Confirming the NRP Findings in 2022

(A condensed report)

Written by Nathaniel Hansford and Joshua King
Reviewed by Rachel Schecter Ph.D., LXD Research

Abstract:

In 2022, a meta-analysis was conducted of 61 language program studies, focusing on phonics and balanced literacy studies, making it the largest meta-analysis done on the topic in the last 10 years. This meta-analysis was conducted 20 years after the National Reading Panel (NRP) meta-analysis and included studies from up to 2022,almost double the number of studies originally included in the NRP. Results suggest an identical .45 mean effect size for phonics.

This report also includes a secondary meta-analysis of 13 phonics meta-analyses conducted over the last 25 years, which found a mean effect size of phonics, of .43. All of this helps to show that despite the fact that the NRP meta-analysis was conducted 20 years ago, its findings are still valid today. Moreover, not only did I find the same mean effect size, but I found the same general trends about phonics education held true as well. Moreover, while the NRP meta-analysis did not look at Balanced Literacy Outcomes, both the 2022 meta-analysis and the secondary meta-analysis did.

This research showed that phonics focused programs consistently outperformed Balanced Literacy programs across all demographics. Indeed, phonics on average doubled the effects found for Balanced Literacy for grades 1-2, for class based instruction, and for at risk learners.

Phonics interventions showed efficacious results, both for early primary instruction, and for older students with reading deficits. This suggests that students should receive phonics instruction during their foundational education years and that if they miss this instruction that they benefit from getting it later on.

Methodology

Finding the Studies

Searches of studies were conducted on Education Source, Sage Pub, Google, and company websites of the most popular language programs. Programs were selected based on polling by teachers interested in the science of reading. All studies found with a control group, sufficient reporting data (meaning either effect sizes or sufficient data for me to calculate a proper effect size), and total samples above 10 were included in this analysis. No time restrictions were placed on this inclusion criteria. In total 61 studies were included.
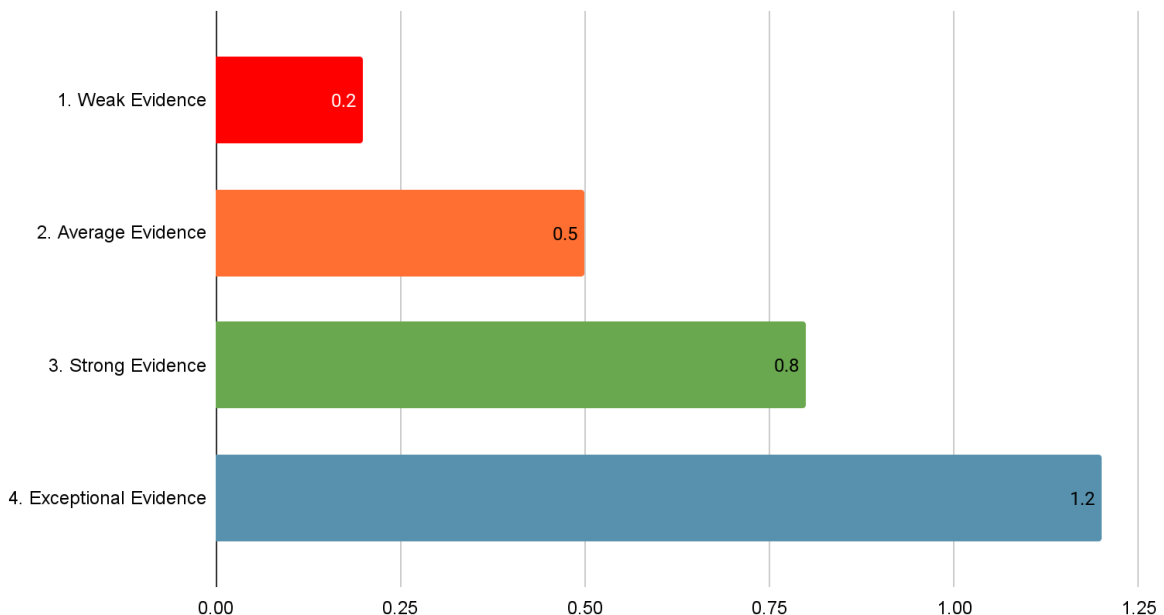
Excluding Studies

Three studies were initially included but later excluded (after review) due to non-equivalent pretest scores. In all cases, effect sizes based on the post-test results were extremely misleading. One study was by Ring, Et, al on Take Flight. The second study was by Stuart, Et, al, on Jolly Phonics. In the cases of the studies on Take Flight and Jolly Phonics, the Cohen's D formula provided an effect size that was exaggerated in comparison to the actual gains made. For the Take Flight study, the gains were statistically insignificant. For the Jolly Phonics study the claimss were positive and described a high product impact, but there were in fact uneven or mixed results across measures or nonequivalent groups at pretest that should have impacted the overall findings." The third study excluded was on Remediation Plus, by Corcoran, Et al. The gains found by the treatment group were very significant with an effect of over .50, however, the nonequivalent groups, resulted in much smaller Cohen's d effect sizes.

Effect Size Calculations

Meta-analyses typically report on results with effect sizes. These effect sizes are meant to be interpreted in the following way:

Cohen's Guide to Interpreting Effect Sizes

Category	Value
1. Weak Evidence	0.2
2. Average Evidence	0.5
3. Strong Evidence	0.8
4. Exceptional Evidence	1.2

Effect sizes for studies with sample sizes above 50 per group were calculated with a Cohen's d calculation. Effect sizes for studies with sample sizes below 50 were
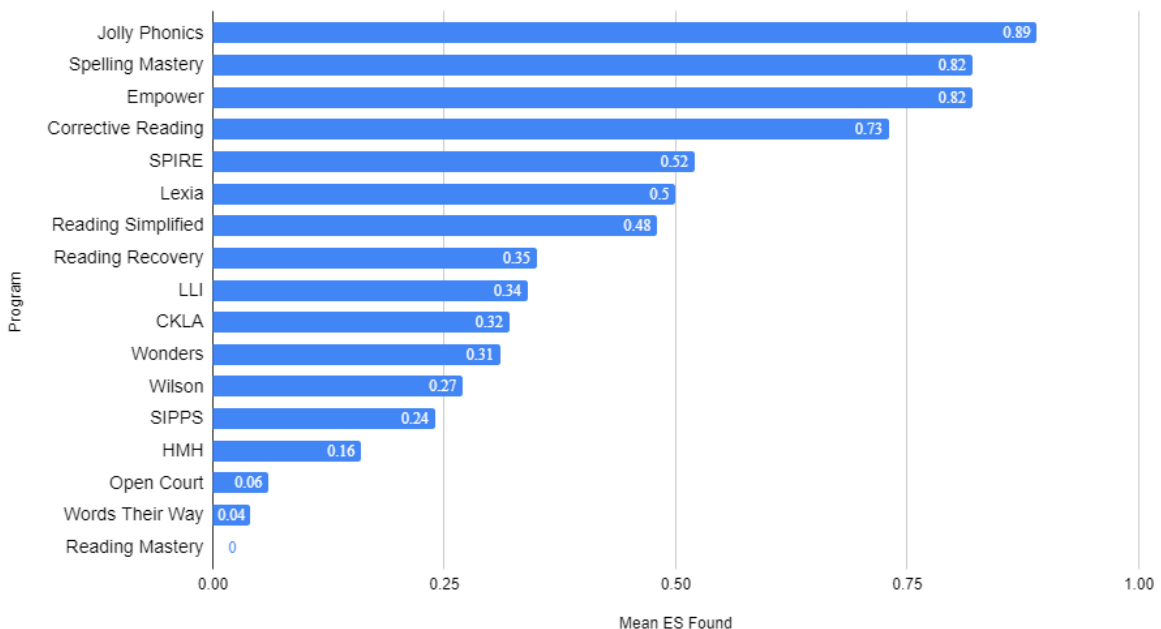
calculated with Hedge's g. When effect sizes were already calculated by the peer-reviewed author, these effect sizes were usually accepted as is, unless there was probable cause for an improperly calculated effect size, in which case, the effect size was double-checked. If the effect sizes did not match, I used my own calculation.

All effect size calculations were independently duplicated by a second analyst, to support the reliability of the results, in the case of a disagreement by the second analyst, both persons re-did their calculations, and discussed to reach a consensus. Effect sizes were not rounded, they were cut off at the hundredth place value. When the effect size calculation found the control group to outperform the treatment group, the effect size was entered into the meta-analysis as a negative effect size.

Grouping of Studies

Effect sizes tend to be context specific. For example, studies on reading tend to show lower results, whereas studies on math tend to show higher results. For this reason, effect sizes for single outcomes, programs, or pedagogical ideas are less helpful for teachers. It is more useful to compare the effect sizes of similar factors, so that teachers can see which factors have a stronger research base. For this purpose, research has been compiled by language programs into the following graph:

Phonics Studies Ranked by Effect Size

Program	Mean ES Found
Jolly Phonics	0.89
Spelling Mastery	0.82
Empower	0.82
Corrective Reading	0.73
SPIRE	0.52
Lexia	0.5
Reading Simplified	0.48
Reading Recovery	0.35
LLI	0.34
CKLA	0.32
Wonders	0.31
Wilson	0.27
SIPPS	0.24
HMH	0.16
Open Court	0.06
Words Their Way	0.04
Reading Mastery	0

These results found that programs considered Balanced Literacy (Levelled Literacy Instruction, Reading Recovery, Wonders, and HMH) are not found in the top 7 programs. Moreover, all Balanced Literacy programs showed a mean effect size that
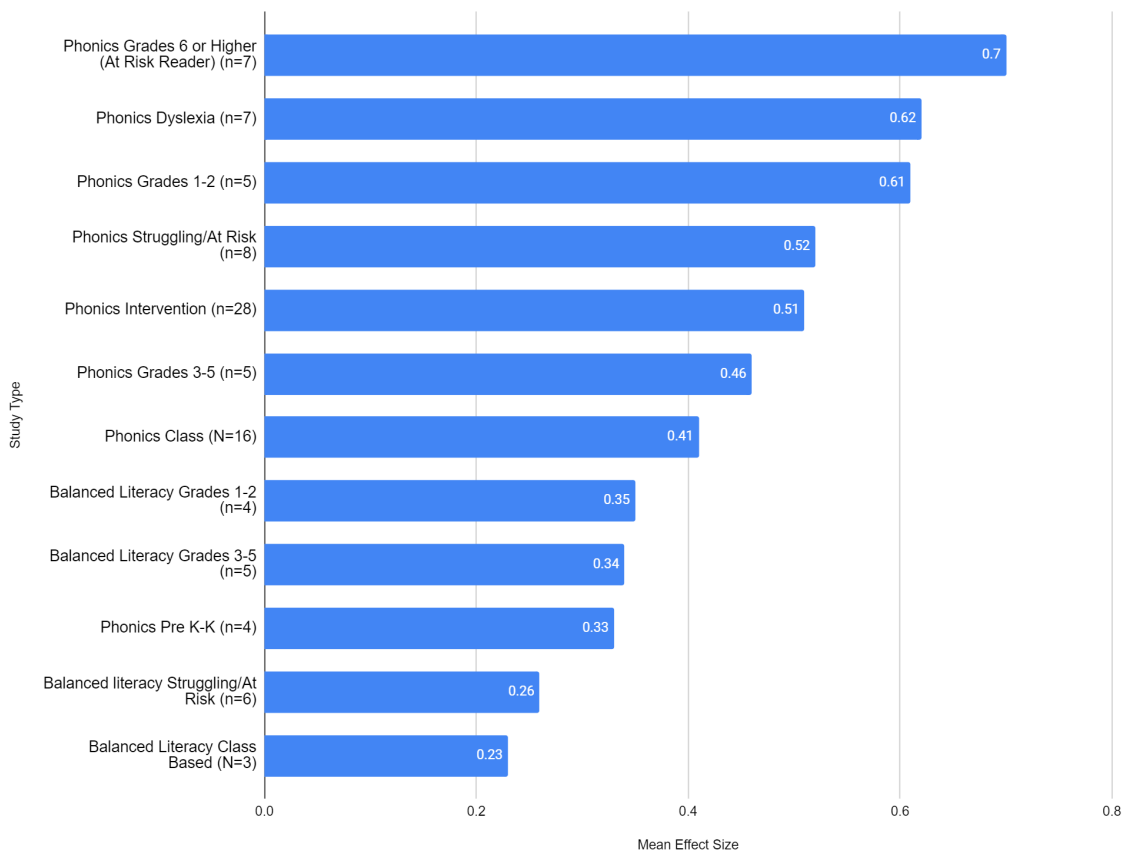
was statistically low. In order to find larger trends in this data, outcomes were also broken down according to moderator variables, sample type, and program type. Of course, different programs had studies that fit into multiple criterion according to this graph.

Phonics heavy programs outperformed Balanced Literacy programs across all grades, and sample types. Indeed phonics programs showed roughly double the impact for grades 1-2, at risk learners, and class based instruction. This research does not show support for the use of Balanced Literacy programs, over phonics focused programs, in any context.

Language Program Moderator Effects Ranked

www.teachingbyscience.com

Study Type	Mean Effect Size
Phonics Grades 6 or Higher (At Risk Reader) (n=7)	0.7
Phonics Dyslexia (n=7)	0.62
Phonics Grades 1-2 (n=5)	0.61
Phonics Struggling/At Risk (n=8)	0.52
Phonics Intervention (n=28)	0.51
Phonics Grades 3-5 (n=5)	0.46
Phonics Class (N=16)	0.41
Balanced Literacy Grades 1-2 (n=4)	0.35
Balanced Literacy Grades 3-5 (n=5)	0.34
Phonics Pre K-K (n=4)	0.33
Balanced literacy Struggling/At Risk (n=6)	0.26
Balanced Literacy Class Based (N=3)	0.23

Phonics heavy programs outperformed Balanced Literacy programs across all grades, and sample types. Indeed phonics programs showed roughly double the impact for grades 1-2, at risk learners, and class based instruction. This research does not show support for the use of Balanced Literacy programs, over phonics focused programs, in any context.
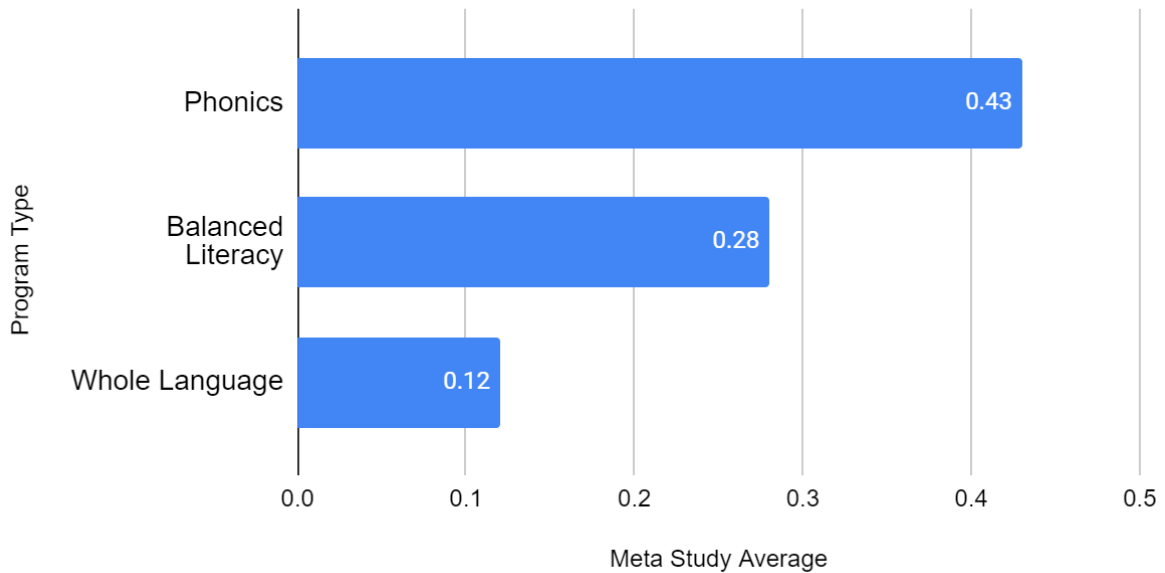
All other meta-analyses on the topic of phonics, whole language, and balanced literacy were examined, in order to cross reference our results. This included 14

other phonics meta-analyses, 2 other balanced literacy meta-analyses, and 4 whole language meta-analyses.An average of each program type was taken, across these 21 meta-analyses (mine included), the results of which can be seen below.

A Secondary Meta-Analysis of Language Program Types

www.teachingbyscience.com

Program Type	Meta Study Average
Phonics	0.43
Balanced Literacy	0.28
Whole Language	0.12

Meta Study Average

For a full copy of the original article/meta-analyses and references, go to:
https://www.teachingbyscience.com/a-meta-analysis-of-language-programs