Journal of Research on Educational Effectiveness

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

The Meta-Analytic Rain Cloud Plot: A New Approach to Visualizing Clearinghouse Data

Kaitlyn G. Fitzgerald & Elizabeth Tipton

To cite this article: Kaitlyn G. Fitzgerald & Elizabeth Tipton (2022): The Meta-Analytic Rain Cloud Plot: A New Approach to Visualizing Clearinghouse Data, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2022.2031366

To link to this article: https://doi.org/10.1080/19345747.2022.2031366

Published online: 14 Mar 2022.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2022.2031366

METHODOLOGICAL STUDIES

The Meta-Analytic Rain Cloud Plot: A New Approach to Visualizing Clearinghouse Data

Kaitlyn G. Fitzgerald[a] and Elizabeth Tipton[b]

[a]Azusa Pacific University, Azusa, California, USA; [b]Northwestern University, Evanston, Illinois, USA

ABSTRACT

As the body of scientific evidence about effective policies and practices grows, so does the need to effectively communicate that evidence to policy-makers and practitioners. Clearinghouses have emerged to facilitate the evidence-based decision-making process for education practitioners. While the results and methods for developing and analyzing the data in clearinghouses are based upon rigorous and scientific study, there has been little rigor or empirical effort to determine effective ways of presenting that evidence to practitioners. In this paper, we present a new visualization for clearinghouse data, called a Meta-Analytic Rain Cloud (MARC) Plot, designed based on evidence from the data visualization and statistical cognition literatures. We evaluate the efficacy of this visualization in a statistical cognition experiment and find that compared to three other visualizations used in practice, the MARC Plot is more effective in helping participants correctly interpret evidence (0.76, 0.43, and 0.43 standard deviation improvements respectively; each $p < 0.05$, corrected for multiple comparisons). To our knowledge, this is one of the first studies providing evidence regarding how to best present the type of information found in clearinghouses.

ARTICLE HISTORY
Received 24 May 2021
Revised 13 September 2021
Accepted 28 October 2021

As the body of scientific evidence about what works in education grows, so does the need to effectively communicate that evidence to policy-makers and practitioners. Websites and clearinghouses such as the *What Works Clearinghouse* (WWC), *Evidence for ESSA*, *Blueprints for Healthy Youth Development*, and the *Education Endowment Foundation* (EEF) have emerged to facilitate the evidence-based decision-making process for these policy-makers and practitioners. These clearinghouses have taken on the non-trivial task of distilling often complex research findings to non-researchers. Among other things, this often involves reporting effect sizes, statistical uncertainty, and meta-analytic results, and this information is often reported visually. To do so, for example, the WWC provides readers with bar plots that indicate the direction and size of the effect, as well as the sample size and an overall summary statistic.

Clearinghouses contain hundreds and thousands of results from research using rigorous methods for determining the effectiveness of educational curricula and programs. These methods are rigorous in their ability to discern causality, as well as in terms of

CONTACT Kaitlyn G. Fitzgerald ✉ kfitzgerald@apu.edu Azusa Pacific University, Azusa, CA, USA

their statistical properties (e.g., bias, statistical power). Similarly, the coding protocols and methods for extracting effect sizes and standard errors from these studies for inclusion in clearinghouses have also been rigorously defined and studied. For example, this includes methods for creating comparable effect sizes from different reported data, as well as methods for appropriately accounting for clustering in standard errors (e.g., Hedges, 2007).

It is interesting, then, that while the results and methods for developing and analyzing the data in clearinghouses is based upon rigorous and scientific study, that the methods for presenting and visualizing this data for policymakers and practitioners are typically not. This is not to say that research on data visualization does not exist, but instead that very little of such research has been conducted in the policymaker and practitioner communities, in general, and in education specifically. This paper seeks to provide one of the first studies in this area. Our hope is that in addition to providing a new visualization, our paper also encourages others to develop and study visualizations as well.

While there is little research in the education context regarding visualization, there is a substantial literature in psychology demonstrating cognitive biases in information visualization (Dimara et al., 2020) and that poor statistical reasoning and statistical misconceptions are widespread and persistent among students, lay people, and researchers (Belia et al., 2005; Garfield & Ahlgren, 1988; Kühberger et al., 2015; Tversky & Kahneman, 1974). Furthermore, in computer science and psychology, there is a substantial literature regarding best practices for data visualization (Cleveland & McGill, 1984; Correll & Gleicher, 2014; Franconeri et al., 2021; Qiao & Hullman, 2018; Schild & Voracek, 2015). Based upon these best practices, this paper proposes a new visualization—a Meta-Analytic Rain Cloud (MARC) Plot—intended to communicate the type of evidence found in clearinghouses to education practitioners' for the purposes of decision-making.

In this paper, we begin by discussing a framework for visualizing meta-analytic findings for decision-making in education and a review of forest plots and their variations, with a particular focus on use in education research, in Section I. In Section II, we introduce the newly proposed MARC plot. We then evaluate the efficacy of this plot in a statistical cognition experiment. We explain the methods for this study in Section III and the results in Section IV. We conclude with a discussion of the findings and questions for future research in Section V.

Criteria for Evaluating Visualizations

When considering what data visualization(s) may be most appropriate for communicating education research findings, it is important to consider the *purpose* of the visualization and what information we hope it conveys. We argue that one of the primary functions of providing evidence from education research via clearinghouses and other forums is to facilitate *decision-making* by stakeholders and practitioners. That is, most people engaging with the evidence need to use that evidence to make some type of decision, often regarding whether or not to purchase or adopt a given curriculum or program. Certainly, statistical evidence is only one factor in these decisions, with others including cost, feasibility, values, priorities and the constraints of local contexts.

However, this fact does not negate the necessity of effectively and accurately communicating the statistical evidence, so that users can weigh it appropriately with these other forms of evidence in their decision-making process.

To that end, we begin with the normative question of how people *should* evaluate the statistical evidence, holding all other factors constant. That is, given a set of meta-analytic data—e.g., clearinghouse data—what is the statistically appropriate way to judge the evidence? Here we can turn to statistical research on meta-analysis to guide us. We know that when meta-analyzing results from multiple studies, if we are interested in a summary effect, the most precise summary effect is an average effect calculated based on inverse variance weights (Hedges, 1985). Thus, this average effect size should be the basis for decision-making, above and beyond each of the individual studies.

Third, decisions should consider the magnitude of the true effect. That is, when the effect size δ is large (say greater than some c) then the intervention is considered worthwile, whereas otherwise it is not. Note that the threshold c is user-defined and often subjective. In some cases for low-cost interventions it may simply be that $c = 0$; that is, the user would decide to implement so long as the treatment can be shown to simply not be harmful.[1] In other more cost- or time-intensive interventions, c is likely to be larger. We are not concerned here with defining or modeling c; we are simply making the point that normative decision-making considers the *magnitude of the true effect* in relation to *some threshold.*

Finally, decisions should consider the statistical uncertainty of the summary effect estimate. While there are many statistically valid ways to quantify uncertainty, because the magnitude of the effect is pertinent to decision making, we assert that uncertainty should be reported in such a way that provides intuition about the range of plausible values of the true effect. For example, a confidence interval is more informative than a p-value because it communicates the uncertainty that remains *around the point estimate.*

In combination, this suggests that visualizations intending to communicate the kind of meta-analytic data found in clearinghouses ought to visually emphasize:

1. Precise studies over imprecise studies,
2. Summary effects over individual effects,
3. The magnitude of the summary effect, and
4. The uncertainty of the summary effect via a range of plausible values.

While education policy-makers and practitioners cannot be expected to be familiar with meta-analytic methods and normative statistical decision-making—and in fact, precisely *because* they may not know, for example, that they should weight studies inversely proportional to their variances—it is important to provide visualizations that can intuitively guide people toward this type of meta-analytic thinking.

Finally, we expect that some readers will take issue with our focus on the summary effect in decision making. While we agree that ideally decision-making could be locally contextualized—e.g., based off of predictions about how effective the treatment is

[1] Note this is for the case where 0 corresponds to "no effect," such as when the treatment effect of interest is a standardized mean difference. Other effect sizes may have a different "no effect" threshold.

expected to be in *their context* (i.e. school, district, etc.)—the current landscape of available evidence does not yet allow for this type of predictive decision-making. Many clearinghouses only provide individual effect sizes or vote counting based on these effect sizes. The WWC, for example, has only recently shifted from vote-counting to use of fixed-effects meta-analysis, producing an average effect size for each intervention. Here the use of fixed-effects meta-analysis is necessitated by the small number of effect size estimates provided for most interventions (most with fewer than 5 studies). Thus, while we are hopeful that advances in statistical and data science methodology and improved research practice will facilitate locally contextualized decision-making in the future, we focus for now on decisions that can be made from the evidence that is readily available: that is, evidence about the *average treatment effect*.

Visualizations of Clearinghouse Data

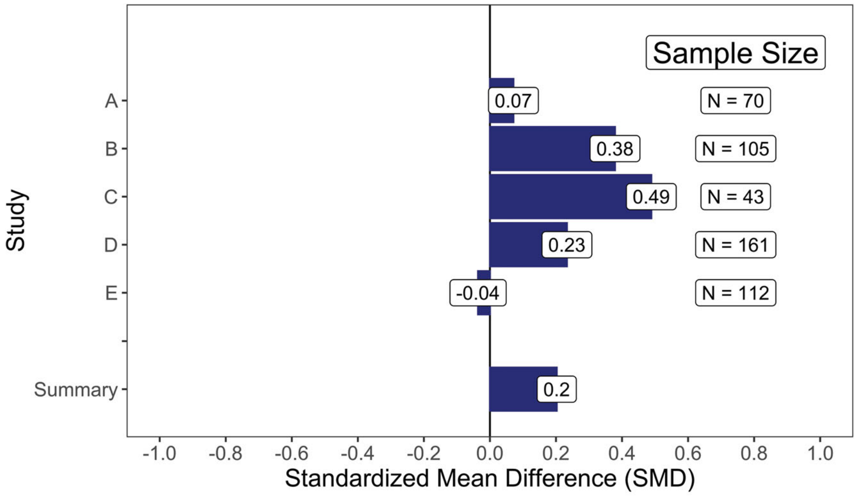
In this section, we begin by providing an overview of the current visualizations available to or used by clearinghouses when conveying evidence to policy-makers and practitioners. We break this apart into bar plots (Section A), forest plots (Section B), and rain-forest plots (Section C). Then, based upon principles of data visualization, we introduce a new plot—the Meta-Analytic Rain Cloud (MARC) plot (Section D). Throughout, we provide figures illustrating these plots. Importantly, the same underlying data is included in each plot, so that only the visualization itself changes.[2]

Bar Plots

The What Works Clearinghouse (WWC) appears to be one of the only clearinghouses that uses visualizations to display meta-analytic data on their dashboard. Figure 1 provides a version of the type of bar plot used by the WWC. This figure depicts the results of five different studies, as well as the summary effect (at the bottom). Here the magnitude of an effect size is depicted by the length of the bar and the *x*-axis position of the end of the bar. The sample size for each study is also included. This type of bar plot is likely familiar to education practitioners and relatively simple to read.

However, given the four goals outlined in Section I, we anticipate several concerns. First, effect sizes with the largest magnitude draw the most visual attention, regardless of whether or not they were precisely estimated. For example, Study C should be given the least consideration among the studies in Figure 1 because it is the least precisely estimated, but it draws the most visual attention in this case. Similarly, the summary effect is not visually emphasized or distinguished. Third, while the sample size is given, this is not always an accurate means of understanding precision. For example, if the effect sizes arises from a cluster-randomized trial, displaying total sample size can be a very misleading proxy for precision, as precision depends more on the total number of

[2] Throughout our paper and experiment, we display effect sizes as standardized mean differences. Some clearinghouses use other metrics such as the Improvement Index (used by WWC) or months of learning (used by EEF). While choosing a readily interpretable metric is no doubt important for effective communication, the task of empirically determining which metric is most effective for communication to lay audiences is a research question in and of itself and one outside the scope of the present study (Lortie-Forgues et al., 2021).



This figure displays SMD estimates from 5 studies.
The summary SMD is a weighted average of the evidence from the 5 studies.

SMD > 0 indicates that the new curriculum increased student scores, SMD < 0 indicates it decreased scores, and SMD = 0 indicates it had no effect on scores.

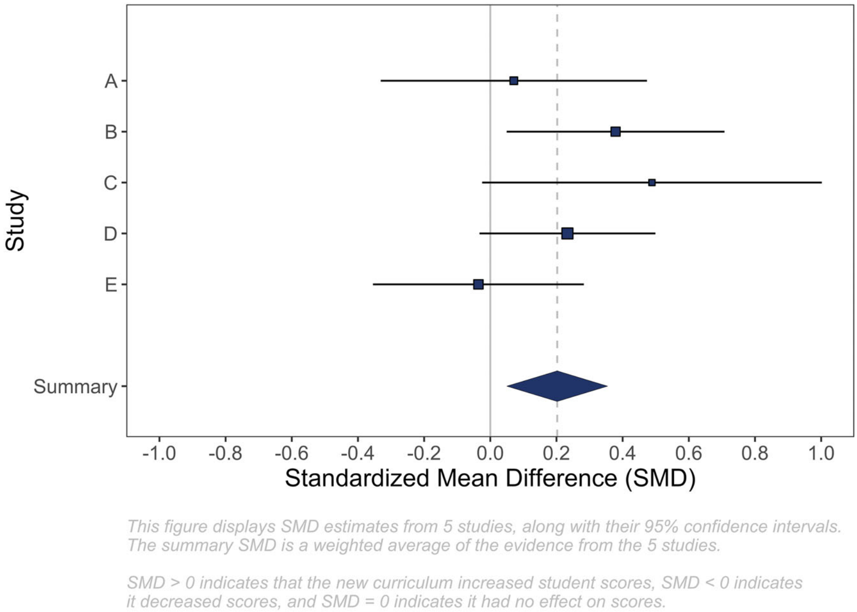
Figure 1. Bar Plot (BP).

clusters as opposed to the total number of *people*. Finally, it is well-established in the data visualization literature that bar plots are subject to "within-the-bar bias," where users tend to think a value that falls within the area of the bar is more likely than a value that falls beyond it, even if those values are equidistant from the estimated effect and therefore equally likely (Newman & Scholl, 2012). This type of bias is particularly problematic when decision-making is an end-goal of the information you are providing.

Forest Plots

While not typically displayed in dashboards, some clearinghouses such as the Campbell Collaboration use another visualization—the forest plot (FP)—to depict meta-analytic evidence in their full reports. Figure 2 provides a visualization of a forest plot using the same data as found in Figure 1. These are similar in that there are results of each of the five studies indicated first, with a summary effect at the bottom. In the FP, each effect size is represented by a square, where the magnitude of the estimate is encoded as the x-axis position of the center of the square, and the area of each square is proportional to the study's weight in the meta-analysis. These weights are set to be inversely proportional to the within-study sampling variance of the effect estimate $\hat{\delta}_j$, such that for study j, $w_j = 1/V(\hat{\delta}_j)$. A study with a smaller variance (i.e. higher precision) thus receives a larger weight and thus is represented by a larger square.

The bars on each effect size estimate represent its 95% confidence interval. Since both the weight and the confidence interval are functions of within-study variance, a study's precision is encoded as both the size of the square and the width of the bars. Statistically, a study with the largest weight will also inherently have the narrowest confidence interval, so the most precisely estimated study is visually identifiable by having



This figure displays SMD estimates from 5 studies, along with their 95% confidence intervals. The summary SMD is a weighted average of the evidence from the 5 studies.

SMD > 0 indicates that the new curriculum increased student scores, SMD < 0 indicates it decreased scores, and SMD = 0 indicates it had no effect on scores.

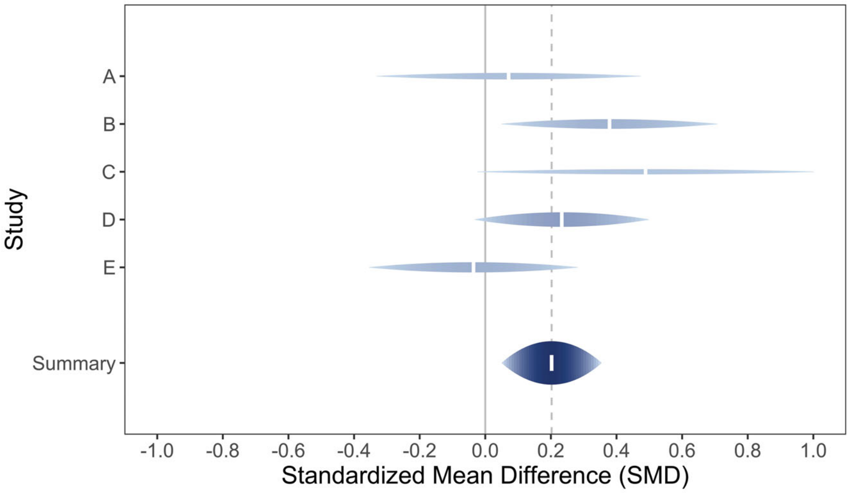
Figure 2. Forest Plot (FP).

the largest square and the shortest bars among the set of studies being meta-analyzed. The summary effect, computed as a precision-weighted average, is depicted as a diamond in a FP, with the magnitude of the effect encoded as the center of the diamond and its 95% confidence interval as the width of the diamond.

There have been repeated criticisms of FPs (Anzures-Cabrera & Higgins, 2010; Barrowman & Myers, 2003; Jackson, 2008; Schild & Voracek, 2013; Schriger et al., 2010) which Schild and Voracek (2015) summarize as including three primary faults. First, imprecise studies receive un-due visual attention because their wide confidence intervals take up more space in the visualization. Second, encoding the square size proportional to study weight makes both heterogeneity between effect sizes and the exact (center) value of individual effect sizes hard to discern. For example, the horizontal distance between two small squares will appear misleadingly larger (i.e. more heterogeneous) than the distance between two large squares, even if their centers are equally far apart in both cases.

Third, displaying a confidence interval with bars does not provide any visual indication that the likelihood of values is not constant across the confidence interval. When effect size estimators are approximately normally distributed (which is true of standardized mean differences and many others), considering the conditional likelihood of parameter values over the confidence interval constructed from the observed data indicates that values toward the center of the interval are more likely than those toward the lower and upper bounds. However, simple confidence interval bars can misleadingly make this likelihood function appear to follow a uniform distribution and suggest all values within the interval are equally likely[3].

[3] Note we are using the frequentist notion of likelihood here, where given the observed data $\boldsymbol{x}$, $L(\delta_1|\boldsymbol{x}) > L(\delta_2|\boldsymbol{x})$ indicates that the sample we observed was more likely to have occurred if $\delta = \delta_1$ than if $\delta = \delta_2$; that is, δ_1 can be



This figure displays SMD estimates from 5 studies, along with their 95% confidence intervals. The summary SMD is a weighted average of the evidence from the 5 studies.

SMD > 0 indicates that the new curriculum increased student scores, SMD < 0 indicates it decreased scores, and SMD = 0 indicates it had no effect on scores.

Figure 3. Rain Forest Plot (RFP).

Finally, to these criticisims, we add one more. While bar plots are a common data visualization, which policy-makers and practitioners are likely to have prior knowledge and experience with, FPs are not. This means that reading a FP correctly likely requires training. We anticipate, therefore, that as a result of the *curse of expertise* (Xiong et al., 2020) the ability to extract data from a FP depends upon one's training in research and experience with meta-analysis.

Rainforest Plots

Schild and Voracek (2015) proposed an alternative to FPs, which they name rainforest plots (RFPs).[4] Although RFPs are not yet widely used in practice, Schild & Voracek's statistical cognition experiment suggests they are an improvement over FPs, so we include them here. Figure 3 provides a rainforest plot that displays the same underlying data as Figures 1 and 2. Note that again, the data is ordered in the same way, with each of the five study findings followed by the summary effect at the bottom. Instead of squares, however, each study's findings are represented by a raindrop (proposed by Barrowman & Myers, 2003), where the magnitude of the effect size is depicted by the *x*-axis position of the white line at the center of the raindrop. The raindrop is composed by drawing the likelihood curve for the effect size parameter over its 95% confidence interval constructed from the observed data. The curve is then mirrored across

interpreted as being a more plausible value of δ than δ_2. This should not be confused with the Bayesian interpretation that treats δ as random and considers the *probability* that it takes on certain values.

[4] They also proposed thick forest plots, which performed comparably to RFPs, but we choose to use only the rainforest plots because we find them to be better suited for the decision-making described above that occurs in the education context.

an invisible horizontal axis, giving the appearance of a raindrop. The height at any given point of a raindrop is therefore proportional to the likelihood of that value relative to all other values in the interval. The effect size estimate itself is the most likely value in the interval, and therefore the highest point of the raindrop occurs at the center of a symmetric interval. Drawing from the concept of density strips proposed by (Jackson, 2008) to display precision via shading, the likelihood values are also encoded as color saturation, so the most likely values appear darker than less likely values, again with the darkest shading occurring at the center of the interval.

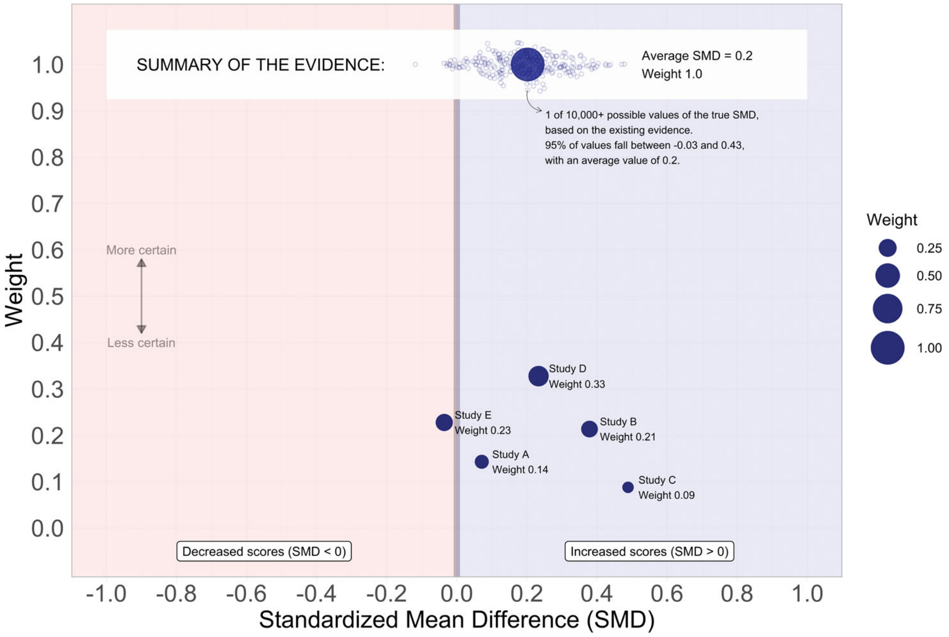
Additionally, when comparing individual-study raindrops relative to all other raindrops, both the relative height and shading is scaled proportional to study weight. That is, more precise studies will have "thicker" and darker raindrops, whereas imprecise studies will have "thinner" and lighter raindrops. This feature is intended to address the first criticism of FPs by making more precise studies more visually striking and less precise studies less so. Clearly marking the center of each interval with a white strip addresses the second criticism, and visually encoding the likelihood by raindrops and shading addresses the third concern.

While the rainforest plot proposed by Schild and Voracek (2015) still uses the traditional diamond from FPs to display the summary effect, the version we propose depicts the summary effect with a raindrop instead, as shown in Figure 3. We think this is important particularly in a decision-making context; because the summary effect is the primary piece of data we hope people base their decisions off of, its visual depiction should contain at least as much information as the individual studies. We also think it is easier to miss the encoding of the summary effect's confidence interval when it is encoded as the width of the diamond than if it is encoded the same way as the individual studies. Furthermore, representing the summary effect by a raindrop results in it being the darkest and largest object on the visualization, drawing attention to it over individual studies, as desired.

While we agree with Schild and Voracek (2015) that RFPs seem to address many of the common critiques of FPs, we expect this advantage may not carry over for audiences beyond trained meta-analytic researchers. Not only is this another instance to be wary of the curse of expertise, often adding more sophisticated and repetitive encodings to further contextualize the data can actually serve as a distraction and confuse the user more than it clarifies.

Meta-Analytic Rain Cloud Plot

Based upon these previous plots and upon the statistical congnition literature, we developed a new plot for displaying clearinghouse evidence; we call this the Meta-Analytic Rain Cloud (MARC) plot. Figure 4 shows the new proposed visualization, which displays the same underlying data as the other three figures. This plot diverges from the previous three in several regards. First, in developing this plot, we noted that all three previous plots use the y-axis to order the studies first (often alphabetically) and then provide the summary effect at the bottom. However, it is well-established from a cognition perspective that visual encodings of *position* (e.g. position on the x- and y-axes) are the easiest for our brains to decode, process, and draw comparisons from (Cleveland &



This figure displays SMD estimates from 5 studies and a corresponding weighted average that summarizes the evidence. Other plausible values for the true SMD - simulated based on the existing evidence and uncertainty - are also displayed.

SMD > 0 indicates that the new curriculum increased student scores, SMD < 0 indicates it decreased scores, and SMD = 0 indicates it had no effect on scores.

Figure 4. Meta-Analytic Rain Cloud (MARC) Plot.

McGill, 1984). For example, our brains are able to more easily determine how two positions relate to one another than we are able to determine how two sizes or two hues or transparencies of color relate to one another. Therefore, it is advantageous to leverage both *x*- and *y*-axis positions to make important features of the data salient. In order to leverage the visual power of a positional encoding, we therefore choose to use the *y*-axis as a continuous scale for studies' relative weights (0.0−1.0). As such, studies with the largest weights will appear higher on the *y*-axis, and the summary effect will appear at the top with a corresponding weight of 1.0.

Second, we add to the visual emphasis on precise studies by also encoding weight as dot size. We intentionally choose not to depict the uncertainty for individual study effects, but only for the summary effect. Not only is this visually preferable because it simplifies the encodings (Qiao & Hullman, 2018), but we also argue that it promotes better meta-analytic thinking; depicting the confidence intervals of each individual effects may lead to the undesirable behavior of "vote-counting," where people evaluate the evidence by considering how many of the individual effects were statistically significant from zero. Vote-counting leads to poor meta-analytic reasoning and decision-making (Hedges, 1985). While confidence intervals of individual effects may be of interest to researchers who want to discern if variation in observed effects is due to sampling variance or true heterogeneity, this level of statistical nuance is unnecessary and in fact distracting for lay people (Qiao & Hullman, 2018). Therefore, we choose to depict an individual study's effect size magnitude and relative weight only, and do not display confidence intervals or other forms of precision except on the summary effect.

Third, we also seek to make the process of *how* meta-analytic data is synthesized more salient. For example, it is not obvious in any of the three previous approaches *how* the studies are combined to obtain the summary diamond or raindrop respectively. In our new MARC plot, we make this explicit and show exactly how each study is weighted. We do this by encoding the weight directly—as *y*-axis position, as dot size, and with a label for the actual study weight. Furthermore, we add the informal legend "More certain ←→ Less certain" to draw out the meta-analytic idea that studies receive higher weights *because* they are more certain. We use "certain" instead of "precise" here because we expect the common language use of "certain" may be more likely to evoke intuitive interpretations that align with the desired statistical interpretation. Depicting weight directly in this way prevents the pitfall mentioned previously where studies with the largest sample sizes may not necessarily be most precise (e.g. in a cluster-randomized trial).

Fourth, taking pointers from data journalists and graphic design experts, we included highlights and annotations to guide the user as much as possible (Qiao & Hullman, 2018). For example, we shaded the region of negative *x*-values red and the region of positive *x*-values as blue and provided labels to aid in the interpretation of positive SMD values as desirable. We also placed visual emphasis on the summary effect by making it the largest dot (with its relative weight of 1.0) and by outlining and labeling it at the top of the visualization. These principles can be found, too, in the depiction of precision. Since the precision of the summary effect is key to normative decision-making, we therefore also depicted it in the MARC plot. However, we avoid using traditional confidence interval bars, because it is well-documented that they are poorly understood and elicit widespread misconceptions (Belia et al., 2005; Correll & Gleicher, 2014). Instead, we depict the precision of the summary effect using an adaptation of 538's 2020 Election ball swarm plot instead of a confidence interval (Wiederkehr, 2020).

This "cloud" featured at the top of the new visualization is created by generating random values from a normal distribution with mean and standard deviation equal to the mean and standard error of the summary effect.[5] Values generated in this way have convenient asymptotic properties where the mean is equivalent to the meta-analytic mean, and 95% of the values fall in the 95% confidence interval of the mean. Similar to the summary raindrop in the RFP, the cloud makes visible the statistical property that values toward the center of the interval are more likely. However, we expect the MARC plot to be a more intuitive depiction of this property by displaying individual potential values rather than a likelihood curve. The literature suggests that visualizations that display individual potential outcomes as opposed to aggregate probabilities or densities are better at helping people grasp uncertainty (Fernandes et al., 2018; Hullman et al., 2015). Both the depiction of the dots and the annotation describing what one dot in the rain cloud represents aim to provide a more intuitive and concrete interpretation of the

[5] We generate 10,000 $\hat{\delta}_i \sim N(\hat{\delta}, SE(\hat{\delta}))$, where $\hat{\delta}$ is the meta-analytic estimate based on the existing evidence. The *y*-axis position of each simulated dot is jittered by an amount proportional to the density for that corresponding *x*-value. This results in the pattern of the dots roughly mimicking the Normal pdf, mirrored across the line $y = 1$. See code in Appendix for details.

uncertainty without relying on precise (and often convoluted) statistical language of a confidence interval.[6]

Finally, another potential advantage to this type of display is that it is less likely to evoke dichotomous thinking of statistical significance. Even though the raindrops in the RFP may illuminate the statistical property that values toward the boundaries of the confidence interval are less likely, the raindrops are still prone to being interpreted dichotomously based on whether or not they overlap zero. In a MARC plot, however, even when an effect is significant at the 95% level, a cloud may still have some dots that fall below zero, which is more reflective of the statistical reality that there is still a small likelihood that the true effect is negative.

Methods

In order to evaluate the efficacy of each of these four plots (Section II) at meeting the normative statistical goals (Section I), we conducted a statistical cognition experiment. The study was conducted online via Qualtrics. Our research questions and analysis plan were pre-registered with the Open Science Foundation (OSF) prior to data collection.

Research Questions

In this study, we ask four questions:

1. Are education practitioners able to accurately interpret each of these plots?
2. Which type of visualization leads to most accurate understanding among education practitioners?
3. Do user beliefs about strength of evidence and effectiveness of curriculum vary across types of visualization?
4. Do education practitioners and education researchers interpret forest plots differently?

We hypothesized that education practitioners may have difficulty interpreting the forest plots (both FPs and RFPs), but that the new visualization (MARC) would lead to the most accurate understanding. We anticipated that user beliefs about strength of evidence and effectiveness of curriculum may vary across types of visualization. In particular, we hypothesized that users would rate the evidence most highly when they are not shown uncertainty (i.e. in the bar plots). We hypothesized that education researchers would be able to interpret forest plots better than education practitioners overall.

[6] We acknowledge that generating “possible values” is perhaps more akin to a Bayesian framework, which as a whole can lend itself to more natural interpretations. However, we still think it is advantageous to avoid even the precise language of a Bayesian credible interval, which still relies on notions of aggregate probabilities. For consistency throughout the paper we maintain frequentist language, however we don’t think the plot itself necessitates a strictly frequentist or Bayesian interpretation, and the distinction is largely unecessary when communicating to practitioners. Further, plain language and simplified interpretation is recommended as best practice for communicating to public audiences (Qiao & Hullman, 2018).

Table 1. Participant demographics.

	Practitioner ($n = 83$)	Researcher ($n = 94$)
Gender		
Female	0.627	0.585
Male	0.349	0.287
Other	0.024	0.128
Highest degree*		
Bachelor's	0.060	0.011
Master's	0.675	0.191
Doctorate or professional degree	0.265	0.798
Role in PreK12†		
Teacher	0.747	
Principal	0.120	
Curriculum specialist	0.205	
District superintendent	0.024	
Other	0.590	
Age		
18–24	0.012	0.011
25–34	0.337	0.245
35–44	0.361	0.447
45–54	0.193	0.074
55–64	0.072	0.074
65 or older	0.000	0.043
Did not respond	0.024	0.106

Note. Numbers given as proportions.
*Researchers whose highest degree was Bachelor's or Master's were currently enrolled in a doctoral program.
†Role in PreK12 asked only of practitioners and was select all that apply, so proportions don't add to 1.

Participants

Two samples were recruited for this study: education practitioners and education researchers. We defined education practitioners as those who had at least a Bachelor's degree and self-identified as "educators or education decision-makers employed in the US PreK-12 education system." Education researchers are those who indicated they are employed by a university, think tank, or research organization, have a doctorate degree or are currently enrolled in a doctoral program, and conduct research relevant to education.

Based on a power analysis informed by pilot data[7], we sought to recruit up to 265 education practitioners and up to 80 education researchers, with a minimum requirement of 50 and 12 per sample respectively. We recruited primarily on Twitter and through listservs of professional organizations including the American Educational Research Association (AERA), Harvard University's Strategic Data Project, and networks of alumni, mentor teachers and principals through Northwestern University's School of Education and Social Policy.

Our analysis is based on 177 eligible people who consented and completed the study online via Qualtrics – 83 practitioners and 94 researchers. In total, 270 people consented to the study, but 41 were screened out for not meeting inclusion criteria, and an additional 52 people were excluded for not participating in the experiment portion of the survey ($n = 49$), completing the survey in under 3 min ($n = 1$), or failing an attention check ($n = 2$).[8] Table 1 provides demographic data on the 177 people included in the

[7] Details of power analysis can be found in Online Appendix B.
[8] Details of the exclusion criteria can be found in Online Appendix D.

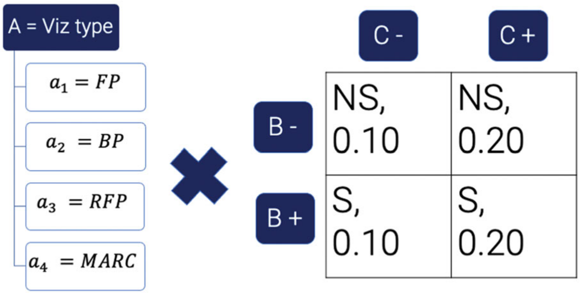


Figure 5. Experimental design.

analysis. The primary sample of interest was the practitioners, and research questions 1–3 were answered based on this sample alone. The researchers served as a point of comparison and were used only to answer research question 4 and in exploratory analyses. Note that a primary limitation of our study is that these 177 respondents are a convenience sample and caution should be taken when generalizing beyond it.

Of the 83 practitioners in our sample, about two-thirds (63%) identified as female and one-third (35%) identified as male. The populations of US public school teachers and princpals have similarly higher proportion of females (76% and 54%, respectively, National Center for Education Statistics, 2020). The practitioners in our sample were asked to indicate any roles they had held in the US PreK-12 education system. Three-quarters of respondents had served as teachers (75%), 21% as curriculum specialists, and 12% as principals.[9] Fifty-nine percent also indicated they had held other roles such as research & evaluation or data analysts, instructional coaches, district administrators, and other support staff. A majority of the practitioners in our sample had a Master's degree (68%), and an additional 27% had a doctorate or professional degree. This is a higher rate of education compared to the population of US public school teachers (58% with a graduate degree) but similar to the rate among principals (98% with a graduate degree) (National Center for Education Statistics, 2020).

Experimental Design

We used a $4*2^2$ blocked factorial design with the following factors and levels:

- **Factor A**: Visualization type (4 levels: $a_1 = FP,\ a_2 = BP,\ a_3 = RFP,\ a_4 = MARC$)
- **Factor B**: Statistical significance of meta-analytic summary effect (2 levels: significant (+) and non-significant (−))
- **Factor C**: Magnitude of meta-analytic summary effect (2 levels: $\delta_{(+)} = 0.20$ and $\delta_{(-)} = 0.10$)

This design amounts to crossing the 4 visualizations with the 2×2 $B \times C$ grid as shown in Figure 5, resulting in 16 total treatment combinations, which are displayed in Table 2.

[9] Respondents could select all that applied, so percentages do not add up to 100.

Table 2. Experimental design—$4{*}2^2$ in 16 runs.

ID	Experimental Condition	Treatment Combination	A	B	C
1	(1)	$a_1(1)$	a_1	$-$	$-$
2		$a_2(1)$	a_2	$-$	$-$
3		$a_3(1)$	a_3	$-$	$-$
4		$a_4(1)$	a_4	$-$	$-$
5	b	a_1b	a_1	$+$	$-$
6		a_2b	a_2	$+$	$-$
7		a_3b	a_3	$+$	$-$
8		a_4b	a_4	$+$	$-$
9	c	a_1c	a_1	$-$	$+$
10		a_2c	a_2	$-$	$+$
11		a_3c	a_3	$-$	$+$
12		a_4c	a_4	$-$	$+$
13	bc	a_1bc	a_1	$+$	$+$
14		a_2bc	a_2	$+$	$+$
15		a_3bc	a_3	$+$	$+$
16		a_4bc	a_4	$+$	$+$

Having participants answer questions about all 16 visualizations would be burdensome, so we confounded the factorial design into 4 blocks, so that each participant was randomized to a block and answered questions about 4 visualizations in total. We ran replicates of the full 16-run design, so we employed 3 different confounding patterns in order for each factor and interaction to be estimable in at least some of the replicates. Because visualization type was the primary factor of interest, we chose confounding patterns where Factor A was fully estimable within people. That is, each person viewed all four visualization types. Details of the confounding patterns can be found in Online Appendix A.

Generating Meta-Analytic Data

We generated four meta-analytic data sets corresponding to each of the conditions in the 2×2 grid in Figure 5, which are also represented in the "Experimental Condition" column of Table 2. Throughout, we chose values for each of these parameters to be somewhat realistic for the education context. We set δ equal to 0.10 and 0.20 for the "low $(-)$" and "high $(+)$" levels of factor C, respectively. While Schild and Voracek (2015) used forest plots with 10 studies, we chose the number of studies for each meta-analytic data set to be $k = 5$, since meta-analytic evidence presented in education clearinghouses tend to be based on a small number of studies (Institute of Education Sciences, 2020).

In order to generate the data from each of the individual studies, we assumed that each study was a cluster randomized trial (CRT) and that our effect size parameter of interest within the study was the standardized mean difference $\delta_j = \frac{\mu_{Tj} - \mu_{Cj}}{\sigma_{[\text{Total}]j}}$, $j = 1, 2, \ldots 5$, where $\sigma^2_{[\text{Total}]j}$ is the total variance in study j, incorporating within- and between-cluster variance. We denote our effect size estimates as $\hat{\delta}_j$, which are normally distributed with mean δ_j and sampling variance $V(\hat{\delta}_j)$. For simplicity of data generation, we assume a fixed effects model for the meta-analytic estimate $\hat{\delta}$, so that the weight for a study j is simply given by $\frac{1}{\mathrm{V}(\hat{\delta}_j)}$. We use Equation (16) from Hedges (2007) to compute $V(\hat{\delta}_j)$,

Table 3. Meta-analytic data by experimental condition.

Experimental condition	$\hat{\delta}$	$SE(\hat{\delta})$	Criteria
(1)	0.098	0.061	Non-significant
b	0.100	0.039	Significant
c	0.197	0.122	Non-significant
bc	0.202	0.078	Significant

which is a function of δ, the intra-class correlation ρ, and total, within-treatment groups, and within-cluster sample sizes.

For the purposes of computing a reasonable value of $V(\hat{\delta}_j)$ with which to generate $\hat{\delta}_j$ values, we chose the total sample size within studies to be $N = 492$ and total number of clusters within studies to be $M = 20$, as these are the median values from the WWC database of education research studies (Institute of Education Sciences, 2020), and assumed an intraclass correlation of $\rho = 0.2$ (Hedges & Hedberg, 2016). Therefore for each of the four meta-analytic data sets, we generated 5 effect size estimates by $\hat{\delta}_j \sim N(\delta, V(\hat{\delta}_j)),\ j = 1, 2, ..., 5.$[10] We then adjusted the $V(\hat{\delta}_j)$ values to achieve a $V(\hat{\delta})$ value that meets the significance criteria determined by Factor B.[11]

Table 3 shows the summary $\hat{\delta}$ values and corresponding $SE(\hat{\delta})$ values for each experimental condition generated according to the procedure described above. Further details on the meta-analytic data generation can be found in Online Appendix C, and R code and final meta-analytic data can be found in the supplementary material.

Creating the Visualizations

The `metaviz` package (Kossmeier et al., 2019) in R has functions `viz_forest` and `viz_rainforest` for creating FPs and the RFPs proposed by (Schild & Voracek, 2015). We adapted their source code to suit our purposes. Changes included plotting the summary effect as a raindrop in the RFPs, creating a reference line for the summary effect, and other minor aesthetic changes. We created the BPs and MARC plots using the `ggplot2` package in R (Wickham et al., 2020). Each of the 4 meta-analytic data sets were plotted as a FP, BP, RFP, and MARC plot for a total of 16 visualizations. For each visualization, we randomized the order of the 5 studies. All 16 visualizations used in the experiment as well as the *R* code used to create them can be found in the supplementary material.

Questionnaire Design

Questionnaire development was informed by "think-aloud" interviews in focus groups and refined based on data from a pilot study. Table 4 shows the 9 questions that were asked each time a participant viewed a visualization. The full Qualtrics survey including

[10] We used a while loop to repeat this process until each fixed effects meta-analytic summary effect was within 5% of its target δ value. This ensured that the visualizations were reflective of the intended experimental conditions of $\delta_{(-)} = 0.10$ and $\delta_{(+)} = 0.20$.

[11] We use a single value $V(\hat{\delta})\ \forall j$ for the purposes of generating the $\hat{\delta}_j$ values and then adjust the $V(\hat{\delta}_j)$s to achieve weights necessary for meeting experimental conditions. See Online Appendix C for details.

Table 4. Visualization survey questions.

ID	Question Text
Q1	Which study's findings do you trust the most?
Q2	Which study was given the *most* weight in determining the summary SMD?
Q3	Which study's results are the *least* certain?
Q4	Which study found that the new curriculum improved student scores by the *greatest* amount?
Q5	On average, how much did the new curriculum increase or decrease student scores?
Q6	Which of the following provides the best (i.e. the most certain) estimate of the true SMD?
Q7	Is there sufficient evidence to conclude that the new curriculum improves scores (i.e. that the true SMD > 0)?
Q8	Assuming this curriculum is affordable and feasible to implement, how likely are you to purchase this curriculum?
Q9	Based on this information, please provide an effectiveness rating of this curriculum.

answer choices can be found in the supplementary material. Note that participants answered these questions a total of 4 times, once for each type of visualization.

As discussed previously, the decision-making process in education is complex, as practitioners rightfully weigh the statistical evidence with many other factors including the values, priorities, and constraints of their local contexts. Therefore rather than trying to model this complexity to replicate the full decision-making context in the experiment, we instead focus on whether or not participants can decode the information necessary to make a normative statistical judgment about the evidence (Q1–Q7) and whether or not their subjective beliefs about the strength of the evidence and the effectiveness of the curriculum differ by visualization type (Q8–Q9).

Q1–Q7 have objectively correct answers determined by statistical norms in meta-analysis and are therefore recoded as 0 or 1 for being incorrect or correct, respectively. All questions except Q5 are multiple choice. Q5 was asked as a continuous slider scale with range from -0.5 to 0.5 and is coded as correct if $|SMD_{obs} - SMD_{actual}| < 0.05,$ where SMD_{obs} is the respondent's answer and SMD_{actual} is the actual value of the meta-analytic average displayed in the visualization. This decision rule was determined *a priori* based on pilot data. Q8 was asked as a 5 point Likert scale (Extremely Unlikely to Extremely Likely), and Q9 used the same scale used by the WWC: Negative ($--$); Potentially negative ($-$); No discernible evidence (0); Mixed ($+ \, -$); Potentially positive ($+$); Positive ($+$ $+$). We also collected demographic data including age, gender, highest level of education, current occupation, level of formal statistical training, and familiarity with meta-analysis.

Analysis Plan

In the Appendix, we provide the analysis plan for each of the four research questions. All models and parameters are explicitly defined in the preregistration, which can be found along with R code for these analyses in the supplementary material. As determined in the pre-registration, we control the Type I error rate for each analysis at the level $\alpha = 0.05$.

Results

Research Question 1—Are Education Practitiones Able to Accurately Interpret Each of These Plots?

Table 5 shows the observed proportion of practitioners that answered each question (Q1–Q7) correctly for each visualization type. Across all visualizations, participants

Table 5. Proportion of participants who answered correctly by question and visualization type (RQ1).

Visualization	*n*	Q1	Q2	Q3	Q4	Q5	Q6	Q7
FP	82	0.512	0.573	0.683	0.866	0.780	0.598	0.476
BP	83	0.759	0.554	0.687	0.904	0.831	0.663	0.458
RFP	81	0.580	0.580	0.617	0.827	0.802	0.667	0.420
MARC	82	0.720	0.951	0.890	0.866	0.805	0.610	0.500

RQ1: Are education practitioners able to accurately interpret each of these plots?
Note. These proportions are estimates subject to statistical uncertainty. Their standard errors range from 0.02 to 0.06.

performed best on questions that required them to extract information from positional encodings; roughly 80% of participants were able to identify the study with the largest SMD (Q4) and to identify the average SMD (Q5), which were both encoded as *x*-axis position. MARC plots are the only design that takes advantage of *y*-axis position and explicit annotations, and as such, this design substantially outperforms the other three visualizations on Q2 and Q3. When weight is encoded directly in the *y*-axis as in the MARC plot, 95% of participants were able to correctly identify which study received the most weight in determining the summary effect (Q2), compared to less than 6 in 10 participants that were able to intuit this important meta-analytic information from the other three visualizations. Similarly, the new MARC plot enabled 89% of participants to identify the least certain study (Q3), compared to less than 70% who could do so on the other three visualizations. While on the one hand it may seem obvious that participants were able to extract information correctly when it is depicted explicitly (e.g. in the MARC plot), the fact that participants were *not* able to easily answer these questions with the three existing visualizations is actually *more* interesting and important. That is, the fact that participants were not able to intuit weight or certainty from even sample size or seemingly simple encodings such as square size offers compelling evidence that practitioners may not interpret visualizations as we intend and that we need to guide them more explicitly with our visualization design choices.

All else constant, we might expect people to be able to extract information from the *x*- and *y*-axis positions with comparable ease; the fact that MARC plots provided even higher gains on Q2 and Q3 (∼90% correct, compared to the ∼80% on Q4 and Q5) suggest that there is perhaps some added benefit to the explicit annotations of certainty and redundant (but simple) encodings of weight.

The barplot (BP) and MARC plot perform comparably on Q1, with over 7 in 10 participants correctly selecting the most precise study as the one they trust the most, a marked improvement over both FPs and RFPs. Overall, practitioners struggled to determine whether or not there was sufficient evidence to conclude the new curriculum improves scores (as determined by statistical significance), with only roughly half answering correctly (Q7). Despite the new design features, the MARC plot did not offer any advantages over the other visualizations in making this determination. Similarly, the MARC plot did not offer any advantages in helping participants identify that the summary effect provides the best (i.e. the most certain) estimate of the true SMD (Q6).

Research Question 2—Which Type of Visualization Leads to Most Accurate Understanding among Education Practitioners?

ANOVA results for Research Question 2 are given in Table 6, which indicate that visualization type (Factor A) is a significant source of variation in participant scores. The

Table 6. ANOVA Results (RQ2).

Source of Variation	Sum of Squares	Degrees of Freedom	F	p-Value
A	38.259	3	10.171	0.000
B	0.019	1	0.015	0.903
C	0.050	1	0.040	0.842
AB	3.794	3	1.009	0.390
AC	9.638	3	2.562	0.056
BC	1.213	1	0.968	0.326
ABC	2.853	3	0.758	0.519
PersonID	523.735	82	5.094	0.000
Intercept	6968.625	1	5557.531	0.000
Residuals	288.399	230		

RQ2: Which type of visualization leads to most accurate understanding among education practitioners? This model uses visualization type (A) and the statistical significance (B) and magnitude (C) of the summary effect to explain variation in participant scores.

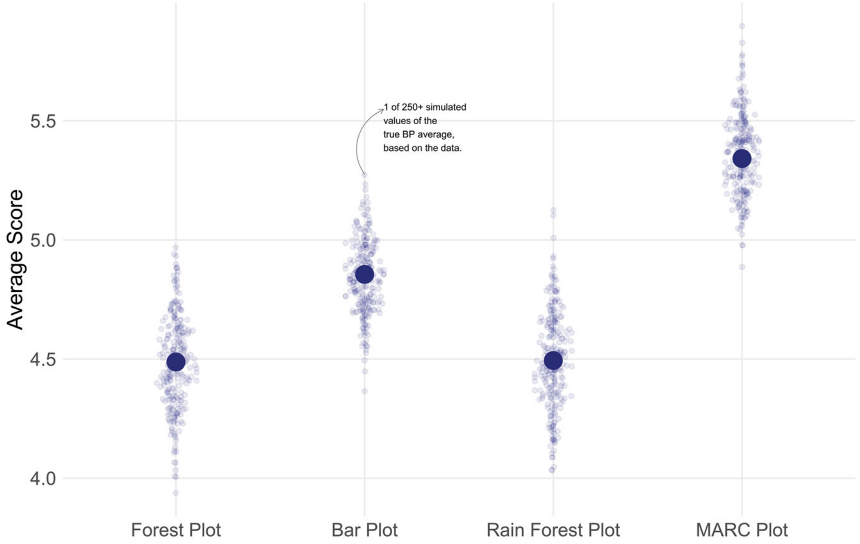


Figure 6. Average practitioner score by visualization type.

average score for each visualization is displayed in Figure 6. In the same spirit as the MARC design, we also display the uncertainty of each mean estimate by including a cloud of plausible values of the true mean, simulated based on the existing data. As Figure 6 indicates, the MARC plot performed best overall, with participants able to answer 5.3 out of 7 questions correctly, on average.

The six Tukey's pairwise comparisons of these four means are given in Table 7. Here we can see that the proposed MARC plot performed better than all three other visualizations, with pairwise differences that were all statistically significant. The MARC plot performed better than BP with an estimated difference of 0.49 and better than FP and RFP with an estimated difference of 0.85 each. Converting these to Cohen's d effect sizes,[12] the new MARC plot offers a 0.43 standard deviation unit improvement over the

[12] We use the MSE from Table 6 as an estimate of the within-person variance in scores (σ_y^2) and standardize these differences to convert to Cohen's d effect size estimates.

Table 7. Tukey's pairwise comparisons of visualization type (RQ2).

Contrast	Difference Estimate	Lower	Upper	Adjusted *p*-value
BP-FP	0.368	−0.084	0.819	0.153
RFP-FP	0.006	−0.448	0.460	1.000
MARC-FP	0.854	0.401	1.306	0.000
RFP-BP	−0.362	−0.814	0.091	0.167
MARC-BP	0.486	0.035	0.937	0.029
MARC-RFP	0.848	0.394	1.302	0.000

RQ2: Which type of visualization leads to most accurate understanding among education practitioners?

Table 8. ANOVA Results (RQ3).

Source of variation	Sum of squares	Degrees of freedom	*F*	*p*-Value
A	7.232	3	2.564	0.055
B	0.203	1	0.216	0.642
C	13.889	1	14.773	0.000
AB	0.341	3	0.121	0.948
AC	3.158	3	1.120	0.342
BC	9.313	1	9.907	0.002
ABC	2.464	3	0.874	0.455
PersonID	362.692	82	4.705	0.000
Intercept	13904.021	1	14789.649	0.000
Residuals	216.227	230		

RQ3: Do user beliefs about strength of evidence and effectiveness of curriculum vary across types of visualization? This model uses visualization type (A) and the statistical significance (B) and magnitude (C) of the summary effect to explain variation in participant's subjective rating of the evidence.

BP, and a 0.76 improvement over both the FP and RFP. BP performs better than FP and RFP by an effect size of approximately 0.32, though these differences are not statistically significant, and FP and RFP perform almost identically to one another.

Research Question 3—Do User Beliefs about Strength of Evidence and Effectiveness of Curriculum Vary across Types of Visualization?

Table 8 gives the ANOVA results for research question 3. Visualization type (Factor A) is a not a significant source of variation in participants' subjective ratings (from Q8—Q9) at the specified alpha level of 0.05, so we do not proceed with any pairwise comparisons.

Research Question 4—Do Education Practitioners and Education Researchers Interpret Forest Plots Differently?

Figure 7 compares the mean score between researchers and practitioners for each visualization. The practitioner data is the same as what is displayed in Figure 6, and we now have researcher means and uncertainty clouds analogously represented with lighter dots. As the figure indicates, researchers score higher than practitioners on all four visualization types, although researchers have the largest gains over practitioners on FPs and RFPs. It's also worth noting that while practitioners always have more difficulty than researchers in accurately interpreting forest plots, the new MARC plot design brings their scores to a level that is comparable to researchers' scores on FPs and RFPs.

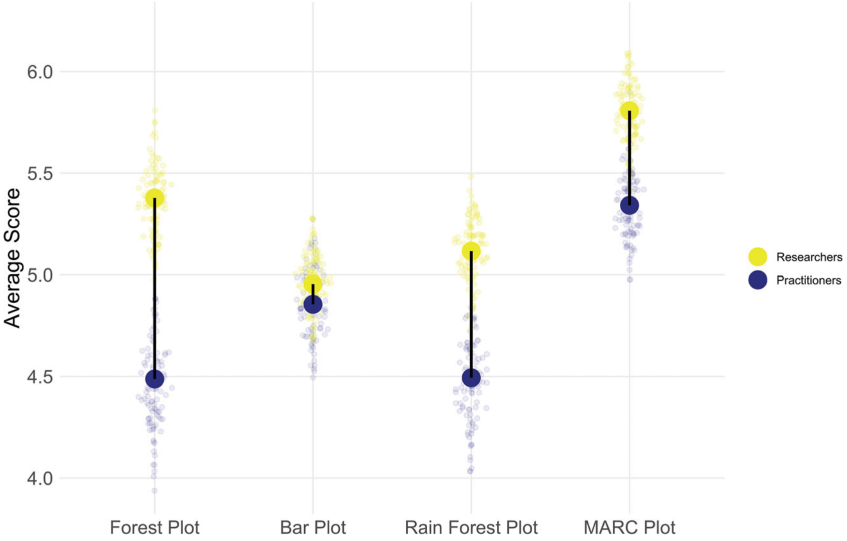


Figure 7. Average practitioner and researcher score by visualization type.

Table 9. *t*-Test Comparisons of Practitioner vs. Researcher Scores by Visualization Type (RQ4).

Visualization Type	Practitioner mean	Researcher Mean	Difference estimate	*t*-statistic	Degrees of Freedom	*p*-Value
FP	4.488	5.378	−0.890	−3.474	159.877	0.000
BP	4.855	4.955	−0.100	−0.477	163.804	0.317
RFP	4.494	5.116	−0.622	−2.450	154.886	0.008
MARC	5.341	5.807	−0.466	−2.090	159.475	0.019

RQ4: Do education practitioners and education researchers interpret forest plots differently?.

Table 9 presents the t-test results for comparing each of these practitioner-researcher differences. The differences between practitioners and researchers are significant at the pre-determined $\alpha = 0.05/4 = 0.0125$ level for both FP and RFP, which can also be seen in Figure 7.

Exploratory Analyses

Common Misconceptions in Meta-Analytic Reasoning by Visualization Type

Investigating incorrect answers sheds light on some potential misconceptions in practitioners' meta-analytic reasoning. In most cases, "I don't know" was the most common incorrect answer, suggesting a general lack of familiarity with how to extract information encoded in the visualizations or perhaps a difficulty in understanding the questions. Consistent with the overall results from Table 5, participants were most likely to answer "I don't know" on FPs and RFPs and least likely to do so on MARC plots.

On the bar plot, 10% of practitioners in our sample ($n = 8$), chose the study with the largest effect size as the one they thought received the most weight. This gives some credence to the critique of bar plots discussed in Section II; representing effect size magnitude via bar length gives undue visual attention to the largest effect sizes and can lead to poor meta-analytic reasoning, where more consideration is given to large effects rather than precisely estimated ones. This misconception did not occur for MARC plots or RFPs and occurred for only 2% of cases for FPs.

Similarly, the anticipated pitfalls of FPs also occurred; 9% of practitioners in our sample ($n = 7$) chose the study with the widest confidence interval bars as the one that receives the most weight (Q2). This is a particularly costly misconception because it leads to conclusions that the *least precise* study should receive the most weight, which is the exact opposite of normative meta-analytic reasoning. This was also the most common misconception on Q2 for RFPs, although it only occurred in 5% of cases ($n = 4$).

Participants also had difficulty with identifying the least certain study (Q3) on FPs. Seventeen percent of practitioners answered "I don't know," but an additional 15% ($n = 12$) got it wrong by selecting another incorrect study. Comparable numbers got Q3 wrong for BPs and RFPs, but the vast majority of wrong answers were "I don't know" in these two cases. This suggests that FPs might be particularly ill-suited because not only do they have encodings that practitioners may not understand, but the encodings may lead to confidence in the *wrong* interpretation.

In all four visualizations, among the roughly 30–40% of practitioners who answered Q6 incorrectly, the most common misconception was that the most precise study—rather than the summary effect—offered the best estimate of the true SMD. This suggests that users may require more explicit guidance about this important meta-analytic feature: that the summary effect represents our best guess at the true SMD and is a more trustworthy estimate than any of the individual studies. Of the four visualizations, this misconception occurred the least for RFPs (12% of cases), perhaps because it is the only visualization where the summary effect is encoded in exactly the same way as the individual studies, making it easier to directly compare its importance to the individual studies. This was an intentional design feature we chose to change from the original RFP in Schild and Voracek (2015) where the summary effect was still represented by a diamond. We tried to facilitate a similar consistency between individual and summary effect encodings in the new MARC design by depicting the summary effect as a dot with weight 1.0. However, because the summary effect is the only one that displays precision in the MARC plot, this may have diluted some of the benefits and made the summary effect appear less precise in relation to the individual studies.

Practitioner Evaluation of Meta-analytic Evidence

In our data, practitioners were curiously more likely to rate the evidence as "positive" (the highest rating on the 5-point scale) when the summary effect was large ($\delta = 0.2$) and *non-significant* (experimental condition *c*) than they were when the summary effect was large and *significant* (experimental condition *bc*). Investigating further, we realized that this is likely due to the fact that even though the overall evidence was significant, experimental condition *bc* included one study with a negative effect, whereas experimental condition *c* did not. This suggests that participants may be exhibiting a variation of vote-counting behavior wherein they pay attention to *any* negative results, regardless of whether or not that negative finding was imprecisely estimated. In exploratory analysis, including a binary variable for whether or not the visualization included a study with a negative result turned out the be a significant predictor of practitioners' ratings of the evidence. Further studies should be conducted to investigate this more directly.

We originally hypothesized for Research Question 3 that subjective evaluation of the evidence—i.e. ratings of curriculum effectiveness (Q9) and stated likelihood of purchasing (Q8)—would differ by visualization type and that participants might be more likely to rate the curriculum highly when the evidence is presented to them in BPs, because BPs do not communicate uncertainty. However, this did not end up being the case. Relatedly, we also anticipated that participants may be more likely to erroneously indicate that there was sufficient evidence for non-significant results when viewing BPs that did not display uncertainty. Instead, people were most likely to answer this way on MARC plots. We wonder if this is suggestive of people having a higher tolerance for uncertainty when individual plausible values are displayed as in the MARC plot, rather than as dichotomous thresholds. For example, even for the non-significant results, there are relatively few simulated outcomes (i.e. ∼5%) that fall in the negative region, which users may still consider "sufficient evidence."

This illustrates one of potential benefits of displaying precision via a cloud of plausible values: it allows users to make their own judgments about how much uncertainty they are willing to tolerate. In traditional displays of confidence intervals such as in FPs, the confidence interval is displayed at a single analyst-specified level of confidence that may or may not align with the user's own tolerance. For example, if the analyst displays a 95% confidence interval that overlaps zero, this is likely to evoke dichotomous thinking at that threshold, whereas using the same evidence to display a 90% confidence interval that does not overlap zero may evoke a different evaluation of the evidence. The MARC plots, however, implicitly allow for judgments to be made at any confidence level by considering the proportion of dots that fall in a range. More work should be done to investigate people's ability to reason appropriately about the cloud of plausible values, the extent to which people's perception of "sufficient evidence" is influenced by the depiction of the cloud of plausible values, how that changes for varying levels of "statistical insignificance," and whether their thresholds for error align with traditional statistical norms (e.g. 5%).

Discussion and Conclusion

This paper has provided some of the first empirical evidence on how education practitioners reason about meta-analytic evidence, with a particular focus on the role that visualization information plays in this evaluation. In this section we discuss both the implications of this study for practice and areas in which future research is needed.

Recommendations for Practice

First, this study serves as a good reminder of the curse of expertise and that care should be taken not to assume consumers will interpret information the way we (researchers) intend. Consistent with the literature, practitioners in our study had great difficulty interpreting forest plots, and the more complex encodings in rainforest plots did little to mitigate this difficulty. It is worth noting that the bar plots currently in use by the

WWC performed better than both types of forest plots and comparably to the MARC plot in several regards. We think this is likely due to the bar plot's simple encodings and is a demonstration that the value of simplicity should not be underestimated when communicating to practitioners. However, we still recommend against the use of bar plots for communicating meta-analytic evidence in particular, because they are prone to misconceptions and cognitive biases.

In all four visualizations, we found that practitioners were not as successful as researchers at correctly interpreting the meta-analytic visualizations. However, the new Meta-Analytic Rain Cloud plot brought practitioner scores to a level comparable to researcher scores on the visualizations that are currently being used in practice. The primary merit of the new design appears to be the use of y-axis position to explicitly encode an attribute key to meta-analytic reasoning: the meta-analytic weight. While other visualizations may have seemingly intuitive encodings from which you can intuit weight and precision (e.g. sample size or square area), both the data visualization literature and the present results suggest there is a clear advantage to using positional encodings (e.g. x- and y-axis) and adding annotations that make the intended take-aways as explicit as possible. This design—with its explicit y-axis encoding of weight and corresponding annotation for how to interpret the y-axis in relation to certainty—enabled nearly *all* practitioners in the sample to discern which studies received the most weight and which ones were the least certain.

Care should be taken to guide users more explicitly toward the key meta-analytic idea that the summary effect provides the best (i.e. most precise) estimate of the treatment effect and should be considered over and above the individual study results. This may involve adding additional annotations and/or creating data-journalism style explainer videos to accompany the visualization and that teach users how to interpret it. Relatedly, it is worth investigating in future empirical work whether it would be advantageous to *only* visualize the summary effect rather than the summary effect and the individual study effects, as in the forest plots considered in this experiment. There is likely a tradeoff between transparency and simplicity here, and we should further investigate the utility of and user preference for each.

Based both on existing literature and the present results, we recommend avoiding confidence interval bars to display precision, particularly when communicating to non-researchers. The "cloud" of plausible values in our newly proposed MARC plot offers an alternative way to display uncertainty. The literature suggests several reasons why such a design may be advantageous, but further work is needed to investigate this feature of the design more directly. Specifically, the literature suggests that displaying individual potential values promotes better statistical reasoning, particularly about uncertainty, and may be less likely to lead to dichotomous thinking as compared to traditional confidence interval displays.

Based on the theoretical reasons outlined in Section I and the results of the present study, we caution against using bar plots, forest plots, and rainforest plots when communicating meta-analytic evidence to education practitioners. In particular, the visual encodings are prone to bring undue visual attention to imprecise studies, which is in contradiction to normative meta-analytic reasoning. We recommend the use of the

MARC Plots as an alternative to those currently used in practice. More generally, we recommend further development and use of visualizations of this type: ones that use simple encodings, provide annotations to guide the user as much as possible, and utilize positional encodings to make key features salient.

Directions for Future Research

While the findings of this study suggest that the MARC plot offers many advantages, it is important to keep in mind the conditions actually studied. First, this study focuses on the development and testing of a visualization aimed at improving *practioner* understanding and interpretation of the *average* effect size across studies. Experts (researchers) were included in the study only to serve as a comparison. It is possible that the MARC plot may also offer advantages for experts as well, particularly for identifying heterogeneity of effects. However, this hypothesis would need to be tested in a future study.

Second, this study focused on the case most relevant to clearinghouses like the WWC—when there are a small number of studies (∼5) of single intervention. In more general meta-analytic contexts—and in multi-site randomized trials—the number of studies or effect sizes to be visualized may be much larger than the conditions studied here. In these situations, as the number of studies increases, the weight attributable to each study decreases and often becomes more equal. In the MARC plot, this could result in 'puddles'—flattened, overlapping circles clustered at the bottom of the figure. On the one hand, these 'puddles' may actually be useful for conveying the scope of evidence and the primacy of the averarage effect size at the top of the plot. On the other hand, this may simply be confusing. It would be interesting to test the MARC plot in cases with a large number of studies, but also to compare different versions—for example, one option may be to rescale the study weights to be relative to the average weight, another option may be to display the plot interactively (allowing studies to be clicked on by a mouse), and yet another option is to not display individual study findings at all. Since we have not yet conducted any of these studies, for now we cannot yet recommend the use of the MARC plot more generally.

Third, this study has focused on methods for visualing study findings at a single moment in time. However, over time the evidence base for an intervention is likely to increase, which results in the inclusion of additional data points in both the summary measure and visually. Since the weights in a meta-analysis are relative to one another, with the inclusion of an additional study, the weights of all other studies are reduced. In the MARC plot, this means that the study data points would move toward the bottom of the plot and that each data point may itself be interpreted as less 'certain.' Again, since this situation was not the focus of this study, it is impossible to know for sure how well interpretations of the MARC plot would hold up in comparisons over time.

Finally, in addition to developing and testing the MARC plot, to our knowledge this is the first study to develop a measure of meta-analytic reasoning. Unfortunately, given the sample sizes within subgroups of participants in this study, it was not possible to assess the reliability of the developed measures. Our hope is that the measures we have developed can be further refined and validated in future studies, both of the MARC plot and of other visualizations and reporting methods.

Conclusion

We hope this study serves as an example of the type of empirical studies that can and should be conducted to generate evidence specific to the the translation of findings to decision-making in education research. Furthermore, this study illustrates the importance of being thoughtful about our choice of data visualizations when communicating statistical evidence and demonstrates the utility of turning to evidence from the existing data visualization and cognitive science literatures for best practices.

Finally, we recommend carrying these principles over into all forms of communication. We should be ever-cognizant of the curse of expertise and not anticipate that users will automatically interpret information the way we, as researchers, intend. Rather, we as an education research community should strive to be more rigorous and evidence-based in our efforts to effectively disseminate research findings to practitioners. We hope this study serves as an example of the type of *translation science* that will move us toward this goal.

Acknowledgements

We would like to thank Steven Franconeri for his initial design of the MARC plot and for offering his data visualization expertise throughout.

Funding

This work was supported by UBS Optimus Foundation on behalf of the Yidan Prize Foundation [Award ID SP0054403] and by U.S. Department of Education, Institute of Education Sciences, Multidisciplinary Program in Education Sciences [Grant Award R305B140042]. This study was approved by Northwestern University's Institutional Review Board and was *pre-registered* with the Open Science Foundation prior to data collection.

ORCID

Kaitlyn G. Fitzgerald http://orcid.org/0000-0001-6569-4494
Elizabeth Tipton http://orcid.org/0000-0001-5608-1282

References

Anzures-Cabrera, J., & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods*, *1*(1), 66–80. https://doi.org/10.1002/jrsm.6

Barrowman, N. J., & Myers, R. A. (2003). Raindrop plots: A new way to display collections of likelihoods and distributions. *The American Statistician*, *57*(4), 268–274. https://doi.org/10.1198/0003130032369

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*(387), 531–554. https://doi.org/10.2307/2288400

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A., & Dragicevic, P. (2020). A task-based taxonomy of cognitive biases for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, *26*(2), 1413–1432. PP. https://doi.org/10.1109/TVCG.2018.2872577

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty *displays using quantile dotplots* or CDFs *improve transit decision-making* [Paper presentation]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI, Montreal QC, Canada. 18, 1–12. ACM Press. https://doi.org/10.1145/3173574.3173718

Fox, J., & Weisberg, S. (2019). *car: Companion to applied regression (Version 3.0-11)*. https://cran.r-project.org/web/packages/car/index.html

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

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, *19*(1), 44–63. https://doi.org/10.2307/749110

Hedges, L. V. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hedges, L. V., & Hedberg, E. C. (2016). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. http://journals.sagepub.com/doi/10.3102/0162373707299706

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), e0142444. https://doi.org/10.1371/journal.pone.0142444

Institute of Education Sciences (2020). *Data from individual studies*. Retrieved May 6, 2020, from What Works Clearinghouse website: https://ies.ed.gov/ncee/wwc/StudyFindings

Jackson, C. H. (2008). Displaying uncertainty with shading. *The American Statistician*, *62*(4), 340–347. https://doi.org/10.1198/000313008X370843

Kossmeier, M., Tran, U. S., & Voracek, M. (2019). *metaviz: Forest plots, funnel plots, and visual funnel plot inference for meta-analysis (version 0.3.0)*. https://CRAN.R-project.org/package=metaviz

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes*, *8*(1), 84. https://doi.org/10.1186/s13104-015-1020-4

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

National Center for Education Statistics (2020). *The condition of education—preprimary, elementary, and secondary education—teachers and staff—characteristics of public school teachers—indicator May (2020)*. Retrieved May 18, 2021, from https://nces.ed.gov/programs/coe/indicator_clr.asp

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*(4), 601–607. https://doi.org/10.3758/s13423-012-0247-5

Qiao, X., & Hullman, J. (2018). Translating scientific graphics for public audiences. In Proceedings of the VisGuides: 2nd Workshop on the Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization. IEEE VIS 2018.

Schild, A. H. E., & Voracek, M. (2013). Less is less: A systematic review of graph use in meta-analyses. *Research Synthesis Methods*, *4*(3), 209–219. https://doi.org/10.1002/jrsm.1076

Schild, A. H. E., & Voracek, M. (2015). Finding your way out of the forest without a trail of bread crumbs: Development and evaluation of two novel displays of forest plots. *Research Synthesis Methods*, *6*(1), 74–86. https://doi.org/10.1002/jrsm.1125

Schriger, D. L., Altman, D. G., Vetter, J. A., Heafner, T., & Moher, D. (2010). Forest plots in reports of systematic reviews: A cross-sectional study reviewing current practice. *International Journal of Epidemiology*, *39*(2), 421–429. https://doi.org/10.1093/ije/dyp370

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New York, N.Y.)*, *185*(4157), 1124–1131.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C, … RStudio. (2020). *ggplot2: Create elegant data visualisations using the grammar of graphics (Version 3.3.0)*. https://CRAN.R-project.org/package=ggplot2

Wiederkehr, A. (2020, August 13). *How we designed the look of our 2020 forecast | FiveThirtyEight*. Retrieved January 9, 2021, from https://fivethirtyeight.com/features/how-we-designed-the-look-of-our-2020-forecast/

Xiong, C., van Weelden, L., & Franconeri, S. (2020). The curse of knowledge in visual data communication. *IEEE Transactions on Visualization and Computer Graphics*, *26*(10), 3051–3062. https://doi.org/10.1109/TVCG.2019.2917689

Appendix

The following decribes the analysis plan for each research question, all of which was pre-registered with OSF prior to data collection, including choice of multiple-comparison correction procedures for each analysis.

Research Question 1—*Are education practitioners able to accurately interpret each of these plots?*

For RQ1, we descriptively consider the proportion of respondents that are able to answer each Question $1-7$ correctly for each of the 4 visualizations. That is, we consider a 4×7 table of estimated proportions. We did not conduct any hypothesis tests to evaluate this question.

Research Question 2—*Which type of visualization leads to most accurate understanding among education practitioners?*

Let y_{irst} be a sum score (range $0-7$) for the number of Questions $1-7$ individual i ($i=1,2,\dots,n$) answered correctly when viewing the visualization condition *rst*, where Factors A, B, and C are indexed by $r=1,2,3,4$; $s=1,2$; and $t=1,2$ respectively. We model the outcome y_{irst} with an ANOVA model that includes an overall mean, main effects for factors A, B, and C, all two- and three-way interactions between factors, a blocking factor for individuals, and an individual error term. The ANOVA model was fit using the `car` package in R (Fox & Weisberg, 2019) in order to use Type 3 sums of squares, which are the more conservative choice and allow for interaction between the three factors. In order to answer this research question, pairwise comparisons are of interest, so we use Tukey's test for the six pairwise comparisons between the 4 levels of Factor A to answer RQ2. Tukey's test procedure controls the experiment or "family-wise" error rate, which we set to be $\alpha=0.05$ (*Tukey, 1953*).

Research Question 3—*Do user beliefs about strength of evidence and effectiveness ratings vary across types of visualization?*

We define the sum score z_{irst} (range 2 to 10) of the Likert scale responses to Q8 and Q9 for individual i when viewing visualization *rst*. Higher sum scores correspond to higher ratings of the curriculum and stated likelihood of purchasing. We use the same ANOVA model described for RQ2, but with z_{irst} as the outcome. If Factor A is significant in the ANOVA, we conduct one-sided hypothesis tests on the contrasts between BP and each of the three other visualizations to investigate whether the BP—which does not display uncertainty—results in higher ratings of the evidence. In our pre-registration, we hypothesized that users would rate the evidence most highly when they were not shown uncertainty (i.e. in the bar plots), so we were primarily interested in comparisons between the bar plot with each of the three other visualizations. We therefore use Dunnett's test procedure here, with BP as the reference group. This choice was made a priori based on the analysis of interest and was pre-registered prior to data collection. Similar to Tukey's test, Dunnett's test is a multiple comparison procedure that controls the family-wise error rate.

Research Question 4—*Do education practitioners and education researchers interpret forest plots differently?*

Here we conduct a series of 4 one-sided two-sample t-tests to determine if education practitioners score lower on average than researchers for each visualization type. The level of each test is set to $\alpha = 0.05/4 = 0.0125$.