Highlights

- Communicating effect sizes visually in the form of half-eye plots to teachers is associated with higher accuracy.
- Teachers perceive verbal Cohen's U_3 effect sizes as more informative and easier to understand than half-eye plots.
- Teachers can generally distinguish between large and small effect sizes.

MAKING EFFECT SIZES UNDERSTANDABLE FOR TEACHERS

Abstract

Background: Teachers are encouraged to consider relevant findings from educational research to improve the quality of their instruction and enhance learning. This method requires them to have an accurate estimate of the effectiveness of educational interventions to select the most effective one. Currently prevalent approaches of communicating effect sizes, however, often lead to imprecise estimations or even misconceptions among laypersons and teachers.

Aims: To investigate whether it is accurate and effective to communicate effect sizes verbally using a common language effect size (Cohen's U_3) or visually by showing half-eye plots.

Pilot Sample: $N = 60$ English-speaking teachers, $M_{\text{age}} = 42.3$ years, recruited on Prolific.

Planned Sample: XXX

Planned Methods: Randomized, 2x6x4 within-between-person trial. The primary outcome of interest is perceived effect size value. Secondary outcomes are relevance, perceived informativeness and difficulty.

Pilot Results: Teachers displayed less accuracy in the verbal than in the visual condition (Cohen's $d = -0.52$, 95% $CI = [-0.81, -0.23]$). Conversely, teachers perceived verbal Cohen's U_3 expressions as more informative ($r_{\text{rb}} = -0.27$, 95% $CI = [-0.40, -0.13]$), and less difficult ($r_{\text{rb}} = -0.31$, 95% $CI = [-0.44, -0.17]$) to understand than effect sizes shown as half-eye plots. Bayesian estimation for the outcomes above demonstrated, at minimum, evidence of substantial effects.

Conclusions: Communicating effect sizes via half-eye plots seems to be superior in terms of

accuracy, but inferior regarding perceived informativeness and difficulty.

Keywords: Evidence-informed practice, Effect sizes, Visualization, Science communication

MAKING EFFECT SIZES UNDERSTANDABLE FOR TEACHERS

1. Introduction

> When researchers do draw implications from effect sizes, the interpretations they offer are, more often than not, superficial, uninformative, misleading, or completely wrong. In sum, effect sizes are widely unappreciated and often misunderstood, even by professional researchers (Funder & Ozer, 2019, p.156).

Imagine that you want to share the consistent findings of your research on a learning or instruction method, for example the benefits of using formative assessment elements in the classroom, within a professional development workshop for teachers. How would you present the information on the intervention's effect size? By plotting the raw data? By using verbal expressions of (common language) effect sizes?

Effect sizes quantify core statistical information from empirical research. As such, they fulfil an important purpose of a research discipline, that is, to inform answers to practical questions (Szollosi & Donkin, 2021). So far, communicating effect sizes successfully has been proven various times to be a challenge. Research on laypeople (Brooks et al., 2014; Hanel & Mehler, 2019; Hofman et al., 2020; Kim et al., 2022; Lortie-Forgues et al., 2021; McPhetres & Pennycook, 2020; Michal & Shah, 2024; Schmidt et al., in press; Schmidt et al., 2023; Xiong et al., 2023) and experts (O'Keefe, 2017; Schuetze & Yan, 2023; Zhang et al., 2023) alike suggests that communicating effect sizes

often leads to inaccurate estimations or even misconceptions. Thus, it is still unclear how effect sizes should be communicated to teachers and the broader public. There are different approaches to overcome this problem regarding the measure (e.g., Cohen's *d*, Cohen's U_3) of an effect size or its forms of presentation (verbal, visual). Furthermore, there are common language (e.g., the probability of superiority), and in educational research converted effect sizes, like percentile gain or months of learning, which can help to express the findings of a study in a more intuitive way (see Grissom, 1994; Kraft, 2020; Von Hippel, 2024). Presenting effect sizes using verbal representations like Cohen's *d* is a conventional, frequently used way in the social sciences, and particularly in empirical educational research (e.g. Hattie, 2009, 2023; Kraft, 2020). Given that both in-service teachers and trained psychologists struggle to grasp the effect size metric Cohen's *d* (Schmidt et al., 2023; Schuetze & Yan, 2023), using this specific measure in a workshop, as described in the introductory example, for teachers may not be advisable. Instead, one could use visually depicted effect sizes to inform the attending teachers about the effectiveness of one's educational intervention. Unfortunately, this way of displaying effect sizes also has its drawbacks, as the visualization format can heavily influence effect size perception (e.g. Kale et al., 2021; Kim et al., 2022; Zhang et al., 2023). There are clearinghouse institutions in the educational environment [1] that help disseminate scientific findings to a broader public. In their role as knowledge brokering agents, some of them use both verbal and visual representations of effect sizes to inform their audience about the effectiveness of a specific learning or instruction method. Similarly, Hattie also uses both a verbal effect size

[1] in the US: https://ies.ed.gov/ncee/wwc, in Germany https://www.tuedilb-tuebingen.de/tuedi-base.html or https://www.clearinghouse.edu.tum.de/.

(Cohen's d or Pearson's correlation coefficient r), and a visual one (either a speedometer or a barometer depicting the Cohen's d value) to express the findings of his prominent meta-meta-analysis (Hattie, 2009, 2023). According to Thomm et al. (2021), making empirical research available and accessible to teachers, either through clearinghouses or through books such as Visible Learning, is a prerequisite for so-called evidence-informed practice (EIP). Many stakeholders in education have been calling for an increased use and consideration of this approach to improve school quality and student achievement (Bauer & Kollar, 2023; Brown et al., 2017; European Commission, 2007; Ferguson, 2021; Nelson & Campbell, 2017; Pellegrini & Vivanet, 2021; Slavin, 2020). As more and more clearinghouses are being developed to help to encourage and establish the implementation of EIP (Wadhwa et al., 2024), the goal from a research perspective is to identify ways to improve the communication of effect sizes. Our paper addresses this research gap by comparing verbal with visual effect sizes. In the following, we will shortly reflect on the relevance of communicating effect sizes accurately and effectively to teachers, and then briefly illustrate the two common ways of presenting them: as verbal and visual entities.

2. Effect Sizes

2.1. Communicating Effect Sizes

A modern way of presenting scientific results not only involves communicating inferential uncertainty, but also information about the size of effects (Coe, 2002; Funder & Ozer, 2019; Santesso et al., 2020; Stukas & Cumming, 2014). As a researcher, you want to make sure that you communicate the results of your study in a way that helps your audience fully understand your findings. Successfully implementing this intention comprises two distinct aspects: First, if a teacher reads your article, they should ideally perceive the presented effect size in a comparable magnitude. This is a necessary, though not sufficient, condition for using scientific results in practice. If teachers mistakenly perceive the effectiveness of an intervention as low, there is a serious risk that teachers will not consider the described intervention, although it is, in fact, worth considering. Conversely, teachers could overestimate the interventions' effect size, and consequentially be later disappointed when there is only little learning progress in the classroom (Lortie-Forgues et al., 2021). Second, it is important that educational researchers convey effect sizes in an informative, easy to understand and relevant manner to teachers, or otherwise they will not consider the results of a study. Here, we call this style an effective way of communicating effect sizes. In summary, it can be stated that teachers' effect size perception should be as accurately as possible, but at the same time, it is also critical that teachers value and appreciate the presented effect size of a result in a way that they are willing to spend their (limited) time engaging with it.

2.2. Verbal Effect Sizes

Presenting effect sizes on a verbal basis, like Cohen's d or Cohen's U_3, has been used in primary (Lortie-Forgues et al., 2021; Schmidt et al., in press; Schmidt et al., 2023) and secondary (Hattie, 2009, 2023) research to communicate effect sizes to teachers. Verbal effect sizes can be split up into three different families with distinct properties: the difference, correlation, and ratio family (Rosnow & Rosenthal, 2003, p.222). Consequently, they can either reflect the magnitude of the difference between groups, highlight the strength of the relationship between different variables, or represent the efficacy of a treatment as quantified by a count-based metric (Rosnow & Rosenthal, 2003). Verbal effect sizes offer the advantage over visual effect sizes of being concise and quick to process. Further, Hearst argues that "language should be considered as co-equal with visualization" when communicating findings from research, as some people seem to prefer verbal representations to visualizations (Hearst, 2023, p.68). At the same time, however, they have the disadvantage that readers could be under the illusion of explanatory depth when perceiving verbal effect sizes (Rozenblit & Keil, 2002). This means that recipients might have the impression that they fully understand the meaning of a presented effect size, but in reality they only have a superficial or flawed comprehension. Research shows that even experienced individuals with an academic background in psychology have problems estimating Cohen's d values accurately (Schuetze & Yan, 2023). It can therefore be assumed that teachers as laypeople do so as well. Another problematic aspect of verbal effect sizes, besides being deceivingly easy to read and understand, might be that they represent aggregated statistical information. This entails that they convey very little

information about the underlying structure of the data (e.g., distribution, skewness etc.). Even though data distributions can vary greatly in shape, verbal effect sizes can produce values of identical magnitude (Matejka & Fitzmaurice, 2017), thereby reducing the inherent informativeness of verbal effect sizes.

To prevent inaccurate estimates or misconceptions from occurring (e.g., Michal & Shah, 2024; Schmidt et al., 2023), researchers have proposed so-called common language effect sizes that aim to overcome some shortcomings of traditional effect sizes (Brooks et al., 2014; McGraw & Wong, 1992; Vargha & Delaney, 2000). So far, their use has been limited due to their relative recent discovery [2], and the fact that their usage is not yet recommended broadly (Mastrich & Hernandez, 2021).

2.3. Visual Effect Sizes

Communicating effect sizes visually by displaying graphs or plots is another common way of illustrating results from empirical research (e.g. Kim et al., 2022). Visually shown effect sizes capitalize on the speed at which information is processed by the visual system (Thorpe et al., 1996). Under certain circumstances, visual representations of effect sizes are credited with facilitating rapid and accurate extraction, as well as the comparison of statistical information (Franconeri et al., 2021). Merk et al. (2023) found that teachers display high accuracy when extracting information, such as mean differences, from graphs. In two studies, they examined teachers' abilities to interpret mean differences sensitively, efficiently, and accurately using various graph types. Note that they only examined raw

[2] To put it into perspective: Traditional effect sizes took about 60 years to be incorporated into empirical sciences (Huberty, 2002)

mean differences, but not standardized effect sizes. Results showed that while teachers can perceive mean differences with high accuracy, they only consider moderate to large effects to be meaningful, showing low sensitivity. In addition, they discovered that teachers' efficiency in interpreting graphs rapidly increases when they are repeatedly exposed to the same type of graph, with stacked or clustered bar graphs being the most effective for sensitivity (Merk et al., 2023, p.8). Nevertheless, choosing the right type of plot or graph from the many available, and in doing so, preventing potential misunderstandings on the reader's part, is not an easy decision (see Wilmer & Kerns, 2022). When creating and modifying visualizations, there are a plethora of layout options to choose from. Some of these options, for example truncating the y-axis, or not showing the underlying distribution of the data, can influence or bias information perception and processing (Franconeri et al., 2021). Contrary to the abovementioned observation regarding the speed of information extraction, obtaining information from a complex visualization might take longer than from a verbal effect size. This is especially the case when the visualization of an effect and the task at hand do not align (Padilla et al., 2018). Finally, there are numerous empirical studies that indicate that visual representations of effect sizes can lead to an under- or overestimation of the effects shown, which in turn might create false expectations and facilitate potentially spurious inferences (Hofman et al., 2020; Kale et al., 2021; Zhang et al., 2023).

3. The Current Study

The effects of communicating effect sizes verbally versus visually to teachers have, to the best of our knowledge, hardly been investigated in direct comparison. In other words, there is little evidence as to which presentation mode (visual or verbal) is superior in the presentation of effect sizes and should therefore be preferred when it comes to science communication[3]. Against this backdrop, we investigate and compare the accuracy and effectiveness of communicating effect sizes verbally and visually to teachers. We plan to collect data on the following four dependent variables:

- *Accuracy*: How accurately do teachers perceive effect sizes? That is, how closely do they get in their ratings to the true, given effect size value?
- *Relevance*: How relevantly do teachers regard effect sizes? Are they willing to spend more money on an intervention with a higher effect size value compared to one with a lower effect size value?
- *Perceived informativeness*: How informatively do teachers perceive verbal and visual effect sizes?
- *Perceived difficulty*: How difficult is it for teachers to understand verbal and visual effect sizes?

We choose the two independent variables *effect size* and *presentation mode* for our study to (a) examine the impact of effect size value (e.g., smaller vs. higher effect size

[3] For related findings in the realm of risk communication, see (Stone et al., 1997) and (Chua et al., 2006)

values) and (b) measure any possible differences between the two presentation modes regarding our outcome variables. We decided to use the Cohen's U_3 measure in the verbal condition as a means of expressing the independent variable *effect size* because two studies have indicated that teachers and laypeople assign high values of perceived informativeness to this specific effect size metric (Hanel & Mehler, 2019; Schmidt et al., 2023). Further, we determined to build upon previous research by Schmidt et al. (in press) who explored 44 plot types for successfully communicating effect sizes to teachers in the context of two studies. Their comparisons revealed that half-eye plots emerged as the most accurate, efficient, and valued among the options considered. We therefore chose half-eye-plots as a way of visually depicting effect sizes. We conclude with the following two research questions:

Research question 1: How accurately (dependent variable 1), relevantly (dependent variable 2), informatively (dependent variable 3), and with what difficulty (dependent variable 4) are verbally and visually presented effect sizes perceived by teachers?

Research question 2: Are there any differences between the presentation modes regarding DV1-DV4?

As there is little empirical evidence from previous research, we came up with the hypothesis that there are no differences between the two presentation modes:

Null Hypothesis (H_0): There are no significant differences between the two presentation modes regarding the four dependent variables.

4. Methods

In the following section, we lay out the experimental design for our registered report.

4.1. Experimental Procedure and Design

In a 2x6x4 randomized within-between-person trial, participants will first receive a short research report on an educational science topic (e.g., the use of an AI reading tutor in a classroom setting), which is randomly selected from four different topics (between-person factor 4; topic 1: feedback from an AI tutor vs. feedback from a teacher on reading fluency, topic 2: the impact of 3D videos vs. 2D videos on knowledge retention, topic 3: the impact of an activity tracker-based program vs. normal school attendance on average daily steps, topic 4: the effectiveness of a science kit intervention vs. traditional science program on improving content knowledge). The research report contextualizes several subsequently presented fictitious results (see Figure A1), which vary randomly in the presented effect size value (within-person factor 6), Cohen's $d \in \{-.8, -.5, -.2, .2, .5, .8\}$. Note that negative values represent beneficial effects of the intervention, and that positive Cohen's d values signal that the control group is superior. Further, the presentation mode (verbally vs. visually) is also randomized (within-person factor 2). Participants will either first see six different effect sizes visualized as half-eye plots, and then another six effect sizes depicted as Cohen's U_3 phrases, or vice versa (cf. Figures A2 and A3).

4.2. Measures

After each of the 12 fictitious research results presented verbally and visually, we ask participants to answer four measures. For the dependent variable *accuracy*,

participants will have to give an estimate of their perceived effect size value: "*How many times out of 100 do you estimate that a randomly selected member of the AI tutor group would have a higher score in the reading test than a randomly selected person from the teacher feedback group?*" This item is based on the common language effect size probability of superiority, which was developed to be an easy-to understand and intuitive effect size measure (McGraw & Wong, 1992; Ruscio, 2008). Furthermore, we measure the *relevance* assigned by teachers by asking them how much they would be willing to spend on a specific educational intervention, e.g., : "*How much money are you willing to spend on an AI reading tutor license for a class of 30 students for one year?*" For the remaining two dependent variables, *perceived informativeness* and *perceived difficulty*, participants will state how informative or difficult they perceive the presented effect size value on a Likert scale from 1 (*extremely uninformative/very difficult*) to 7 (*extremely informative/very easy to understand*).

4.3. Analysis Plan

4.3.1. Preprocessing of the Data

Data Exclusion Criteria: Participants who are either exceptionally fast, which we define as three standard deviations below the mean survey duration, who fail to pass both attention checks in the survey, or who show no variation in their responses ("straightlining") are excluded from the analysis (Stosic et al., 2024). Additionally, we will also exclude data points if the participant's effect size estimate indicates the wrong direction of effect, assuming that the participant mixed up the treatment and the control group. Moreover, we will exclude participants who finish the survey prematurely.

Outliers: For the relevance variable, we will mark values as outliers if a person's centered absolute z-scores are greater than 3 (Freedman et al., 2007, p.102). Outliers for the other three dependent variables are rather unlikely due to their limited scales, but we will consider values as outliers if their grand mean centered absolute z-scores are greater than 3.

Missing Values: Missing data will not occur, as all answers in the online survey framework will be mandatory. Additionally, we will only include data from participants who completed the survey.

Variable transformations: We will transform the obtained probability of superiority (PoS) values into Cohen's d values to ensure the linearity of this dependent variable using the following formula $\delta = qnorm(PoS) \cdot \sqrt{2}$. This transformed rating of the perceived effect size will be subtracted from the true effect size, resulting in a variable which we call "rating error". Additionally, we will center values for the relevance variable at the person (cluster) mean, and then divide them by the person-specific standard deviation so that we can operationalize the intraindividual differences in willingness to pay on a z-scale. We will call this newly created variable "relevance_pstand", for personalized and standardized relevance. Moreover, we will also create a variable called "objective relevance" which equals zero if the intervention is detrimental (which corresponds to the positive Cohen's d values of 0.2, 0.5 and 0.8), and which equals the true, given effect size, if the treatment is beneficial (equaling Cohen's d values of -0.2, -0.5 and -0.8). In essence, this variable only contains values for the willingness to pay item if an intervention has an advantageous effect; that is, if it is superior to the "business as usual" (control) group.

4.3.2. Parametrization of the models

Accuracy (transformed to rating error; see 4.3.1) is modeled using random intercept models assuming a heteroscedastic, normally distributed outcome variable. We use dummy variables for the independent variable presentation mode to predict differences in the mean rating error per condition, as well as its spread.

We model relevance (transformed to standardized deviations from the person's mean; see 4.3.1) assuming a normally distributed outcome variable and, again, using dummy variables for the independent variable presentation mode after adjusting for the objective evidence.

To account for the discrete nature of the Likert item measuring the outcomes of the two variables perceived informativeness and perceived difficulty, we use cumulative link models (Bürkner & Vuorre, 2019). These models assume a latent normally distributed variable with non-equidistant thresholds. Exceeding these thresholds leads to the next step on the Likert scale. Within these models, we use random intercepts (Tutz & Hennevogl, 1996) to account for the within-person design.

The R code for all these models is documented in the Reproducible Documentation of Analysis (RDA) file, using the newest available R version (R Core Team, 2023) .

4.3.3. Estimation procedure

To get point estimates along with measures of their uncertainty for the specified models above, we leverage Bayesian estimation techniques implemented in the probabilistic programming language Stan (Stan Development Team, 2024) using the interface of the R package brms (Bürkner, 2017). As for the present research questions, no solid expert

knowledge is available, and therefore we use the brms default uninformative priors for all parameters to ensure that the point estimates are completely driven by the data. Convergence and stability of the Hamiltonian Monte Carlo sampling algorithm is achieved by using four different chains with at least 1000 warm-up iterations, 2000 iterations and 4000 post-warmup draws. We monitor this behavior using the criteria $\hat{R} < 1.05$ (Vehtari et al., 2021) and $ESS > 1000$ (Bürkner, 2017).

4.3.4. Inference criteria for rejecting or accepting parameter values

We will use Bayesian estimation to approximate point estimates and their 95% credibility intervals (CI) within random intercept models to account for the within-person design. To gather evidence for either substantial or negligible effects, we define a region of practical equivalence (ROPE) around the null value (Kruschke, 2018). The limits of the ROPE for mean differences will be set at ± 0.1 around zero, referring to a value of half of Cohen's convention for a small effect (Cohen, 1988; Kruschke, 2018). For distributional effects, we will derive the ROPE similarly: If the second standard deviation deviates less than 10% from the other, then we will consider this as a null effect.

We can derive from the science communication literature that effect size communication can be improved, but only in a limited way. Schmidt et al. (in press) found that providing enrichment options (signaling or benchmarking) next to half eye plots only leads to negligible effects in terms of accuracy, perceived value and informativeness (Schmidt et al., in press). Other empirical research in the visualization research realm indicates that design changes, influencing the way uncertainty is depicted, also only create small effects (Hullman et al., 2015; Kale et al., 2021). In other words, based on previous

research, we expect to find rather small effect size values, and therefore we set our ROPE in narrow boundaries are the null. For the HDI + ROPE decision rule, there are three outcomes possible (cf. Kruschke, 2018): First, the CI can be inside the ROPE, which would provide evidence for a negligible effect. Second, the CI can be completely outside the ROPE, which would provide evidence for an at least substantial effect. Third, the CI can be partially inside the ROPE. In this case, the results are inconclusive.

5. Pilot Data

We conducted a pilot study with $N = 20$ English-speaking teachers from the UK and USA, using participants from the online panel provider Prolific ($M_{\text{age}} = 42.3$ years). We ran this pilot study to test the validity of our measurements in a rather under-explored population (teachers), and to gain insights into the distribution of our dependent variables to deduct strategies for the data analysis in the main study.

5.1. Results

Teachers in the visual condition (see Figure 1, first violin plot in the top-left corner) were on average quite accurate in their effect size estimations ($M = 0.08$, $SD = 0.65$), whereas teachers in the verbal condition showed on average substantial error ($M = 0.37$, $SD = 0.43$; Rank-biserial correlation $= -0.44$, 95% $CI = [-0.56, -0.30]$)). The difference between the two conditions regarding the outcome rating error was also reflected in the estimation of the unstandardized regression coefficient (Presentation mode (Text) = .32, $CI = [0.21, 0.43]$, see Table 1). As 100% of the 95% Credibility Interval (CI) was outside the ROPE, the model provides strong evidence for an at least substantial effect.

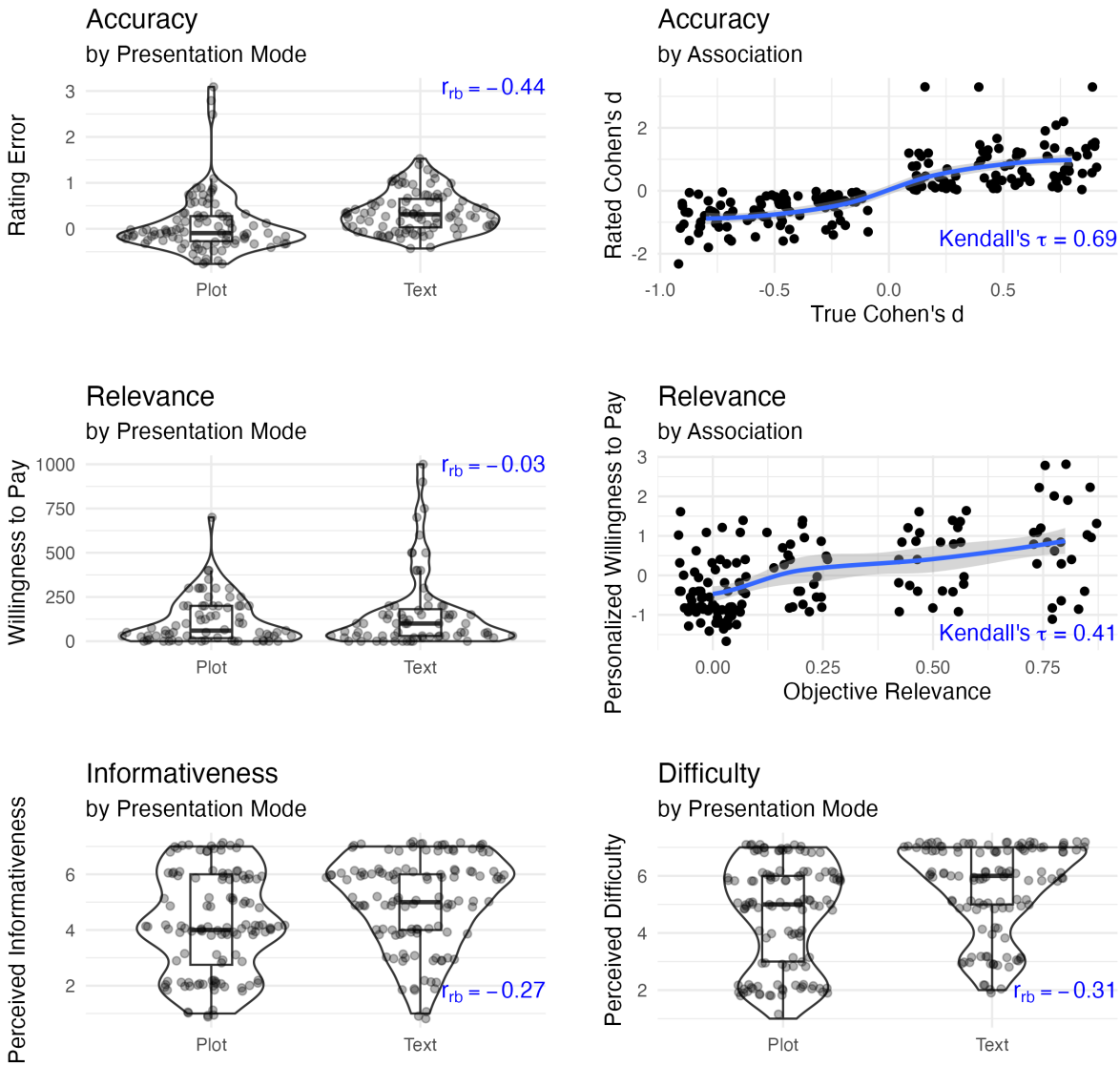

Figure 1

Overview of the four dependent variables

Across both conditions, teachers showed that they can distinguish between small and large effect size values, as depicted in the first jitter plot in the top-right corner ($\tau = 0.69$). The prediction of the standard deviation of the rating error (Difference in SD, that is; the noisiness of teachers' estimations) with dummy coded variables was inconclusive, though: The model implies a -0.02 smaller standard deviation in the verbal group ($SD_{verbal} = 0.38$, $SD_{visual} = 0.40$), but 74.96% of the 95% CI of $[-0.12, 0.08]$ was within the ROPE.

Regarding the relevance variable, it can be stated that teachers were inclined to spend slightly more money for an educational intervention when it was depicted verbally ($M = 147$, $SD = 269$) than visually ($M = 121$, $SD = 233$; Rank-biserial correlation = -0.03, 95% $CI = [-0.17, 0.12]$)), as shown in Figure 1, second violin plot. This relationship was also reflected in the estimation of the unstandardized regression coefficient (Presentation mode (Text) = .39, $CI = [.13, .65]$, Table 1). Additionally, as depicted in the second jitter plot in Figure 1, there was a positive correlation of $\tau = .41$ between the objective relevance and the personified relevance, indicating that respondents were increasingly willing to spend money on an intervention as the true effect size value increased. This association was also reflected in the estimation of the unstandardized regression coefficient (Objective Relevance = 1.68, $CI = [1.25, 2.09]$, Table 1). As 100% of the 95% CI was outside the ROPE for both findings, the two models for the dependent variable relevance provide strong evidence for an at least substantial effect.

As shown in the third violin plot in Figure 1, teachers perceived verbal effect sizes ($M = 5.09$, $SD = 1.61$) as more informative than visual ones ($M = 4.24$, $SD = 1.81$; Rank-biserial correlation = -0.27, 95% $CI = [-0.40, -0.13]$). This difference between the two presentation modes was also mirrored in the estimation of the unstandardized

regression coefficient (Presentation mode (Text) = .98, $CI = [0.68, 1.28]$, Table 1). As 100% of the 95% CI was outside the ROPE, this model, likewise, provides strong evidence for an at least substantial effect.

As illustrated in the last violin plot of Figure 1, in the bottom-right corner, teachers perceived verbal effect sizes ($M = 5.58$, $SD = 1.58$) as less difficult to understand than visual ones ($M = 4.63$, $SD = 1.86$; Rank-biserial correlation = -0.31, 95% $CI = [-0.44, -0.17]$). This difference between the visual and verbal presentation mode can, once again, be depicted in the estimation of the unstandardized regression coefficient (Presentation mode (Text) = 1.29, $CI = [0.96, 1.61]$, Table 1). As 100% of the 95% CI was outside the ROPE, this model provides strong evidence for an at least substantial effect.

Table 1

Overview of the four specified Models

Variable	Accuracy	Relevance	Informativeness	Difficulty
Presentation mode (Text)	0.32	0.39	0.98	1.29
	[0.21, 0.43]	[0.13, 0.65]	[0.68, 1.28]	[0.96, 1.61]
Difference in SD	-0.02	—	—	—
	[-0.12, 0.08]	—	—	—
Objective Relevance	—	1.68	—	—
	—	[1.25, 2.09]	—	—
Num.Obs.	189	144	240	240
R^2	0.366	0.328	0.675	0.708

Note. Intercept parameters for the ordinal models for the outcome variables perceived informativeness & perceived difficulty are not shown. Brackets indicate 95% Credibility Intervals. Cells without values inform that the metric is not applicable to the specific variable.

6. Main Study

7. Sampling plan

The sample will consist of English-speaking teachers from the UK and the USA, and participants will again be recruited via Prolific, as this online panel seems to warrant high data quality (Douglas et al., 2023). The sample size is determined by calculating a Power (Precision) Analysis within the R package brms. According to our analysis, we should include $N = 60$ participants in our study to detect at least small (> 0.1 SD) effects. The calculation and derivation of the sample size and the corresponding R code can also be found in the RDA file.

8. Analysis Plan

We will use the same procedures, identical R code, for data preprocessing, model parametrization, parameter estimation and inference criteria for rejecting/accepting parameter values as described in the pilot study. Therefore, we will pre-register the code from the pilot study as the analysis code for the main study.

9. References

Bauer, J., & Kollar, I. (2023). (Wie) kann die Nutzung bildungswissenschaftlicher Evidenz Lehren und Lernen verbessern? Thesen und Fragen zur Diskussion um evidenzorientiertes Denken und Handeln von Lehrkräften [(How) can the use of educational evidence improve teaching and learning? Theses and questions on the discussion about evidence-orientated thinking and action by teachers]. *Unterrichtswissenschaft*, *51*(1), 123–147. https://doi.org/10.1007/s42010-023-00166-1

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, *99*(2), 332–340. https://doi.org/10.1037/a0034745

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/https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/https://doi.org/10.1177/2515245918823199

Chua, H. F., Yates, J. F., & Shah, P. (2006). Risk avoidance: Graphs versus numbers. *Memory & Cognition*, *34*(2), 399–410. https://doi.org/10.3758/BF03193417

Coe, R. (2002, September 12-14). *It's the effect size, stupid. What effect size is and why it is important* [Paper presentation]. British Educational Research Association Annual Conference, University of Exeter, England.

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

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), Article e0279720. https://doi.org/10.1371/journal.pone.0279720

European Commission. (2007). *Improving the quality of teacher education.* https://op.europa.eu/en/publication-detail/-/publication/1470f875-50bb-4331-a41d-9f1783d1b09c/language-en

Ferguson, L. E. (2021). Evidence-informed teaching and practice-informed research. *Zeitschrift für Pädagogische Psychologie*, *35*(2-3), 199–208. https://doi.org/10.1024/1010-0652/a000310

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

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). W. W. Norton & Company.

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*(2), 314–316. https://doi.org/10.1037/0021-9010.79.2.314

Hanel, P. H., & 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/https://doi.org/10.1177/0963662519834193

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Hearst, M. A. (2023). Show it or tell it? Text, visualization, and their combination. *Communications of the ACM*, *66*(10), 68–75. https://doi.org/10.1145/3593580

Hofman, J. M., Goldstein, D. G., & Hullman, J. (2020, April 25-30). *How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results* [Paper presentation]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA. https://doi.org/10.1145/3313831.3376454

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*(2), 227–240. https://doi.org/10.1177/0013164402062002002

Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS ONE*, *10*(11), Article e0142444. https://doi.org/10.1371/journal.pone.0142444

Kale, A., Kay, M., & Hullman, J. (2021). Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 272–282. https://doi.org/10.1109/TVCG.2020.3030335

Kim, Y.-S., Hofman, J. M., & Goldstein, D. G. (2022, April 29-May 5). *Putting scientific results in perspective: Improving the communication of standardized effect sizes* [Paper presentation]. CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA. https://doi.org/https://doi.org/10.1145/3491102.3502053

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

Kruschke, J. K. (2018). Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

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/https://doi.org/10.3102/0013189X20987856

Mastrich, Z., & Hernandez, I. (2021). Results everyone can understand: A review of common language effect size indicators to bridge the research-practice gap. *Health Psychology*, *40*(10), 727–736. https://doi.org/10.1037/hea0001112

Matejka, J., & Fitzmaurice, G. (2017, May 06-11). *Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing* [Paper presentation]. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, Colorado, USA. https://doi.org/https://dl.acm.org/doi/10.1145/3025453.3025912

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361

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

Merk, S., Groß Ophoff, J., & Kelava, A. (2023). Rich data, poor information? Teachers' perceptions of mean differences in graphical feedback from statewide tests. *Learning and Instruction*, *84*, Article 101717. https://doi.org/10.1016/j.learninstruc.2022.101717

Michal, A. L., & Shah, P. (2024). A practical significance bias in laypeople's evaluation of scientific findings. *Psychological Science*, Article 09567976241231506. https://doi.org/10.1177/09567976241231506

Nelson, J., & Campbell, C. (2017). Evidence-informed practice in education: Meanings and applications. *Educational Research*, *59*(2), 127–135. https://doi.org/10.1080/00131881.2017.1314115

O'Keefe, D. J. (2017). Misunderstandings of effect sizes in message effects research. *Communication Methods and Measures*, *11*(3), 210–219. https://doi.org/10.1080/19312458.2017.1343812

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, *3*(1), 1–25. https://doi.org/10.1186/s41235-018-0120-9

Pellegrini, M., & Vivanet, G. (2021). Evidence-based policies in education: Initiatives and challenges in europe. *ECNU Review of Education*, *4*(1), 25–45. https://doi.org/10.1177/2096531120924670

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *57*(3), 221–237. https://doi.org/10.1037/h0087427

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Santesso, N., Glenton, C., Dahm, P., Garner, P., Akl, E. A., Alper, B., Brignardello-Petersen, R., Carrasco-Labra, A., De Beer, H., Hultcrantz, M., Kuijpers, T., Meerpohl, J., Morgan, R., Mustafa, R., Skoetz, N., Sultan, S., Wiysonge, C., Guyatt, G., & Schünemann, H. J. (2020). Grade guidelines 26: Informative statements to communicate the findings of systematic reviews of interventions. *Journal of Clinical Epidemiology*, *119*, 126–135. https://doi.org/10.1016/j.jclinepi.2019.10.014

Schmidt, K., Schneider, J., Bohrer, K., & Merk, S. (in press). Communicating effect sizes to teachers: Exploring different visualizations and their enrichment options. *Zeitschrift für Psychologie*.

Schmidt, K., Edelsbrunner, P. A., Rosman, T., Cramer, C., & Merk, S. (2023). When perceived informativity is not enough. How teachers perceive and interpret statistical results of educational research. *Teaching and Teacher Education*, *130*, Article 104134. https://doi.org/https://doi.org/10.1016/j.tate.2023.104134

Schuetze, B. A., & Yan, V. X. (2023). Psychology faculty overestimate the magnitude of cohen's d effect sizes by half a standard deviation. *Collabra: Psychology*, *9*(1), Article 74020. https://doi.org/10.1525/collabra.74020

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, *55*(1), 21–31. https://doi.org/10.1080/00461520.2019.1611432

Stan Development Team. (2024). *Stan modeling language users guide and reference manual*.

Stone, E. R., Yates, J. F., & Parker, A. M. (1997). Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied*, *3*(4), 243–256. https://doi.org/10.1037/1076-898X.3.4.243

Stosic, M. D., Murphy, B. A., Duong, F., Fultz, A. A., Harvey, S. E., & Bernieri, F. (2024). Careless responding: Why many findings are spurious or spuriously inflated. *Advances in Methods and Practices in Psychological Science*, *7*(1), Article 25152459241231581. https://doi.org/10.1177/25152459241231581

Stukas, A. A., & Cumming, G. (2014). Interpreting effect sizes: Toward a quantitative cumulative social psychology. *European Journal of Social Psychology*, *44*(7), 711–722. https://doi.org/10.1002/ejsp.2019

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, *16*(4), 717–724. https://doi.org/10.1177/1745691620966796

Thomm, E., Sälzer, C., Prenzel, M., & Bauer, J. (2021). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Zeitschrift für Pädagogische Psychologie*, *35*(2-3), 173–184. https://doi.org/10.1024/1010-0652/a000301

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. https://doi.org/10.1038/381520a0

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, *22*(5), 537–557. https://doi.org/https://doi.org/10.1016/0167-9473(96)00004-7

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. https://doi.org/10.2307/1165329

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R^ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). https://doi.org/https://doi.org/10.1214/20-BA1221

Von Hippel, P. (2024). Multiply by 37 (or divide by 0.027): A surprisingly accurate rule of thumb for converting effect sizes from standard deviations to percentile points. *Educational Evaluation and Policy Analysis*, Article 01623737241239677. https://doi.org/10.3102/01623737241239677

Wadhwa, M., Zheng, J., & Cook, T. D. (2024). How consistent are meanings of "evidence-based"? A comparative review of 12 clearinghouses that rate the effectiveness of educational programs. *Review of Educational Research*, *94*(1), 3–32. https://doi.org/https://doi.org/10.3102/00346543231152262

Wilmer, J., & Kerns, S. (2022). How bar graphs deceive: Readout-based measurement reveals three fallacies. *Journal of Vision*, *22*(14), Article 3968. https://doi.org/10.1167/jov.22.14.3968

Xiong, C., Stokes, C., Kim, Y.-S., & Franconeri, S. (2023). Seeing what you believe or believing what you see? Belief biases correlation estimation. *IEEE transactions on visualization and computer graphics*, *29*(1), 493–503. https://doi.org/10.1109/TVCG.2022.3209405

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, *120*(33), Article e2302491120. https://doi.org/10.1073/pnas.2302491120

Appendix

Stimulus Examples

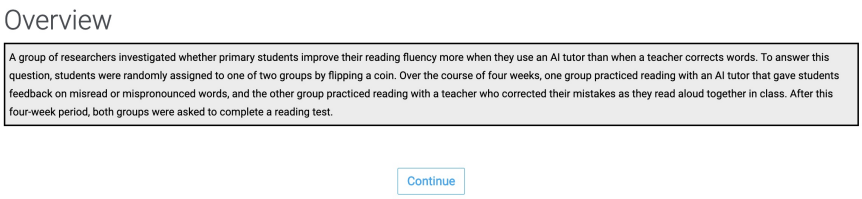

Figure A1

Example of a research report: Topic 1 AI feedback

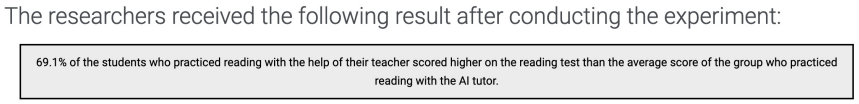

Figure A2

Example of a Cohen's U3 phrase: Topic 1 AI feedback

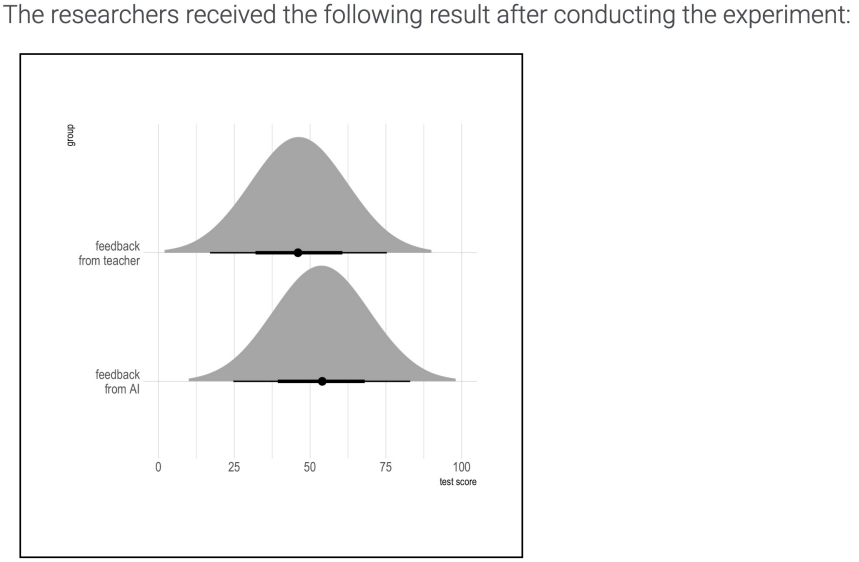

Figure A3

Example of a half-eye plot: Topic 1 AI feedback