Introduction

As an applied science, education research should be closely integrated into school and district decision-making. Yet, despite longstanding knowledge mobilization efforts, education leaders still report relying on social connections, intuition, and Google at higher rates than research evidence (Farley-Ripple, E., 2018). Education leaders indicate two main reasons for low use of research evidence: (1) Challenges accessing strong evidence and (2) Difficulty understanding the upshot of the evidence (Shewchuck, S., & Farley-Ripple, E., 2022; Chuter, C., 2022). School and district leaders often find it overwhelming to sift through numerous studies of questionable quality themselves (Cooper et al., 2018). Given these challenges, meta-analysis is uniquely positioned to summarize the results of well-designed and conducted causal studies.

Unfortunately, traditional meta-analysis is both costly and time-consuming, often exceeding $1 million and taking three or more years to complete. Additionally, due to rapid changes in policy and demography, and technological advancement, a systematic review may become outdated within just a decade of its completion. Recently, “living” systematic reviews have emerged to resolve this problem, but these remain time and resource-intensive. Furthermore, if a particular question they are interested in has not been reviewed, education leaders are simply out of luck. Traditional meta-analysis is not affordable or timely enough for many local education leaders.

Emergence of Affordable AI

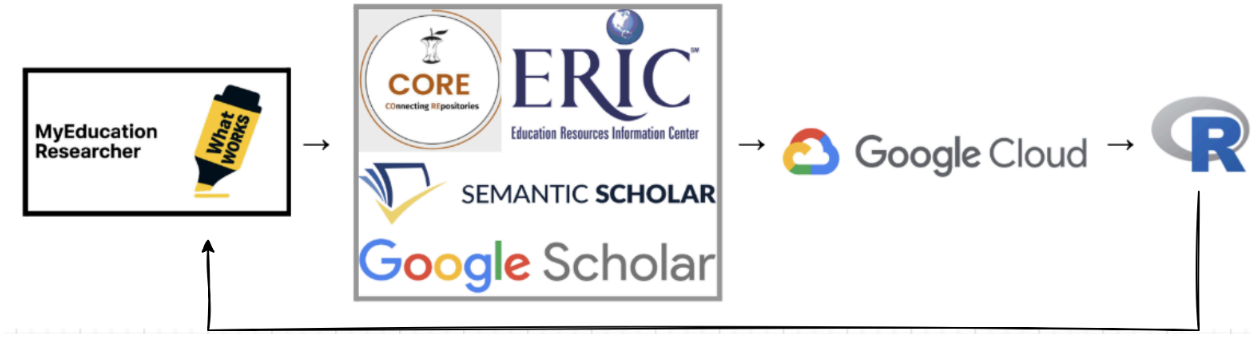
To effectively bridge the research-practice gap, the field requires a tool that can perform meta-analysis more swiftly and affordably, without sacrificing validity. Recent advancements in large language models' ability to identify and extract data from large amounts of text brings the possibility of such a tool into view. Currently, some AI-assisted tools provide partial automation—such as title and abstract screening (e.g., Abstrakr), and data extraction (e.g., MetaMate). However, these systems do not facilitate comprehensive end-to-end meta-analysis.

Systematic Review & Meta-Analysis Web Application

MyEducationResearcher (MER; patent pending; CLHA.001.PPA - Provisional Application - US Application No. 63/929,668) is a free web application that runs meta-analyses on any causal research question, and summarizes the quantitative outcomes. Specifically, it produces an average Hedges' *g* effect size of the specified intervention, and prediction interval across included studies. To complete its review, MER follows this process:

Figure 2

MyEducationResearcher Process

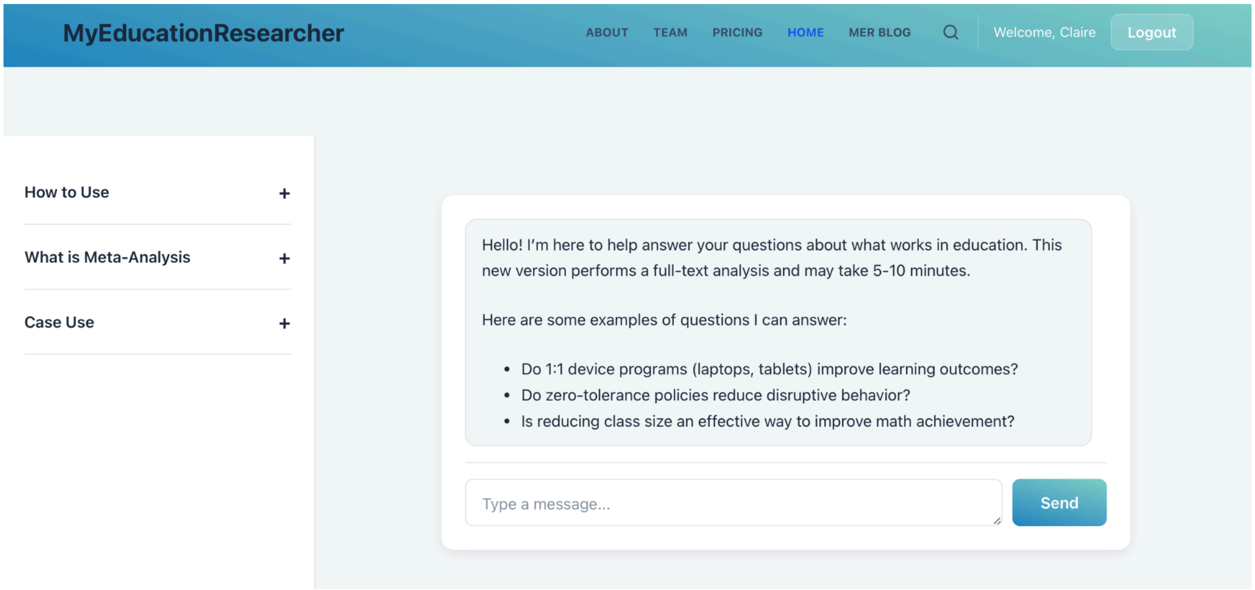


Step 1: Data Collection: Based on a user's research question, MER builds multiple search strings. The User Interface (UI) is built to understand what kinds of questions it can answer. For example, if a user were to input the question, "What is a chair?" MER would output: "That is not a relevant question."

Once it receives a causal research question, MER initiates the data collection process. First, it throws a request to multiple open access sites including Semantic Scholar, Google Scholar, and CORE (see Figure 1). This works the same as if someone were to type the question into the search bar on those sites. From each of these sites, the first 10 full-text articles are retrieved and dropped into a Google Cloud temporary storage space. Following review, these articles are automatically deleted from storage.

Figure 1

MyEducationResercher User Interface



Hello! I'm here to help answer your questions about what works in education. This new version performs a full-text analysis and may take 5-10 minutes.

Here are some examples of questions I can answer:

- Do 1:1 device programs (laptops, tablets) improve learning outcomes?
- Do zero-tolerance policies reduce disruptive behavior?
- Is reducing class size an effective way to improve math achievement?

For long-term use of MER, we plan to increase full-text retrieval substantially. However, because this project is currently unfunded it is currently set at a lower rate of retrieval to save on server space.

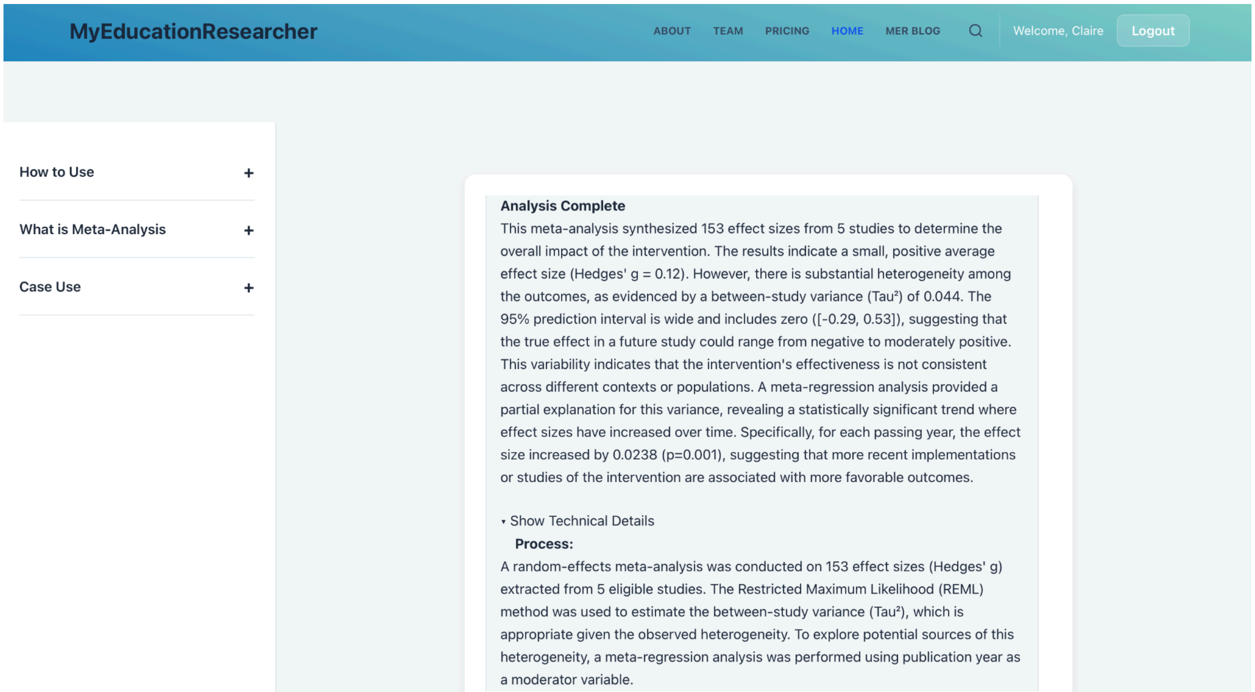
Step 2: Screening: Full-text studies are reviewed according to What Works Clearinghouse Procedures and Standards Handbook, Version 5.0 (WWC, 2022; Appendix A). These include inclusion criteria for study design, baseline equivalence, and confounding factors. We are in the process of testing code that will screen for compositional change and outcome measures as well.

Step 3: Data Extraction: Any statistical information that may be used to derive Hedges' *g* and its standard error for each effect size, is extracted from studies in the final inclusion pool (Appendix B). The information provided in Appendix B encompasses all data MER can use to derive Hedges' *g* for the ith effect size of each study (g_i) and the standard error of the effect, for the *i*th effect size of each study ($SE[g_i]$).

Step 4: Analysis: MER uses its own code within *R* to convert extracted statistical data into effect size (*g*) and standard error information (The application only supports continuous outcomes at the moment). Whenever possible, MER adjusts for pretest differences between conditions using a difference-in-difference effect size estimation procedure (WWC, 2022). When multiple effect size data are provided MER prefers adjusted to unadjusted effects sizes, unless goodness-fit is smaller in the more adjusted model. MER prefers regression coefficients over F and t-statistics, which it prefers over means and standard deviations.WWC statistical conversion formulas are built into MER's code (in Python) to calculate Hedges' g for each relevant outcome (see Appendix C for full list of formulas).

This data table is then passed into *R*, which uses the *metafor* package to conduct a linear regression model. Within *R*, MER conducts an average, random-effects model with robust variance estimation (Hedges et al., 2010) to produce a weighted average of the effect sizes. We are in the process of adding code to translate the overall average effects into the probability of positive impact (PPI), which some suggest is an easier metric to understand (Mathur & VanderWeele, 2020).

Step 5: Reporting: Back in the UI, the user sees the average effect size and the prediction interval in response to their research question. For an example of reporting within MER, see Figure 2).



Analysis Complete

This meta-analysis synthesized 153 effect sizes from 5 studies to determine the overall impact of the intervention. The results indicate a small, positive average effect size (Hedges' g = 0.12). However, there is substantial heterogeneity among the outcomes, as evidenced by a between-study variance (Tau²) of 0.044. The 95% prediction interval is wide and includes zero ([-0.29, 0.53]), suggesting that the true effect in a future study could range from negative to moderately positive. This variability indicates that the intervention's effectiveness is not consistent across different contexts or populations. A meta-regression analysis provided a partial explanation for this variance, revealing a statistically significant trend where effect sizes have increased over time. Specifically, for each passing year, the effect size increased by 0.0238 (p=0.001), suggesting that more recent implementations or studies of the intervention are associated with more favorable outcomes.

▾ Show Technical Details

Process:

A random-effects meta-analysis was conducted on 153 effect sizes (Hedges' g) extracted from 5 eligible studies. The Restricted Maximum Likelihood (REML) method was used to estimate the between-study variance (Tau²), which is appropriate given the observed heterogeneity. To explore potential sources of this heterogeneity, a meta-regression analysis was performed using publication year as a moderator variable.

Note. This project is ongoing. Functionalities of the application may have developed since this report was updated (12.7.2025).

Planned Improvements

This project is currently unfunded. With adequate funding, we intend to implement improvements at multiple points of MER’s review process:

Data Collection

- **Increased Search Capacity**: With additional funding, we plan to expand our search capabilities to retrieve the first 500 potentially relevant full texts instead of just 50.
- **Access to Educational Databases**: Funding would allow us to subscribe to educational databases, reducing our reliance on free full texts and minimizing potential biases in our results.

- **Reference Harvesting**: We would implement code for both forward and backward reference harvesting, as well as scrape common gray literature sites for potentially relevant studies.

Screening

- **Expanded Screening**: We aim to broaden our screening process to include assessments of compositional changes in quasi-experimental studies.

Analysis

- **User-Defined Covariates**: We plan to enable users to incorporate covariates based on available data, increasing the precision of focal outcomes.
- **User-Defined Moderator Analysis**: Users will have the option to select moderator analyses to investigate outcome heterogeneity.
- **Probability of Positive Impact (PPI):** We are in the process of adding code that will compute this metric.

Reporting

- **Education Practitioner-Led Discussions:** Conversations with practitioners to improve comprehension and usability of reporting
- **Enhanced Process Efficiency**: Purchasing increased memory would also significantly expedite the full MER process, which currently takes about 10 minutes per search.

Appendix A

Full-Text Study Screening Criteria

Outcome Measure Standards

Studies are excluded if they *do not* have each of the following:

- **Face validity**: the outcome measure appears measure what it claims to measure
- **Reliability**: The outcome measure must produce consistent findings
- **Not overaligned with the intervention**: The outcome measure must give unfair advantage to participants in the treatment or control condition
- **Consistent data collection procedures:** The outcome measure must be measured consistently for the groups or participants being compared

[Outcome measure standards are not yet included in MER].

Confounding Factors

Studies are excluded if they *do* have one of the following:

- The intervention or comparison group contains a single unit, such as a teacher, classroom, school, or district, that is aligned perfectly with either the comparison or intervention group. For example, if a study assigns one school to the intervention condition and two schools to the control condition. It is impossible to disentangle the intervention effect from the school effect in this case.
- The characteristics of the students or teachers differ systematically in the control versus the treatment group, in ways that are associated with the outcome. For example, if students with disabilities are given a reading intervention and students without disabilities are given the traditional reading, the study would be excluded.

- If data from intervention and control groups are collected across different semesters, quarters, or years, then time is a confounder and the study is excluded. For example, if an intervention group of students in grade 3 and a comparison group is given the traditional curriculum the following year, the two groups cannot be compared. A study must have a control group and a treatment group to be included.

Assignment to Conditions

Included studies must use one of the following designs:

- Randomized-controlled trials
- Quasi-experimental designs

Randomized controlled trials, quasi-experimental designs, and regression discontinuity designs are group designs in which participants, usually students, are assigned to intervention and comparison conditions.

Compositional Change of Intervention and Comparison Groups

Compositional change is the movement of participants into or out of the intervention or comparison groups after assignment. There are two types of compositional change: attrition and sample joining. Attrition is the lack of outcome data for participants who were assigned to either the intervention or comparison group. Sample joiners are individuals who enroll in the intervention or comparison condition after researchers have assigned students, teachers, classrooms, or schools to conditions. If sample joiners are not included in the study's analysis, then the study is still eligible. If sample joiners are included in the study's analysis they may pose a high or low risk of bias. For acceptable levels of compositional change, we use the conservative table set by WWC 5.0.

Randomized-controlled trials that do not produce evidence that compositional change has risked biasing the study are retained in the review. Quasi-experimental studies that do not produce such evidence are removed. **[Compositional change screening is not yet included in MER].**

Baseline Equivalence

If a study is a randomized controlled trial, it has assumed overall baseline equivalence. If a study is a quasi-experimental design, it *must* demonstrate baseline equivalence between treatment and control groups.

Quasi-experimental studies can satisfy baseline equivalence in one of three ways:

- Baseline differences on a student or teacher outcome of interest is equal to or less than 0.05 standard deviations
    - Demographic variables are rarely the main outcome of interest, and so can be greater than .05 standard deviations without excluding the study
    - Baseline differences on the main outcome of interest are between .05 and .25 standard deviations and apply an acceptable adjustment for baseline differences
    - An adjustment is acceptable if the pretest measure is included in the regression model, analysis of covariance, or if the authors used weighting or matching techniques to account for pretest baseline inequivalence

If there are no studies fitting this criteria, then the output is: “there is not strong enough evidence to respond confidently to this research question.”

Appendix B

Table 2

Effect Size Data that may be Extracted from Included Studies

The following data are extracted from each included study, as available.

Symbol	Description
β	Standardized regression coefficient for intervention impact estimate
ϱ_{ICC}	Intraclass correlation coefficient
$\hat{\sigma}^2$	Estimated level-1 variance for variability within individuals
τ^2	Estimated level-2 variance for variability across individuals
ω	Small sample bias correction term
$\bar{A}_{\mathrm{c}}$	Attrition rate in the comparison group
$\bar{A}_{\mathrm{i}}$	Attrition rate in the intervention group
b	Unstandardized intervention-comparison group mean difference of the outcome
$d\mathrm{f}$	Degrees of freedom
e_{ij}	Level-1 error term for case *i* at time j in a multilevel model
F	F-statistic
G	Number of mutually exclusive subgroups
g	Hedges' *g* standardized effect size
g_{i}	Effect size for the *i*th main finding in a study
$\bar{g}_{\mathrm{s}}$	Domain-level composite effect size for study s
J	Total number of studies

M	Total number of clusters in a cluster-level assignment study
n	Total number of individuals in the analytic sample
n_{c}	Number of individuals in the comparison group analytic sample
n_{i}	Number of individuals in the intervention group analytic sample
R^2	Multiple correlation between the covariates and the outcome
SD_{bc}	Baseline standard deviation for the control group
SD_{bi}	Baseline standard deviation for the intervention group
SD_{pc}	Posttest standard deviation for the control group
SD_{pi}	Posttest standard deviation for the intervention group
$S\mathrm{E}[b]$	Standard error of the unstandardized mean difference
$S\mathrm{E}[g]$	Standard error of the Hedges' g effect size
$S\mathrm{E}[g_{\mathrm{i}}]$	Standard error of the ith effect size in a study
$S\mathrm{E}[y_{\mathrm{i}}]$	Standard error of the intervention group mean
$S\mathrm{E}[y_{\mathrm{c}}]$	Standard error of the comparison group mean
t	*t* test statistic for group mean difference
u_{i}	Level-2 error term for case *i* in a multilevel model
$\bar{x}_{\mathrm{bc}}$	Baseline mean for the comparison group
$\bar{x}_{\mathrm{bi}}$	Baseline mean for the intervention group
$\bar{x}_{\mathrm{pc}}$	Posttest mean for the control group
$\bar{x}_{\mathrm{pi}}$	Posttest mean for the intervention group

Appendix C

Formulas for Effect Size Conversions

Formulas for Studies with Individual Assignment and Continuous Outcomes

*These formulas are largely pulled from the What Works Clearinghouse Procedures and Standards Handbook, Version 5.0 (2022).

Hedges' g is computed by dividing an estimate of the unstandardized mean difference *b,* by the pooled within-group standard deviation (SD_p). For the unstandardized mean difference *b,* **MER** prefers covariate-adjusted Hedges' g over non-adjusted, but unadjusted standard error of the effect size is preferred over adjusted. The most important covariate to adjust for is baseline score.

The Hedges' g computation also includes multiplication by a small-sample correction factor ω. This correction factor is needed to produce unbiased estimates of the population effect size.

Below is the formula to calculate Hedges' g:

[E.1]

$$g = \frac{\omega b}{SD_p}$$

Below are formulas for the pooled standard deviation SD_p and small-sample correction factor ω:

[E.2]

$$SD_p = \sqrt{\frac{(n_i - 1)SD_i^2 + (n_c - 1)SD_c^2}{n_i + n_c - 2}}$$

[E.3]

$$\omega = 1 - \frac{3}{4df - 1}$$

where n_i and n_c are the sample sizes for the intervention and comparison groups, *SD*i and *SD*c are the intervention and comparison standard deviations, and *d*f is the degrees of freedom. The following formula calculates the degrees of freedom for individual-level assignment studies:

[E.4]

$$df = N - 2$$

If the covariate-adjusted mean difference is not available, then **MER** uses the raw means and standard deviations for randomized-controlled studies. The formula to calculate Hedges' g using the unadjusted mean difference is:

[E.5]

$$g = \frac{\omega(\bar{y}_i - \bar{y}_c)}{\sqrt{\frac{(n_i - 1)SD_i^2 + (n_c - 1)SD_c^2}{n_i + n_c - 2}}}$$

The following formula provides the standard error for effect sizes calculated using unadjusted means (Borenstein & Hedges, 2019):

[E.6]

$$SE[g] = \omega\sqrt{\frac{n_i + n_c}{n_i n_c} + \frac{g^2}{2(n_i + n_c)}}$$

This following formulas calculates Hedges' g from the t-value:

[E.7]

$$g = \omega t\sqrt{\frac{n_i + n_c}{n_i n_c}}$$

Hedges' g is calculated in the following way for the F-value, from an ANOVA test:

[E.8]

$$g = \omega\sqrt{\frac{F(n_i + n_c)}{n_i n_c}}$$

An exact two-tailed p-value can be transformed into a t-statistic, which can then be applied in equation E.7. For example, a p-value of 0.077 for a total sample size of 80 students (refer to table E.1) corresponds to a t-statistic of 1.79, using 78 degrees of freedom (see equation E.4). One-tailed p-values can be converted to two-tailed p-values by simply multiplying by two.

However, **MER** does not perform this p-to-t conversion for inexact p-values, such as "$p < 0.05$," since these values can represent a range of effect sizes.

However, these formulas cannot be used for studies that require baseline adjustment, such as for quasi-experimental designs (QEDs) and high-attrition RCTs because they are based on unadjusted mean comparisons . Effect sizes from those studies must instead use covariate-adjusted statistics that control for baseline differences.

For its calculations, **MER** requires the standard error of Hedges' g - which is standardized. If the study instead provides the standard error for the unstandardized mean difference, **MER** calculates SE[g] as follows:

[E.9]
$$SE[g] = \omega\sqrt{\left(\frac{SE[b]}{SD_p}\right)^2 + \frac{g^2}{2(n_i + n_c)}}$$

If the t-value for an unstandardized regression coefficient is reported, the standard error for the unstandardized mean differences is calculated by: SE[b] = b/t. SE[b] is then included in E.9 to arrive at SE[g].

If studies only report the intervention mean standard error $SE[y_i]$ and comparison mean standard error $SE[y_c]$, then SE[g] is calculated using the following:

[E.10]
$$SE[b] = \sqrt{SE[y_i]^2 + SE[y_c]^2}$$

A study might not provide enough information to calculate the covariate-adjusted standard error of the mean difference, making equation E.9 impractical. However, if the authors reported the multiple correlation R^2 between the covariates and the outcome, **MER** calculates the covariate-adjusted standard error for the effect size using the following method:

[E.11]
$$SE[g] = \omega\sqrt{\frac{n_i + n_c}{n_i n_c}(1 - R^2) + \frac{g^2}{2(n_i + n_c)}}$$

In contrast to unadjusted R^2 values, the **MER** favors adjusted R^2 values that consider the number of predictors included and correct for potential overconfidence in model predictions. The **MER** considers negative adjusted R^2 values as 0 percent.

If an t-test or F-test is adjusted and includes an R^2 value, then this should be added to the calculation for *g:*

[E.12]
$$g = \omega t\sqrt{\frac{n_i + n_c}{n_i n_c}(1 - R^2)}$$

[E.13]
$$g = \omega\sqrt{\frac{F(n_i + n_c)}{n_i n_c}(1 - R^2)}$$

For these formulas, the adjusted R^2 must be used.

References

Chuter, C., & Neitzel, A. J. (2022, April 7). Research Use and Knowledge Mobilization in K-12 Decision-Making. Retrieved from osf.io/bkwam

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.

Mathur, M. B., & VanderWeele, T. J. (2020). Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology, 31*(3), 356-358.

Shewchuk, S., & Farley-Ripple, E. N. (2022). Understanding Brokerage in Education: Backward Tracking from Practice to Research. *Center for Research Use in Education*.