Journal of Research on Educational Effectiveness

Routledge
Taylor & Francis Group

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

# A Knowledge Mobilization Framework: Toward Evidence-Based Statistical Communication Practices in Education Research

Kaitlyn G. Fitzgerald & Elizabeth Tipton

**To cite this article:** Kaitlyn G. Fitzgerald & Elizabeth Tipton (2023): A Knowledge Mobilization Framework: Toward Evidence-Based Statistical Communication Practices in Education Research, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2023.2209082

**To link to this article:** https://doi.org/10.1080/19345747.2023.2209082

Published online: 22 May 2023.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2023.2209082

METHODOLOGICAL STUDIES

# A Knowledge Mobilization Framework: Toward Evidence-Based Statistical Communication Practices in Education Research

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

[a]Mathematics, Physics, and Statistics, Azusa Pacific University, Azusa, California, USA; [b]Statistics, Northwestern University, Evanston, Illinois, USA

## ABSTRACT

The evidence-based decision-making movement often assumes that once evidence is available (e.g., via the What Works Clearinghouse), decision-makers will integrate it into their practice. Research-practice partnership studies have shown this is not always true. In this paper, we argue that instead of assuming research will be useful and used, we should directly study strategies for disseminating evidence and mobilizing knowledge. We present a framework for organizing knowledge mobilization research into three facets: (1) examining *norms* embedded in evidence we communicate, (2) *descriptively* understanding how decision-makers reason about this evidence as well as their varied decision-making needs, and (3) *prescriptively* developing and evaluating communication strategies that facilitate better use of evidence by decision-makers. We delineate this three-faceted framework—*normative, descriptive, prescriptive*—and demonstrate how it considers the perspectives and priorities of both researchers and decision-makers. Focusing on a case study—of how statistical evidence is conveyed by clearinghouses—we point to existing evidence in education and other fields such as data visualization and cognitive psychology that should inform our communication practices and identify areas where further knowledge mobilization research is needed.

**ARTICLE HISTORY**
Received 22 November 2021
Revised 12 April 2023
Accepted 17 April 2023

**KEYWORDS**
Knowledge mobilization; statistical cognition; communicating evidence

Since the founding of the Institute of Education Sciences (IES) in 2002, the field of education research has seen impressive progress in its efforts to understand which interventions can effectively improve student outcomes. This success can be seen in the rise of high-quality causal studies—including both randomized trials and strong quasi-experiments—in the field. Over the past twenty years, for example, IES alone has funded over 400 efficacy and effectiveness studies testing interventions in schools in the United States. Internationally, funders like the Education Endowment Foundation (EEF), the World Bank, The Abdul Latif Jameel Poverty Action Lab (J-PAL), and others have funded hundreds more such studies. As Judy Singer noted, high-quality causal inference studies have become the "normal science" in much of education research (Singer, 2019).

**CONTACT** Kaitlyn G. Fitzgerald ✉ kfitzgerald@apu.edu Mathematics, Physics, and Statistics, Azusa Pacific University, Azusa, CA, USA.

Over time, researchers focused on Research-Practice Partnerships (RPPs) have called in to question if this evidence is useful and used in actual education decision-making. In a nationally representative sample of school district leaders, Penuel et al. (2017) showed that over 50% of leaders never or only rarely consulted either the *What Works Clearinghouse* or *Regional Education Laboratories* (both housed in NCEE within IES) when making curricular decisions in schools. Instead, those studying RPPs have shown that decision-makers are often in need of information that is more robust and detailed than that provided by clearinghouses, thus leading them to turn to original research articles instead. For example, in addition to effect size and cost, decision-makers need information on the comparison conditions studied, how well the intervention is able to be implemented, if the intervention can be adapted to local contexts, and the underlying logic and theory of the intervention (e.g., Coburn et al., 2021; Farley-Ripple et al., 2018; Farrell et al., 2022; Penuel et al., 2018). More generally, this and other research suggests that relative to researchers, education decision-makers put far higher weight on issues of external validity than internal validity (Nakajima, 2021; Vivalt et al., 2022).

This disconnect between research and practice has not gone unnoticed by the causal community in education, including IES and related organizations. For example, over time IES has increased requirements for both generalizability and dissemination (e.g., RFA, Institute of Education Sciences, 2022). The EEF has focused efforts on improving how clearinghouses convey evidence (e.g., Teaching and Learning Toolkit, Education Endowment Foundation, n.d.). These concerns can also be seen in the themes of SREE's annual meetings, which between 2016 and 2021 focused on knowledge mobilization, translation, relevance, and practice.[1] During this same period, SREE also increased efforts to incorporate education practitioners and decision-makers into the conference with the addition of practitioner co-chairs and new conference tracts (Society for Research on Educational Effectiveness, n.d.). Similarly, AERA's annual convention has convened sessions on the barriers to data-based decision-making in education, resulting in enhanced efforts to build capacity for data literacy and data use by educators (Mandinach & Gummer, 2016; Mandinach & Schildkamp, 2021)

In response to these same concerns with the "use and usefulness of education research," the 2022 National Academies of Science, Engineering, and Medicine report on the "Future of IES" proposed *knowledge mobilization* as one of five types of research necessary for the field to become more equitable and responsive to the needs of US educational organizations and decision-makers (National Academies of Sciences et al., 2022). The report highlights questions that knowledge mobilization studies should attempt to answer, including:

> how schools and decision-makers identify problems and develop solutions; which interventions, curricula, and programs are currently used in schools; how to get promising evidence into their hands; **how educational leaders harness that evidence to guide action**;

[1] Recent conference themes: *Lost in Translation: Building Pathways from Knowledge to Action* (2016); *Expanding the Toolkit: Maximizing Relevance, Effectiveness and Rigor in Education Research* (2017); *The Evidence Behind Evidence Use: When Does Education Research Inform Practice?* (2018); *Tensions and Trade-offs: Responding to Diverse Demands for Evidence* (2019); *Practical Significance and Meaningful Effects: Learning and Communicating What Matters* (2020); *The Fierce Urgency of Knowledge: Education Evidence for Reimagining and Reckoning* (2021).

> **and what conditions support educational leaders to use research more centrally and substantively in their decision making**. ([Bolding added]; National Academies of Sciences et al., 2022; Farley-Ripple et al., 2018; Jackson, 2022)

Importantly, the National Academies report makes clear that its call is not just for a renewed focus on knowledge mobilization efforts but for an investment in knowledge mobilization as a program of research in and of itself, proposing that "strategies to mobilize knowledge be studied directly" and asserting a need for "developing and testing robust strategies to foster the use of research in varied contexts." The report echoes a similar call in Conaway (2021) stating that "IES should prioritize research on research use itself."

But how might this field of knowledge mobilization studies be structured? In this paper, we argue that the field might benefit from structuring itself around a decision-making taxonomy put forth by Bell et al. (1988) and adapted to statistical cognition by Beyth-Marom et al. (2008). In this paper, we adapt the taxonomy further to delineate three facets of knowledge mobilization research in education—normative, descriptive, and prescriptive—and discuss their ideal interdependence on one another. We identify areas of research where normative, descriptive, and prescriptive work relevant to knowledge mobilization in education is currently being conducted and where there are gaps. Similar to the field of statistical cognition (Beyth-Marom et al., 2008), often these three facets are being studied in disparate places (in our case both inside and outside of education), and by different researchers in different fields, who publish in different journals.

In order to situate and illustrate the usefulness of this framework, we focus on one narrow aspect of knowledge mobilization in education: the communication of statistical evidence. We do so not because we believe this to be the only or even most important part of decision-making in education—though we do believe it is essential—but because it is our own area of research. Here we heed the National Academies call for knowledge mobilization studies to be conducted and "owned" by all topic areas and fields (National Academies of Sciences et al., 2022). As such, we seek to develop a framework for answering questions such as, "*How should statistical evidence be reported and conveyed to facilitate evidence-based decision-making by education practitioners and policy makers?*" Answering this question, we show, will require connecting research across a range of fields, from statistics to cognitive psychology to human-computer interaction.

In the following section we introduce the framework and the types of questions the three facets attempt to answer. In the Section "Normative, Descriptive, and Prescriptive Research," we point to other literatures the education research community can draw from (e.g., cognitive science, data visualization) in establishing this area of knowledge mobilization research and further illustrate the framework by identifying existing normative, descriptive, and prescriptive research relevant to communicating statistical evidence in education. We then discuss the integration of the three facets and how to create healthy feedback loops in which the mobilization of knowledge is not a linear process but rather each facet is mutually informed by the others. We conclude with a discussion.

## A Framework for Statistical Decision Making

Recall our question: *How should statistical evidence be reported and conveyed to facilitate evidence-based decision-making by education practitioners and policy makers?* To provide a robust framework for tackling this question, we utilize a decision-making taxonomy proposed by Bell et al. (1988). Beyth-Marom et al. (2008) adapted the taxonomy for statistical cognition, and we adapt it further for decision-making from statistical evidence in education research more specifically. The taxonomy can be used to divide research on knowledge mobilization in education into three types:

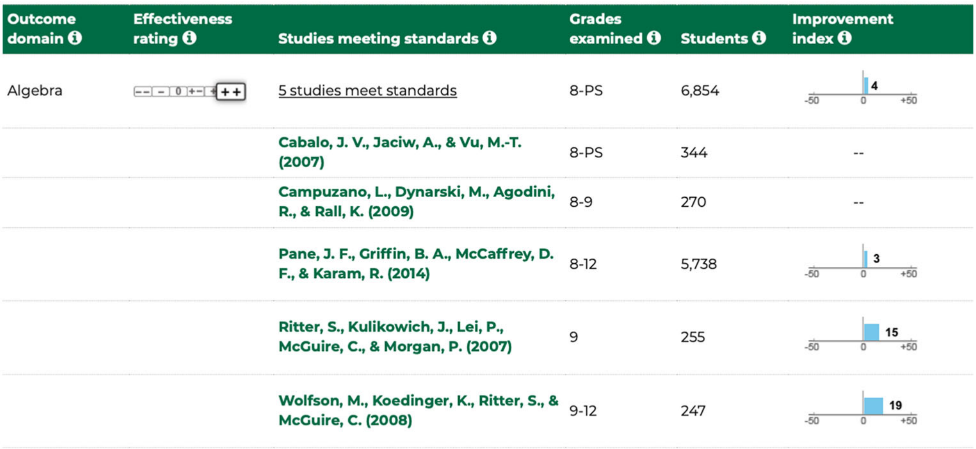
- **Normative**: How *should* education decision-makers evaluate statistical evidence?
- **Descriptive:** How *do* education decision-makers evaluate statistical evidence?
- **Prescriptive**: How can we *help* education decision-makers make *better* evaluations of statistical evidence?

We begin here with the *normative* component since it is how researchers naturally approach the dissemination and reporting of research. Statistical methods bring with them encoded logics and definitions (norms) regarding the appropriate way people *should* evaluate the statistical evidence presented. These norms guide what is included in the usual quantitative summaries of research findings, such as effect sizes, standard errors, $p$ values, and/or confidence intervals. For example, the findings of intervention studies are typically reported in terms of effect sizes, since doing so allows comparisons across studies and places less emphasis on statistical significance ($p$ values).

*Descriptive* questions focus their attention not on norms or intentions of researchers generating the evidence but on users of such evidence. That is, descriptive research seeks to understand how education decision-makers make sense of the statistical evidence researchers provide for them. For example, how do decision-makers interpret effect sizes and statistical significance? How do effect sizes and $p$ values affect how they interpret the efficacy of an intervention? Often, there is a gap between the descriptive and normative use of statistical evidence, between how people *do* reason about the evidence and how experts believe they *should* reason about the evidence. We will demonstrate several examples of this type of gap in reasoning throughout the paper.

Finally, *prescriptive* research is concerned with developing and testing strategies and mediums of statistical communication that would close that gap and help people reason well (i.e., normatively) about statistical evidence. This includes—but is not limited to—effective data visualizations and dashboards as well as design of training programs, courses, and workshops to enhance data and statistical literacy among education decision-makers.

To further illustrate the normative, descriptive, and prescriptive facets in the education research context, let us consider a concrete example regarding clearinghouses. Here we have selected an intervention, *Cognitive Tutor Algebra I*, which has been evaluated by five studies that meet WWC standards. The summary of this evidence as presented by the WWC can be seen in Figure 1 (Institute of Education Sciences, What Works Clearinghouse, 2021). There are three ways of approaching and studying this particular



| Outcome domain | Effectiveness rating | Studies meeting standards | Grades examined | Students | Improvement index |
|---|---|---|---|---|---|
| Algebra | −− − 0 +− + ++ | 5 studies meet standards | 8-PS | 6,854 | 4 |
| | | Cabalo, J. V., Jaciw, A., & Vu, M.-T. (2007) | 8-PS | 344 | -- |
| | | Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009) | 8-9 | 270 | -- |
| | | Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014) | 8-12 | 5,738 | 3 |
| | | Ritter, S., Kulikowich, J., Lei, P., McGuire, C., & Morgan, P. (2007) | 9 | 255 | 15 |
| | | Wolfson, M., Koedinger, K., Ritter, S., & McGuire, C. (2008) | 9-12 | 247 | 19 |

**Figure 1.** Example evidence from What Works Clearinghouse.

knowledge mobilization problem—how *should* decision-makers reason about the evidence from the five studies (normative); how *do* decision-makers reason about the evidence presented (descriptive); and how can the evidence be communicated more effectively (prescriptive)?

Normative questions in this context ask—*how should people reason about a collection of studies? What's the appropriate way to make sense of the 6 lines of research presented in Figure 1?* It is important to keep in mind that Figure 1 was itself created by researchers and reflects many *normative* ideas about how these findings should be interpreted. At the top, a meta-analytic effect size (4 points) is provided that summarizes the five studies below it. In doing so, the WWC is following norms from statistical methodology that have established meta-analysis as the gold-standard for evidence synthesis. Since this summary statistic is inverse variance weighted, statistical theory establishes that it is the most precise and accurate estimate of the average effect of the intervention. Thus, statistical norms indicate that the appropriate way to reason about the evidence is by trusting the meta-analytic summary effect more than any single estimate below it. An expert might also note that each of the effect sizes below (3, 15, 19 points) are all in the same direction, indicating that the effects vary some, but only in magnitude not direction. Additional normative ideas are embedded in the effect size metric reported—the improvement index—which is based upon an argument (by researchers) that this index is easier to interpret than other indices of effect size. Finally, normative ideas can also be found in what is *not* reported here. For example, Figure 1 only includes studies that have met criteria for high quality evidence as determined by the WWC; these criteria focus on ensuring high internal validity. While sample size is reported, other features of the study sample are not, which reflects the fact that external validity often matters less to the researchers behind this design than internal validity. All of these decisions about what and how statistical evidence gets reported reflect the implicit norms researchers hold regarding the appropriate way to reason about the evidence.

In contrast, *descriptive* research here would be concerned with understanding how education decision-makers reason about and interpret Figure 1. How do they reason about a collection of studies? How do they make sense of the 6 lines of research presented here? Do they primarily consider the summary of the evidence (Index of 4), or do they make ad hoc judgments and syntheses of the 5 individual studies? How do they consider studies not included here? How do they interpret the improvement index? Do they favor more recent studies or authors with whom they are familiar? Do they search for ways to examine sample characteristics to determine contexts "like" theirs? If so, do they ignore the rest of the evidence in favor of only the study most like theirs, or do they integrate all the information in some way? Importantly, notice here that "decision-makers" may be a heterogeneous group and that the answers to these questions might be different for government officials than for school district curriculum experts, for example.

One of the advantages of this framework is considering the normative and descriptive facets in concert with one another, so that a natural question then becomes: is there a gap between the two? For example, researchers may be emphasizing effect sizes—because they find them more interpretable and valid—yet decision-makers may not, and thus may be relying on *p* values. Or alternatively, decision-makers may be informally trying to determine variation in effect sizes—because they worry about how much an intervention effect depends upon context—and yet no such summary statistic is provided by researchers. Here it is important to keep in mind that these differences between intended use and actual use of an information display are not always a result of decision-maker misunderstanding. Such gaps can also result when researchers misunderstand the information needed for decision-making.

The linchpin of an effective knowledge mobilization enterprise is both to identify these gaps and then to conduct *prescriptive* research that asks: what are effective strategies and means of communication to bridge the gap between the normative and the descriptive? What dashboards and visualizations might facilitate better decision-making by practitioners? What information should be included in those dashboards and visualizations, and how should it be presented? Importantly, bridging this gap is not simply about moving decision-makers in line with the priorities of researchers. For example, if the judgments and priorities of researchers (e.g., internal validity) are not in line with those of decision-makers (e.g., external validity) it may be that an effective intervention changes practices for both groups. It may be that sample size and effect size reporting (as in Figure 1) does not fully attend to heterogeneous decision-making needs, and instead communication strategies should be developed and evaluated that illuminate who comprised a sample and to what extent the results are relevant to a particular subgroup.

Underlying the call to knowledge mobilization research is a desire—as an evidence-based field—to be more evidence-based in our own practices as education researchers. We believe the normative, descriptive, prescriptive taxonomy provides a useful framework for highlighting what evidence should inform how we collectively set to the task of communication of statistical evidence, and what further evidence is needed to establish best practices. In the following sections, we further illustrate and give examples of existing normative, descriptive, and prescriptive research relevant to communicating statistical evidence in education and point to the other literatures the education research community can draw from to establish a robust evidence-base in this area of knowledge mobilization.

## Normative, Descriptive, and Prescriptive Research

### Normative

As described by Beyth-Marom et al. (2008), “the normative rules, theories, and models of the science of statistics are the standards recommended for summarizing data, interpreting it, and evaluating hypotheses.” These include long-standing basic norms about how to compute means, regression coefficients, standard errors, and confidence intervals as well as how to interpret them. Normative rules determine formulaic relationships between sample size and statistical uncertainty. Statistical evidence in education certainly relies on these long-standing norms and many more, but the normative facet in education also includes the creation of new methods or adaptations of old ones to answer education-specific questions.

For example, causal inference and experimental design methods have been developed and adapted to offer norms for how to design and power (often cluster-randomized) studies to appropriately estimate treatment effects in education (e.g., Hedges & Rhoads, 2010; Raudenbush, 1997; Rubin, 1974; Schochet, 2008); generalizability methods offer norms for how to select samples and what generalizations can appropriately be claimed to what populations (e.g. Stuart et al., 2001*;* Tipton & Olsen, 2018); meta-analysis methods offer norms for how to convert effect sizes and synthesize them across studies (e.g., Hedges, 1985, 2007). The methodological papers published in *JREE* or other social science methods journals (e.g., *JEBS, Psych Methods*) are often good examples of normative research relevant to education, whereby researchers are establishing methods (norms) for generating and analyzing statistical evidence.

Here we want to highlight that we are using the word “norms” and not “facts,” which may be surprising to some readers. We do so, though, because not all the statistical decisions and interpretations provided by researchers are based solely on mathematics, but instead are based on the assumptions and models underlying the estimators provided. For example, while it is a fact that clustering affects certainty, *how* this clustering is accounted for involves the choice of frameworks (e.g., frequentist, Bayesian) and models (e.g., ANOVA, HLM, finite-population, robust standard errors), each with their own interpretation. Furthermore, while bias and variance are properties of all estimators, our choice of a preferred estimator often depends upon priorities determined by researchers. For example, in the field of causal inference, the focus is often on finding an unbiased estimator, whereas those focused on prediction find it reasonable to accept some bias in exchange for improved precision (e.g., a focus on mean-squared-error).

The result is that the statistical evidence presented is a combination of norms—modeling choices, goals and priorities made by researchers—and facts—mathematical and statistical definitions that govern proper implementation of methods and interpretation of findings. For example, if a study reports an effect size of 0.25 (SE $= 0.10$, $p < 0.05$), it is clear that this estimate uses a frequentist framework and that within this framework the chance is less than 5% that if in fact the intervention had no effect at all that we would see an estimated effect this large in our sample. Furthermore, we can interpret this effect size of 0.25 as meaning, for example, that if the treatment effect is constant, this amounts to increasing everyone’s score by 0.25*SD units where SD is the standard deviation of the outcome of interest. Again, though, here our interpretation of if this as a “large” or “small” effect moves beyond statistical definitions and has to do with norms that situate this intervention relative to others of the same scope.

Notably, so far, we have only addressed the reporting and interpretation of statistical summary statistics. In practice, however, evidence is often presented visually, especially when being conveyed to policymakers and decision-makers (e.g., Figure 1). While there are innumerable ways to visualize data, some visualizations are common to statistical experts and scientists. For example, researchers in meta-analysis are used to examining forest plots, which convey findings of multiple studies in terms of small boxes (for the effect sizes) and error bars (for confidence intervals). Often, researchers’ use of particular visualizations are so commonplace that they fail to recognize the statistical norms embedded in them and their audience’s potential lack of familiarity with those norms. This is true even of the “basic” effect size and uncertainty reporting included above (i.e., *“0.25 (SE = 0.10, $p < 0.05$)”*). Thus, when presenting statistical evidence for the purpose of communicating to non-researchers, it is important to consider: *what statistical norms are inherent in the information I am trying to communicate? What norms govern a statistically appropriate evaluation of the evidence?* It is then appropriate to turn to *descriptive* evidence regarding how consumers of evidence reason about those norms.

## Descriptive

If we think of knowledge mobilization in part as a communication process between researchers and decision-makers, it is important to consider that the “message sent” may not always be the “message received.” The normative facet illuminates what norms are infused in the “message sent” by researchers, and the descriptive facet illuminates the “message received” by education decision-makers. When trying to effectively communicate information of any kind, it is crucial to be conscious of the *curse of expertise*, which renders experts oblivious to the fact that other people often do not know what they know or see what they see. This is a well-established and wide-spread psychological phenomenon wherein it is nearly impossible for people who possess some form of knowledge to take on the perspective of a naïve person who does not hold that same knowledge (Birch & Bloom, 2007; Camerer et al., 1989; Xiong et al., 2020). For example, once someone identifies a pattern in a visualization, it cannot be unseen, and they severely overestimate the extent to which others will see the same pattern. This is a key motivating factor for why it is necessary to conduct descriptive work to understand how education decision-makers reason about evidence in ways that may be different from trained researchers.

Research in cognitive science brings much to bear on this conversation, as descriptively understanding how people’s brains process information and visualizations can go a long way in effective communication. For example, there is substantial literature demonstrating that in general, people have poor statistical reasoning skills and that statistical misconceptions are widespread and persistent among researchers and lay people alike (Badenes-Ribera et al., 2017; Bar-Hillel & Neter, 1993; Garfield, 2002; Garfield & Ahlgren, 1988; Tversky & Kahneman, 1974). The reader is perhaps familiar with misuse and long-standing critique of *p* values and statistical significance (e.g., Cohen, 1994; Haller & Krauss, 2002; Rosenthal & Gaito, 1963; Wasserstein & Lazar, 2016), but there is also evidence to suggest reasoning about confidence intervals is similarly fraught. For example, Belia et al. (2005) indicate that reasoning about confidence intervals is error-prone, particularly when the task involves using confidence intervals to make a judgment about the difference between two means. Coulson et al. (2010) found that people tended to invoke the Null Hypothesis Significance Testing (NSHT) framework when interpreting confidence intervals, and that those who did were more likely to incorrectly evaluate the evidence than those who did not invoke the NHST framework. In general, people's difficulty reasoning about randomness and uncertainty is well-documented, which poses inherent challenges to the communication of statistical evidence (e.g., Castro Sotos et al. 2007; Falk & Konold, 1997; Garfield & Ahlgren, 1988; Goldstein & Rothschild, 2014; Joslyn & LeClerc, 2013; Tversky & Kahneman, 1974, 1982). Furthermore, research on data use and data-based decision-making in education suggests that overall data literacy among educators remains low, as curriculum in colleges of education provide limited training on data use (Mandinach & Gummer, 2016; Mandinach & Schildkamp, 2021).

Not only do we assert that the education research community should look to the statistical cognition literature and pay heed to the statistical reasoning of education decision-makers, it is also advantageous to turn to the data visualization and human computer interaction literatures. These offer a robust evidence-base demonstrating how people interpret data visualizations and cognitive pitfalls in reasoning visually about data (Cleveland & McGill, 1984; Correll & Gleicher, 2014; Dimara et al., 2020;; Franconeri et al., 2021; Heer & Bostock, 2010; Hullman et al., 2019; Qiao & Hullman, 2018; Xiong et al., 2020). For example, choice of y-axis scale, use of three-dimensional visualizations, and visually encoding size by scaling it to area as opposed to length can all lead to distortions and perceptual errors when mapping visual ratios back to numbers (Franconeri et al., 2021; Huff, 1954; Tufte, 1983). Researchers have also demonstrated optical illusions in line graphs with multiple lines that make it difficult to visually estimate the differences between lines, especially lines with steep slopes (Cleveland & McGill, 1984; Franconeri et al., 2021).

There is a growing body of literature on visual depictions of uncertainty and demonstrations that common visual displays such as error bars are prone to misconceptions (Belia et al., 2005; Correll & Gleicher, 2014; Franconeri et al., 2021; Hullman et al., 2015, 2019). In Figure 2, we provide a series of examples of error bars and related misunderstandings. Here we focus on bar charts, where the mean or other point estimate is represented by the height of the bar, and the error is represented by whiskers. In practice, error bars are often not well-labeled and are visually depicted in the same way regardless of whether the error bars represent standard errors or 80%, 90%, 95%, or 99% confidence intervals, or even other statistical measures of spread such as standard deviations and interquartile range. For example, in Figure 2, the first plot (1) shows error bars that represent +/- one standard error, and the second (2) shows the same data where the error bars represent 95% confidence intervals. As Correll and Gleicher (2014) point out, each relies on a different visual heuristic for "inference by eye." It should also be noted that standard deviations and interquartile ranges are measures of spread in raw data and should not be conflated with measures of uncertainty used for inference. Even if the error bars were clearly labeled or visually distinguished, they

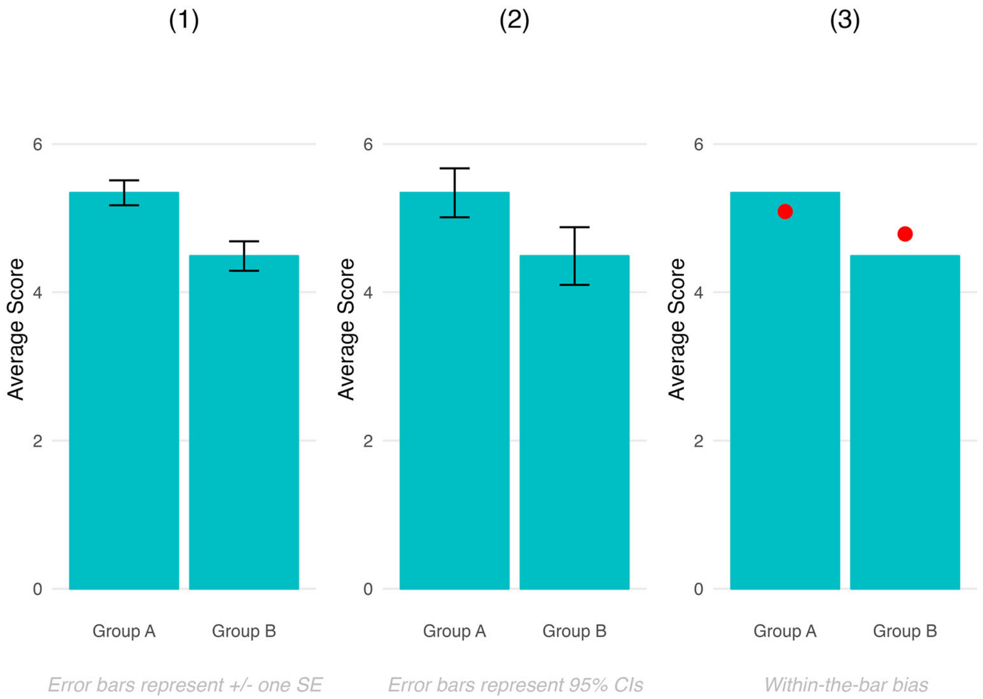


**Figure 2.** Error bars and within-the-bar bias.

would still suffer from the fundamental cognitive difficulties people have with these statistical ideas and their relation to one another. For example, Belia et al. (2005) find that even researchers have a difficult time discerning that standard error bars and confidence intervals depict two very different levels of precision.

Error bars added to bar charts also suffers from "within-the-bar bias," which is a phenomenon in which people think values within the area of the bar are more likely to occur than those that fall beyond the height of the bar. For example, in the third plot in Figure 2 (3), people tend to judge the red dot in Group A as a likely outcome more so than they judge the red dot in Group B as a likely outcome, because the former falls within the bar, whereas the latter falls beyond it. In reality, both fall 1.5 standard errors away from their respective means. Newman and Scholl (2012) found that this phenomenon affects people's statistical interpretation of the data as well as their decision-making.

We highlight these examples from the cognitive science and data visualization literatures not as an exhaustive list of the descriptive work on statistical and visual cognition (far from it) but as a general caution that there are inherent cognitive difficulties that complicate the "message sent" to "message received" pipeline, and effective communication of education research must attend to these challenges. For the most part, the education research community is not in close conversation with the statistical cognition or data visualization literatures. This may be partly because these literatures do not focus explicitly on education decision making; indeed, much of the research in both cognitive science and data visualization focus on cognition among researchers or undergraduates. Thus, we know little about how decision-makers in general, and education practitioners in particular, understand the statistical evidence presented to them. Additionally, what literature there is has focused on misconceptions regarding null-hypothesis significance testing, *p* values, and confidence-intervals, and less so on the concepts essential here, such as effect sizes, meta-analysis, causality, and validity.

We are aware of several recent studies that offer a starting point for the type of descriptive research needed in education. Bowers and colleagues investigate how practitioners and education decision-makers make sense of their own institutional data, providing important descriptive insight into how practitioners reason about data more broadly (e.g., Bowers, 2021; Bowers & Krumm, 2021). For example, they suggest that educators do not often engage with data dashboards despite their availability, and there is often a disconnect between what educators want and what data scientists create: as a whole, educators "wanted to discuss what data were most important for their current problems of practice, and how they could access useful summaries, metrics, comparisons, and visualizations that help support actions and next steps for instructional and organizational improvement" (Bowers, 2021). Part of the role of descriptive work is to not only illuminate gaps in how people understand the evidence presented but also gaps in the evidence provided versus evidence desired.

Fitzgerald and Tipton (2022) conduct a statistical cognition experiment to evaluate how education practitioners (versus researchers) make sense of meta-analytic findings via data visualizations. They find that practitioners had difficulty interpreting traditional displays of meta-analytic data such as forest plots, which make use of confidence interval bars. In general, practitioners did not exhibit consistent intuition regarding meta-analytic weight or the fact that a meta-analytic summary offers a more precise estimate of the average treatment effect than individual studies. Exploratory analysis also indicated practitioners might engage in "vote-counting" behavior whereby they judge evidence by counting the number of positive and negative effects; vote-counting is known—from a normative perspective—to be a poor way to synthesize evidence (Hedges & Olkin, 1980). Fitzgerald and Tipton (2022) also find evidence that using bar plots (e.g., Figure 1) to display effect sizes can lead to poor statistical reasoning in the meta-analytic context; the largest effect sizes are depicted by the longest bars and therefore receive the most visual attention, regardless of whether they were precisely estimated, leading users to give large effect sizes undue emphasis in their reasoning.

Nakajima (2021) and Vivalt et al. (2022) both provide results of experiments focused on examining policymakers' preferences when decision-making. In different audiences (one domestic, the other international), these studies both show that decision-makers prefer large studies in contexts similar to their own, while having little to no preference for experimental over observational studies. In particular, Vivalt et al. (2022) shows that decision-makers show a preference for study characteristics associated with external validity (e.g., location, recommendation by local expert) rather than internal validity (e.g., study design, narrow confidence interval), whereas the reverse is true for researchers. Nakajima (2021) also finds that policymakers do update their beliefs in response to research evidence, but that the effect is large and persistent only when the explanation provided for how the evidence was generated is brief and accessible.

Lortie-Forgues et al. (2021) provide an empirical study of how teachers understand different effect size metrics such as months of progress (used by EEF) and percentile gain (used by WWC). They find that the metric on which evidence is presented greatly influences teachers' level of engagement with the evidence as well as their perception of the effectiveness of the intervention. For example, teachers perceived an intervention to be most effective when the effect was presented as months of progress, and lowest when presented in terms of change in test score. Lortie-Forgues et al. point out that the dearth of research regarding how teachers perceive effect size metrics is particularly surprising when contrasted with the plethora of analogous research in the health sciences that evaluates how clinicians and patients perceive various ways of communicating treatments' outcomes. Placing their work in the language and framework of the present paper, we believe they have identified an important knowledge mobilization problem, and we echo their calls that much more (*descriptive*) work is needed to understand how practitioners reason about and interpret effect sizes and to (*prescriptively*) establish best practices for communicating them.

## Prescriptive

The previous two sections indicate multiple places for disconnects between researcher norms and decision-maker practices. Between these two sits the prescriptive facet, which asks, "*What numerical, graphical, or other information should be presented so that target readers will understand most accurately what was found and what conclusions are justified?* (Beyth-Marom et al., 2008). In education research, the "target readers" may be a variety of education decision-makers such as school district staff, policy-makers (local, state, federal), principles, teachers, or other school administrators. In short, the prescriptive facet can be thought of as interventions—that is, changes to our communication practices—to narrow the gap between the normative and the descriptive—between the message sent and the message received.

Perhaps the most helpful literature here is that in data visualization and Human Computer Interaction (HCI). For an excellent review of data visualization best practices, we point the reader to a recent review paper by Franconeri et al. (2021), *The Science of Visual Communication: What Works.* They outline many common visual illusions and misperceptions to avoid as well as offer recommendations such as prioritizing positional (i.e., x- and y-axis) encodings; utilizing effective color pallets to both design for color-vision impairments and facilitate more effective group comparisons; and guiding users to desired comparisons via annotations, visual grouping cues, and directly depicting deltas.

Beyond looking to the visualization and HCI literature for best communication practices, their studies also serve as a useful model for *how* to study such problems and establish evidence-based prescriptive advice. Typically, these researchers begin by drawing on descriptive work regarding what is known about visual cognition, statistical reasoning, and decision-making. Next, they consider any known cognitive biases or misconceptions (discrepancies in comparison to some norm), and design new visualizations to overcome these misconceptions (prescriptive). Importantly, the design of new visualizations is not the end of the prescriptive work; instead, such new visualizations are then *evaluated* empirically via small randomized experiments. Evidence from these experiments in turns adds to descriptively what is known as well as serves as the basis for prescriptive establishment of best practices (e.g., Correll & Gleicher, 2014;

Padilla et al., 2020). For example, Correll and Gleicher (2014) use the known biases regarding error bars to inform the design of new visual encodings of uncertainty, which they then evaluate empirically; the empirical evidence leads them to prescribe encodings that are symmetric and continuous such as gradient plots and violin plots to protect against both within-the-bar-bias as well as dichotomous interpretation. Many studies of this type can be found in the journal *IEEE Transactions on Visualization and Computer Graphics* and in proceedings of the related VIS Conferences.

Fitzgerald and Tipton (2022) is an example of this type of study in education research. The authors designed a new visualization for communicating meta-analytic data to education decision-makers to overcome known cognitive biases with confidence interval bars and bar plots. They then evaluated this plot with both experts and non-experts and found that it was effective at improving meta-analytic reasoning.

The National Academies report on the future of IES highlights data visualization research as one important aspect of knowledge mobilization research needed in education, stating that "research comparing different types of visualizations, both static and dynamic, are themselves worthy of scientific study." Not only does the education decision-making context include particularities not yet studied in the data visualization and HCI, but existing literature suggests that visualizations and data exploration tools that work in one field may not necessarily be appropriate in another field (Padilla et al., 2018; National Academies of Sciences et al., 2022). We strongly echo the call that there is a "pressing need for studies regarding data visualization" in education (National Academies of Sciences et al., 2022).

## Integration of the Normative, Descriptive, and Prescriptive

While we have introduced each of these facets separately, ideally there is an interplay between the normative, descriptive, and prescriptive, with each having the potential to mutually inform the others. In fact, we argue that any effective knowledge mobilization enterprise—for example, to improve communication of statistical evidence—in education is best served by a robust integration of the three facets, where none gets siloed, and in which each are considered equally important. This integration with healthy feedback loops may include a single researcher thinking holistically and attending to all three facets within their own program of research or letting others' work in one facet inform their work in another. In this section, we discuss the mutual paths of influence between each of the facets in education research.

### *Descriptive ↔ Normative*

We have already discussed one aspect of the interplay between descriptive and normative, whereby misconceptions are identified by comparing how decision-makers *do* reason about the evidence (descriptive) to the normative standard of how they *should* reason about the evidence. An example is when the descriptive work in Fitzgerald and Tipton (2022) described above found that when presented with meta-analytic data in a forest plot, education practitioners often identify the effect size with the *widest* confidence interval as the one they think should receive the most weight in the meta-analytic summary (the opposite of what researchers intended). By empirically evaluating a visualization used to display evidence, they were able to identify that there is a gap between how education practitioners do and how they should reason about the evidence. Similarly, the descriptive work in Vivalt et al. (2022) illuminated that due to their high-value of external validity, policy-makers were willing to adopt a program that had a 6.3% points lower effect (outcome: enrollment rates) if that program was recommended by a local expert, and a 4.5% point lower effect if it had been evaluated in their own country; when using normative knowledge about typical effects, researchers point out that these tradeoffs are in some cases larger than the effect of the program.

Descriptive work can also enlighten education researchers to the needs, considerations, and constraints faced by education decision-makers which can illuminate the need for new methodological developments. For example, Nakajima (2021) found via a discrete choice experiment that education policy-makers show preference for studies in contexts like theirs. This is in line with the National Academies recommendations, which highlight that

> Decision-makers are rarely interested in the average impact of an intervention; instead, they want to understand the projected effect in *their* local context, often for a specific student population. This suggests that the primary focus on "the effect" of an intervention – at any stage of research – is likely inappropriate. (National Academies of Sciences et al., 2022)

Together these suggest that decision-makers seek localized predictions of treatment effects—information on if, how, and to what extent an intervention could work in their context. Meeting this need would require further methodological (normative) work in areas such as experimental design, machine learning, and Bayesian Additive Regression Trees.

## *Descriptive ↔ Prescriptive*

Descriptive work that illuminates how education decision-makers reason about evidence can and should inform prescriptive work that recommends better forms of communication. For example, the descriptive work of Lortie-Forgues et al. (2021) regarding teachers' perceptions of effect size metrics led them to prescriptive recommendations; they caution that both percentile gain and months of progress should be avoided as they may lead teachers to have higher expectations for improvements on students' raw scores than those metrics actually imply. Notably, a first implication of descriptive work is often to illuminate what *not* to do, as well as highlight the need for further prescriptive work to develop alternatives that do better.

Importantly, there is a necessary feedback loop between the descriptive and the prescriptive. New prescriptive ideas (e.g., a new metric, or a new visualization) that are informed by descriptive illuminations must then in turn be evaluated. It is not uncommon for researchers to informally identify perceived gaps between the descriptive and the normative and offer prescriptive recommendations, but it is less common to empirically evaluate those prescriptions to determine if they are effective in closing that gap. For example, effect sizes in education are often reported statistically as standardized mean differences on the metric of Cohen's *d*, with many arguing that this effect size metric is more natural than others (e.g., Cooper et al., 2019). However, researchers have surmised that it is unlikely that the average lay person has any intuition about how a Cohen's $d$ of 0.2 should be practically interpreted in terms of student outcomes. Here researchers have informally identified a gap between the descriptive and the normative. To remedy this problem, other metrics have been proposed: the WWC converts effect sizes to an Improvement Index (which functions as percentile gain), and the Education Endowment Foundation (EEF) converts to "months of learning." However, it wasn't until Lortie-Forgues et al. (2021) empirically evaluated how teachers perceive these different metrics that it was determined that they might not be all that useful after all.

This feedback loop is at the heart of what we hope this framework can offer the knowledge mobilization enterprise. Education researchers should look to the descriptive evidence that exists to identify needs and inform our development of prescriptive strategies, and those proposed strategies should in turn be evaluated, adding to the descriptive evidence-base of how people reason about evidence. Importantly, we are not suggesting every strategy or recommendation must be evaluated by a randomized experiment, but we are asserting that we as education researchers should be in the business of collecting data and evaluating the effectiveness of our own practices. But in general, making our own practices the object of study and evaluation is an important and necessary component of improving the use and usefulness of education research.

Again, cognitive science and HCI offer many useful examples for how to utilize this feedback loop well. Researchers in those fields have developed a robust evidence base about cognitive load, working memory limits, and how the brain's visual system processes information; ongoing studies then use that descriptive knowledge to intentionally design visualizations such that extracting information will involve tasks that are easier for the visual system to process and will lead to less cognitive errors. For example, an understanding that the visual system is slow at processing comparisons (e.g., "this bar is taller than that bar") has led to the prescriptive advice to "directly depict your deltas," and that advice has been empirically evaluated and found to be a more effective way of visualizing differences (Franconeri et al., 2021; Nothelfer & Franconeri, 2020). Importantly, this robust feedback loop is not established within a single study or even by a single researcher, but can be achieved through researchers thinking holistically about the normative-descriptive-prescriptive throughout their program of research.

## Normative ↔ Prescriptive

What is normative influences the prescriptive in that prescriptive work is concerned with developing communication strategies that help facilitate normative evaluations of the evidence. For example, in any attempt to communicate statistical evidence, education researchers should consider the statistical processes that generated that evidence and what would constitute a statistically valid interpretation of that evidence. Communication strategies should then intentionally be chosen (or developed) to facilitate valid interpretations by education decision-makers. An example of this process can be found in the development of a new visualization for meta-analytic data by Fitzgerald and Tipton (2022). They found that by intentionally implementing visual encodings and annotations the resulting plot facilitated a more normative evaluation of the evidence by education decision-makers, compared to other visualizations used in practice.

Importantly, the goal should not be to ensure education decision-makers have a robust understanding of each relevant statistical norm. Nevertheless, these norms do have implications for appropriate interpretation of the statistical evidence education decision-makers consume. For example, because clustering is common in education research, and it is the number of clusters not the number of students, that drives precision of effect size estimates, statistical norms from meta-analysis and hierarchical linear modeling tell us that simply giving the most weight to the study with the largest number of *students* may be an inappropriate way to synthesize and reason about the evidence. Paying explicit attention to these norms when designing visualizations or dashboards alerts us to the fact that displaying only the number of students in a study, as in the WWC example in Figure 1, can be a misleading proxy for precision and potentially lead to poor evaluation of the evidence. Care should be taken to communicate statistical evidence in a way that facilitates evaluations of the evidence that align with statistical norms. It is the role of the prescriptive facet—informed by the normative—to develop such communication strategies.

Importantly, the prescriptive facet in some instances can and should challenge the norms of researchers. For example, when there is a misalignment between the priorities of decision-makers and researchers, prescriptive work should privilege norms that are attentive to decision-maker priorities and constraints as well as spur additional normative work when existing methods do not meet decision-makers' needs. For example, because it is clear decision-makers value external validity, prescriptive work must determine how best to embed external validity into visualizations and dashboards. This process will require researchers to consider the perspective of decision-makers more deeply, and doing so may illuminate methodological gaps in facilitating localized decision-making.

## Discussion

This paper has presented a knowledge mobilization framework—of normative, descriptive, and prescriptive research—and applied it to the problem of communicating statistical information to facilitate evidence-based decision-making. Our framework and the logic of the present paper draws heavily from the Beyth-Marom et al. (2008) adaptation of Bell et al.'s taxonomy for decision-making. As we end this paper, we wish to elevate three findings.

***First, the norms, values, and expertise that researchers apply to their own decision-making may not apply to those making decisions in schools.*** The framework proposed in this paper highlights these preconceptions held by researchers and makes explicit how these norms and priorities are embedded into the evidence researchers generate and communicate (or don't). As a result, the "message sent" may not be the "message received". Some of this mismatch has to do with the "curse of expertise"—researchers simply cannot unknow what they know. But some of this mismatch also has to do with an understanding of the decision-making process that is more hypothetical and stylized than rooted in empirical practice. One of the strengths of the framework proposed here is that it does not treat this mismatch as a unidirectional problem, in which decision-makers misunderstand researchers. Instead, the framework provides a system of normative and descriptive knowledge that places researchers and decision-makers on equal footing and illuminates discrepancies between the two. Importantly, failure to examine embedded norms can lead to other, less benign, consequences beyond a simple mismatch in “message sent” ≠ “message received”. When norms are set solely by researchers who may be construed as “elites,” the resulting priorities and research agendas are likely to disenfranchize some groups and privilege others. To combat this, we see promise in participatory research methods that engage community stakeholders in co-creating research norms (e.g., Peko-Spicer, 2023; Salimi et al., 2012; Swartz & Nyamnjoh, 2018).

***Second, education decision-makers are not all the same, nor are the decisions, resources, and constraints they face.*** This is important since it suggests that simply including one or two “practitioners” on research panels or as dissemination experts is not enough to close the normative—descriptive gap. Large urban school districts have different resources than the more common, smaller districts. In the former, it may be district level curriculum experts—some with PhDs—making decisions, whereas in the latter, this is less likely. Future prescriptive work should consider how different communication practices may differentially land for people with different educational backgrounds, for example. Schools serving diverse bodies of students—with multiple nationalities, languages, and backgrounds—face different opportunities and constraints than those serving primarily White or minoritized students. Evidence syntheses and clearinghouses that focus only on internal validity simply do not address these well-founded concerns regarding if interventions can, and will, work in local conditions. Here our framework provides strategies for better describing these different education decision-makers, as well as prescribing and testing strategies for mobilizing knowledge across them.

***Third, to mobilize knowledge, we need to develop an integrated science.*** Until now, the translation and dissemination strategies developed and implemented have often been disconnected from the scientific disciplines—data visualization, cognitive psychology, human-computer interaction—that are most relevant to effective communication. Treating these problems as a science allows principled, careful study—specific to the education context—and the development of general theories and strategies. By doing so, we can ensure that the science of learning that our field has so robustly developed over the past two decades can make it to the schools and students that need it.

Finally, as we close this paper, we want to remind readers that while we intentionally kept a relatively narrow scope by focusing on the communication of statistical evidence from education research studies, the lessons involved apply more broadly. The recent calls for knowledge mobilization research, reflected in the National Academies report and elsewhere, assert that strategies for mobilizing knowledge should be studied directly and that the education research community should be in the business of collecting data and evaluating our own practices. As we have shown here, any effective knowledge mobilization enterprise must consider the needs of a broad range of educational decision-makers and communities, question the norms of and stylized facts about decision-makers held by researchers, and subject approaches to knowledge mobilization with the same scrutiny as we do interventions themselves. The normative, descriptive, prescriptive framework offers a way forward for establishing this integrated science.

## Funding

This work was supported by U.S. Department of Education, Institute of Education Sciences, Multidisciplinary Program in Education Sciences [Grant Award R305B140042].

## ORCID

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

## References

Badenes-Ribera, L., Frías-Navarro, D., & Bonilla-Campos, A. (2017). Effect size and meta-analysis in Spanish professional psychologists. *European Journal of Investigation in Health, Psychology and Education*, *7*(2), 111–122. https://doi.org/10.3390/ejihpe7020008

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology*, *65*(6), 1119–1131. https://doi.org/10.1037/0022-3514.65.6.1119

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

Bell, D. E., Howard, R., Amos, T. (1988). Descriptive, normative, and prescriptive interactions in decision making. In D. E. Bell, H. Raiffa, and A. Tversky (Eds.), *Decision making: Descriptive, normative, and prescriptive interactions*. Cambridge University Press. https://doi.org/10.1017/CBO9780511598951.003

Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal*, *7*(2), 20–39. https://doi.org/10.52041/serj.v7i2.468

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382–386. https://doi.org/10.1111/j.1467-9280.2007.01909.x

Bowers, A. J. (2021). *Data visualization, dashboards, and evidence use in schools: Data collaborative workshop perspectives of educators, researchers, and data scientists*. https://doi.org/10.7916/d8-jj2g-e225.

Bowers, A. J., & Krumm, A. (2021). Supporting the initial work of evidence-based improvement cycles through a data-intensive partnership. *Information and Learning Sciences*, *122*(9/10), 629–650. https://doi.org/10.1108/ILS-09-2020-0212

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*(5), 1232–1254. https://doi.org/10.1086/261651

Castro Sotos, A. E., Vanhoof, S., Noortgate, W. V. d., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98–113. https://doi.org/10.1016/j.edurev.2007.04.001

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

Coburn, C. E., Penuel, W. R., & Farrell, C. C. (2021). Fostering educational improvement with research-practice partnerships. *Phi Delta Kappan*, *102*(7), 14–19. https://doi.org/10.1177/00317217211007332

Cohen, J. (1994). The earth is round ($p < 05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Conaway, C. (2021). Funding research that is useful and used. *Education Next* (blog). https://www.educationnext.org/funding-research-that-is-useful-and-used-20-years-institute-education-sciences/

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

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*, 26. https://doi.org/10.3389/fpsyg.2010.00026

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. https://doi.org/10.1109/TVCG.2018.2872577

Education Endowment Foundation. (n.d.). *Teaching and learning toolkit*. EEF. https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301–318. https://doi.org/10.1037/0033-295X.104.2.301

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, *47*(4), 235–245. https://doi.org/10.3102/0013189X18761042

Farrell, C. C., Penuel, W. R., & Davidson, K. (2022). What counts' as research? Comparing policy guidelines to the evidence education leaders report as useful. *AERA Open*, *8*(January), 233285842110731. https://doi.org/10.1177/23328584211073157

Fitzgerald, K. G., & Tipton, E. (2022). The meta-analytic rain cloud plot: A new approach to visualizing clearinghouse data. *Journal of Research on Educational Effectiveness*, *15*(4), 848–875. https://doi.org/10.1080/19345747.2022.2031366

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.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3), 2. https://doi.org/10.1080/10691898.2002.11910676

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

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1–14. https://doi.org/10.1017/S1930297500004940

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*(January), 1–20.

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., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*(2), 359–369. https://doi.org/10.1037/0033-2909.88.2.359

Hedges, L. V., & Rhoads, C. (2010). *Education power analysis in education research*. https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. *CHI*. http://vis.stanford.edu/files/2010-MTurk-CHI.pdf

Huff, D. (1954). *How to lie with statistics*. W. W. Norton & Company.

Hullman, J., Qiao, X., Correll, M., Kale, A., & Kay, M. (2019). In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

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. (2022). *Request for applications: Education research grants program 84.305A*. Indexes. https://ies.ed.gov/funding/22rfas.asp.

Institute of Education Sciences, What Works Clearinghouse. (2021). *WWC summary of evidence for this intervention: Cognitive Tutor® Algebra I*. IES > WWC What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Intervention/818.

Jackson, C. (2022). Democratizing the development of evidence. *Educational Researcher*, *51*(3), 209–215. https://doi.org/10.3102/0013189X211060357

Joslyn, S., & LeClerc, J. (2013). Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, *22*(4), 308–315. https://doi.org/10.1177/0963721413481473

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

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*(November), 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Mandinach, E. B., & Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation, from Data-Driven to Data-Informed Decision Making: Progress in the Field to Improve Educators and Education*, *69*(June), 100842. https://doi.org/10.1016/j.stueduc.2020.100842

Nakajima, N. (2021). Evidence-based decisions and education policymakers. *Unpublished Paper*. https://nozominakajima.github.io/files/nakajima_policymaker.pdf.

National Academies of Sciences, Engineering, and Medicine. (2022). *The future of education research at IES: Advancing an equity-oriented science*. The National Academies Press. https://doi.org/10.17226/26428

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

Nothelfer, C., & Franconeri, S. (2020). Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 311–320. https://doi.org/10.1109/TVCG.2019.2934801

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

Padilla, L. M. K., Powell, M., Kay, M., & Hullman, J. (2020). Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, *11*(January), 579267. https://doi.org/10.3389/fpsyg.2020.579267

Peko-Spicer, S. (2023). *Lessons learned from a participatory evidence synthesis.* https://www.srma-sig.org/seminar/2022-23-seminars/

Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., Farrell, C., & Allen, A.-R. (2017). How school and district leaders access, perceive, and use research. *AERA Open*, *3*(2), 233285841770537. https://doi.org/10.1177/2332858417705370

Penuel, W. R., Farrell, C. C., Allen, A.-R., Toyama, Y., & Coburn, C. E. (2018). What research district leaders find useful. *Educational Policy*, *32*(4), 540–568. https://doi.org/10.1177/0895904816673580

Qiao, X., & Hullman, J. (2018). *Translating scientific graphics for public audiences* [Paper presentation]. Proceedings VisGuides: 2nd Workshop on the Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization, 4.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, *55*(1), 33–38. https://doi.org/10.1080/00223980.1963.9916596

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Salimi, Y., Shahandeh, K., Malekafzali, H., Loori, N., Kheiltash, A., Jamshidi, E., Frouzan, A. S., & Majdzadeh, R. (2012). Is community-based participatory research (CBPR) useful? A systematic review on papers in a decade. *International Journal of Preventive Medicine*, *3*(6), 386–393.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87. https://doi.org/10.3102/1076998607302714

Singer, J. D. (2019). Reshaping the arc of quantitative educational research: It's time to broaden our paradigm. *Journal of Research on Educational Effectiveness*, *12*(4), 570–593. https://doi.org/10.1080/19345747.2019.1658835

Society for Research on Educational Effectiveness. (n.d.). *Conferences*. https://www.sree.org/conferences.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2001). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Swartz, S., & Nyamnjoh, A. (2018). Research as freedom: Using a continuum of interactive, participatory and emancipatory methods for addressing youth marginality. *HTS Teologiese Studies/Theological Studies*, *74*(3), 11. https://doi.org/10.4102/hts.v74i3.5063

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524. https://doi.org/10.3102/0013189X18781522

Tufte, E. (1983). *The visual display of quantitative information*. Graphics Press.

Tversky, A., & Kahneman, D. (1982). Belief in the law of small numbers. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 23–31). Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vivalt, E., Coville, A., & Kc, S. (2022). *Weighing the evidence: Which studies count?* https://evavivalt.com/wp-content/uploads/Weighing-the-Evidence.pdf

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on P-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

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