INTERNATIONAL JOURNAL OF MODERN EDUCATION STUDIES

| Volume 7 - No 1 - June 2023 - ISSN 2618-6209 |

Research

Challenges and Opportunities of Meta-Analysis in Education Research

Nathaniel Hansford [1] **Rachel L Schechter** [2]

Article Type
Theoretical Article

International Journal of Modern Education Studies
2023
Volume 7, No 1
Pages: 219-231
http://www.ijonmes.net
dergipark.gov.tr/ijonmes

Article Info:

Received : 02.05.2023
Revision : 08.06.2023
Accepted : 15.06.2023

Abstract:

Meta-analyses are systematic summaries of research that use quantitative methods to find the mean effect size (standardized mean difference) for interventions. Critics of meta-analysis point out that such analyses can conflate the results of low- and high-quality studies, make improper comparisons and result in statistical noise. All these criticisms are valid for low-quality meta-analyses. However, high-quality meta-analyses correct all these problems. Critics of meta-analysis often suggest that selecting high-quality RCTs is a more valid methodology. However, education RCTs do not show consistent findings, even when all factors are controlled. Education is a social science, and variability is inevitable. Scholars who try to select the best RCTs will likely select RCTs that confirm their bias. High-quality meta-analyses offer a more transparent and rigorous model for determining best practices in education. While meta-analyses are not without limitations, they are the best tool for evaluating educational pedagogies and programs.

Keywords: Meta-Analysis, RCTs, Best Practice, Evidence-Based, Evaluating Pedagogy

Citation:
Hansford, H & Schechter, R., E. (2023). Challenges and Opportunities of Meta-Analysis in Education Research. *International Journal of Modern Education Studies, 7*(1), 218-231. https://doi.org/10.51383/ijonmes.2023.313

[1] Nathaniel Hansford, Canada, nathaniel.hansford@gmaill.com, Orcid ID: 0009-0002-2873-8149

[2] Dr. Rachel Schechter, LXD Research, USA Rachel@LXDresearch.com, Orcid ID: 0000-0001-9061-2892

INTRODUCTION

It is a common public/layman conceptualization for science that research results consistently show contradictory findings. This might partly stem from the media's poor reporting on new research. The media tends to report on each new landmark study, as if it stands in a vacuum, as the sole edict, to what science proves. This is problematic because it assumes the newest study is always the most correct, rather than looking to see what the majority of research shows. In the past, researchers would complete systematic literature reviews to discover the scientific consensus on a topic. With this approach, a researcher reads all the studies on a topic and then writes about their findings. This can be problematic because it tends to be purely qualitative, and the researcher gets to present their interpretation, without being beholden to any quantitative data.

A meta-analysis is similar to a literature review, except the authors also find the average statistical result for studies on a topic. Typically, meta-analysis results are displayed in effect sizes, an equation that seeks to create a standardized mean difference, so we can compare multiple studies. Looking at research through meta-analysis is the most systematic way of examining research. The author must review all studies, and then systematically synthesize quantitative results. Ideally, this removes as much bias as possible and provides an interpretation of the most normalized result on a topic. Meta-analysis also serves a fundamental scientific principle, replication. A scientific finding is only truly valid if it can be consistently replicated. Using meta-analysis, we can be sure whether a finding has been well replicated. Replication is especially important in education research because scientific results tend to be more variable, and experiments are often carried out by those selling pedagogical products.

Over the last two decades, meta-analyses have been crucial in helping to determine what best practice in literacy instruction is. Most famously, the National Reading Panel conducted multiple meta-analyses, including one that compared systematic phonics and whole language instruction. Their research showed systematic phonics has a mean effect size of .44 (NRP, 2001). This is why many reading researchers today recommend systematic phonics instruction as part of a comprehensive literacy program.

Some scholars object to meta-analysis, and they usually cite three main arguments:

1. Meta-analysis ignores study quality.
2. Meta-analysis makes apples-to-oranges comparisons.
3. Meta-analysis tends to show random statistical results, but not meaningful results.

1. Quality

There are typically 4 main types of studies included in a meta-analysis.

1. Case studies: studies without control groups or done retrospectively
2. Correlation studies: studies that look at the correlation between two datasets
3. Quasi-experimental studies: studies that have a non-randomized treatment and control group
4. Randomized Control Trial (RCT): studies with randomized treatment and control groups.

Typically, an RCT is seen as a higher quality study than a quasi-experimental study, and a quasi-experimental study is seen as higher quality than a case study. Sample size, duration, fidelity tracking, attrition, and measurement also affect the quality of a study. Typically, higher-quality studies show, on average, lower results. For example, a large sample size, long duration RCT with standardized measurements, is far more accurate than a small, short-duration case study that uses researcher-designed assessments.

Meta-analyses that do a poor job of controlling quality will typically include studies with varying levels of quality, such as case studies and RCTs, and report on one mean effect size. A well-done meta-analysis will either exclude low quality studies or show the difference in results for high versus low quality studies. For example, look at this result section from a fantastic meta-analysis by (Fritton, 2018).

Table 1

Fritton 2018 Sensitivity Analysis

Overall effect size estimation and sensitivity analyses

Constraints	Effect size	*SE*	*t*	*df*	*p* value	CI lower bound	CI upper bound	*n*	*k*	τ^2	Assumed ρ	Q_E	I^2
Full sample (*r* = .70)	0.28	0.05	5.69	53	<.001	0.18	0.38	226	54	0.09	.70	253.83	79.12
No within-group comparisons	0.42	0.16	2.62	17	.018	0.08	0.76	65	18	0.38	.70	136.66	87.56
Only BAU control conditions	0.27	0.11	2.54	12	.026	0.04	0.51	45	13	0.10	.70	31.82	62.29
No studies without RA	0.47	0.30	1.56	8	.158	−0.22	1.16	30	9	0.64	.70	115.93	93.10
Full sample (*r* = .80)	0.30	0.05	6.00	53	<.001	0.20	0.40	226	54	0.10	.70	343.14	84.55
Full sample (*r* = .90)	0.32	0.05	6.14	53	<.001	0.22	0.44	226	54	0.11	.70	666.17	92.04
Only standardized outcome measures (*r* = .70)	0.17	0.04	4.15	38	<.001	0.09	0.25	143	39	0.04	.70	100.23	62.23
No within-group comparison	0.35	0.13	2.61	13	.022	0.06	0.64	44	14	0.25	.70	75.00	82.67
Only BAU control conditions	0.37	0.11	3.18	10	.010	0.11	0.62	36	11	0.09	.70	24.28	58.81
No studies without RA	0.31	0.30	1.02	4	.366	−0.53	1.14	13	5	0.50	.70	50.35	92.06
r = .80	0.17	0.04	4.62	38	<.001	0.10	0.25	143	39	.04	.70	116.49	67.38
r = .90	0.19	0.04	5.31	38	<.001	0.12	0.26	143	39	.04	.70	174.99	78.28
Only experimenter-created outcome tools (*r* = .70)	0.34	0.09	3.81	26	.001	0.16	0.53	83	27	0.14	.70	206.61	87.42
No within-group comparison	0.34	0.45	0.76	5	.481	−0.82	1.50	21	6	0.83	.70	96.94	94.84
Only BAU control conditions	−0.04	0.16	−0.26	2	.821	−0.73	0.65	9	3	0.07	.70	3.95	49.37
No studies without RA	0.40	0.55	0.72	4	.509	−1.13	1.94	17	5	0.98	.70	96.00	95.83
r = .80	0.37	0.09	4.05	26	<.001	0.18	0.56	83	27	0.14	.70	285.36	90.89
r = .90	0.43	0.10	4.16	26	<.001	0.22	0.65	83	27	0.16	.70	574.41	95.47

Note. *SE* = standard error; *df* = degrees of freedom; CI = 95% confidence interval; *n* = effect sizes; *k* = studies; τ^2 = estimated variance in effect sizes across studies; ρ = assumed correlation between scores in within-group designs; Q_E = weighted sum of squares on a standardized scale; I^2 = index of the magnitude of heterogeneity between studies; RA = random assignment; BAU = business-as-usual.

In this study, the authors used the above sensitivity analysis to show the mean effect size changed across varying levels of quality. Interestingly, the highest quality studies showed a similar effect size (.31) to the overall mean for the study (.28), suggesting that quality did not significantly impact results, an unusual finding.

2. Apples to Oranges

"Apples to Oranges" is often used as a metaphor for comparisons that are too dissimilar to be meaningful. Within the context of meta-analysis, an example could be drawn from trying to find the mean effect of comprehension instruction and including multiple types of comprehension instruction together, as if they were the same thing. For example, vocabulary and strategy instruction are used to teach comprehension, but they are very different approaches. That said, good meta-analyses control for this by separating the results as moderator variables, as can be seen in Table 2 (Filderman, 2022). Moderator analysis can show what is the mean effect size for different types of studies, or outcomes. For example, a moderator analysis could differentiate between the effect sizes of RCTs, quasi-experimental studies, and case studies. In contrast, multilevel modeling and regression analysis can be used to estimate the impact of multiple moderator variables at once. As can be seen in Table 3.

Table 2

Filderman Sensitivity Analysis

Variable	No. of ES	No. of studies	*df*	*g*	95% CI	Between-study sampling variance (τ^2)	*p*
Participant characteristics							
Upper elementary	126	25	45.96	0.47	[0.31, 0.63]	0.17	<.001
Secondary	185	30	43.14	0.67	[0.46, 0.87]	0.43	<.001
Intervention characteristics							
Text type							
Narrative	40	7	11.83	0.31	[−0.01, 0.62]	0.23	.06
Expository	168	25	41.08	0.72	[0.51, 0.93]	0.30	<.001
Both	59	13	14.19	0.34	[0.16, 0.53]	0.08	.001
Interventionist							
Researcher	171	27	50.26	0.53	[0.35, 0.70]	0.30	<.001
Teacher	120	25	37.88	0.69	[0.44, 0.93]	0.34	<.001
Instructional approach							
Strategy instruction							
Main idea	215	37	53.2	0.72	[0.54, 0.89]	0.30	<.001
Inferencing	73	15	22.54	0.56	[0.32, 0.81]	0.18	<.001
Text structure	99	17	27.44	0.47	[0.24, 0.70]	0.25	<.001
Retell	54	11	18.95	0.59	[0.29, 0.90]	0.31	<.001
Predicting	86	19	28.1	0.60	[0.39, 0.81]	0.25	<.001
Strategy only	173	33	50.48	0.69	[0.47, 0.90]	0.39	<.001
Multiple strategies	164	30	45.48	0.59	[0.41, 0.76]	0.24	<.001
Background knowledge (BK)							
Vocabulary	126	21	31.15	0.39	[0.26, 0.51]	0.10	<.001
Content knowledge	92	14	23.62	0.64	[0.35, 0.93]	0.37	<.001

Table 3

Regression Analysis Example

Cognitive Strategy/Skill Regression Analysis:

Number of Effects	Mean Effect Size [95% CI]	Fixed Effect	Standardized Assessment	Randomized
7	.50 [-.47, .49]	x	x	x
17	.20 [.05, .36]		x	x
4	.53 [-.39, 1.46]	x		x
		x	x	
17	1.72 [1.0, 2.40]			x
3	.04 [-1.15, 1.24]	x		
2	.33 [-2.14, 2.81]		x	
3	.37 [-.26, 1.01]			

2. Statistical Noise

One less common criticism of meta-analysis is that the authors capture random effects and averages, not meaningful trends. Let's make a hypothetical example. Say we have 10 studies, and they show the following effects: .10, .20, .30, .40, .50, .60, .70, .80, .90, 1.0, you will find a mean effect size of .50, which is quite significant. However, there is no average discernible trend within those studies. So by taking the mean, we have actually made the data less meaningful, as opposed to more meaningful. Of course, there are multiple tools for addressing this issue. Most typically, meta-analyses will use confidence intervals, which show the likely range of results between effect sizes, and or p values, which display the likelihood that a statistic is random, alongside their mean effect sizes so that readers can discern if the mean effect found was meaningful or random noise. Indeed, if you look back at the two graphics from well-done meta-analyses, they both included confidence intervals and p values alongside their effect sizes.

So, Are these Criticisms Valid?

All three of these criticisms are valid. However, they also only really apply to a poorly done meta-analysis. Meta-analysis is a relatively new technique for reviewing research, and it has evolved over the last 20 years. If you read meta-analyses done in the late 90's, they often combine multiple poor-quality studies to produce one mean effect size. While more modern meta-analyses tend to be much more sophisticated, there is a lack of consistency within the field of education for meta-analysis methodology. For example, we reviewed meta-analyses on the topic of ESL education. We found 12 meta-analyses on ESL education

research, dating back to 2009 (all of which can be found in the references section). Of these 12 meta-analyses, 6 included studies without control groups and did not use moderator analysis to compare the impact of studies with and without control groups. The 6 meta-analyses that did not control quality were not rigorous and, therefore, cannot be used as a definitive proof for the scientific consensus.

THE ALTERNATIVE

Those who criticize meta-analysis often claim we should rely on high-quality RCTs instead. This is a problematic solution for two reasons. Firstly, researchers independently decide which RCTs are the most rigorous using complex processes. For example, many scholars have cited Balanced Literacy as the gold standard of reading instruction, based on a handful of RCTs reviewed by WWC (Hechinger, 2022). This suggestion was made in comparison to the findings of the NRP meta-analysis, which recommended systematic phonics instruction, based on dozens of studies.

Secondly, this methodology is based on the belief that well-done RCTs show precise outcomes and therefore do not need replication. But within the field of education, this is undoubtedly false. Let's look at some of the findings from the (Hansford, 2022) meta-analysis on language programs. There were 20 identified RCTs that looked at structured literacy phonics programs. The mean effect size was .48, and the 95% confidence intervals were [.31, .66]. We can expect results of .31 to .66 in 95% of structured literacy RCT studies. This is a pretty wide range. .66 is a moderate to high effect size, and .31 is low. The lowest study showed an effect size of -.11 (Vaden-Kiernan, 2008). And the highest effect size was 1.16 (Farokhbakht, unlisted date). Neither effect size is particularly representative of the normal effect of a phonics intervention. However, a scholar with an agenda could point to either study to make a case for or against structured literacy.

The Vaden-Kiernan study is of far higher quality than the Farokhbakht study. If we examine the highest quality RCT studies, in this case, longitudinal RCTs with standardized assessments. We get 3 studies: (Vaden-Kiernan, 2008), (Torgesen, 2007), and (Bratsch, 2020). These studies showed a mean effect size of .22, with 95% confidence intervals of [-.50, .95], suggesting a high degree of variability. The lowest study showed a mean effect size of -.11, and the highest study showed a mean effect size of .43 (Bratsch, 2020). Again, a biased academic could pick any of those three studies and argue for or against phonics/structured literacy.

All of these studies could also be apple-to-oranges comparisons, as each study looked at different demographics, programs, and styles of approaches. One study looked at a scripted DI approach (Vaden-Kiernan, 2008). One study looked at an Orton Gillingham approach (Torgesen, 2007). And one study looked at a speech-to-print approach (Bratsch, 2020).

However, we also see very different results even if we only look at RCT studies on the same program. For example, let's look at Read 180. In 2022 Hansford and Mcglynn identified 12 RCTs on Read 180, with a mean effect size of .11 and 95% confidence intervals of [.04, .19]. Here the confidence intervals suggest a very narrow range. However, the highest effect size study (Interactive Inc, 2002) showed a mean effect size of .41, and the lowest effect size study (Fitzgerald, 2008) showed a mean effect size of 0 (for longitudinal outcomes). If we remove all the lowest quality studies and only include those that used standardized measurements, were longitudinal, and controlled for fidelity, we get 4 studies, (Interactive Inc 2002), (Fitzgerald, 2008), (Meisch, 2011), and (Sprague, 2012). Together the studies show a mean effect size of .16, but the confidence intervals are much wider than when all 12 RCTs are included, [-.12, .40].

Moreover, both the (Fitzgerald, 2008) study and the (Interactive Inc, 2002) study were within the highest quality category. Hence, the range of effect sizes was still 0-.41. If we look at both quasi-experimental and RCT studies, 13 out of 19 mean effect sizes were between 0 and .29. With 95% confidence intervals of [0, .19]. While looking at all the studies together suggested a very consistent trend of a low effect, looking at only the highest quality studies made the found effect appear more random, and difficult to find a meaningful trend.

That said, the Read 180 studies, covered multiple grades and used different designs. Reading Recovery might be a better example. Within the (Hansford, 2022) meta-analysis of Language programs, we identified 11 RCT studies on Reading Recovery, all of which looked at the identical grade. Moreover, all but two used the same basic design, comparing 1-on-1 intensive reading instruction for 20 weeks, to a no-treatment control group. These 11 studies showed a mean effect size of .38, with 95% confidence intervals of [-.99, 1.24] (outliers included). All these studies are RCTs on the same grade and same program. All but 2 of these studies compared no treatment to treatment. And yet, a large range of effect sizes were found. The largest impact was found in (Iverson, 1999), with a mean effect size of 2.59, and the lowest was (Schmitt, 2004), with a mean effect size of -.50. Again, any scholar with an agenda could pick either RCTs and make the opposite arguments. Even if we take the two highest quality studies, in this case (Holliman, 2013) and the (Center for Research in Education and Social Policy, 2022), we still get opposite results. Both studies were large-scale longitudinal RCTs. (Holliman, 2013) showed a mean effect size of .48, and the (CRESP, 2022) study showed a mean effect size of -.19.

Inconsistent findings among RCTs create a sentiment that education science produces inconsistent findings and cannot be trusted. Again, any scholar with an agenda could pick either of these RCTs and make completely opposite arguments. Scholars on either side of the reading wars debate will likely want to point to the flaws in either study as a defense for their perspective. Indeed, pro-Reading Recovery scholars frequently point to the (Holliman, 2013) study as evidence that Reading Recovery works, and pro-structured literacy advocates frequently point to the (CRESP, 2022) study, including Emily Hanford. Both the Holliman and CRESP study have weaknesses. The Holliman study had poor fidelity controls in the

control group, and the CRESP study had high attrition rates. Both studies compared intensive 1-1 reading instruction to no additional instruction, which is not an ideal study design. That said, the studies are both of higher-than-average quality compared to other studies in the Hansford 2022 meta-analysis.

Of course, instructional programs include multiple variables at once and are often compared to business-as-usual control groups. For this reason, it might be easier to isolate the fixed effect of a pedagogy than a program. (Bakken, 1997) and (Boyle, 1993) conducted RCTs, on the effects of cognitive strategies on reading comprehension for intermediate students with learning disabilities. Both studies used active control groups, in which instruction was the same as in the treatment group, minus the instruction on cognitive strategies. Both studies used standardized assessments. Both studies had a sample of between 30-40 students. Both studies were short and lasted less than a month. However, in the Bakken study we found an effect size of 2.71 for the use of cognitive strategies and in the Boyle study we found an effect size of .15. Both studies were of extremely high quality and similar, and yet, they yielded completely different results. The above-discussed anomalies suggest that even the highest quality RCTs do not lead to precise or consistent results and that even a very high-quality RCT cannot be considered reliable evidence of efficacy in isolation.

Do Meta-Analyses Provide More Homogenous Effects?

While the above research suggests that RCTs do not provide homogenous results, this does not necessarily mean that meta-analyses do. Indeed, many meta-analyses at face value appear to show very different results, for similar research questions. As of 2023, we can identify at least 14 peer-reviewed and experimental meta-analyses, conducted on English phonics instruction. (Camilli, 2003) identified the lowest effect size of .24. Conversely, (Weiser, 2011) found the highest effect size of .78. These differences are seemingly very different; however, these meta-analyses examined very different questions. (Camilli, 2003), attempted to identify the fixed effect of systematic phonics versus unsystematic phonics, and (Weiser, 2011) was attempting to identify the random effect of encoding instruction. These research questions are fundamentally different. Comparatively, (Steubings, 2008), which had the same research question as Camilli, found a mean effect size of .31, which is statistically comparable. Similarly, (Hansford, 2022), (Piasta, 2011), (Ehri, 2001), and the (NRP, 2001) all looked at the random effect for general phonics instruction and found a mean effect size of between .40 and .45 for phonics, suggesting a very homogenous effect. While, meta-analyses often produce heterogeneous effects, these differences usually have to do with the research question and methodology used.

Meta-analyses also have unique advantages for detecting outlier data. It is impossible to tell if the results from a single RCT represent outlier data when taken in isolation. However, tools like funnel plots, trim and fill, and IQR analysis, can be used within a meta-analysis to identify if a single study is an outlier (Terrin, 2003). Funnel plots

can be especially useful in visualizing whether there is outlier data related to sample size. Small sample size studies often have larger effect sizes, partly because it is harder to effectively implement a new pedagogy with a larger group of teachers. Smaller sample size studies can also produce more random results, as individual outliers can have a greater impact (IntHout, 2015). Lastly, smaller sample size studies can be more easily replicated and used to "fish" for better results (Lee, 2012). Funnel plots are commonly used to compare the results of studies with the sample sizes of studies, to test whether smaller sample size studies increase heterogeneous effects. To help illustrate this point, we created a scatter plot of RCT studies on Read 180, based on the (Hansford, 2022) meta-analysis of Read 180. The results can be seen in Figure 1.

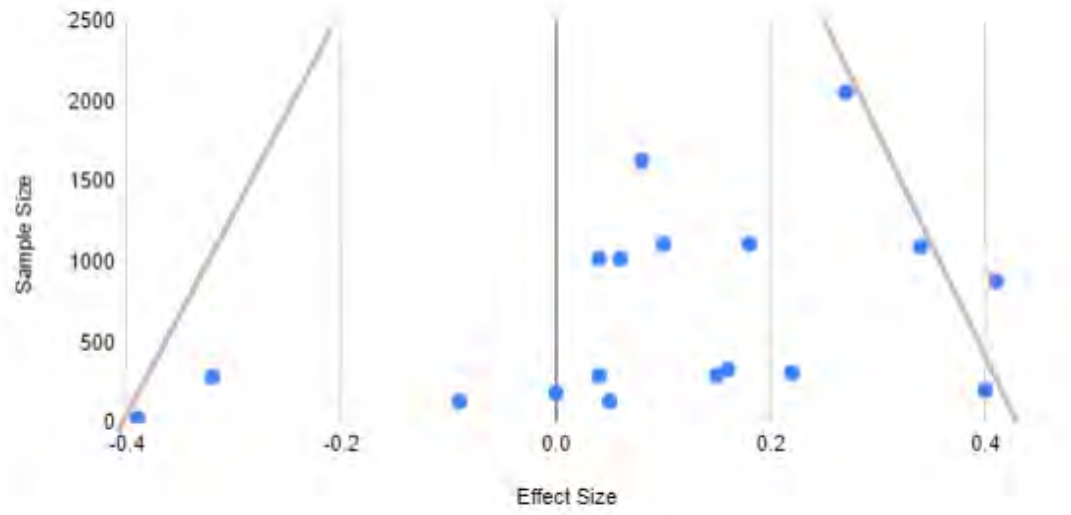

Figure 1: *Read 180 Funnel Plot*

Figure 1 shows that the negative effect sizes were associated with low sample size studies, suggesting that these low sample sizes led to more random and, thus, more heterogeneous results. Similarly, the two highest effect sizes found were both associated with studies that also had study samples below the median sample size. According to Cohen's guide, most studies with a sample size above 1000 fell between the effect size range of 0-.20, suggesting a negligible outcome. This meta-analysis tool allows readers and researchers to better understand what an expected outcome might be for a pedagogy or program than any single RCT could provide.

IMPLICATIONS FOR PRACTICE

Whether trying to measure the efficacy of a principle, or a program, it is incredibly difficult to find a consistent effect found across multiple RCTs. This difficulty stems from the fact that education is not a hard science, it is a social science. There is a large degree of variability in research results. Teacher quality, student motivation, demographics, study

design, and study quality will all impact the effect size. Controlling all these variables consistently is nearly impossible. You, therefore, cannot expect results to be static across multiple studies. It is not rational to expect individual RCT studies to produce results that do not vary.

Even if high-quality RCTs did show consistent results, isolating the highest-quality RCTs is very difficult and requires people to make unbiased judgments. People are likely to be more critical of the studies that do not confirm their biases and less critical of the ones that do. We can only truly avoid cherry-picking results to support our biases by reviewing all of the relevant studies on a topic. This does not mean viewing all studies uncritically, instead a good meta-analysis uses objective criteria to identify how effect sizes varied according to study quality. Moreover, when factors limit the validity of a meta-analysis, such as a lack of studies with control groups, the authors should identify it as a limitation. Using methodologies like moderator variable analysis, regression analysis, and multilevel modeling, with meta-analysis is a far more transparent process than simply trying to select the most valid study.

REFERENCES

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2011). Pedagogical strategies for teaching literacy to ESL immigrant students: a meta-analysis. *The British journal of educational psychology*, *81*(Pt 4), 629–653. https://doi.org/10.1111/j.2044-8279.2010.02015.x

Bratsch-Hines, M., Vernon-Feagans, L., Pedonti, S., & Varghese, C. (2020). Differential effects of the Targeted Reading Intervention for students with low phonological awareness and/or vocabulary. *Learning Disability Quarterly*, *43*(4), 214-226. https://doi.org/10.1177/0731948719858683

Camilli, G., Vargas, S., & Yurecko, M. (2003). Teaching Children to Read: The Fragile Link Between Science & Federal Education Policy. *Education Policy Analysis Archives*, *11*, 15. https://doi.org/10.14507/epaa.v11n15.2003

Center for Research in Education and Social Policy. (2022). Reading Recovery: Long-term effects and cost-effectiveness. Under the investing in innovation (i3) scale-up. *University of Delaware.* https://www.cresp.udel.edu/research-project/efficacy-follow-study-long-term-effects-reading-recovery-i3-scale/

Clougherty, L. (2019). Emergent bilinguals and academic language acquisition through the use of sentence frames. *Cal Poly Humboldt theses and projects*, 247. https://digitalcommons.humboldt.edu/cgi/viewcontent.cgi?article=1304&context=etd

Fitton, L., McIlraith, A. L., & Wood, C. L. (2018). Shared book reading interventions with English learners: A meta-analysis. *Review of Educational Research*, 88(5), 712-751. https://doi.org/10.3102/0034654318790909

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, 71(3), 393-447. https://doi.org/10.3102/00346543071003393

Farokhbakht, L. (n.d.). The effect of using multisensory-based phonics in teaching literacy on EFL young female/male learners' early reading motivation. *University of Isfahan.* https://jolly2.s3.amazonaws.com/Research/The%20Effect%20of%20Using%20Multisensorybased%20Phonics%20in%20Teaching%20Literacy%20on%20EFL%20Young%20FemaleMale%20Learners'%20Early%20Reading%20Motivation.pdf

Filderman, M. J., Austin, C. R., Boucher, A. N., O'Donnell, K., & Swanson, E. A. (2022). A meta-analysis of the effects of reading comprehension interventions on the reading comprehension outcomes of struggling readers in third through 12th grades. *Exceptional Children*, 88(2), 163-184. https://doi.org/10.1177/00144029211050860

Fitzgerald, R., & Hartry, A. (2008). What works in afterschool programs: The impact of a reading intervention on student achievement in the Brockton Public Schools (Phase II). *Berkeley, CA: MPR Associates, Inc. and the National Partnership for Quality Afterschool Learning at SEDL*. https://ies.ed.gov/ncee/wwc/Study/82599

Hansford, N & King, J. (2022). A meta-analysis and literature review of language programs. *Teaching by Science*. https://www.teachingbyscience.com/a-meta-analysis-of-language-programs

Hansford, N & McGlynn, S. (2022). Read 180. *Teaching by Science*. https://www.pedagogynongrata.com/read-180

Hechinger Report. (2022). Opinion: A call for rejecting the newest reading wars. https://hechingerreport.org/opinion-a-call-for-rejecting-the-newest-reading-wars/?hsCtaTracking=87e509b0-21a9-40e1-aa17-2d072a8d37f6%7C4aa6fa0e-fb2a-45c3-866c-05165077f3fd

Holliman, A., & Hurry, J. (2013). The effects of Reading Recovery on children's literacy progress and special educational needs status: A three-year follow-up study. *Educational Psychology*, 33, 10.1080/01443410.2013.785048 https://www.tandfonline.com/doi/abs/10.1080/01443410.2013.785048

Interactive, Inc. (2002). An efficacy study of READ 180: A print and electronic adaptive intervention program, grades 4 and above. *Scholastic*. http://teacher.scholastic.com/products/research/pdfs/ER_Council_Great_Schools.pdf

IntHout, J., Ioannidis, J., Borm, G., & Goeman, J. (2015). Small studies are more heterogeneous than large ones: A meta-meta-analysis. *Journal of Clinical Epidemiology*, 68(8), 860-869. https://doi.org/10.1016/j.jclinepi.2015.03.017

Jia. (2021). Toward a set of design principles for decoding training: A systematic review of studies of English as a foreign/second language listening education. *Educational Research Review*, 33, N.PAG. https://www.sciencedirect.com/science/article/abs/pii/S1747938X21000154

Lang, L., Torgesen, J., Vogel, W., Chanter, C., Lefsky, E., & Petscher, Y. (2009). Exploring the relative effectiveness of reading interventions for high school students. *Journal of Research on Educational Effectiveness*, 2, 149-175. https://doi.org/10.1080/19345740802641535

Lee, W., & Hotopf, W. (2012). Funnel plot. *Science Direct*. https://www.sciencedirect.com/topics/medicine-and-dentistry/funnel-plot#:~:text=The%20funnel%20plot%20is%20a,this%20difference%20which%20is%20detectable

Li, R. (2022). Effects of blended language learning on EFL learners' language performance: An activity theory approach. *Journal of Computer Assisted Learning*, 38(5), 1273-1285. https://doi.org/10.1111/jcal.12697

Lv, X., Ren, W., & Xie, Y. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. *Asia-Pacific Education Researcher*, 30(6), 643-653. https://doi.org/10.1007/s40299-021-00594-6

Meisch, A., Hamilton, J., Chen, E., Quintanilla, P., Fong, P., Gray-Adams, K., & Thornton, N. (2011). Striving Readers study: Targeted and whole-school interventions-year 5. *Rockville, MD: Westat.* https://eric.ed.gov/?id=ED601086

NRP. (2001). Teaching children to read: An evidence-based assessment of the scientific literature on reading instruction. *United States Government*. Retrieved from https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

Piasta, S. B., & Wagner, R. K. (2010). Developing early literacy skills: A meta-analysis of alphabet learning and instruction. *Reading Research Quarterly*, 45(1), 8-38. https://doi.org/10.1598/RRQ.45.1.2

Roessingh, H. (2004). Effective high school ESL programs: A synthesis and meta-analysis. *Canadian Modern Language Review*, 60(5), 611-636. https://doi.org/10.3138/cmlr.60.5.611

Rui Li. (2022). Effects of mobile-assisted language learning on EFL/ESL reading comprehension. *Journal of Educational Technology & Society*, 25(3), 15-29. https://www.jstor.org/stable/48673721

Schenck, A. (2020). Using meta-analysis of technique and timing to optimize corrective feedback for specific grammatical features. *Asian-Pacific Journal of Second and Foreign Language Education*, 5, 1-20. https://doi.org/10.1186/s40862-020-00097-9

Sprague, K., Zaller, C., Kite, A., & Hussar, K. (2012). Springfield-Chicopee School Districts Striving Readers program final report Years 1-5: Evaluation of implementation and impact. *Providence, RI: The Education Alliance at Brown University.*

Stuebing, K. K., Barth, A. E., Cirino, P. T., Francis, D. J., & Fletcher, J. M. (2008). A response to recent reanalyses of the National Reading Panel report: Effects of systematic phonics instruction are practically significant. *Journal of Educational Psychology*, 100(1), 123-134. https://doi.org/10.1037/0022-0663.100.1.123

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22, 2113-2126. https://doi.org/10.1002/sim.1461

Thompson, C. (2020). Video-game-based instruction for vocabulary acquisition with English language learners: A Bayesian meta-analysis. *Educational Research Review*, 30, N.PAG. https://www.researchgate.net/publication/340640024_Video-game_based_instruction_for_vocabulary_acquisition_with_English_language_learners_A_Bayesian_meta-analysis

Unkyoung Maeng. (2014). The effectiveness of reading strategy instruction: A meta-analysis. *English Teaching*, 69(3), 105-127. https://doi.org/10.15858/engtea.69.3.201409.105

Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., Sullivan, K., Ruiz de Castilla, V., Fleming, G., Rodriguez, D., Henry, C., Long, T., & Hughes Jones, D. (2018). Findings from a multiyear scale-up effectiveness trial of Open Court Reading. *Journal of Research on Educational Effectiveness*, 11(1), 109-132. https://doi.org/10.1080/19345747.2017.1342886

Biographical notes:

Nathaniel Hansford: Nathaniel Hansford is a teacher of 11 years. He holds a specialist qualification in both reading and special education. He is the author of two education books and is the lead writer for Pedagogy Non Grata. Contact Nathaniel: Nathaniel.hansford@gmail.com

Dr. Rachel Schechter: Is the founder of (LXD) Research and has led education research teams since 2007; she was previously the Vice President of Learning Sciences at HMH. Dr. Schechter holds a doctorate in Child Development from Tufts University and a Master from Harvard University.

Author(s)' statements on ethics and conflict of interest

Ethics statement: We hereby declare that research/publication ethics and citing principles have been considered in all the stages of the study. We take full responsibility for the content of the paper in case of dispute.

Statement of interest: We have no conflict of interest to declare.

Funding: None

Acknowledgements: None