"Books in the Home" Effect Size Calculations

Vinay Tummarakota

November 2024

1 Context

This document summarizes the effect size calculations used in the article "Do books in the home really improve academic achievement?". Each effect size is initially calculated as Cohen's d (adjusting for pre-test scores) and then converted to Hedge's g in R.

2 Melosh (2003)

Unfortunately, I was only able to obtain a preview of this doctoral thesis, and thus, do not have access to the results.

3 Kim (2006)

The effect size on the Iowa Test of Basic Skills (ITBS) is reported as a semi-standardized regression coefficient ($B = 0.08 \pm 0.04$) in Table 4.

TABLE 4
Ordinary Least Squares Models Predicting Treatment Effect on ITBS (Total Reading Scores)

	All	White	Black	Latino	Asian
Variables	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Treatment	0.08~	0.11	0.22*	0.14~	–0.17
	(0.04)	(0.09)	(0.09)	(0.08)	(0.11)
Spring ITBS	0.87**	0.84**	0.83**	0.77**	0.88**
	(0.02)	(0.04)	(0.05)	(0.05)	(0.07)
(Constant)	–0.07	–0.03	–0.17~	–0.12	0.07
	(0.05)	(0.10)	(0.09)	(0.09)	(–0.13)
R^2	0.76	0.71	0.76	0.69	0.71
N	486	160	93	125	85

Note. All models include fixed effects for the randomization block. Standard errors in parentheses.
The model for "other ethnic students" (21 multiethnic, 2 Native American) revealed nonsignificant treatment effects.
~$p < .10$, *$p < .05$, **$p < .01$.

To estimate Cohen's d, we must also know the sample sizes in the treatment and control groups. This information is described on page 344:

As a result of attrition during the summer, the final analytic sample included 486 students, including 252 students in the treatment group and 234 students in the control group.

By 1.31 of [1], Cohen's d and its corresponding standard error can be estimated like so:

$$\begin{aligned} d &= \frac{B}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\ &= \frac{-0.08}{\sqrt{\frac{1(252+234-1)-\frac{0.08^2(252*234)}{252+234}}{252+234-2}}} \\ &\approx \frac{-0.08}{1} \\ &\approx 0.08 \end{aligned}$$

$$\begin{aligned} SE(d) &= \frac{SE(B)}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\ &\approx \frac{0.04}{1} \\ &= 0.04 \end{aligned}$$

The effect size on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), however, is reported as an unstandardized regression coefficient ($B = -2.09 \pm 1.50$) in Table 5.

TABLE 5
Ordinary Least Squares Models Predicting Treatment Effect on Oral Reading Fluency (WCPM)

	All	White	Black	Latino	Asian
Variables	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Treatment	–2.09	–2.83	–1.79	–2.21	–0.41
	(1.50)	(2.73)	(3.31)	(2.81)	(3.95)
Spring WCPM	0.83***	0.86***	0.83***	0.77***	0.77***
	(0.02)	(0.04)	(0.05)	(0.05)	(0.05)
(Constant)	11.08***	7.24	12.21~	17.00**	20.14*
	(–3.07)	(–5.51)	(–6.72)	(–6.37)	(8.54)
R^2	0.80	0.80	0.81	0.73	0.75
N	450	150	85	116	80

Note. All models include fixed effects for the randomization block. Standard errors in parentheses.
Sample sizes for OLS models predicting fluency are not equal to the ITBS analysis because of missing data on the fall fluency assessment.
~$p < .10$, *$p < .05$, **$p < .01$.

To estimate Cohen's d, we may once again use 1.31 of [1]. However, because the

DIBELS outcome was not already standardized, we have to specify an appropriate value of s_y. Unfortunately, the study does not report the standard deviation of the post-test DIBELS score, so the best approximation of s_y available is the standard deviation of the pre-test DIBELS score listed in Table 1.

TABLE 1
Characteristics of Students at the Beginning of the Study (N = 552)

Variable	%	Min	Max	*M*	*SD*
Female	0.47				
White	0.33				
Black	0.19				
Latino	0.26				
Asian	0.17				
Other	0.05				
Free-reduced lunch	0.39				
Limited English proficiency	0.38				
Title I school	0.26				
Age (months)		108	140	123.45	4.74
Iowa Test of Basic Skills (DSS)		142	263	202.78	24.08
Iowa Test of Basic Skills (NPR)		1	99	51.97	28.08
Oral-reading fluency (WCPM)		6	242	120.27	37.83
Elementary Reading Attitude Survey (Total)		23	80	58.45	11.12

Note. DSS = Developmental Standard Score; NPR = National Percentile Rank; WCPM = words correctly read per minute.

The study also does not report the # of students in each group who completed the DIBELS. I approximate these quantities by assuming that the proportion of students in each group remains equal to the proportion observed among the total sample of students who took the post-test ITBS.

$$\begin{aligned} n_1 &= 450 \cdot \frac{252}{486} \\ &\approx 233 \\ n_2 &= 450 \cdot \frac{234}{486} \\ &\approx 217 \end{aligned}$$

We then estimate Cohen's d and its corresponding standard error like so:

$$\begin{aligned} d &= \frac{B}{\sqrt{\frac{s_y^2(n_1+n_2-1) - \frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\ &= \frac{-2.09}{\sqrt{\frac{37.83^2(233+217-1) - \frac{2.09^2(233*217)}{233+217}}{233+217-2}}} \\ &= \frac{-2.09}{37.856} \\ &\approx -0.0552 \end{aligned}$$

$$
\begin{aligned}
SE(d) &= \frac{SE(B)}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\
&= \frac{1.50}{37.856} \\
&= 0.04
\end{aligned}
$$

4 Kim (2007)

The effect size on the Stanford Achievement Test 10 (SAT10) is reported as Hedge's g ($g = 0.04$) in Table 5 with no accompanying standard error.

Table 5
Means and Standard Deviations for the Pre- and Posttest Reading Scores and Effect Sizes, By Grade Level and Free-Lunch Status

	Treatment		Control		
Student characteristic	*M*	*SD*	*M*	*SD*	Effect size (Hedges's *g*)
All grades (n = 279)					
Pretest	618.86	51.06	614.89	53.08	
Posttest	624.54	53.18	618.58	52.82	0.04
Grade 1 (n = 60)					
Pretest	567.81	36.59	558.50	40.37	
Posttest	567.56	41.03	558.82	39.09	−0.01
Grade 2 (n = 47)					
Pretest	605.52	28.71	593.04	38.27	
Posttest	608.48	33.55	593.67	29.06	0.07
Grade 3 (n = 65)					
Pretest	624.06	48.88	616.50	41.25	
Posttest	634.84	41.61	622.38	36.26	0.13
Grade 4 (n = 41)					
Pretest	641.79	38.64	629.55	39.52	
Posttest	642.32	33.14	643.00	43.00	−0.34
Grade 5 (n = 66)					
Pretest	659.58	37.90	667.18	31.76	
Posttest	671.06	38.80	667.21	35.87	0.31
Free lunch (no; n = 216)					
Pretest	627.32	49.17	624.36	51.47	
Posttest	632.98	52.71	628.37	50.66	0.03
Free lunch (yes; n = 63)					
Pretest	590.84	47.72	581.26	44.98	
Posttest	596.56	45.10	583.84	45.79	0.07

By 1.1 of [1], we can estimate the standard error with two additional pieces of information: (1) the # of students in treatment and control (2) the corresponding estimate of Cohen's d. Using Table 4 and its accompanying caption, we know that 138 students belonged to the treatment group and 141 students belonged to the control group.

Table 4
Average Number of Books Read During Summer Vacation by the Treatment and Control Groups, By Grade Level and Free-Lunch Status

	Treatment			Control		
Student characteristic	*M*	*SD*	*n*	*M*	*SD*	*n*
All grades	4.72	1.68	137	3.45	1.89	139
Grade						
1	5.16	1.57	32	3.39	2.06	28
2	5.00	1.54	23	3.64	2.08	22
3	4.77	1.74	30	3.41	1.76	34
4	4.00	1.80	19	4.05	2.06	22
5	4.45	1.68	33	3.00	1.56	33
Free lunch						
No	4.82	1.64	105	3.56	1.87	108
Yes	4.38	1.79	32	3.03	1.92	31

Note. One student in the treatment group and 2 students in the control group did not complete the item "During summer vacation, about how many books (picture books and chapter books) did you read at home?" on the fall reading survey.

We can then estimate Cohen's d like so:

$$\begin{aligned} d &= \frac{g}{1 - \frac{3}{4(n_1+n_2-2)-1}} \\ &= \frac{0.04}{1 - \frac{3}{4(138+141-2)-1}} \\ &= \frac{0.04}{0.9973} \\ &\approx 0.04 \end{aligned}$$

By 1.1 of [1], the Cohen's d standard error can be estimated like so:

$$\begin{aligned} SE(d) &= \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}} \\ &= \sqrt{\frac{138+141}{138*141} + \frac{0.04^2}{2(138+141)}} \\ &= 0.12 \end{aligned}$$

Thus, we have $d = 0.04 \pm 0.12$ in the overall sample.

The same procedure can be employed to estimate the effect size in the low-income sample (where $g = 0.07$, $n_1 = 32$, and $n_2 = 31$).

5 Kim and White (2008)

The adjusted post-test means, standard deviations, and sample sizes of each group are reported in Table 2.

TABLE 2

Posttest Means and Standard Deviations for ITBS and DIBELS

Experimental Condition	Unadjusted Mean	Adjusted Mean	*SD*	*n*
ITBS (Total Reading)				
Books Only	200.92	203.57	28.67	93
Books with Oral Reading Scaffolding	205.39	204.83	26.53	100
Books with Oral Reading and Comprehension Scaffolding	207.23	207.00	28.57	100
Control Group	204.63	203.07	28.01	107
DIBELS Oral Fluency (WCPM)				
Books Only	114.52	116.07	37.97	81
Books with Oral Reading Scaffolding	123.20	120.18	38.43	89
Books with Oral Reading and Comprehension Scaffolding	121.36	121.00	39.93	89
Control Group	118.12	120.05	35.54	91

By 1.1 of [1], we can estimate Cohen's d and the corresponding standard error using the "Books Only" arm:

$$\begin{aligned} d &= \frac{203.57 - 203.07}{\sqrt{\frac{28.67^2(93-1)+28.01^2(107-1)}{93+107-2}}} \\ &= \frac{0.5}{27.63} \\ &\approx 0.018 \end{aligned}$$

$$\begin{aligned} SE(d) &= \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}} \\ &= \sqrt{\frac{93+107}{93*107} + \frac{0.018^2}{2(93+107)}} \\ &= 0.142 \end{aligned}$$

Thus, the effect size of the "Books Only" arm is $d = 0.02 \pm 0.142$.

The same procedure can be employed to estimate the effect size of all other arms on both outcomes.

6 Kim and Guryan (2010)

Table 3 reports the post-test mean and standard deviation in each group.

Table 3
Means and Standard Deviations for GMRT Pretest and Posttest, by Condition

	Control			Treatment			Family literacy			Nonattending			Attending		
Measure	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Pretest															
Total reading	26.38	20.67	107	23.30	17.86	102	23.52	18.27	109	22.23	16.65	164	27.57	21.89	47
Comprehension	26.03	21.49	108	25.50	17.68	102	24.95	19.49	110	24.12	16.88	164	29.02	23.35	48
Vocabulary	29.10	21.92	108	24.10	19.32	102	24.63	18.77	110	23.33	18.07	164	27.94	21.71	48
Posttest															
Total reading	19.40	19.29	106	16.64	16.66	101	17.32	19.01	110	15.71	15.69	162	21.24	23.45	49
Comprehension	21.07	19.58	108	19.78	18.05	101	18.98	20.90	110	18.31	17.30	162	22.84	25.53	49
Vocabulary	20.96	21.54	108	16.87	17.67	103	18.29	19.71	110	16.30	16.58	164	21.96	24.28	49

Note. The nonattending group includes children originally assigned to the treatment group and children originally assigned to the family literacy group who attended zero family literacy events. The attending group includes children originally assigned to the family literacy group who attended one or more family literacy events. GMRT = Gates–MacGinitie Reading Test (national percentile rank).

Unfortunately, the adjusted post-test means are not reported. However, we may approximate them by leveraging two pieces of additional information. First, we may note that there was not a significant difference between the adjusted post-test means.

> An ANCOVA on the total reading scores revealed a nonsignificant main effect of condition, $F(2, 307) = 0.40$, ns, suggesting that there was no difference in covariate-adjusted total reading scores among children in the three experimental conditions.

Second, we may note that the correlation between pre-test and post-test score should be approximately equal to the test-retest reliability in the case where the intervention has an approximately null impact.

> Recently normed in 2005, the GMRT includes a total reading score based on a 48-item comprehension subtest and a 45-item vocabulary subtest. The Kuder–Richardson Formula 20 reliability coefficient for the GMRT Level 4 is .96, and test–retest reliability is .92.

By Equation 3a in Section 7.5 of [5], we may calculate the adjusted post-test means like so:

$$\begin{aligned}
\mu_{1,adjusted} &= \mu_{1,unadjusted} - \beta(\mu_{1,pre} - \mu_{pre}) \\
&= 16.64 - 0.92(23.30 - 24.88) \\
&\approx 18.09 \\
\mu_{2,adjusted} &= \mu_{2,unadjusted} - \beta(\mu_{2,pre} - \mu_{pre}) \\
&= 19.40 - 0.92(26.38 - 24.88) \\
&\approx 18.02
\end{aligned}$$

By 1.1 of [1], we estimate Cohen's d and the corresponding standard error like so:

$$\begin{aligned} d &= \frac{18.09 - 18.02}{\sqrt{\frac{16.66^2(101-1)+19.29^2(106-1)}{101+106-2}}} \\ &= \frac{0.07}{18.06} \\ &\approx 0 \end{aligned}$$

$$\begin{aligned} SE(d) &= \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}} \\ &= \sqrt{\frac{101+106}{101*106} + \frac{0^2}{2(101+106)}} \\ &= 0.139 \end{aligned}$$

Thus, the effect size of the intervention is $d = 0 \pm 0.14$.

7 Allington et al (2010)

The effect size in the overall population is reported like so:

> A t-test found statistically significant differences (t = 2.434, df = 1,328, p = .015) in the performance of the treatment and control students on the FCAT administered after three consecutive summer book distributions.

By 1.9 of [1], we can estimate the effect size like so:

$$\begin{aligned} d &= t\sqrt{\frac{n_1+n_2}{n_1 n_2}} \\ &= 2.434\sqrt{\frac{852+478}{852*478}} \\ &\approx 0.1391 \end{aligned}$$

By 1.1 of [1], we can estimate the standard error like so:

$$\begin{aligned} SE(d) &= \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}} \\ &= \sqrt{\frac{852+478}{852*478} + \frac{0.1391^2}{2(852+478)}} \\ &= 0.0572 \end{aligned}$$

Thus, the effect size of the intervention in the overall sample is $d = 0.14 \pm 0.06$.

The same procedure can be used to estimate the effect size in the sample of free/reduced-price lunch students. However, because the sample size in each group is not directly listed for this subset of the data, we must do two calculations.

First, note that the degrees-of-freedom in the low-income sample t-test is 1088.

> A t-test again found statistically significant differences (t = 3.280, df = 1088, p = .001) in the performance of the free lunch–eligible students in the treatment and control groups on the FCAT administered after three consecutive summer book distributions.

Because only 2 groups were compared, this implies that the total low-income sample size is 1090 students.

To then compute the # of low-income students in each group, we assume that a similar proportion of students are in treatment vs control in the overall population:

$$\begin{aligned} n_1 &= 1090 \cdot \frac{852}{1330} \\ &\approx 698 \\ n_2 &= 1090 \cdot \frac{478}{1330} \\ &\approx 392 \end{aligned}$$

8 Pagan (2010)

The means and standard deviations of the pre-test and post-test outcomes are reported in Table 10.

Table 10

Study 2: Pre-test, Post-test, and Gains Scores for Reading Ability According to Group

	Pre-test		Post-test		Gains Score	
Group/Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*
Book Reading Group (N = 28)						
Passage Comprehension						
Raw score	30.57	7.72	32.29	9.39	1.71	4.40
Standard score	95.86	10.58	94.57	10.69		
Word Identification						
Raw score	59.14	11.70	60.82	15.39	1.69	7.57
Standard score	97.50	10.15	95.39	10.64		
Oral Reading Fluency						
Total raw score	86.21	41.70	91.86	42.82	4.19	16.03
Reading Ability Composite[a]					2.53	7.20
Control Group (N = 29)						
Passage Comprehension						
Raw score	32.07	5.17	31.66	6.65	-0.41	4.42
Standard score	97.07	8.07	94.17	9.10		
Word Identification						
Raw score	60.72	12.02	60.28	12.01	-0.45	5.13
Standard score	98.28	11.78	95.62	10.47		
Oral Reading Fluency						
Total raw score	80.24	37.47	78.62	38.00	-1.62	17.26
Reading Ability Composite[a]					-0.83	6.67

[a]Reading Ability Composite score was computed using average gains scores for passage comprehension, word identification, and oral reading fluency.

Unfortunately, the adjusted post-test means are not reported and, given that the intervention actually impacted the outcomes of interest, we cannot approximate the adjusted post-test means using the test-retest reliability as was done in Kim and Guryan (2010). Thus, we cannot estimate the effect size.

9 Wilkins et al (2012)

The effect size is reported as an unstandardized regression coefficient ($B = 4.89 \pm 9.83$) in Table 4-1.

Table 4-1. Summer reading program impact estimate with 95 percent confidence interval, 2009 (*n*=1,571)

Outcome measure	*Treatment group mean (n=791)*	*Control group mean (n=780)*	*Estimated intent-to-treat impact coefficient*[a]	p-*value*	*95% confidence interval*	*Estimated impact (effect size*[b]*)*
Scholastic Reading Inventory Scores	330.10 (238.94)	321.49 (228.89)	4.89 (9.83)	.62	–14.39, 24.17	0.02

Note: Numbers in parentheses are standard deviations for treatment and control group means and the standard error for estimated intent-to-treat impact. The means and standard deviations are unadjusted.

a. Results are analyzed using covariate-adjusted ordinary least squares regression.

b. Calculated using Hedges' *g*.

Source: Scholastic Reading Inventory data collected September 2009–December 2009.

Before estimating Cohen's d, it's worth noting at the outset that the standard deviations associated with the treatment and control group scores are unexpectedly high. To contextualize this variability, note that the pre-test variable has a much lower standard deviation as shown in Table 2-5.

Table 2-5. Baseline characteristics for treatment and control groups

Characteristic	*Treatment group mean*[a] *(n=896)*	*Control group mean*[a] *(n=889)*	*Difference in means*[b]	*Test statistic*[c]	p-*value*
Age in years	9.41 (0.47)	9.43 (0.50)	–0.02	–0.70	.48
Female	0.51 (0.50)	0.50 (0.50)	0.01	0.59	.56
Race/ethnicity[d]				6.39	.17
American Indian	0.00 (0.05)	0.00 (0.05)	0.00		
Asian	0.03 (0.18)	0.05 (0.21)	–0.01		
Black	0.23 (0.42)	0.18 (0.39)	0.05		
Hispanic	0.66 (0.47)	0.69 (0.46)	–0.03		
White	0.07 (0.26)	0.08 (0.27)	–0.01		
With Individualized Education Program (IEP)	0.05 (0.22)	0.04 (0.19)	0.01	1.48	.14
English language learner student	0.47 (0.50)	0.49 (0.50)	–0.02	–0.92	.36
Baseline Lexile measure	400.95 (139.60)	398.19 (140.73)	2.76	0.42	.68

Though the pre-test variable and post-test variable are measured using different assessments, both are converted to the Lexile scale as noted on page 32. Thus, it is surprising that the post-test variable has much more variability relative to the pre-test variable.

Alternative covariate for baseline reading. Because the outcome variable in this study is reported on the Lexile scale, a baseline variable also measured on the Lexile scale was used. However, converting the TAKS scaled scores to Lexile measures could have resulted in a loss of information about reading ability. Therefore, the results of the analysis were also examined using an alternative baseline variable—TAKS scaled scores—to assess the robustness of the results to potential error in linking scaled scores to Lexile measures.

Nevertheless, we may proceed with estimating Cohen's d. By [2], we can esti-

mate the pooled standard deviation of the post-test score like so:

$$\begin{aligned} s_y &= \sqrt{\frac{s_1^2(n_1-1)+\mu_1^2 n_1 + s_2^2(n_2-1)+\mu_2^2 n_2 - (n_1+n_2)\mu^2}{n_1+n_2-1}} \\ &= \sqrt{\frac{238.94^2(791-1)+330.10^2\cdot 791+228.89^2(780-1)+321.49^2\cdot 780-(791+780)325.825^2}{791+780-1}} \\ &\approx 233.821 \end{aligned}$$

By 1.31 of [1], we can estimate Cohen's d and its corresponding standard error like so:

$$\begin{aligned} d &= \frac{B}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\ &= \frac{4.89}{\sqrt{\frac{233.825(791+780-1)-\frac{4.89^2(791*780)}{791+780}}{791+780-2}}} \\ &= \frac{4.89}{233.886} && \approx 0.02 \end{aligned}$$

$$\begin{aligned} SE(d) &= \frac{SE(B)}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\ &= \frac{9.83}{233.886} && = 0.042 \end{aligned}$$

Thus, the effect size of the intervention is $d = 0.02 \pm 0.04$.

10 White et al (2014)

The covariate-adjusted means and standard deviations are reported in Table 3. By 1.1 of [1], we can compute Cohen's d and its corresponding standard error like so:

$$\begin{aligned} d &= \frac{184.01-184.95}{\sqrt{\frac{28.88^2(397-1)+29.73^2(395-1)}{395+397-2}}} \\ &\approx -0.0321 \end{aligned}$$

$$\begin{aligned} SE(d) &= \sqrt{\frac{397+395}{397*395}+\frac{-0.0321^2}{2(397+395)}} \\ &= 0.071 \end{aligned}$$

Thus, the intervention had an effect size of $d = -0.03 \pm 0.07$ in the overall sample.

11 Kim et al (2016)

The effect size is reported as a semi-standardized coefficient ($B = 0.039 \pm 0.019$) in Table 2.

Table 2. Delayed intent-to-treat effects on spring 2014 EOG posttest reading scores.

	(1)	(2)	(3)
Variable	All schools	High-poverty schools	Moderate-poverty schools
READS	0.039*	0.052*	0.012
	(0.019)	(0.022)	(0.035)
ITBS Pretest (std)	0.741***	0.752***	0.722***
	(0.012)	(0.016)	(0.018)
N	5569	3815	1754
R^2	0.626	0.618	0.617

Note. EOG and ITBS are standardized within grade. Standard errors clustered at school-level in parentheses. All models control for fixed effects of spring 2013 homeroom.
$\sim p < 0.10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

To compute Cohen's d, we need to know the # of students in treatment and control. Unfortunately, however, the study only reports the total sample size $n = 5569$. Because the study is a randomized experiment, we assume that the sample is split 50-50, yielding the following sample sizes: $n_1 \approx 2784$ and $n_2 \approx 2785$. Then, by 1.31 of [1], Cohen's d and its corresponding error can be estimated like so:

$$
\begin{aligned}
d &= \frac{B}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\
&= \frac{0.039}{\sqrt{\frac{1(2784+2785-1)-\frac{0.039^2(2784*2785)}{2784+2785}}{2784+2785-2}}} \\
&\approx \frac{0.039}{1} \\
&\approx 0.039
\end{aligned}
$$

$$
\begin{aligned}
SE(d) &= \frac{SE(B)}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\
&\approx \frac{0.019}{1} \\
&\approx 0.019
\end{aligned}
$$

Thus, the effect size of the intervention is $d = 0.039 \pm 0.019$.

12 Stein (2016)

The effect size is reported as an unstandardized regression coefficient in Table 2.

Table 2. Estimated Effect of SummerREADS Participation on MSA Reading Scale Scores and Proficiency Categories.

	Scale score	Basic[a]	Proficient or advanced[a]
Rising third graders	3.47	0.77*	1.30*
	(2.70)	(0.12)	(0.20)
Rising fourth graders	7.00**	0.75*	1.32*
	(3.17)	(0.13)	(0.20)

Unfortunately, it is not possible to convert this effect size into Cohen's d because the study does not report the standard deviation of the outcome variable. Two additional points are worth noting:

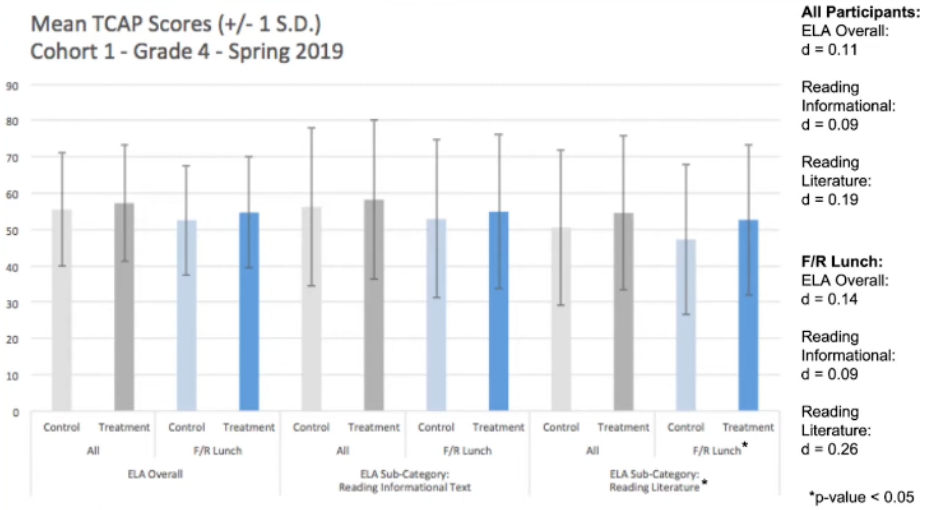
1. The study does not report the effect size on oral reading fluency, but states that the null results are available upon request.

2. The study pools together students who have experienced the intervention for 2 years with students who have experienced the intervention for 1 year, complicating interpretation.

Unfortunately, I was unable to contact the author to retrieve this information given that they left their university position in August 2023 (so presumably, they do not have access to their university email) and their personal website does not list any contact information.

13 McGill-Franzen, Allington, and Ward (2020)

The effect size is reported as Cohen's d in the online seminar.



However, no corresponding standard error is reported. To estimate the standard error, we can make use of the sample sizes reported in the online seminar.

PARTICIPANTS

		Treatment	Control	Total
Cohort 1	Baseline Sample (Fall 2016)	541	503	1,044
	Attrition	130	112	242
	Final Sample (Spring 2019)	411	391	802
	Attrition Rate	27%	25%	26%
Cohort 2	Baseline Sample (Fall 2017)	233	245	478
	Attrition	52	42	94
	Final Sample (Spring 2019)	181	203	384
	Attrition Rate	25%	19%	22%
Cohort 1 & 2 Pooled*	ELA Overall	519	545	1,064
	ELA Informational	519	545	1,064
	ELA Literature	579	585	1,164

*Cohort 1 (Spring 2018) & Cohort 2 (Spring 2019)

By 1.1 of [1], the standard error can be estimated like so:

$$\begin{aligned} SE(d) &= \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}} \\ &= \sqrt{\frac{411 + 391}{411 * 391} + \frac{0.11^2}{2(411 + 391)}} \\ &= 0.0707 \end{aligned}$$

Thus, the effect size of the intervention in the overall population is $d = 0.11 \pm 0.07$.

To estimate the standard error of the effect size in the low-income population, we need the sample size of the low-income students in the Cohort 1 treatment and control groups. Unfortunately, these #'s are not reported in the paper. However [3] reports the % of low-income students in this sample as 75%. Thus, we can approximate the sample sizes like so:

$$\begin{aligned} n_1 &= 0.75 \cdot 411 \\ &\approx 308 \\ n_2 &= 0.75 \cdot 391 \\ &\approx 293 \end{aligned}$$

Using these sample size estimates and the same procedure as before, $SE(d) = 0.0817$. Thus, the effect size of the intervention in the low-income sample is $d = 0.14 \pm 0.08$.

Two points are worth noting about this study:

1. The effect size on the "Reading Information" subtest decreased in Cohort 1 likely because, in 2019, Tennessee modified the reading curriculum to include more time dedicated to reading informational texts in classroom. The qualitative divergence in effect size between the "Reading Information" and "Reading Literature" subtests is consistent w/this curriculum change.

2. The effect size in Cohort 2 did not reveal any differences between treatment and control because (1) Tennessee introduced an early-grade summery literacy program which impacted the targeted sample of students (2) Save the Children's Action Network provided tutoring and books throughout the school year and summer to the targeted sample of students. In other words, control effectively received the same intervention as experiment, so it is not surprising that both groups performed equally well on standardized reading tests. For this reason, I do not include the Cohort 2 effect size in the meta-analysis.

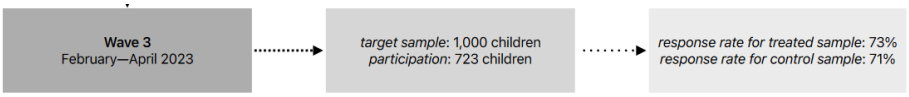
14 Anger et al (2024)

The effect size is reported as a semi-standardized regression coefficient in Table 5.

Table 5: Effect of e-book reading intervention on components of summary indices

	Intention-to-treat (ITT)	p-value of randomized inference	Treatment-on-the-treated (TOT)	p-value of randomized inference	Mean
Reading behavior					
At least one e-book	0.165	0.000	0.244	0.000	0.125
At least one printed book	0.062	0.040	0.071	0.062	0.764
At least one book[a]	0.084	0.004	0.102	0.004	0.795
At least two days/week[b]	0.079	0.012	0.091	0.012	0.741
Academic achievement					
Reading comprehension	0.114	0.162	0.156	0.086	-0.050
German	0.013	0.752	0.050	0.294	0.494
Mathematics	0.084	0.038	0.109	0.018	0.467
Aspirations	0.002	0.984	0.016	0.734	0.620

To estimate Cohen's d, we need to know the sample size of the treatment and control group. We can retrieve these sample sizes by multiplying the attrition rates documented in Figure 2 by the sample size at baseline.



Note that we use Survey Wave 3 attrition rates as the caption below Table 5 states that reading comprehension assessments were administered 1 year after the intervention, ruling out the possibility that tests were administered during Survey Wave 2. These attrition rates yield sample sizes equal to $n_1 = 365$ and $n_2 = 355$, respectively.

Notes: This table shows the impact of the e-book treatment on components of the four indices. The variables in the index *Reading behavior* are binary, taking the value one if the child has read at least one book in the last four weeks or reported reading on at least two days a week. Reading comprehension is a standardized score of the reading test a year after the intervention. Math and German grades are dichotomized to take on the value one if the student has a good grade (1 or 2 on the German grading scale), and zero otherwise. Components of SDQ are standardized with a mean of zero and a standard deviation of one. All regressions control for strata (randomization block) fixed effects. Only observations with observed outcomes are included.

By 1.31 of [1], we can estimate Cohen's d like so:

$$\begin{aligned} d &= \frac{B}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\ &= \frac{0.114}{\sqrt{\frac{1(365+355-1)-\frac{0.114^2(365*355)}{365+355}}{365+355-2}}} \\ &\approx \frac{0.114}{1} \\ &\approx 0.114 \end{aligned}$$

Because Table 5 does not report the standard error of the regression coefficient, we have to approximate it using the corresponding p-value ($p = 0.162$). Under the assumption that the non-parametric (i.e. randomization inference) p-value is equivalent to the parametric p-value, $p = 0.162$ should map to $Z = 1.398$ (see [4]). Because $Z = \frac{B}{SE(B)}$, $SE(B) = \frac{0.114}{1.398} = 0.0815$.

We can then estimate the standard error corresponding to Cohen's d like so:

$$\begin{aligned} SE(d) &= \frac{SE(B)}{\sqrt{\frac{s_y^2(n_1+n_2-1)-\frac{B^2(n_1*n_2)}{n_1+n_2}}{n_1+n_2-2}}} \\ &\approx \frac{0.0815}{1} \\ &\approx 0.0815 \end{aligned}$$

Thus, the effect size of the intervention is $d = 0.11 \pm 0.08$.

15 References

1. Wilson (2023). "Practical Meta Analysis Effect Size Calculator". Link.
2. Wikipedia. "Standard deviation [Sample-based statistics]". Link.
3. Arnold Ventures. "Replication Randomized Controlled Trial of Annual Book Fairs to Promote Summer Reading in Grades 1-3, Conducted in Rural, High-Poverty Elementary Schools". Link.
4. GIGA Calculator. "P-Value to Z-Score Calculator". Link.
5. Boomsma 2012. "Analysis of Covariance in R". Link.