Communicating Effect Sizes to Teachers: Exploring Different Visualizations and Their Enrichment Options

Abstract

This study explores how various visualizations impact teachers' understanding of effect sizes and their relevance to practice. In Study 1, 40 teachers assessed four visualization types displaying six effect sizes. Half-eye plots were perceived as less challenging and more informative compared to raincloud and Gardner–Altman plots. However, misconceptions of U3 and overlap were prevalent across all types. In Study 2, enrichment options (signaling, benchmarking) integrated into half-eye plots were examined for their impact on, among others, understanding and misconception reduction. While benchmarks increased time needed for decision-making, they did not improve accuracy compared to a control group, and neither did signaling. However, both signaling options substantially reduced misconceptions. Furthermore, one signaling option increased participants' sensitivity to group differences. This study offers first indications of a promising visualization type, the half-eye plot, and an enrichment option to decrease misconceptions when considering visualization design in science communication efforts aimed at supporting educators.

Keywords. evidence-informed educational practice, science communication, effect sizes, clearinghouses, visualizations

Introduction

Teachers are encouraged to consider scientific evidence in their professional practice to enhance school quality, teaching quality, and student learning (Brown et al., 2017; Slavin, 2020). To engage with scientific evidence, teachers need to go through a complex process. They have to access, comprehend, and critically reflect on the evidence, combine it with their own knowledge and experiences, and then find ways to integrate it with their teaching practice by, for example, developing new practices (Brown et al., 2022) or solving problems that repeatedly emerge (Kiemer & Kollar, 2021). However, external barriers such as limited resources (e.g., time) and teachers' own attitudes and motives can keep them from successfully implementing this process (e.g., van Schaik et al., 2018). At the same time, teachers are trained as experts in teaching but not as experts in science; teacher education often does not include the development of skills or knowledge to engage with scientific evidence (e.g., van Schaik et al., 2018). Thus, they can be regarded as scientific laypersons for whom the realization of evidence-informed practice is challenging.

There are several approaches to overcoming these barriers by improving one or both sides of the equation: the reception and communication of scientific evidence (Neal et al., 2015). As to communication, making scientific evidence accessible to teachers through user-friendly delivery of information is regarded as important (Higgins et al., 2022). In this context, clearinghouses are receiving increased attention (Hedges, 2018); they represent interfaces between educational research and educational practice with the goal of supporting evidence-informed practice in the classroom. They offer information and explanations clustered by topic on web-based platforms. Examples include the What Works Clearinghouse in the United States (https://ies.ed.gov/ncee/wwc) and the Clearinghouse Unterricht (https://www.clearinghouse.edu.tum.de/) in Germany. To help teachers make decisions, clearinghouses report well-established research findings, often drawing on results from empirical studies or syntheses thereof (Higgins et al., 2022; Knogler et al., 2022). One key piece of information conveyed is effect size (American Psychological Association, 2002). In addition to size, other aspects determine the interpretation and practical relevance of

intervention effects (Anvari et al., 2023) and are also included in the information provided by clearinghouses. These include aspects of sampling, measurements, design, cost-effectiveness, and scalability (Kraft, 2020). In the context of this paper, we focus exclusively on the effect size to be able to study the phenomenon in isolation and draw causal inferences.

User-Friendly and Successful Communication of Effect Sizes

When key challenges in science communication (Jensen & Gerber, 2020) are addressed, user-friendly and successful communication involves producing an understanding of effect sizes in teachers, taking into account features of cognitive processing and eliciting perceptions of relevance of that information to teaching practice.

First, producing an understanding of effect sizes requires homomorphic interpretations by researchers and teachers: for example, if researchers consider one effect size to be substantially larger than another, teachers should come to the same conclusion (*accuracy*). Similarly, if researchers judge an effect as not negligible, this should also be the case in teachers (*sensitivity*). While there is evidence that teachers interpret effect sizes with high accuracy (Author et al., 2023), sensitive interpretations appear to be particularly challenging because laypersons tend to judge intervention effects of typical size in educational science as negligible (McPhetres & Pennycook, 2020).

Second, as science communication specifically addresses scientific laypersons, particular attention should be paid to their prior knowledge and skills such as the cognitive processing limitations rooted in teachers' layperson status (Mayer, 2014). Due to the fact that teachers are generally not trained in statistics and methodology, effect sizes need to be presented in a way that does not overload the limits of the human cognitive processing system (Mayer & Moreno, 2003). One common approach to achieving this aim is reducing the *task difficulty* of the information presented (intrinsic load). Furthermore, to avoid taking too much valuable time away from teachers' core responsibilities, effect sizes should also be presented as resource-efficiently as possible; that is, they should be readily accessible at first glance (*efficiency*).

Third, to avoid communicated knowledge becoming inert, user-friendly and successful science communication considers relevance for the practice of the target audience. The *perceived informativity* and *perceived value* of the communicated information influence how relevant evidence appears for school practice and influences its actual use (e.g., Lortie-Forgues et al., 2021).

Communicating Effect Sizes via Visualizations

When reporting effect sizes, scientists typically use standardized metrics such as the standardized mean difference (Cohen, 1988). While standardized metrics may enable comparability between results, they convey little inherent meaning (Lipsey et al., 2012). When a teacher reads about an effect size of $d = .68$ of using augmented reality for learning (Garzón & Acevedo, 2019), that standardized metric does very little to help understand how the effect's magnitude translates into practice. Initial evidence supports the assumption that textual information about standardized effect size metrics is not especially intuitive and that alternatives such as Cohen's U_3 may perform better (Hanel & Mehler, 2019). However, the results are not yet consistent—for example, Cohen's U_3 has also been reported to be less informative than other metrics such as months of progress or percentage of students reaching a certain threshold (Lortie-Forgues et al., 2021) and to evoke misconceptions (Author et al., in press).

One promising alternative may involve designing intuitive *visual* representations of effect sizes for statistical novices, as the visual perceptual system is considered very powerful in terms of the processing speed and precision of estimating statistical information such as effect sizes (Franconeri et al., 2021). Although there is initial evidence that visualizations can elicit accurate interpretation of research findings in laypersons (e.g., Hanel et al., 2019) and specifically in teachers (Author et al., 2023), whether teachers prefer visualizations over textual representations of research results remains unexplored.

At the same time, previous studies indicate that design choices from a plethora of options can heavily influence the processing and estimation of effect sizes, even if the underlying

data and amount of information remain the same (Correll et al., 2020; Hanel, et al., 2019; Witt, 2019). Accurate estimation of statistical information is needed for the evaluation of effect sizes (e.g., minimum, maximum, mean), and the success of decision making based on this information varies depending on the type of visualization (Albers et al., 2014; Kale et al., 2020). For example, Pierce and Chick (2013) found that teachers had difficulties making data-based statements when presented with boxplots.

For science communication by clearinghouses, it is therefore important to know whether visualizations could be an effective alternative to inform teachers about research results or, more precisely, whether designing visualizations of effect sizes affects their understanding, cognitive processing, and perceptions of relevance for teaching practice of these effect sizes. In two studies, we investigate different types of visualizations as well as enrichment options that could improve teachers' understanding and cognitive processing of effect sizes and thus their perceptions of effect size relevance.

Study 1: Types of Visualizations

In Study 1, we address this research gap by exploring the effects of different *types of visualizations*: do visualization types make a difference and, if so, which one has the greatest potential for user-friendly and successful communication of effect sizes?

RQ1: Does visualization type affect measures of accuracy (1), sensitivity (2), perceived task difficulty (3), efficiency (4), perceived informativity (5), and perceived value for practice (6) when assessing effect sizes?

RQ2: Which visualization types produce the most desirable results in these measures?

Method

Ethical approval for all studies was obtained from the [blinded for peer review] University (reference number: A2.5.4-174_ns).

Sample

Using Bayesian updating (see section 1.3 in the Electronic Supplementary Material [ESM] for details), we collected a sample of $N = 40$ teachers via the online panel provider Prolific (www.prolific.co). Participants had an average age of 37.83 ($SD = 12.05$) and an average of 10.15 ($SD = 8.66$) years of experience in teaching. The share of female teachers was 67.5%. The United Kingdom was the most common country of residence (35%). There was considerable variance in school type (e.g., nine taught primary school, and seven taught secondary school) and subjects taught (e.g., six in the aesthetic domain and 20 in the humanities and social sciences; multiple answers were accepted).

Design and Procedure

We applied a rotated 4 x 6 within-person design. The first experimental factor visualization type randomly varied within person and consisted of half-eye plots with groups on the x- and y-axes, raincloud plots with groups on the y-axis, and Gardner–Altman plots with groups on the x-axis (see section 1.1 in the ESM for the preceding Delphi study on selecting suitable effect size visualizations for teachers). The *effect size* depicted in the visualization constituted the second randomly varied within-factor using the thresholds established by Cohen (1988): $d \in \{-.80; -.50; -.20; .20; .50; .80\}$. Each participant thus evaluated 24 visualizations. To avoid carry-over effects between the sensitivity and accuracy ratings, a rotated design approach was employed so that half the visualizations were displayed with sensitivity measures and the other half with accuracy measures. All other items were the same for all 24 visualizations (see Figure 1 in the ESM).

Further, we contextualized the data depicted in the visualizations by using a total of four short descriptions of sample studies. These vignettes described studies comparing scores of two groups on a metric variable for different educational science topics (e.g., reading on tablet vs. reading on paper). The topics were taken from original studies, and we ensured

that the descriptions were comparable with respect to several surface, syntactic, and morphological linguistic features (Berendes et al., 2018). We varied topics between-person to avoid additional cognitive load. All materials can be found at https://doi.org/10.5281/zenodo.7437831.

Measures

Data were collected via an online questionnaire using the *formr* survey framework (Arslan et al., 2020). We piloted all measures with $N = 10$ participants (see section 1.2 in the ESM).

Accuracy. We used three topic-specific items to measure accuracy. The first item reflected an *abstract metric* with a slider ranging from −1 to +1, an example item labeled "The group that reads on…," and these response options at the two ends of the slider: *"tablet is entirely superior to the one with paper"* and *"paper is entirely superior to the one with tablet."* The second item corresponded to Cohen's U_3 *metric*, where we asked a topic-specific question: "Look at the mean test score of the group reading on paper: How much percent of the group that reads on tablet has a higher test score than this value?" Participants could indicate their assessment in an open-ended response that accepted numerical values from 0 to 100. The third item reflected the *overlap* of the two groups in terms of the Grice and Barrett (2014) metric with the question, "How much do the groups overlap on the test score?"; responses were numerical values from 0 to 100. To see how accurate teachers' estimations were, we calculated the differences of the participants' ratings and the true effect size shown in the plot. Negative and positive values indicated inaccurate and 0 accurate estimations.

Sensitivity. To assess sensitivity, we adapted the following item from Author et al. (2023): "Is one group superior to the other or are they approximately the same?" with three topic-specific answer options (e.g., *"reading on tablet is superior," "both groups are approximately the same," "reading on paper is superior"*). Answers were recoded as "perceiving difference" if participants regarded the groups as different and indicated the correct direction. We excluded by recoding as missing values the 5.6% of the answers that

classified the groups as different but pointed in the wrong direction. Answers signaling equality were coded as “perceiving equivalence.”

Perceived Task Difficulty. To measure perceived task difficulty, we adopted the single-item instrument from Marcus et al. (1996), which was developed to measure intrinsic cognitive load (“How difficult was it for you to understand the graph?”: 1 = *very difficult*; 7 = *very easy*).

Efficiency. We operationalized efficiency in terms of the time it took for a participant to make an initial rating of accuracy or sensitivity (dwell time in milliseconds).

Perceived Informativity. Adapted from Lortie-Forgues et al. (2021), we measured perceived informativity using a seven-point Likert item “How informative do you perceive the way the information is presented in the graph?” with the response options at the extremes 1 = *extremely uninformative* and 7 = *extremely informative*.

Perceived Value. We asked participants to answer “To what extent is the result valuable for your teaching?” using a seven-point Likert scale, with responses ranging from 1 = *not at all* to 7 = *to a great extent*.

Attention Check. To ensure data quality from the online panel survey provider, we administered an attention check (Agley et al., 2022) after 12 of 24 plots (see section 1.5 in the ESM for details). We only included participants that passed the attention check in the analysis.

Statistical Analysis

To answer RQ1, we estimated Bayesian multilevel models to obtain evidence for null effects (van Doorn et al., 2023) and simultaneously account for the nested data structure (Gelman & Hill, 2007). For accuracy, perceived task difficulty, perceived informativity, and perceived value measures, we specified random intercept models with effect size as a predictor, a random intercept to control for the within-person design, and visualization types as dummy-coded predictors. We then compared the predictive power of this model to the model without visualization types using Bayes factors (BFs) based on bridge sampling (Gronau et al., 2017). Regarding sensitivity, we used the same approach but specified logit

links to account for the binary dependent variable. Efficiency was again modeled with Bayesian multilevel regression but using 5% right-winsorized log-transformed data to account for skewness and outliers in the dwell time data.

To answer RQ2, we estimated highest density intervals (HDIs) for six dummy variables indicating the visualization types and standardized the dependent variables of perceived task difficulty, informativity, value and efficiency to interpret the slopes and HDIs as Cohen's *d*. For all analyses, we specified critical BF thresholds of 5 and 1/5, respectively.

There were no missing values other than those missing by design. In all analyses we used listwise complete observations; a detailed analysis can be found in the Reproducible Documentation of Analysis (RDA) together with the open data publicly available under https://doi.org/10.5281/zenodo.7437831.[1]

Results

Misconceptions of Effect Sizes

The descriptive data revealed high accuracy when teachers assessed the presented effect sizes in abstract metrics ($0.37 \leq \tau_b \leq 0.55$), whereas we found unexpected distributions of the accuracy U_3 and overlap items with, among others, a mode around zero in each distribution (see Figure 2 in the ESM). Mapping these data to participants, we identified misconceptions, as some participants, for example, consistently gave implausibly small U_3 ratings, whereas others indicated plausible U_3 ratings throughout (see Figure 3 in the ESM). Hence, we decided to operationalize misconceptions of effect sizes as follows. Participants were classified as Cohen's U_3 misconceptualizers if their median U_3 rating was lower than the smallest plausible value; that is, the smallest plotted value of $d = -.8$, which corresponds to a Cohen's U_3 of 21.2%. Participants were classified as overlap

1 Correction: In the accepted Stage 1 version, there was an issue in the operationalization of the accuracy variables. This issue has been corrected and thus all accuracy results have been adjusted. The original and corrected RDA are both publicly available.

misconceptualizers if their median overlap rating was smaller than the smallest plausible value due to the largest presented effect size (d = .8 and d = −.8, respectively, which in our study context was an overlap of 68.9%). Based on these operationalizations, we found n = 22 U_3 misconceptualizers and n = 9 overlap misconceptualizers. Given these results, we excluded misconceptualizers from the analyses of Cohen's U_3 and overlap.

RQ1: Does Visualization Type Affect Understanding, Cognitive Processing, and Perceptions of Relevance?

The descriptive results indicate a difference between visualization types in favor of both half-eye plots, especially in perceived task difficulty, informativity, and value (see Figure 4 in the ESM). Table 1 presents the means and standard deviations in the raw metrics of all dependent variables across all visualizations and for each visualization on its own. It is worth highlighting that the average perceptions of task difficulty, informativity, and value lie in the middle range of the respective scales. Thus, the different visualizations are, for example, neither particularly informative nor particularly uninformative.

Based on inferential statistics, we found evidence for no influence of visualization type on accuracy using an abstract metric (BF_{10} = 0.04), Cohen's U_3 metric (BF_{10} = 0.03), and overlap metric (BF_{10} = 0.01). On the other hand, we did find evidence that visualization type makes a difference in the sensitivity ratings (BF_{10} = 16.25), perceived task difficulty (BF_{10} > 100), efficiency (BF_{10} > 100 for the first measurement time and BF_{10} > 100 for the last three measurement times), perceived informativity (BF_{10} > 100), and perceived value (BF_{10} > 100).

Table 1. Descriptive Statistics in Study 1

Variables	***M (SD)***				
	total	Half-eye plots (groups on x-axis)	Half-eye plots (groups on y-axis)	Raincloud plots	Gardner–Altman plots
Accuracy					
Abstract metric	-0.06 (0.76)	-0.06 (0.53)	0.07 (0.76)	-0.12 (0.87)	-0.12 (0.86)
U3 metric	-0.55 (1.01)	-0.58 (1.03)	-0.41 (1.01)	-0.59 (1.01)	-0.60 (0.98)
Overlap metric	-0.57 (1.22)	-0.54 (1.30)	-0.68 (1.27)	-0.45 (0.99)	-0.59 (1.29)

Sensitivity	0.68	0.72	0.72	0.63	0.66
Perceived task difficulty	4.05 (1.71)	4.23 (1.54)	4.55 (1.58)	3.83 (1.68)	3.59 (1.87)
Efficiency	28.82 (67.78)	27.18 (62.05)	23.95 (44.39)	28.40 (52.25)	35.74 (99.04)
Perceived informativity	4.11 (1.45)	4.19 (1.34)	4.40 (1.35)	4.04 (1.48)	3.8 (1.57)
Perceived value	4.13 (1.49)	4.31 (1.40)	4.38 (1.39)	4.03 (1.58)	3.8 (1.51)

Note. Accuracy: Reported as the difference of the rating and the true effect size shown in the plot. Negative and positive values indicate inaccuracy. Sensitivity (binary): Proportion of judgements where participants perceived a difference and indicated the correct direction. Efficiency: Measured in milliseconds but reported in seconds. Perceived task difficulty, informativity, and value: Measured using seven-point Likert scales.

RQ2: Which Visualization Types Produce the Most Desirable Ratings?

Based on Bayesian random intercept models using dummy-coded predictors to compare the different visualization types, we found substantial differences in the analyzed dependent variables between visualization types in favor of the two half-eye plots. Compared to Gardner–Altman plots, we found evidence for higher sensitivity of half-eye plots with groups on the x- and y-axis ($1.73 \leq OR \leq 1.85$). In line with this, we found small to medium effects ($.26 \leq d \leq 0.54$) that half-eye plots with groups on the x- and y-axis were perceived as less difficult, more informative, and valuable for practice than Gardner–Altman plots. Both half-eye plots were also perceived as less difficult ($.25 \leq d \leq 0.43$) than raincloud plots, while half-eye plots with groups on the y-axis also outperformed raincloud plots in perceived informativity ($d = .25$) and value for practice ($d = .24$). Rather negligible differences were found between both half-eye plots with respect to perceived informativity and value for practice ($0.05 \leq d \leq 0.16$). Concerning efficiency, half-eye and raincloud plots induced substantially shorter dwell times at first glance than Gardner–Altman plots ($-.79 \leq d \leq -.69$), whereas they were similarly efficient when compared to each other ($-.23 \leq d \leq .15$; see

Figure 4 in the ESM). The exact estimates of the corresponding models can be found in the RDA.

Conclusion

Although we examined only high-potential visualizations in Study 1 based on expert opinion, we found considerable differences between the four plots in most of our measures. Both half-eye plots outperformed the Gardner–Altman and raincloud plots, especially regarding perceived task difficulty, informativity, and value for practice. However, there was little difference between the two half-eye plots (see General Discussion for details). Hence, we will include both half-eye plots in Study 2 for further investigation.

We also found comparatively accurate estimates by teachers when interpreting effect sizes in abstract metrics, but a substantial number of misconceptions regarding Cohen's U_3 and overlap estimates. Initial findings from prior research indicate that misconceptions are fairly common among laypersons when dealing with statistical terms and information (Author et al., 2023). We will additionally investigate these misconceptions in Study 2.

Study 2: Enrichment Options

Because visualizations are rarely used as "naked" plots in practice, we further investigate whether two types of enrichment options (benchmarks and signaling) positively affect teachers' understanding, cognitive processing, and perceptions of relevance for practice. In this confirmatory study, we focus on the two half-eye plots based on the results of Study 1.

Benchmarks as Enrichment Option for Visualizations

Using *benchmarks* is an established approach to supporting the evaluation of the magnitude of effect sizes as they can serve as abstract points of reference. Commonly employed benchmarks of this kind were established by Cohen (1988), who divided the Cohen's *d* of effect sizes into small (.20), medium (.50), and large (.80). At the same time, Cohen emphasized that these abstract benchmarks should only be used if no better frame of reference is available. Accordingly, Kraft (2020) proposed developing empirically driven benchmarks for specific contexts (e.g., causal research on education interventions with

standardized achievement outcomes). Therefore, reporting an effect size together with other effect sizes from comparable interventions in the same research area (that is, concrete instead of abstract benchmarks) could lead to a more accurate interpretation of effect sizes (Lipsey et al., 2012). Furthermore, prior research suggests that 1) scientific laypersons are often unimpressed by typical effect sizes from empirical research (McPhetres & Pennycook, 2020) and 2) providing different types of benchmarks or comparison plots can result in astonishingly accurate perceptions of effect sizes (Author et al., 2023; Kim et al., 2022). We therefore argue that providing context-specific visual benchmarks alongside effect size visualizations might increase accuracy (in an abstract metric [$H^{1a}_{abstract\ metric}$], Cohen's U_3 [H^{1a}_{u3}], and overlap metric [$H^{1a}_{overlap}$]) and sensitivity judgments of teachers (H^{2a}).

On the other hand, there is an increase in the number of comparisons that teachers must perform while cognitively processing effect sizes in comparison to the presented benchmark visualization. These higher-level visual tasks can heavily tax humans' limited working memory (Baddeley, 1992). With benchmark visualizations, teachers need to make additional comparisons between the benchmark visualization and the visualization to be assessed. For this reason, we assume that task difficulty (H^{3a}) and dwell time (efficiency; H^{4a}) increase with the use of benchmarks. Lastly, since benchmarks provide more information, we assume that more information is also perceived as more informative (H^{5a}) and more valuable for practice (H^{6a}).

Signaling as an Enrichment Option for Visualizations

Another approach to support comprehension processes for visualizations of complex data is to use cues that guide the attention of the viewer to select and organize information, particularly lower-level information (Mayer & Moreno, 2003). This effect, known as *signaling*, has a positive impact on both retention and transfer and reduces cognitive load by directing attention to relevant information (Schneider et al., 2018).

Just as there are different metrics for expressing effect sizes on continuous data (Cohen, 1988), there are ways to signal them by focusing on either the difference or the overlap of

two groups. From research on multimedia learning, we infer that signaling differences in visualizations using Cohen's U_3 may help improve teachers' sensitivity (H^{2b}), task difficulty (H^{3b}), and efficiency (H^{4b}). The same applies when signaling overlap for an overlap metric. Again, we assume that additional information due to signaling is perceived as more informative (H^{5b}) and more valuable for practice (H^{6b}).

Referring to accuracy, we found in Study 1 that more than half our participants indicated effect size misconceptions. Hence, we formulate the following hypotheses. On the one hand, we assume that signaling both Cohen's U_3 and overlap increases accurate interpretations in all three dependent variables (accuracy in abstract metric [$H^{1b}_{abstract\ metric}$], Cohen's U_3 [H^{1b}_{u3}], and overlap metric [$H^{1b}_{overlap}$]) in those teachers who do not indicate any misconception. On the other hand, we assume a smaller number of Cohen's U_3 misconceptualizers when Cohen's U_3 is signaled (H^{7b}_{u3}) and a smaller number of overlap misconceptualizers when overlap is signaled ($H^{7b}_{overlap}$).

Method

Design and Procedure

To test our hypotheses, we used a within-between design with the within-person factors *visualization type* (half-eye plot with groups on the x-axis vs. groups on the y-axis) and *effect size* ($d \in \{-.80; -.65; -.50; -.35; -.20; .20; .35; .50; .65; .80\}$) and the between-person factors *benchmark* (benchmark vs. no benchmark) and *signaling* (signaling difference using U_3 vs. signaling overlap vs. no signaling). This approach resulted in three experimental groups (benchmarks EG, signaling of difference using U_3 EG, signaling of overlap EG) and one control group.

As in Study 1, participants in all groups were randomly introduced to one of four fictitious sample studies (*topic*) in educational science. Afterward, they were presented with a total of 20 visualizations that depicted the possible results of the sample study and were randomly varied by a) type of visualization (both types shown 10 times) and b) the effect size presented.

Additionally, the participants were again asked to answer questions referring to the interpretation of the presented results (accuracy and sensitivity), perceived difficulty, informativity, and value for practice.

The benchmarks EG received the same visualizations as in Study 1, but with an additional visualization of a small effect size for reference and a literal explanation (e.g., “Example for comparison. In a similar intervention, researchers found the results shown in this figure. They judge the difference between the two groups as a *small effect* that is still worth attention.”; https://doi.org/10.6084/m9.figshare.22285834). In the signaling of Cohen’s U_3 condition EG, visual highlighting and textual information (e.g., “Part of the group reading on paper that has a higher test score than the mean test score of the group reading on tablet”; example: https://doi.org/10.6084/m9.figshare.22285828) were integrated into the visualization to draw participants’ attention to the differences between the two groups. By comparison, for the signaling of overlap EG, we used visual highlighting and textual information on the overlap of both presented groups (e.g., “Students (from both groups) who overlap in their test scores”; example: https://doi.org/10.6084/m9.figshare.22285837). As in Study 1, the control group received “naked” visualizations without any of these enrichment options (see Figure 5 in the ESM).

Measures

The measures are generally identical to those in Study 1; we only made minor modifications to the wording in the accuracy and sensitivity items to increase construct validity (see demonstration survey at https://graphs-2t-demo.formr.org/).

Moreover, we administered a treatment check for the two conditions. In the benchmarking condition, participants indicated their agreement with the statement: “I was able to use the top graphs as an example for comparison.” on a six-point Likert scale (1 = *not at all*; 6 = *to a great extent*). In the signaling conditions, we used the same Likert-type item with the statement “The highlighting and labels helped me see where I needed to look on the graphs.”

To check whether our participants generally understand the statistical information presented in the half-eye plots, we gave the participants three more tasks at the end of the survey. First, we added a visualization representing an effect size of d = .80 with single-choice items on assessing the mean value of the different groups (e.g., “What is the mean test score of the group reading on tablets” with the answer options “14,” “29,” “44,” “59,” “74”) and a question on variability, “How do students’ test scores vary?” with the single-choice answer options “Students reading on paper vary more in their test scores,” “Students in both groups vary to about the same degree,” and “Students reading on tablets vary more in their test scores.” Second, following Kaplan et al. (2014), we presented two different half-eye plots simultaneously, one showing a narrow and tall and the other one a flat and wide dispersion (example see https://doi.org/10.6084/m9.figshare.23694417). Afterwards, we again asked the participants how students' test scores vary with the same single answer options as described before. Third, also adapted from Kaplan et al. (2014), we presented a last single half-eye plot to test for participants' axes comprehension. In order to do so, we asked “What does the horizontal axis (from left to right) represent in this graph?” and “What does the vertical axis (from bottom to top) represent in this graph?”. For both questions, answers were given in an open response format. Responses were coded as correct or wrong according to prespecified categories (e.g., for the horizontal axis responses like knowledge test score or synonyms will be coded as correct; see RDA for all categories).

Data Analysis

Exclusion Criteria

For the following statistical analyses, we excluded data from participants who did not pass our attention check. Based on our procedure, we did not expect missing data but planned to delete it listwise if missingness occurs in less than 5% (or multiply impute if it occurs in more than 5%). We checked the classical criteria for detecting outliers and - if present - excluded data points with a z-score higher than 3.29 (Tabachnick & Fidell, 2013). Furthermore, we

added another filter question to the Prolific filter question "You indicated that you worked in the education industry. Does your job involve teaching?" Ours referred to the type of school where the participants teach, with the answer options "primary school," "secondary school," "high school," "comprehensive school," "special school," or "other." Participants who checked "other" were asked to indicate the educational institution in a text field. Those who do not teach at a school were filtered out of further analysis (see also Discussion).

Confirmatory Analysis

As in Study 1, we estimated Bayesian multilevel models using the R package *brms* (Bürkner, 2017). Furthermore, we evaluated the predictive power of the entire models using *Conditional* R^2 (Bürkner, 2017) and compared the predictive power of the models using BFs based on bridge sampling (Gronau et al., 2017).

For all continuous dependent variables (accuracy, perceived informativity, perceived difficulty, perceived value), we first specified Bayesian random intercept models that included effect size and visualization type as the first two predictors and random intercepts to control for the within-person design in each model. To test the hypotheses referring to the different enrichment options in the benchmark condition (H^{1a}, H^{3a}, H^{5a}, H^{6a}), we compared these models to extended models that also included benchmarking as a further dummy-coded predictor. To test the hypotheses regarding the enrichment option signaling (H^{1b}, H^{3b}, H^{5b}, H^{6b}), we compared the models with the first two predictors to the extended models with the dummy-coded predictor signaling. To allow for more nuanced interpretation, we standardized the continuous dependent variables. As in Study 1, in the models where accuracy U_3 and overlap were included as dependent variables, the respective misconceptualizers' data was excluded.

To test H^{2a} and H^{2b}, we used the same approach as previously described, but due to the binary dependent variable (sensitivity), we specified Bayesian generalized multilevel models with logit links. As in Study 1, the efficiency measures (H^{4a}, H^{4b}) were modeled with log transformation and analyzed for the first and last three measurement times. Referring to H_{u3}^{7b}

and $H_{overlap}^{7b}$, we also compared Bayesian multilevel models with logit links with and without predictors for the EGs.

Exploratory Analysis

As an additional exploratory analysis, we also considered the HDIs of all dependent variables to make a statement about the quality of the differences to see, for example, whether signaling overlap is more helpful than signaling difference. Furthermore, we explored whether the different items on understanding the statistical information presented in half-eye plots explain differences in teachers' accuracy interpretations.

Sample

As in Study 1, we recruited English-speaking teachers via Prolific and determined the sample size via Bayesian updating. Based on the heuristics offered by Simmons et al. (2013), we initially stopped data collection after 50 teachers per group (see below for details) and carried out our registered data analyses. Based on our exclusion criteria, we had to exclude a total of 62.00% participants (only 62.47% participants completed the survey, of which 23.19% did not pass the attention check and further 20.79% did not teach in school context). The resulting analysis based on N = 160 teachers did not reveal a satisfactory level of evidence, defined as BF > 5 or BF < ⅕ as in Study 1.

Based on these high and difficult to predict drop-out rates (e.g., number of failed attention checks was considerably higher than in Study 1), we deviated from the pre-registered maximum budget of €2,000 and spent additional roughly €677 to collect further data from 100 participants (Lakens, 2021).

Considering our exclusion criteria, the final sample consisted of N = 220 teachers that were on average 41.80 (SD = 11.40) years old and had, on average, 13.58 (SD = 9.22) years of experience in teaching. Most of the teachers were female (64.5%) and lived in the United Kingdom (55%). As in Study 1, there was a considerable variance in school type (e.g., 38.64% teach at a primary school, and 31.36% at a secondary school) and teaching subjects (e.g.,

54.09% STEM- and 22.27% aesthetic-teachers). All detailed analyses are publicly available https://zenodo.org/doi/10.5281/zenodo.7324339.

Results

As in Study 1, teachers average perceptions of task difficulty, informativity, and value lie in the middle range of the respective scales and they are more accurate in their interpretations when using the abstract metric compared to U_3 and overlap metrics (Table 2). On average U_3 and overlap metrics show biases in the same direction, with the overlap metric showing the biggest bias. As hypothesized, participants in the signaling U_3 group exhibited considerably less U_3 misconceptions, while those in the signaling overlap group showed less overlap misconceptions.

Table 2. Descriptive Statistics in Study 2

Variables	*M (SD)*			
	Control group (*N* = 55)	Benchmarking condition (*N* = 55)	Signaling U_3 condition (*N* = 53)	Signaling overlap condition (*N* = 57)
Accuracy				
Abstract metric	0.05 (0.74)	0.00 (0.72)	0.12 (0.90)	0.10 (0.77)
U3 metric	-0.36 (0.79)	-0.39 (0.81)	-0.25 (0.54)	-0.44 (0.88)
Overlap metric	-0.53 (1.30)	-0.43 (1.36)	-0.46 (1.51)	-0.45 (1.13)
Sensitivity	0.75	0.69	0.80	0.73
Perceived task difficulty	4.41 (1.65)	4.19 (1.68)	4.11 (1.66)	3.95 (1.66)
Efficiency	29.08 (56.06)	40.24 (75.27)	34.40 (10.10)	30.81 (46.98)
Perceived informativity	4.03 (1.65)	4.00 (1.51)	3.94 (1.54)	3.85 (1.52)
Perceived value	4.02 (1.66)	3.97 (1.57)	3.88 (1.54)	3.79 (1.67)
U_3 misconceptualizers	0.29	0.35	0.09	0.37
overlap misconceptualizers	0.29	0.29	0.42	0.21

Note. Accuracy: Difference of participant's rating and the true effect size shown in the plot. Negative and positive values indicate inaccuracy. Sensitivity (binary): Proportion of judgements where participants perceived a difference and indicated the correct direction. Efficiency: Reported in seconds. Perceived task difficulty, informativity, and value: Measured using seven-point Likert scales. U_3 and overlap misconceptualizers (binary): Proportion of participants with a misconcept.

Benchmarks as Enrichment Option for Visualizations

Against our hypotheses summarized as H^{1a}, Bayesian random intercept models reveal moderate to strong evidence for no difference in overlap and U_3 metrics regardless of whether teachers interpret visualization with or without benchmarks ($0.04 \leq BF_{10} \leq 0.10$; $-0.04 \leq$ Cliff's $d \leq -0.02$). According to accuracy judgments based on abstract metrics, the results indicate inconclusive evidence ($BF_{10} = 0.23$), with the tendency towards no difference (Cliff's $d = 0.05$).

Regarding H^{4a}, the evidence strongly supports our hypothesis that teachers require more time to process visualizations with compared to visualizations without benchmarks (efficiency for the first plot: $BF_{10} > 100$; Cliff's $d = -0.31$), even if they interpret them several times in a row (efficiency for the last three plots: $BF_{10} = 6.27$; Cliff's $d = -0.23$).

Results regarding sensitivity (H^{2a}), perceived difficulty (H^{3a}), perceived informativity (H^{5a}) and perceived value (H^{6a}) are inconclusive ($0.45 \leq BF_{10} \leq 3.28$).

Signaling as an Enrichment Option for Visualizations

Again in contrast to our hypotheses on accuracy, Bayesian multilevel models reveal strong evidence for no differences between the signaling and the control groups in all three accuracy measurements ($1/100 \leq BF_{10} \leq 0.04$; $-0.13 \leq$ Cliff's $d \leq 0.01$). In line with H^{2b}, we found weak evidence for a difference in sensitivity between the signaling and the control groups ($BF_{10} = 5.07$) with signaling U_3 increasing sensitivity (OR = 1.11).

Referring to our hypotheses H^{3b} to H^{6b} ($0.20 \leq BF_{10} \leq 4.38$), the analyses yield mostly inconclusive evidence, except for parts of our hypothesis on efficiency (H^{4b}): Based on the three last visualizations, we found—in contrast to H^{4b}—strong evidence for no difference in the signaling and control groups ($BF_{10} = 0.06$; $-0.09 \leq$ Cliff's $d \leq -0.02$).

In line with our hypotheses H_{u3}^{7b} and $H_{overlap}^{7b}$, signaling U_3 reduces the amount of U_3 misconceptualizers (OR = 0.25) and signaling overlap the amount of overlap misconceptualizers (OR = 0.65). The effects are supported by strong evidence ($BF_{10} > 100$).

Explorative Analyses

Since none of the dependent variables showed substantial differences in favor of both tested enrichment options, it did not appear reasonable to explore the question which option is superior any further. Furthermore, descriptively, we could not find any substantial associations between the accuracy ratings and teachers' understanding of the statistical information.

Given the unexpectedly high rate of participants failing our attention check (cf. Douglas et al., 2023), we examined their responses and found them to be comparable to those who passed. Therefore, when including all participants in the analyses, results underpin the previously described trend with signaling U_3 improving sensitivity judgments ($BF_{10} = 55.37$) and additionally reveal strong evidence for no difference in task difficulty ($BF_{10} < 1/100$).

Discussion

In this study, we examined the efficacy of visual benchmarks and signaling strategies in enhancing comprehension. Contrary to expectations, we found evidence that these strategies made no difference in fostering accurate comprehension of effect sizes, with benchmarking even prolonging the decision-making processes. Nevertheless, the results are encouraging regarding signaling strategies, as they demonstrated a significant reduction in misconceptions associated with effect size interpretation.

Overall discussion

Within the context of teacher education, finding a good way of communicating effect sizes is important, as teachers tend to underestimate effect sizes. With a wide range of visualization types to choose from, half-eye plots with groups differentiated on the y-axis are

the most promising visualization type, according to our evidence. Enriching these with signaling strategies further heightens the sensitivity of detecting group differences as well as reduces misconceptions.

Limitations and validity

To contextualize our findings, we examine aspects pertaining to the generalizability and practical relevance of our results. Firstly, our investigation concentrates on communicating the magnitude of an effect,. At the same time, we recognize that in addition to size, other aspects (sampling, measurements, design, cost-effectiveness, scalability; Kraft, 2020) may also determine the interpretation and practical relevance of intervention effects (Anvari et al., 2023).

Secondly, real-world data examination often reveals differences in variances, skewness, and kurtosis between groups. In this study, we prioritize normally distributed data due to limited research on visualization effects. This approach enables cautious investigation with fewer interacting variables, ensuring a level of control while investigating the phenomenon.

Thirdly, we selected Prolific as our panel provider due to evidence indicating its ability to generate high-quality data (e.g., Douglas et al., 2023). Further, we addressed common concerns such as limited attention of study participants with established countermeasures (e.g., attention check). Yet, it can be assumed that our samples show positive selection bias. We argue that employing a within-person design safeguards our results from being influenced by time-invariant confounding variables (Rohrer & Murayama, 2021). Therefore, the internal validity of our study would only be compromised if we presuppose an interaction between factors responsible for our selective sample and our experimental conditions.

One notable limitation of our study lies in the use of the prolific built-in filter to screen for teachers based on a supplementary question in Study 1, which revealed no significant differences between school types. We further refined our screening measures in Study 2, to strengthen the integrity of our study's participant selection process.

Finally, at first glance it may seem that the plots are too complicated for teachers' everyday practice. However, we would like to emphasize that they are not designed to stand alone. Rather, they are intended to be integrated into platforms such as clearinghouse websites. Information from clearinghouses are in turn intended to be brokered in teacher education settings, led by instructors for both students and teachers alike.

Conclusion

In exploring visualizations to convey effect sizes in teacher education, there is a universe of options available for consideration. While research in this domain is still in its early stages, our study offers first indications of a promising visualization type, the half-eye plot, and an enrichment option to decrease misconceptions.

Electronic Supplementary Material

ESM 1. "ESM_Details-on-procedure-measures-design-results.docx"
Provides further details on procedures, measures, design, and results to increase reproducibility.

References

Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods*, *54*(2), 885–897. https://doi.org/10.3758/s13428-021-01665-8

Albers, D., Correll, M., & Gleicher, M. (2014). Task-driven evaluation of aggregation in time series visualization. *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 551–560). https://doi.org/10.1145/2556288.2557200

American Psychological Association. (2002). Criteria for evaluating treatment guidelines. *American Psychologist*, *57*(12), 1052–1059. https://doi.org/10.1037/0003-066X.57.12.1052

Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2023). Not All Effects Are Indispensable: Psychological Science Requires Verifiable Lines of Reasoning for Whether an Effect Matters. *Perspectives on Psychological Science*, *18*(2), 503–507. https://doi.org/10.1177/17456916221091565

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, 52, 376–387. https://doi.org/10.3758/s13428-019-01236-y

Author et al. (2023)

Author et al. (in press)

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., & Trautwein, U. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, *110*(4), 518–543. https://doi.org/10.1037/edu0000225

Brown, C., MacGregor, S., Flood, J., & Malin, J. (2022). Facilitating research-informed educational practice for inclusion: Survey findings from 147 teachers and school leaders in England. *Frontiers in Education, 7,* Article 890832. https://doi.org/10.3389/feduc.2022.890832

Brown, C., Schildkamp, K., & Hubers, M. D. (2017). Combining the best of two worlds: A conceptual proposal for evidence-informed school improvement. *Educational Research*, *59*(2), 154–172. https://doi.org/10.1080/00131881.2017.1304327

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Taylor and Francis. https://doi.org/10.4324/9780203771587

Correll, M., Bertini, E., & Franconeri, S. (2020). Truncating the y-axis: Threat or menace? *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376222

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*, *18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, *22*(3), 110–161. https://doi.org/10.1177/15291006211051956

Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review*, *27*, 244–260. https://doi.org/10.1016/j.edurev.2019.04.001

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Grice, J. W., & Barrett, P. T. (2014). A note on Cohen's overlapping proportions of normal distributions. *Psychological Reports*, *115*(3), 741–747. https://doi.org/10.2466/03.PR0.115c29z4

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Hanel, P. H. P., Maio, G. R., & Manstead, A. S. R. (2019). A new way to look at the data: Similarities between groups of people are large and important. *Journal of Personality and Social Psychology*, *116*(4), 541–562. https://doi.org/10.1037/pspi0000154

Hanel, P. H. P., & Mehler, D. M. (2019). Beyond reporting statistical significance: Identifying informative effect sizes to improve scientific communication. *Public Understanding of Science*, *28*(4), 468–485. https://doi.org/10.1177/0963662519834193

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*(1), 1–21. https://doi.org/10.1080/19345747.2017.1375583

Higgins, S., Katsipataki, M., Villanueva Aguilera, A. B., Dobson, E., Gascoine, L., Rajab, T., Reardon, J., Stafford, J., & Uwimpuhwe, G. (2022). The teaching and learning toolkit: Communicating research evidence to inform decision-making for policy and practice in education. *Review of Education*, *10*(1), Article e3327. https://doi.org/10.1002/rev3.3327

Jensen, E. A. & Gerber, A. (2020). Evidence-based science communication. *Frontiers in Communication,* 4, Article 78. https://doi.org/10.3389/fcomm.2019.00078

Kale, A., Kay, M., & Hullman, J. (2020). *Visual reasoning strategies for effect size judgments and decisions*. arXiv. https://doi.org/10.48550/arXiv.2007.14516

Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014). Investigating Student Understanding of Histograms. *Journal of Statistics Education*, *22*(2), 4. https://doi.org/10.1080/10691898.2014.11889701

Kiemer, K., & Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Zeitschrift für Pädagogische Psychologie*, *35*(2–3), 127–141. https://doi.org/10.1024/1010-0652/a000302

Kim, Y.-S., Hofman, J. M., & Goldstein, D. G. (2022). Putting scientific results in perspective: Improving the communication of standardized effect sizes. *CHI '22: CHI Conference on Human Factors in Computing Systems*, Article 625. https://doi.org/10.1145/3491102.3502053

Knogler, M., Hetmanek, A., & Seidel, T. (2022). Determining an evidence base for particular fields of educational practice: A systematic review of meta-analyses on effective

mathematics and science teaching. *Frontiers in Psychology*, *13,* Article 873995. https://doi.org/10.3389/fpsyg.2022.873995

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lakens, D. (2021). *Sample size justification*. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lipsey, M. W., Puzio, K., Yun, C., Herbert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012, November). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research, Institute of Educational Sciences, U.S. Department of Education. https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf

Lortie-Forgues, H., Sio, U. N., & Inglis, M. (2021). How should educational effects be communicated to teachers? *Educational Researcher*, *50*(6), 345–354. https://doi.org/10.3102/0013189X20987856

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology*, *88*(1), 49–63. https://doi.org/10.1037/0022-0663.88.1.49

Mayer, R. (2014). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 43–71). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

McPhetres, J., & Pennycook, G. (2020). Lay people are unimpressed by the effect sizes typically reported in psychological science. PsyArXiv. https://doi.org/10.31234/osf.io/qu9hn

Neal, J. W., Neal, Z. P., Kornbluh, M., Mills, K. J., & Lawlor, J. A. (2015). Brokering the research-practice gap: A typology. *American Journal of Community Psychology*, *56*(3–4), 422–435. https://doi.org/10.1007/s10464-015-9745-8

Pierce, R., & Chick, H. (2013). Workplace statistical literacy for teachers: Interpreting box plots. *Mathematics Education Research Journal*, *25*(2), 189–205. https://doi.org/10.1007/s13394-012-0046-3

Rohrer, J. M., & Murayama, K. (2023). These Are Not the Effects You Are Looking for: Causality and the Within-/Between-Persons Distinction in Longitudinal Data Analysis. *Advances in Methods and Practices in Psychological Science, 6*(1), 251524592211408. https://doi.org/10.1177/25152459221140842

Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educational Research Review*, *23*, 1–24. https://doi.org/10.1016/j.edurev.2017.11.001

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013, January 17–19). *Life after p-hacking* [Paper presentation]. Meeting of the Society for Personality and Social Psychology, New Orleans, LA, United States. https://doi.org/10.2139/ssrn.2205186

Slavin, R. E. (2020). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2023). Bayes factors for mixed models. *Computational Brain & Behavior*, *6*(1), 1–13. https://doi.org/10.1007/s42113-021-00113-2

van Schaik, P., Volman, M., Admiraal, W., & Schenke, W. (2018). Barriers and conditions for teachers' utilisation of academic knowledge. *International Journal of Educational Research*, *90*, 50–63. https://doi.org/10.1016/j.ijer.2018.05.003

Witt, J. K. (2019). Graph construction: An empirical investigation on setting the range of the y-axis. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.895