpose: the violin plot.

## Framing of the Current Proposal

### Top-Down Versus Bottom-Up Meta-Analysis of SCEDs

When meta-analyzing SCEDs, one can roughly distinguish between top-down and bottom-up models for effect size estimation (Parker & Vannest, 2012). Examples of models for top-down meta-analysis include hierarchical linear or multilevel modelling (Van den Noortgate & Onghena, 2003, 2008) and the between-case standardized mean difference (Shadish et al., 2014). Typically, these models yield an overall effect size that expresses the average treatment effect and can potentially give information about the generality of the effect (Van den Noortgate & Onghena, 2008).

However, these techniques reveal relatively little about the large amount of raw data from which the effect size was computed. In addition, these techniques are not commonly used in meta-analyses (Jamshidi et al., 2022; Natesan, 2019), probably because they are not well understood by practitioners. Specifically, using multilevel models requires making several complex modeling decisions that may have an impact on the validity of the estimates (Baek & Ferron, 2020; Moeyaert et al., 2016). Moreover, such top-down techniques may not align well with the dominant visual analytical approach, and reduce a large amount of raw data to a single omnibus effect size (Barbosa Mendes et al., 2022) that may not thoroughly capture the complexity of the data (Parker & Vannest, 2012).

In contrast, “the term ‘bottom-up’ refers to an analytic strategy that proceeds from visually guided selection of individual phase contrasts (the ‘bottom’) to combining them to form a single (or a few) omnibus effect size representing the entire design (the ‘top’)” (Parker & Vannest, 2012, p. 255). Following this approach, the current proposal is meant to supplement top-down techniques by visually representing the raw data from which effect sizes are calculated. The current proposal may be considered “the bottom” because the graphs visualize the raw data which can inform further decision-making in the analytical process. The visual tools presented here do not yield an effect size

themselves. As such, the presented visual tools do not require any assumptions about the data and can combine data from many different types of SCEDs. At the same time, they can inform about specificities of the data that should be taken into account for further analyses because “analyses prescribed by design-type or template may need to be revised in light of actual data obtained” (Parker & Vannest, 2012, p. 256).

### **Complementarity of Visual and Statistical Analyses for SCEDs**

Following this logic, the current proposal builds on the complementarity of visual and statistical analyses for SCEDs. Visual analysis has a long-standing tradition in SCED data analysis and continues to be the most dominant mode of analysis. At the individual study level, it has been recommended to combine visual and statistical analyses (Harrington & Velicer, 2015; Manolov & Moeyaert, 2017). At the individual study level, visual analysis usually relies on the interpretation of time-series graphs of the raw data (e.g., Kratochwill et al., 2010), although proposals have been made for visually assessing the consistency of summary measures (Manolov & Tanious, 2022). Currently, in the context of SCED meta-analysis, visual representations are restricted to summary measures only. Examples of such plots include forest plots, caterpillar plots, and funnel plots (Fernández-Castilla et al., 2020), as well as the L’Abbé plot (Anzures-Cabrera & Higgins, 2010). To further strengthen the complementarity of visual and statistical analyses in the meta-analytical SCED context, the current proposal extends existing plots by introducing the possibility to visually represent the raw data of different experimental conditions from which effect sizes are calculated through multiple violin plots.

## **Violin Plots**

### **Basic Features**

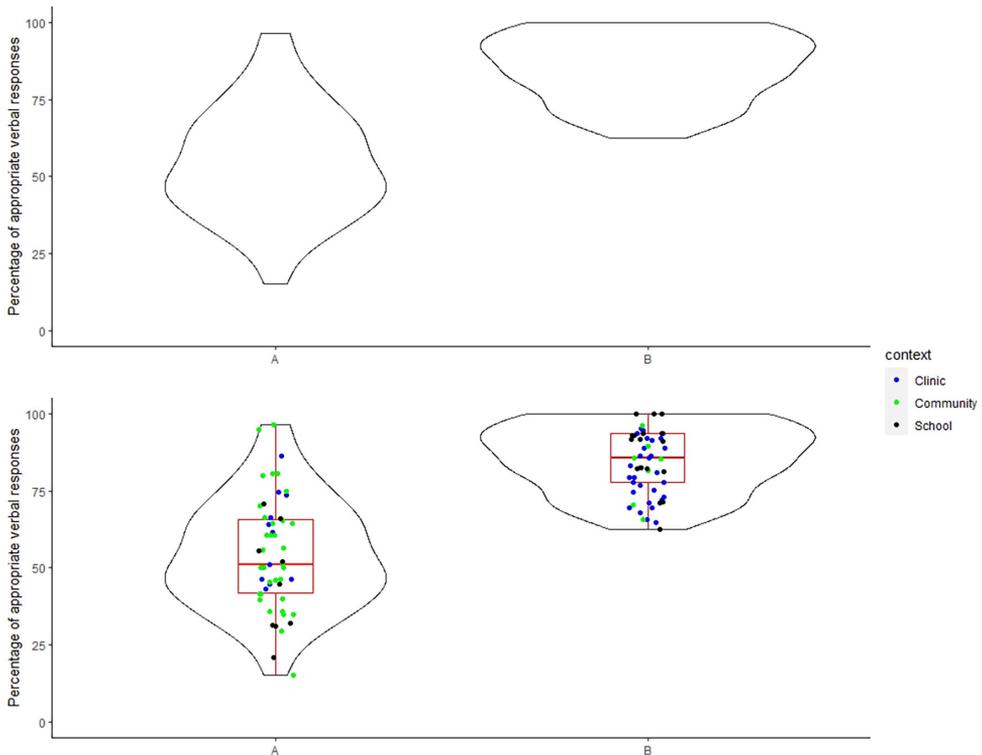
The violin plot was formally introduced by Hintze and Nelson (1998) as an adaptation of Tukey’s (1977) box plot in combination with density traces. In their development of the violin plot, Hintze and Nelson built on previous work by Benjamini (1988) who suggested opening the box of a boxplot to “convey information about the density of the values in a batch” (p. 257). One variation of the boxplot introduced by Benjamini was the vaseplot: “A boxplot where the width of the box at each point is proportional to the estimated density there” (p. 259). Thus, the vaseplot—and its successor the violin plot—do not depict raw data, but rather the estimated density of scores that fall within a given interval. The width of the violin plot on each side may be interpreted as a smoothed histogram of data density (Hu, 2020). The advantage of this is that the graph remains readable with the possibility to include graphical descriptions of summaries as well (Benjamini, 1988). In addition, it is possible to add information about the raw data by jittering data points randomly along the x-axis to avoid cluttering (Benjamini, 1988).

## Initial Illustration of the Basic Features

Violin plots can easily be constructed with very little programming knowledge, for example using R. Consider [Figure 1](#) as an illustrative example. These graphs were created in R using the *ggplot2* package ([Wickham, 2016](#)). A beginner's tutorial for constructing violin plots in R using the alternative *vioplot* package is available in [Hu \(2020\)](#). The data sets and R-code used for creating the figures in the present article are available at the [Supplementary Materials](#) section. [Figure 1](#) shows violin plots that combine the aggregated data of three A–B comparisons. A baseline condition is compared to self-management for a participant called Tony, diagnosed with Autism Spectrum Disorder, with the target behavior being the percentage of appropriate verbal responses in three different contexts (clinic, community, and school). The raw time-series graph is available in the [Supplementary Materials](#). The data were gathered by [Koegel et al. \(1992\)](#) and are part of a multiple-baseline design across participants and across contexts. It should be highlighted that [Figure 1](#) contains the data of only one participant. The violin plots are thus not restricted to the meta-analytic level, but can also be used at the individual level if researchers wish to use an alternative graphical representation next to the traditional time-series graph.

**Figure 1**

Violin Plots for the *Koegel et al. (1992)* Data for Participant Tony



*Note.* The upper panel shows the basic violin plots. The lower panel shows the same violin plots with the addition of a boxplot and jitter of the raw scores. Blue indicates clinical setting, green indicates community setting, and black indicates school setting.

The upper panel of [Figure 1](#) shows the basic violin plots for the two time series. It can clearly be seen that the scores for Phase A are lower, with a greater density of values below 50. In contrast, the data from Phase B show a clear negative skew, as there is a greater density of scores above 75, and there are no values below 50. The lower panel of [Figure 1](#) shows the same violin plots with the addition of the familiar box plots (in red) and the jittered raw scores (blue color is used for measurements from the clinic context, green color for the community context, and black color for the school context). These added graphical features reveal additional information. The boxplot reveals more clearly the positive skew of the central 50% of the data (the length of the box representing the range between the 25th and 75th percentile) in Phase A, with values more tightly grouped between the 25th and the 50th percentile. In contrast, there is a negative skew

in Phase B that is related to the presence of several values close to the third quartile (the upper line of the box marking the 75th percentile). Moreover, it is clear that the central 50% of the Phase B data, marked again by the length of the box, is greater than the 75th percentile of the baseline data (i.e., above the upper limit of the Phase A box). The jitter reveals that Phase A has more scattered measurements. It can also be seen that there are two values below 25 and two values close to 100, with a median approximately equal to 50 (with five values very close to it). Regarding the B values, there are two values exactly equal to the median (around 85) and three values equal to 100. The main difference between settings (marked by the different colors) can be seen in terms of higher (better) baseline values for the community (green) setting, which are all above the 25th percentile of all baseline data considered together. In contrast, the baseline values for the clinic (blue) context are lower (worse), in that they are not very present above the 75th percentile.

## General Considerations for Using the Violin Plot

### Visual Features

As shown in [Figure 1](#), two additional visual features are the boxplot and jitter. We recommend to always add these features to the basic violin plot. Both the boxplot and jitter can be customized. For the box plot, it is mainly the width of the box that can be customized for improving the display, although this dimension—unlike the length of the box—does not convey any information about the data. To increase the readability of the graph, we recommend setting the width of the box plot so that it overlaps as little as possible with the boundaries of the violin plot. In addition, the color of the boxplot can be changed so that it differs from the color of the violin plot and jitter. It is of course also possible to change the color of the violin in case of a large number of data points where the jitter may overlap with the violin.

Similar to the interval width for density estimation in the violin plot, the width of the jitter along the x-axis can be changed. As mentioned previously, the data points are jittered along the x-axis to avoid cluttering of points ([Benjamini, 1988](#)). This entails that data points which are parallel on the x-axis represent the same value on the y-axis. Changing the width of the jitter will spread the jitter more narrowly or widely along the x-axis. If the jitter width is set too narrowly, the data may seem cluttered, but it can also help in identifying clusters. If the jitter is set too widely, it may become difficult to identify which data points belong to the same value on the dependent variable. In addition, it may create overlap between the jitter from different groups and make the underlying box and violin plots difficult to read. Therefore, we recommend to start with a narrow jitter width and increase gradually until a satisfactory visual representation has been found.

## Interval Width and Reproducibility

A crucial consideration when using vase or violin plots is the setting of the interval width  $h$  (cf. Equation 1 in [Hintze & Nelson, 1998](#)). In other words, what is the width of each interval for which the density is estimated? This is analogous to setting the bin width in a histogram. Histograms representing the same data may look completely different depending on the bin width. For example, if the bin width is set too narrow, the histogram will show a spike for each data point, especially in the case of small sample sizes. If the bin width is set too wide, the distribution may appear overly smooth.

Regarding the optimal interval width for violin plots, Nelson and Hintze recommend 15% of the data range as a general rule of thumb. If a violin plot is constructed in R, the default setting for the interval width uses [Silverman's \(1986\)](#) rule of thumb which “defaults to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power” ([R Core Team, 2018](#)). A discussion of the limitations of this approach and a possible alternative approach are discussed in [Hall et al. \(1991\)](#). In R, it is possible to change the interval width manually. We recommend doing so only with sufficient statistical knowledge and if deemed necessary to better represent the data. If this is done, it should be explicitly communicated which density estimation method has been used to ensure reproducibility. In addition, it is recommended to use the `set.seed()` command for setting a fixed seed to make the jittered violin plot reproducible ([Sidiropoulos et al., 2018](#)).

## Requirements for Using Violin Plots in SCED Meta-Analysis

### Measurement Unit of the Dependent Variable

It is important that all data points represented on adjacent violin plots are measured on the same scale. This can be achieved in four ways. First, it is possible to ensure during the study selection process that all included studies used the same measurement unit (e.g., percentage of a certain behavior). If practically feasible, this is the preferred method. Second, it is possible to transform the scores from studies that used a different measurement unit than the majority of studies. For example, if the main interest for a SCED meta-analysis is the percentage of time spent on-task by children with conduct problems, it is possible to transform scores from studies using other measurement units to percentages. If some studies used time spent on-task, it is possible to transform the time spent-on task to percentages by dividing it by the total time and multiplying with 100. However, it should be noted that sufficient background knowledge about the studies using other measurement units is required for this approach. Thus, insufficient reporting ([Barbosa Mendes et al., 2022](#)) may be a limitation for following this approach. Third, it is possible to create separate adjacent violin plots for each different measurement unit. Finally, it is possible to standardize all data as has been suggested in the context of SCED meta-analysis ([Van den Noortgate & Onghena, 2008](#)). In the context of the current

proposal, this approach entails dividing each data point by the within-phase standard deviation.

### Number of Measurements

Hintze and Nelson (1998) recommend that “as a rule of thumb based on practice, the density trace tends to do a reasonable job with samples of at least 30 observations” (p. 183). It should be stressed that both the number of baseline and intervention data points should reach this minimum requirement. However, in the context of SCED meta-analysis this will rarely be a concern given that systematic reviews have found the median number of data points per participant to be 20 (Shadish & Sullivan, 2011) for single designs and between 25 and 67 for embedded designs (Tanious & Onghena, 2022). Even if the number of data points is usually shorter for baseline phases with the modal number being six (Smith, 2012), when aggregating the results from one or several studies, there should generally be a sufficient number of data points.

## Demonstration

### The Losinski et al. Meta-Analysis

To demonstrate the use of violin plots in SCED meta-analysis, we make use of the meta-analysis of SCEDs conducted by Losinski et al. (2014) “focusing on interventions based on the assessment of contextual variables (i.e., circumstances that form the setting for the behaviors)” (p. 407). In total, Losinski et al. included 24 studies. For each of the studies, Losinski et al. calculated three effect sizes: standardized mean difference (SMD) (Busk & Serlin, 1992), percentage of non-overlapping data points (PND) (Scruggs et al., 1987), and improvement rate difference (IRD) (Parker et al., 2009). For effect size calculation, the studies were grouped into three behavior types: disruptive behavior ( $n = 16$ ), on-task behavior ( $n = 10$ ), and stereotypical behavior ( $n = 4$ ). Before proceeding with any further analyses, we assigned each of the included articles a unique identification number ranging from 1 to 24.

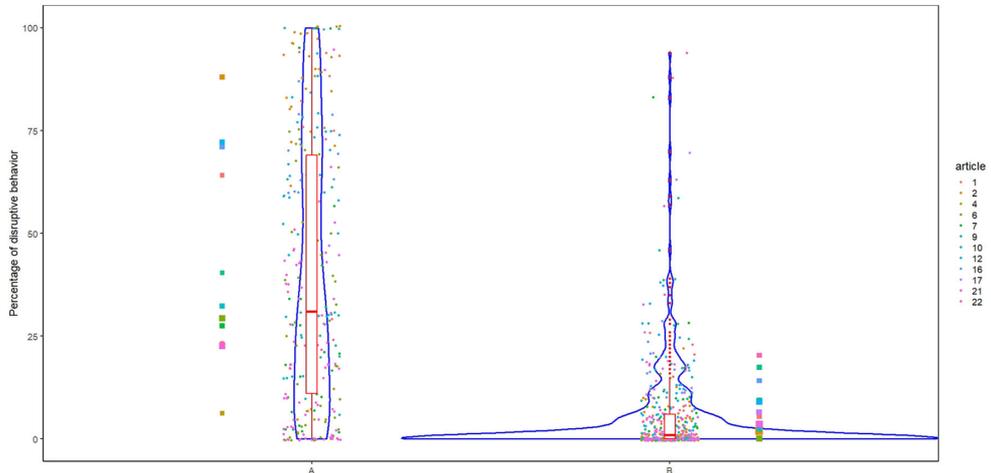
### Percentage of Disruptive Behavior

Figure 2 shows the jittered violin plots with integrated boxplot for the percentage of disruptive behavior. For visual clarity, the violins are shown in blue. The raw data from each study were extracted from the published graphs using GetData Graph Digitizer 2.26 (Fedorov, 2013). Figure 2 contains the raw data from 12 studies with 278 baseline (A) and 502 intervention (B) data points. The data points are colored according to the article from which they were extracted. Four out of the 16 studies measuring disruptive behavior used other measurement units that could not easily be converted to percentages. The data in Figure 2 stem from phase designs ( $n = 5$ ), multiple baseline designs ( $n = 5$ ),

changing criterion designs ( $n = 1$ ), and a reversal with embedded alternating treatments design ( $n = 1$ ).

**Figure 2**

*Violin Plots for Percentage of Disruptive Behavior in the Losinski et al. Meta-Analysis*



As shown by the boxplots, the median percentages of disruptive behavior are 31% for baseline data points and 1% for intervention data points. The third quartile (i.e., 75th percentile) lies at 69% for baseline measurements and 6% for intervention phase measurements. The violin plots show that by far the largest density for intervention measurements lies at 0% whereas the density is distributed more evenly over the whole range for baseline measurements. The density for the intervention data points gets narrower with higher percentages. The jitter clearly shows that there are only a handful of data points above 30%. At the same time, the jitter shows that for the baseline data points, there is also a non-negligible number of data points with 0% disruptive behavior.

To the left of the baseline violin plot and to the right of the intervention violin plot, the mean value per study can be found, following the same color scheme as the raw data. The size of the mean squares is proportionate to the number of measurements from which they were calculated. There are only 11 baseline mean squares because one study did not include baseline data. With the exception of two baseline means, all interventions means are lower than the baseline means. However, one of those two baseline means (pink square) contains the largest number of observations as evidenced by the size of the square. At the same time, the intervention mean is much smaller for that study and the pink intervention data points are mostly clustered around 0%. One study (brown square) already showed a relatively small mean percentage of disruptive behavior during base-

line measures, which was still reduced to some extent during intervention measures. The lowest intervention mean (olive green square) lies at 0%, a reduction of 30% compared to baseline measures.

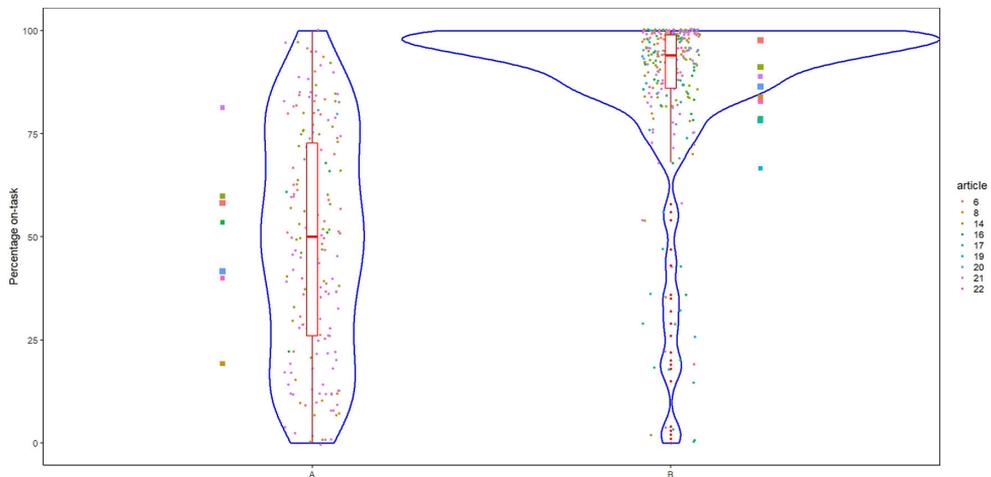
The mean effect sizes (accompanied by the standard deviation in parenthesis) calculated by Losinski et al. for all studies assessing disruptive behavior were SMD:  $-3.08$  (2.18), PND: 79% (34.84%), and IRD: 87% (18.77%). Visual representations (and summary measures) as shown above can be valuable complements to these summary measures by giving more information about the score distributions. While effect size measures are important measures of intervention effectiveness, they reduce a vast amount of raw data to a single number.

### Percentage of On-Task Behavior

Figure 3 shows the jittered violin plots with integrated boxplot for the percentage of on-task behavior. Two studies represented in the violin plots measured the dependent variable in terms of percentage off-task. The data from these studies were transformed with the formula  $100\% - \% \text{ off-task}$ . Figure 3 contains the raw data from nine studies with 186 baseline (A) and 287 intervention (B) data points. One study assessing the on-task behavior used a measurement unit that could not be converted to percentages. The plot represents data from phase designs ( $n = 4$ ), multiple baseline designs ( $n = 3$ ), and alternating treatments designs ( $n = 2$ ). The visual features are the same as in Figure 2.

**Figure 3**

*Violin Plots for the Percentage of On-Task Behavior in the Losinski et al. Meta-Analysis*



The boxplots show a median percentage of on-task behavior of 50% for baseline measurements and 94% for intervention measurements. The third quartile is 73% for baseline data points and 99% for intervention data points. This reveals that the percentage of on-task behavior was already relatively high during baseline measures. However, the boxplots also show that the spread of the scores is much higher for baseline measurements. The violin shows that the density of the baseline scores is relatively even over the whole range of percentages. In addition, the boxplot shows that there are no outliers for the baseline measurements which is supported by the even vertical spread of the jitter. For intervention measurements, the density is much higher on the upper end of the scale and narrows towards 60%. The jitter supports this by showing that there are very few data points below 60%, which are all outliers according to the boxplot.

Two studies did not include baseline data, so that there are seven baseline mean squares and nine interventions mean squares. The mean squares reveal that four intervention phase means are either under or close to the highest baseline mean (light purple square). This highest baseline mean has slightly increased during intervention. The largest baseline square (light blue) lies clearly under 50% of on-task behavior, which has increased to over 75% after the intervention. The two largest intervention squares (salmon pink and olive green) are the ones showing the highest percentage of on-task behavior and both are higher than the highest baseline mean. Four out of the seven baseline mean squares are above 50%, supporting the information from the boxplot that the percentage of on-task behavior was already relatively high during baseline measure.

The mean effect sizes (accompanied by the standard deviation in parenthesis) calculated by Losinski et al. for all studies assessing on-task behavior were SMD:  $-2.93$  (2.32), PND: 84% (30.83%), and IRD: 88% (25.46%).

## Discussion

The current article demonstrated the use of violin plots as visual tools in the meta-analysis of SCEDs. Complementing statistical top-down procedures and graphical representations of summary data (Anzures-Cabrera & Higgins, 2010; Fernández-Castilla et al., 2020), the violin plots represent the raw data and mean scores of the studies included in a meta-analysis. This enables, for instance, exploring whether all intervention means (or raw data points) represent an improvement over the corresponding baseline means (or raw data points) or comparing the best study (mean or raw data) in terms of baseline performance to all studies (or to the best or the worst) in terms of intervention phase performance. In summary, more nuanced comparisons are possible, beyond merely stating whether the overall mean difference represents an improvement or not.

The violin plots can contain data points from different designs. The integration of data points from ABAB designs (and extensions of it) and multiple baseline designs seems straightforward. A-phase data points are represented in one violin and B-phase

data points are represented in another. One issue pertaining to meta-analyses of SCEDs is the question how to integrate data from changing criterion designs and alternating treatments design (Manolov et al., 2022; Shadish et al., 2013). Here, we followed one of the possible approaches for integrating changing criterion designs using the baseline and last intervention phase data points (Manolov et al., 2022; Shadish et al., 2013). While this may lead to a potential loss of information, it may also be argued that the last sub-phase is the one including the final desired level and preceding sub-phases are only used as intermediate steps (Hartmann & Hall, 1976). Moreover, certain sub-phases may include reversal to lower criterion levels to increase experimental control (Klein et al., 2017), which may lead to underestimating the intervention effect. For alternating treatments designs, comparing a baseline and intervention condition, the data can be included in the same way as for ABAB and multiple baseline designs, despite not following a phase structure.

### The Necessity of Data Availability and Operational Definitions

The use of violin plots, just like other SCED meta-analytical tools, requires the availability of the raw data from the included studies. For the sake of transparency and reproducibility, authors would therefore make the raw data available (Tate et al., 2013). If the data are not available in raw form, they can still be extracted from the published graphs if the studies include the original time-series graphs.

In addition, as noted by Losinski et al. (2014), not all researchers are equally objective in operationally defining target behaviors. For example, “‘off task’ could mean doodling, talking to peers, head on desk, and so forth. Yet some studies never go on to objectively define ‘off task’” (p. 411). For the violin plots to be a meaningful representation of the raw data—as well as for the meta-analysis itself—it is important that the included studies represent the same underlying concept. A related point, mentioned previously by Losinski et al., is that measurement units must be the same. This is not only a consideration for the current proposal, but SCED meta-analysis in general: “It is necessary to standardize reported DV [dependent variables] values for analysis by converting them into a percentage and setting the upper and lower limit of the y-axis on all studies to 100 and 0” (Losinski et al., 2014, p. 411). The approach of Losinski et al. to re-scale the original y-axis is less recommended because it may skew the results as well as not aligning well with the original measurement of the dependent variable. If it is not feasible to select studies that report dependent variables in the same measurement unit, the scores can either be converted to the same measurement unit if sufficient information is available, or they can be standardized by dividing by the within-phase standard deviation.

## Limitations and Future Research

The first limitation concerns the density estimation method used for constructing the violin plots. Depending on the density estimation method used, the violin plots may look somewhat different. Nevertheless, when the actually obtained measurements are added (jittered) to the plot, it is still possible to assess their distribution and concentration, independent of the density plot.

The second limitation relates to the number of data points contributed to the violin plots by each study. Studies with a higher number of data points may disproportionately affect the density curve. However, with the colored jittered raw scores and mean squares—whose size is proportionate to the number of data points—this limitation may be mitigated. Moreover, in meta-analysis it is expected (and desired) that studies with more measurements contribute to a greater extent to the overall quantification (i.e., have greater weight). Thus, this relation between number of data points and density is not necessarily a limitation. At the same time, it should be noted that with an increasing number of included articles, it may become difficult for the human eye to distinguish between the many different colors.

The third limitation refers to the need for all data to be expressed in the same measurement units. This may lead to working only with part of the data. However, any kind of summary of the data within a study (e.g., representing the data from a given phase using a mean or trend line or an effect size), entails a loss of data, as well as the assumption that the summary is an adequate representation of the raw measurements. Four ways for addressing this potential limitation have been presented in the text (selecting studies that report the same measurement unit, converting scores to the same measurement unit, standardizing the scores, and presenting separate adjacent violin plots per measurement unit).

Regarding future research, field tests of the conclusions that applied researchers could reach regarding a set of studies included in a meta-analysis, when interpreting independently quantitative summaries (e.g., SMD or non-overlap indices), forest plots, and violin plots may be conducted. It could be evaluated whether these three kinds of information lead to similar conclusions. Moreover, it is possible to study whether the conclusions of applied researchers based on quantitative summaries, change after they are shown graphical summaries (e.g., forest plots) or violin plots and whether any potential changes are mainly attenuating or emphasizing the perceived degree of intervention effectiveness. Additionally, the social validity of the results needs to be taken into consideration (Snodgrass et al., 2018), as it is a relevant piece of information, when expert judgment is applied to assess the impact of interventions (Imam, 2021).

Another avenue for future research is to assess the potential of the violin plots at the individual level, for example in the context of MBDs. In case of a sufficiently high number of data points, it would be possible to make side by side violin plots of baseline versus intervention for each setting. This would allow a quick assessment of the score

distributions as a supplement to the traditional time-series graphs. In that context, the minimum required number of data points for obtaining meaningful violin plots in the SCED context may be further explored as well.

---

**Funding:** The authors have no funding to report.

---

**Acknowledgments:** The authors have no additional (i.e., non-financial) support to report.

---

**Competing Interests:** The authors have declared that no competing interests exist.

---

**Data Availability:** Data is freely available at [Supplementary Materials](#).

---

## Supplementary Materials

The supplementary materials provided are the R code and datasets used in the research and can be accessed in the [Index of Supplementary Materials](#) below.

### Index of Supplementary Materials

Tanious, R., & Manolov, R. (2022). *Supplementary materials to "Violin plots as visual tools in the meta-analysis of single-case experimental designs"* [R code, data sets]. OSF.  
<https://osf.io/z5bya/files/osfstorage>

## References

- Anzures-Cabrera, J., & Higgins, J. P. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods, 1*(1), 66–80. <https://doi.org/10.1002/jrsm.6>
- Baek, E., & Ferron, J. J. (2020). Modeling heterogeneity of the Level-1 error covariance matrix in multilevel models for single-case data. *Methodology, 16*(2), 166–185.  
<https://doi.org/10.5964/meth.2817>
- Barbosa Mendes, A., Jamshidi, L., Van den Noortgate, W., & Fernández-Castilla, B. (2022). Network meta-analysis for single-case design studies: An illustration. *Evaluation & the Health Professions, 45*(1), 66–75. <https://doi.org/10.1177/01632787211067532>
- Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Pearson.
- Becraft, J. L., Borrero, J. C., Sun, S., & McKenzie, A. A. (2020). A primer for using multilevel models to meta-analyze single case design data with AB phases. *Journal of Applied Behavior Analysis, 53*(3), 1799–1821. <https://doi.org/10.1002/jaba.698>
- Benjamini, Y. (1988). Opening the box of a boxplot. *American Statistician, 42*(4), 257–262.  
<https://doi.org/10.2307/2685133>

- Busk, P. L., & Serlin, R. (1992). Meta-analysis for single case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Lawrence Erlbaum Associates.
- Busse, R. T., McGill, R. J., & Kennedy, K. S. (2015). Methods for assessing single-case school-based intervention outcomes. *Contemporary School Psychology, 19*(3), 136–144.  
<https://doi.org/10.1007/s40688-014-0025-7>
- Fedorov, S. (2013). *GetData graph digitizer* [Computer software]. <http://getdata-graph-digitizer.com/>
- Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). Visual representations of meta-analyses of multiple outcomes: Extensions to forest plots, funnel plots, and caterpillar plots. *Methodology, 16*(4), 299–315.  
<https://doi.org/10.5964/meth.4013>
- Hall, P., Sheather, S. J., Jones, M. C., & Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika, 78*(2), 263–269.  
<https://doi.org/10.1093/biomet/78.2.263>
- Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research, 50*(2), 162–183.  
<https://doi.org/10.1080/00273171.2014.973989>
- Hartmann, D. P., & Hall, R. V. (1976). The changing criterion design. *Journal of Applied Behavior Analysis, 9*(4), 527–532. <https://doi.org/10.1901/jaba.1976.9-527>
- Hedges, L. V. (2019). The statistics of replication. *Methodology, 15*(Suppl. 1), 3–14.  
<https://doi.org/10.1027/1614-2241/a000173>
- Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician, 52*(2), 181–184. <https://doi.org/10.1080/00031305.1998.10480559>
- Hu, K. (2020). Become competent within one day in generating boxplots and violin plots for a novice without prior R experience. *Methods and Protocols, 3*(4), Article e64.  
<https://doi.org/10.3390/mps3040064>
- Imam, A. A. (2021). Historically recontextualizing Sidman's *Tactics*: How behavior analysis avoided psychology's methodological Ouroboros. *Journal of the Experimental Analysis of Behavior, 115*(1), 115–128. <https://doi.org/10.1002/jeab.661>
- Jamshidi, L., Heyvaert, M., Declercq, L., Fernández Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2018). Methodological quality of meta-analyses of single-case experimental studies. *Research in Developmental Disabilities, 79*, 97–115.  
<https://doi.org/10.1016/j.ridd.2017.12.016>
- Jamshidi, L., Heyvaert, M., Declercq, L., Fernández-Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2022). A systematic review of single-case experimental design meta-analyses: Characteristics of study designs, data, and analyses. *Evidence-Based Communication Assessment and Intervention*. Advance online publication.  
<https://doi.org/10.1080/17489539.2022.2089334>

- Klein, L. A., Houlihan, D., Vincent, J. L., & Panahon, C. J. (2017). Best practices in utilizing the changing criterion design. *Behavior Analysis in Practice*, *10*(1), 52–61.  
<https://doi.org/10.1007/s40617-014-0036-x>
- Koegel, L. K., Koegel, R. L., Hurley, C., & Frea, W. D. (1992). Improving social skills and disruptive behavior in children with autism through self-management. *Journal of Applied Behavior Analysis*, *25*(2), 341–353. <https://doi.org/10.1901/jaba.1992.25-341>
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*.  
<https://files.eric.ed.gov/fulltext/ED510743.pdf>
- Losinski, M., Maag, J. W., Katsiyannis, A., & Ennis, R. P. (2014). Examining the effects and quality of interventions based on the assessment of contextual variables: A meta-analysis. *Exceptional Children*, *80*(4), 407–422. <https://doi.org/10.1177/0014402914527243>
- Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy*, *48*(1), 97–114. <https://doi.org/10.1016/j.beth.2016.04.008>
- Manolov, R., Onghena, P., & Van den Noortgate, W. (2022). Meta-analysis of single-case experimental designs: How can alternating treatments and changing criterion designs be included? *Evidence-Based Communication Assessment and Intervention*. Advance online publication. <https://doi.org/10.1080/17489539.2022.2040164>
- Manolov, R., & Tanius, R. (2022). Assessing consistency in single-case data features using modified Brinley plots. *Behavior Modification*, *46*(3), 581–627. <https://doi.org/10.1177/0145445520982969>
- Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2016). The misspecification of the covariance structures in multilevel models for single-case data: A Monte Carlo simulation study. *Journal of Experimental Education*, *84*(3), 473–509.  
<https://doi.org/10.1080/00220973.2015.1065216>
- Natesan, P. (2019). Fitting Bayesian models for single-case experimental designs: A tutorial. *Methodology*, *15*(4), 147–156. <https://doi.org/10.1027/1614-2241/a000180>
- Parker, R., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education*, *21*, 254–265. <https://doi.org/10.1007/s10864-012-9153-1>
- Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, *75*(2), 135–150. <https://doi.org/10.1177/001440290907500201>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Foundation for Statistical Computing. <https://www.R-project.org/>
- Schlosser, R. W., & Sigafos, O. (2008). Meta-analysis of single-subject experimental designs: Why now? *Evidence-Based Communication Assessment and Intervention*, *2*(3), 117–119.  
<https://doi.org/10.1080/17489530802520429>
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*(2), 24–33.  
<https://doi.org/10.1177/074193258700800206>

- Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147. <https://doi.org/10.1016/j.jsp.2013.11.005>
- Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*(3), 385–405. <https://doi.org/10.1037/a0032964>
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971–980. <https://doi.org/10.3758/s13428-011-0111-y>
- Sidiropoulos, S., Sohi, S. H., Pedersen, T. L., Porse, B. T., Winther, O., Rapin, N., & Bagger, F. O. (2018). SinaPlot: An enhanced chart for simple and truthful representation of single observations over multiple classes. *Journal of Computational and Graphical Statistics, 27*(3), 673–676. <https://doi.org/10.1080/10618600.2017.1366914>
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510–550. <https://doi.org/10.1037/a0029312>
- Snodgrass, M. R., Chung, M. Y., Meadan, H., & Halle, J. W. (2018). Social validity in single-case research: A systematic literature review of prevalence and application. *Research in Developmental Disabilities, 74*, 160–173. <https://doi.org/10.1016/j.ridd.2018.01.007>
- 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*, 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>
- Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for Single-Case Experimental Designs and N-of-1 Trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23*(5), 619–638. <https://doi.org/10.1080/09602011.2013.824383>
- Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N = 1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 8*(4), 289–309. <https://doi.org/10.3102/10769986008004289>
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
- Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*(3), 325–346. <https://doi.org/10.1521/scpq.18.3.325.22577>
- Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*(3), 142–151. <https://doi.org/10.1080/17489530802505362>
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer.



*Methodology* is the official journal of the European Association of Methodology (EAM).



leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.