Evidence for Action: A Secondary Analysis of the Reading Recovery Scale-Up

Robert M. Schwartz, Oakland University
Richard G. Lomax, The Ohio State University

Purpose

As the world prepares to examine the results of randomized controlled trials (RCTs) to judge the safety and effectiveness of COVID-19 vaccines, the education community should refocus on what research has shown to be effective literacy practices. The Institute of Education Sciences (IES) highest evidence level is "the independent evaluation of a fully-developed education intervention with prior evidence of efficacy, when implemented by the end user under routine conditions" (IES, 2016, p. 5). The May, Sirinides, Gray, and Goldsworthy (2016) evaluation of the Reading Recovery® i3 scale-up provides this type of high-quality evidence.

Both the amount of evidence and the size of the effects shown in the May et al. (2016) evaluation were impressive. The final evaluation report combined data from four independent RCTs conducted in different schools across the 4-year scale-up. These studies included 3,444 matched pairs of low performing first-grade students from 1,122 schools. The 12– to 20–week Reading Recovery intervention produced effects that were 4.6 times larger than studies using similar outcome measures and 3.5 times the average effect of Title I interventions (May et al.; Schwartz, 2016).

The current analysis extends their findings in two ways. First, we examined the intervention's effects on the six subscales of *An Observation Survey of Early Literacy Achievement* (Clay, 2013). The May et al. (2016) analysis only reported gains on the Observation Survey Total Score and Iowa Test of Basic Skills (ITBS) Word Reading and Comprehension subscales. The Observation Survey subscales provide information on components of the reading process that Reading Recovery teachers use to guide instructional decisions. These subscales have previously been reported in the What Works Clearinghouse (WWC) intervention report (2013) to assess their beginning reading domains:

Alphabetics:	Letter Identification, Word Test
Reading Fluency:	Text Reading Level
Reading Ability:	Concepts About Print, Hearing and Recording Sounds in Words, Writing Vocabulary

Second, we conducted a subgroup analysis for those students whose entry Observation Survey Total Scores predicted a severe reading difficulty at the end of first grade (D'Agostino, Rodgers, & Mauck, 2018). The May et al. (2016) report includes a subgroup analysis showing that English learners show as strong growth as the total sample. We conducted this subgroup analysis to address a persistent claim that the Reading Recovery intervention is not effective for the most at-risk beginning readers (Chapman, Greaney, & Tunmer, 2015; Cook, Rodes, & Lipsitz, 2017; Reynolds & Wheldall, 2007; Schwartz, Hobsbaum, Briggs, & Scull, 2009). For example, Reynolds and Wheldall claim that while Reading Recovery "works for many students, it has not demonstrated that it works for the students who are most at-risk of failing to learn to read" (p. 213). Our analysis tested whether the intervention is effective for students predicted to be classified as reading disabled based on the screening scores (D'Agostino, Rodgers, & Mauck; National Center for Intensive Intervention (NCII; 2018).

Methods

The May et al. (2016) data includes four large, independent samples of students taught by different teachers from different schools during each of the 4 years of the scale-up. In total 3,444 pairs of first-grade students were matched on the fall Observation Survey Text Reading Level measure and then randomly assigned to the treatment or control condition. Given this large sample size, almost any difference in scores between the treatment and control groups are likely to be statistically significant.

Therefore, in our secondary analysis we present descriptive statistics and Hedges' *d* effect size calculations consistent with WWC procedures (WWC, 2018, p. 13). "For the WWC, effect sizes of 0.25 standard deviations or larger are considered to be substantively important. Effect sizes at least this large are interpreted as a qualified positive

(or negative) effect, even though they may not reach statistical significance in a given study" (p. 14). Effect sizes greater than 0.80 are generally considered large.

The six Observation Survey subscales are described in Clay's *An Observation Survey of Early Literacy Achievement* (2013). The appendices include U.S. norms for these measures at the beginning, middle, and end of first grade as well as reliability studies and correlation among the subscales. (Also see International Data Evaluation Center [IDEC] publications for the latest versions.)

The Letter Identification (LI) subscale has a maximum score of 54. Children are asked to identify upper- and lowercase letters by providing either the letter name, its sound, or a word beginning with that letter. Two formats of the lowercase *a* and *g* are included in the task.

The Word Test (WT) subscale has a maximum score of 20. Children are asked to read one of three different lists of high-frequency words available for multiple administrations of the task.

The Concepts About Print (C.A.P.) subscale has a maximum score of 24. Children are asked to respond to probe questions during the reading of one of four books designed for this task.

The Hearing and Recording Sounds in Words (HRSW) subscale has a maximum score of 37. Children are asked to listen to and try to represent the sounds in one of five sentences designed for this task. One point is given for each sound represented by an appropriate letter.

The Writing Vocabulary (WV) subscale has no set maximum score. Children are asked to write all the words they know in 10 minutes.

The Text Reading Level (TRL) subscale has a maximum score of 30. A score of 20 demonstrates an ability to read second-grade material with 90% accuracy with reasonable fluency. The analysis for the TRL measure is calculated for both raw scores and scale scores. The scale scores provide an interval measure that is more appropriate for calculations of effect size (D'Agostino, Rodgers, & Mauck, 2018).

Results

Observation Survey subscale analysis

Table 1 shows pre-test means, post-test means, standard deviations, and effect size calculations for the treatment (T) and control (C) groups on the Observation Survey subscales pooled across the 4-year i3 study (n = 3,439 per group). After the intervention, the comparison of the treatment to control groups shows medium to large effects on each of the Observation Survey subscales. These results were replicated in each of the four independent RCTs conducted in the 4-year scale-up evaluation (Appendix, Tables 1A to 6A).

Subgroup analysis: Students predicted to need intensive intervention

Table 2 shows pre-test means, post-test means, standard deviations, and effect size calculations for ITBS and Observation Survey measures pooled across the 4-year i3 study for students predicted to need intensive intervention (Observation Survey Fall Total Score < 419; n = 2,712 per group). After the intervention the comparison of the treatment to control groups shows medium to large effects on each of Observation Survey and ITBS subscales. These results were replicated in each of the four independent RCTs conducted in the 4-year scale-up evaluation. (See Schwartz & Lomax, 2018, Appendix, Tables 7A to 12A.)

Conclusions

In a recent blog (Soldner, 2020), the commissioner of the IES National Center for Education Evaluation and Regional Assistance described the current state of educational research:

> There are currently 10,677 individual studies in the What Works Clearinghouse (WWC) database. Of those, only about 11 percent meet the WWC's internal validity standards. Among them, only 445 have at least one statistically significant positive finding. Because the WWC doesn't consider results from studies that don't have strong internal validity, it isn't quite as simple as saying "only about 4 percent of things work in education." Instead, we're left with "89 percent of things aren't tested rigorously enough to have confidence about whether they work, and when tested rigorously, only about 38 percent do." Between the "file drawer" problem that plagues research generally and our own review of the results from IES efficacy trials, we have reason to believe the true efficacy rate of "what works" in education is much lower.

The May et al. (2016) research is a rare exception to the dismal efficacy rate. Not only does the research meet WWC rigorous standards, but it does so with multiple

Table 1. Pre-Test Means, Post-Test Means, Standard Deviations, and Effect Size Calculations for the Treatment (T) and Control (C) Groups on the Observation Survey (OS) Subscales Pooled Across the 4-Year i3 Study (n = 3,439 per group)

OS Measure	Letter ID		Word Test		C.A.P.		HRSW		WV		TRL Scale Score	
Treatment/Control	T	C	T	C	T	C	T	C	T	C	T	C
Pre-Test Mean	46.4	46.1	3.2	3.1	11.7	11.6	18.0	17.7	9.0	8.9	307.1	305.8
(Standard Deviation)	(8.1)	(8.5)	(3.1)	(3.0)	(3.5)	(3.6)	(9.7)	(9.7)	(6.2)	(6.3)	(76.5)	(76.8)
Post-Test Mean	52.4	51.1	14.7	10.3	18.3	15.7	33.4	29.6	38.8	27.1	489.7	431.0)
(Standard Deviation)	(2.9)	(4.9)	(4.6)	(5.3)	(3.2)	(3.3)	(5.0)	(7.4)	(14.1)	(12.9)	(44.3)	(70.7)
Effect Size	+0.32		+0.89		+0.80		+0.60		+0.87		+0.99	

replications under routine conditions. The extension of these findings to the Observation Survey subscales illustrates some of the core components that need to work together as a teacher helps a child build a literacy processing system. The large effects on these subscales reflect the accelerated growth that most Reading Recovery students are able to make during the intervention as they establish a system that can continue to improve with good classroom instruction.

The current WWC Reading Recovery intervention report (2013) does not include the evidence from the four RCTs included in the May et al. (2016) independent evaluation. WWC has, however, conducted a single study review of this research and found the research to meet their criteria

Table 2. Pre-Test Means, Post-Test Means, Standard Deviations, and Effect Size Calculations for ITBS and Observation Survey (OS) Measures Pooled Across the 4-Year i3 Study for Students Predicted to Need Intensive Intervention (OS Fall Total Score < 419, n = 2,712 per group)

OS Measure	Letter ID		Word Test		C.A.P.		HRSW		WV		TRL	
Treatment/Control	T	C	T	C	T	C	T	C	T	C	T	C
Pre-Test Mean	45.6	45.1	2.5	2.4	11.2	11.1	16.0	15.6	7.4	7.3	0.8	0.8
(Standard Deviation)	(8.5)	(8.9)	(2.4)	(2.3)	(3.4)	(3.5)	(8.9)	(9.0)	(4.5)	(4.7)	(1.0)	(1.1)
Post-Test Mean	52.2	50.9	14.2	9.5	18.0	15.4	32.9	28.9	37.0	25.4	9.8	4.7
(Standard Deviation)	(3.2)	(5.0)	(4.7)	(5.1)	(3.2)	(3.3)	(5.3)	(7.5)	(13.6)	(12.0)	(4.7)	(3.6)
Effect Size	+0.31		+0.96		+0.80		+0.62		+0.90		+1.22	

Measure	OS Ttoal Score		ITBS Word		ITBS Comprehension		OS TRL Scale Score	
Treatment/Control	T	C	T	C	T	C	T	C
Pre-Test Mean	360.4	358.9					295.0	293.5.
(Standard Deviation)	(32.0)	(32.4)					(71.9)	(72.1)
Post-Test Mean	490.7	444.9	139.0	134.9	139.8	135.9	484.5	424.3
(Standard Deviation)	(44.2	(47.3)	(9.2)	(8.5)	(8.6)	(7.6)	(45.8)	(70.4)
Effect Size	+1.00		+.46		+0.48		+1.01	

without reservations. If this evidence and the current secondary analysis were included in the WWC intervention report, Reading Recovery would be rated as showing positive results in all four beginning reading domains (Schwartz, 2018).

The subgroup analysis counters the claim that the intervention is not effective for the lowest-performing first-grade readers. The NCII 2018 review shows that the fall Observation Survey Total Score meets their criteria for an effective literacy screening tool. All of the students in the subgroup analysis fall below the cut point that predicts a severe reading disability at the end of first grade. Despite these low entry scores, the effect sizes shown in Table 2 are as large or larger than those for the total sample. Many of these initially low-performing students are now at grade level after the Reading Recovery intervention.

The May et al. (2016) evaluation and this secondary analysis provide a model for educational effectiveness evidence and research-based early literacy interventions. The replication of substantial effects on the ITBS, the Observation Survey subscales, and with subgroups of English learners (May et al.) and the lowest-performing students demonstrates and underscores the ability of the Reading Recovery network to partner with teachers, principals, and district personnel to implement an effective early intervention. If we are as serious in serving the literacy needs of all students as we are in maintaining their health, then this is the type of evidence we need.

References

Chapman, J. W., Greaney, K. T., & Tunmer, W. E. (2015). Is Reading Recovery an effective early literacy intervention programme for children who most need literacy supports? In W. E. Tunmer & J. W. Chapman (Eds.), *Excellence and equity in literacy instruction: The case of New Zealand* (pp. 41–70). Basingstoke, UK: Palgrave Macmillan.

Clay, M. M. (2013). *An observation survey of early literacy achievement*. (3rd ed.). Portsmouth, NH: Heinemann.

Cook, P., Rodes, D. R., & Lipsitz, K. L. (2017). The reading wars and Reading Recovery: What educators, families, and taxpayers should know. *Learning Disabilities: A Multidisciplinary Journal, 22*(2), 12–23. https://doi.org/10.18666/LDMJ-2017-V22-I2-8391

D'Agostino, J. V., Rodgers, E., & Mauck, S. (2018). Addressing inadequacies of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly, 53*(1), 51–69. https://doi.org/10.1002/rrq.181

International Data Evaluation Center. (2012). *U.S. norms and correlations for an observation survey of early literacy achievement*. Columbus, OH. https://www.idecweb.us/Publications.aspx

Institute of Education Sciences. (2016). *Building evidence: What comes after an efficacy study?* https://ies.ed.gov/ncer/whatsnew/techworkinggroup/pdf/ BuildingEvidenceTWG.pdf

May, H., Sirinides, P., Gray, A., & Goldsworthy, H. (2016). *Reading Recovery: An evaluation of the four-year i3 scale-up*. Philadelphia: Consortium for Policy Research in Education. http://www.cpre.org/reading-recovery-evaluation-four-year-i3-scale. See also Sirinides, P., Gray, A., & May, H. (2018). The impacts of Reading Recovery at scale: Results from the 4-year i3 external evaluation. *Educational Evaluation and Policy Analysis, 40*(3), 316–335. https://doi.org/10.3102/0162373718764828

National Center on Intensive Intervention (2018). *Academic screening tools chart*. https://charts.intensiveintervention.org/chart/academic-screening/observation-survey-early-literacy-achievement#title

Reynolds, M., & Wheldall, K. (2007). Reading Recovery 20 years down the track: Looking forward, looking back. *International Journal of Disability, Development and Education, 54*(2), 199–223.

Schwartz, R. M. (2016). Effective early intervention: Lessons from the i3 evaluation of Reading Recovery. *The Journal of Reading Recovery, 16*(1), 47–54.

Schwartz, R. M. (2018). Reading Recovery: How do we rank? *The Journal of Reading Recovery, 17*(2), 61–65.

Schwartz, R. M., Hobsbaum, A., Briggs, C., & Scull, J. (2009). Reading Recovery and evidence-based practice: A response to Reynolds and Wheldall (2007). *International Journal of Disability, Development and Education, 56*(1), 5–15.

Schwartz, R. M., & Lomax, R. (2018). *Secondary analysis of the Reading Recovery four-year i3 scale-up*. Paper presented at the Society for the Scientific Study of Reading, Brighton, UK. https://www.researchgate.net/publication/344047317_ERIC_Submission-_Secondary_analysis_of_the_Reading_Recovery_four-year_i3_scale-up_Society_for_the_Scientific_Study_of_Readingdocx#fullTextFileContent

Soldner, M. (2020). *"The how" of "What Works:" The importance of core components in education research*. https://ies.ed.gov/blogs/ncee/post/the-how-of-what-works-the-importance-of-core-components-in-education-research

What Works Clearinghouse. (2013). *Reading Recovery intervention report*. https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_readrecovery_071613.pdf

What Works Clearinghouse. (2018). *Procedures handbook, Version 4*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

Appendix: Tables by Year for Full Group and Subgroup Predicted to Need Intensive Intervention

Table 1A. Effect Size Calculations for Text Reading Level (TRL) Raw Scores Across the 4-Year i3 Study

	2011–12		2012–13		2013–14		2014–15		Pooled	
	T	C	T	C	T	C	T	C	T	C
N	429	429	725	725	855	855	1430	1430	3439	3439
TRL Pre-Test										
Mean	1.0	1.0	1.0	0.9	1.0	1.0	1.1	1.1	1.0	1.0
(Standard Deviation)	(1.3)	(1.1)	(1.2)	(1.1)	(1.2)	(1.1)	(1.2)	(1.3)	(1.2)	(1.2)
TRL Post-Test										
Mean	10.6	5.2	10.3	5.1	10.4	5.4	10.5	5.2	10.5	5.2
(Standard Deviation)	(4.8)	(3.7)	(4.7)	(4.2)	(4.9)	(3.9)	(4.9)	(4.0)	(4.9)	(4.0)
TRL Pre-Test Scale Score										
Mean	304.4	302.2	302.1	300.4	308.1	306.2	309.7	309.4	307.1	305.8
(Standard Deviation)	(74.9)	(75.4)	(75.3)	(75.2)	(75.5)	(76.3)	(77.9)	(78.2)	(76.5)	(76.8)
TRL Post-Test Scale Score										
Mean	492.0	435.5	488.7	425.0	489.5	433.4	489.6	431.3	489.7	431.0
(Standard Deviation)	(39.5)	(64.5)	(44.7)	(77.1)	(44.5)	(69.5)	(45.4)	(69.7)	(44.3)	(70.7)
Raw Score Effect Size	+1.46		+1.24		+1.28		+1.33		+1.33	
Scale Score Effect Size	+0.88		+0.83		+0.81		+0.84		+0.83	

Table 2A. Effect Size Calculations for Letter Identification (LI) Scores Across the 4-Year i3 Study

	2011–12		2012–13		2013–14		2014–15		Pooled	
	T	C	T	C	T	C	T	C	T	C
N	429	429	725	725	855	855	1430	1430	3439	3439
LI Pre-Test										
Mean	46.7	46.6	46.4	46.0	46.7	46.5	46.1	45.7	46.4	46.1
(Standard Deviation)	(6.7)	(7.2)	(7.5)	(8.5)	(7.8)	(7.9)	(8.8)	(9.2)	(8.1)	(8.5)
LI Post-Test										
Mean	52.4	51.4	52.3	51.0	52.5	51.4	52.3	50.8	52.4	51.1
(Standard Deviation)	(2.6)	(4.6)	(3.2)	(4.7)	(2.5)	(3.8)	(3.1)	(5.5)	(2.9)	(4.9)
Effect Size	+0.22		+0.28		+0.29		+0.27		+0.26	

Table 3A. Effect Size Calculations for Word Test (WT) Scores Across the 4-Year i3 Study

	2011–12		2012–13		2013–14		2014–15		Pooled	
	T	C	T	C	T	C	T	C	T	C
N	429	429	725	725	855	855	1430	1430	3439	3439
WT Pre-Test										
Mean	3.0	2.8	3.1	3.0	3.3	3.2	3.3	3.2	3.2	3.1
(Standard Deviation)	(3.1)	(2.8)	(3.0)	(3.0)	(3.0)	(3.0)	(3.1)	(3.0)	(3.1)	(3.0)
WT Post-Test										
Mean	14.7	10.4	14.9	10.4	14.8	10.5	14.6	10.0	14.7	10.3
(Standard Deviation)	(4.4)	(4.9)	(4.5)	(5.5)	(4.6)	(5.3)	(4.6)	(5.3)	(4.6)	(5.3)
Effect Size	+0.88		+0.82		+0.81		+0.87		+0.83	

Table 4A. Effect Size Calculations for Concepts About Print (C.A.P.) Scores Across the 4-Year i3 Study

	2011–12		2012–13		2013–14		2014–15		Pooled	
	T	C	T	C	T	C	T	C	T	C
N	429	429	725	725	855	855	1430	1430	3439	3439
C.A.P. Pre-Test										
Mean	11.5	11.5	11.4	11.5	11.9	11.8	11.7	11.5	11.7	11.6
(Standard Deviation)	(3.4)	(3.5)	(3.5)	(3.6)	(3.4)	(3.4)	(3.6)	(3.7)	(3.5)	(3.6)
C.A.P. Post-Test										
Mean	18.3	15.6	18.2	15.6	18.5	15.8	18.3	15.7	18.3	15.7
(Standard Deviation)	(2.9)	(3.3)	(3.2)	(3.4)	(3.1)	(3.4)	(3.2)	(3.3)	(3.2)	(3.3)
Effect Size	+0.79		+0.76		+0.79		+0.79		+0.79	

Table 5A. Effect Size Calculations for Hearing and Recording Sounds in Words (HRSW) Scores Across the 4-Year i3 Study

	2011–12		2012–13		2013–14		2014–15		Pooled	
	T	C	T	C	T	C	T	C	T	C
N	429	429	725	725	855	855	1430	1430	3439	3439
HRSW Pre-Test										
Mean	17.6	17.0	17.0	17.0	18.5	18.2	18.3	17.9	18.0	17.7
(Standard Deviation)	(9.2)	(9.5)	(9.4)	(9.6)	(9.8)	(9.7)	(10.0)	(9.9)	(9.7)	(9.7)
HRSW Post-Test										
Mean	33.6	30.1	33.5	29.2	33.5	30.0	33.2	29.4	33.4	29.6
(Standard Deviation)	(4.5)	(7.1)	(4.8)	(7.9)	(4.9)	(6.8)	(5.3)	(7.4)	(5.0)	(7.4)
Effect Size	+0.49		+0.54		+0.52		+0.51		+0.51	

Table 6A. Effect Size Calculations for Writing Vocabulary (WV) Scores Across the 4-Year i3 Study

	2011–12		2012–13		2013–14		2014–15		Pooled	
	T	C	T	C	T	C	T	C	T	C
N	429	429	725	725	855	855	1430	1430	3439	3439
WV Pre-Test										
Mean	8.5	8.5	8.1	8.4	9.2	9.2	9.5	9.2	9.0	8.9
(Standard Deviation)	(5.6)	(6.1)	(5.5)	(5.8)	(6.2)	(6.5)	(6.6)	(6.5)	(6.2)	(6.3)
WV Post-Test										
Mean	37.7	27.7	38.1	26.8	39.1	27.7	39.3	26.8	38.8	27.1
(Standard Deviation)	(12.9)	(12.5)	(13.2)	(13.3)	(14.4)	(12.6)	(14.6)	(12.9)	(14.1)	(12.9)
Effect Size	+0.80		+0.85		+0.90		+0.97		+0.91	

About the Authors

Dr. Robert Schwartz is an emeritus professor in the Department of Reading and Language Arts at Oakland University in Rochester, MI. He is a past president of and former research consultant for the Reading Recovery Council of North America. His research interests include self-monitoring in beginning reading, early literacy intervention, research design, and professional development for literacy teachers. In the What Works Clearinghouse 2007 review of 887 studies from 153 beginning reading programs, Dr. Schwartz's Reading Recovery research was one of only 27 studies that met WWC's standards without reservations. He can be reached at rschwart@oakland.edu.

Dr. Richard Lomax is the former director of research for Reading Recovery and Descubriendo la Lectura and professor emeritus of educational studies at The Ohio State University, where he was previously associate dean for research and administration in the College of Education and Human Ecology. His research primarily focuses on multivariate analysis and models of literacy acquisition. He has published textbooks and in diverse journals including *Reading Research Quarterly, Parenting: Science and Practice, Understanding Statistics: Statistical Issues in Psychology, Education and the Social Sciences, Violence Against Women, Journal of Early Adolescence, The Journal of Negro Education, International Journal of Computer Science in Sport,* and *International Journal of Sports Medicine.* Named an AERA Fellow, he has served as a Fulbright Scholar on three different occasions; worked on numerous funded projects; and received several teaching, research, service, and book awards.