International Journal of Disability, Development and Education

ISSN: 1034-912X (Print) 1465-346X (Online) Journal homepage: https://www.tandfonline.com/loi/cijd20

Reading Recovery and Evidence-based Practice: A response to Reynolds and Wheldall ()

Robert M. Schwartz , Angela Hobsbaum , Connie Briggs & Janet Scull

To cite this article: Robert M. Schwartz , Angela Hobsbaum , Connie Briggs & Janet Scull (2009) Reading Recovery and Evidence-based Practice: A response to Reynolds and Wheldall (), International Journal of Disability, Development and Education, 56:1, 5-15, DOI: 10.1080/10349120802681564

To link to this article: https://doi.org/10.1080/10349120802681564

International Journal of Disability, Development and Education
Vol. 56, No. 1, March 2009, 5–15

Routledge
Taylor & Francis Group

Reading Recovery and Evidence-based Practice: A response to Reynolds and Wheldall (2007)

Robert M. Schwartz[a]*, Angela Hobsbaum[b], Connie Briggs[c] and Janet Scull[d]

[a]*Reading Department, School of Education, Oakland University, Rochester, Michigan, USA;* [b]*Department of Early Childhood and Primary Education, London, UK;* [c]*Reading Department, College of Professional Education, Denton, Texas, USA;* [d]*Melbourne Graduate School of Education, Melbourne, Victoria, Australia*

Reynolds and Wheldall reviewed research relating to Reading Recovery (RR) and concluded that "RR has provided an excellent model in demonstrating how to plan, promote, and implement an intervention across an educational system and how to design a professional development programme" (2007, p. 218). They balanced this praise with concerns about the research base for RR, its effectiveness for the lowest-performing first-grade students, long-term change in literacy achievement for RR students and RR's cost-effectiveness. This response aims to address these concerns by discussing four central issues of evidence-based practice from their review: evidence of effectiveness; sustained gains; programme evaluation data from a response to intervention perspective; and cost-effectiveness versus cost-benefit.

Keywords: beginning reading; cost-effectiveness; early intervention; meta-analysis; Reading Recovery; response to intervention

Evidence of Effectiveness

A critical element in judging the effectiveness of any form of instruction is whether it produces gains in student performance. Since most children make substantial gains in reading ability across first grade, control group experiments are necessary to provide evidence that an intervention produces gains beyond those expected due to factors such as maturation, regression to the mean, or classroom instruction. Well-designed experimental studies are necessary to establish the causal validity of an intervention (Shadish, Cook, & Campbell, 2002; Whitehurst, 2004). Reynolds and Wheldall suggested that:

> If RR [Reading Recovery] is to maintain its reputation as an effective programme it needs to demonstrate to the general research community that it has a strong research base. To date, despite the collection of a great deal of data, there is limited experimental research to support the efficacy of the programme. Well designed experimental research that uses randomly assigned intervention groups is needed to investigate aspects that have equivocal findings. (2007, p. 217)

In the United States, the US Department of Education has established the What Works Clearinghouse (WWC) to determine which education innovations have scientific, research-based evidence of effectiveness. Their mission is "to provide educators, policymakers, researchers and the public with a central and trusted source of scientific evidence of what works in education" (WWC, 2008). Their procedures use strict design criteria to screen available research to include only studies judged to provide the best evidence of causal

*Corresponding author. Email: rschwart@oakland.edu

ISSN 1034-912X print/ISSN 1465-346X online

DOI: 10.1080/10349120802681564
http://www.informaworld.com

validity (WWC, 2006). This best evidence synthesis identified experimental studies to evaluate the efficacy of Reading Recovery (RR).

Reynolds and Wheldall asserted that "RR has a relatively weak research base" (2007, p. 214). However, of the 153 beginning reading programmes identified by the WWC (2007a), only 24 had one or more studies meeting their evidence criteria; RR had five out of the 51 studies that met these evidence standards.

Based on these five high-quality experimental studies, the WWC concluded that RR has positive effects on students' alphabetic skills and on general reading achievement. The WWC also found potentially positive effects on measures of fluency and comprehension. RR was the only reading intervention reviewed to have one of these top two ratings in all four outcome categories (Schwartz, Askew, & Gómez-Bellengé, 2007; WWC, 2007a).

The WWC also reported an average effect size estimate in each of these outcome categories using an improvement index. "The improvement index represents the difference between the percentile rank corresponding to the intervention group mean and the percentile rank corresponding to the comparison group mean (i.e., 50th percentile) in the comparison group distribution" (WWC, 2007b). An improvement index of 34 percentile points indicates that the mean of the experimental group was one standard deviation higher than the control group mean. The improvement index has a possible range from –50 to +50. RR produced a large average improvement index in each outcome category relative to other programmes reviewed (WWC, 2007a, 2007c): +34 for alphabetics (phonemic awareness, print awareness, letter knowledge, and phonics); +46 for fluency; +14 for reading comprehension (vocabulary and reading comprehension); and +32 for general reading achievement.

The best evidence approach taken by the WWC and meta-analysis procedures can both provide useful information for policy-makers and researchers. Reynolds and Wheldall (2007) included summaries of meta-analyses by D'Agostino and Murphy (2004) and Elbaum, Vaughn, Hughes, and Moody (2000). D'Agostino and Murphy based their analysis on 36 US studies conducted between 1984 and 1995. Almost all of these studies were unpublished evaluation reports by districts implementing RR. Since most of these studies did not include a randomised control group, D'Agostino and Murphy used a combination of meta-analysis and regression procedures to estimate comparison group values and to interpret these non-experimental reports. They also conducted an analysis to determine that the results from the better-designed studies converged with their findings from the other evaluation reports. D'Agostino and Murphy's analyses indicated positive effects for RR although the effect size estimates from these analyses are not as large as those from the high-quality, experimental studies included in the WWC (2007c) report. These analyses support and extend RR's evidence for causal generalisation by providing replication of positive effects across varied setting and student populations (Shadish et al., 2002).

Elbaum et al.'s (2000) meta-analysis was more problematic. This analysis can be viewed as evidence of a strong research base since RR comparisons account for a sizable portion of their data and Elbaum et al. reported the "mean weighted effect size for RR interventions ($d = 0.66$) was significantly higher than that for the other matched interventions ($d = 0.29$)" (2000, p. 615). Still, Elbaum et al. concluded with the statement included in Reynolds and Wheldall (2007) that they did not see RR as superior to other one-to-one interventions. This concern seemed not to be based on their meta-analysis, but rather on issues related to the quality of studies included in their analysis—particularly issues related to outcome measures and questions concerning the inclusion of all participants in post-test analyses. Unlike D'Agostino and Murphy (2004), Elbaum et al. did not compare the high-quality studies in their analysis with the weaker designs to test the validity of this concern.

One of the more problematic designs included in Elbaum et al.'s analysis is the Chapman, Tunmer, and Prochnow (2001) study. This study by Chapman et al. is one of the very few published studies to show a negative effect of RR relative to the comparison group. This is a retrospective matching design; the comparison group was identified after the data collection was complete to provide a group of students matched to the RR students on the end of Year 1 (Kindergarten) assessments.

The flaw inherent in the design of this study is well documented in the research literature (Campbell & Kenny, 1999). The Institute of Education Sciences included a critique of this design category in their publication *Identifying and Implementing Educational Practice Supported by Rigorous Evidence: A User Friendly Guide* (Coalition for Evidence-Based Policy, 2003, see p. 12, item #3). When the results of matching studies contradict results from random assignment studies, researchers should be careful in interpreting the matching study.

This design flaw was particularly problematic in the context of Chapman et al.'s (2001) study. The only children available for assignment to the retrospective matched comparison group were children who performed poorly on the end of Year 1 test and who were then judged to be making adequate progress in learning to read in Year 2, and so did not receive RR later in the year. In a school system that has fully implemented RR there *is* no appropriate comparison group for a retrospective matching study.

Published meta-analyses and best-evidence reports, like those produced by the WWC (2007a), can appear objective and impartial but may rest on a number of more subjective decisions that are not apparent in the final reports. Best-evidence analyses rely on a small number of high-quality studies that meet specific selection criteria. Researchers may argue over the nature of the criteria, their application to particular studies, or the sufficiency of the evidence to support claims of effectiveness (Slavin, 1986, 2008; Slavin, Cheung, Groff, & Lake, 2008).

Meta-analyses have similar subjective elements. For example, Elbaum et al. reported that "small-group intervention … achieved outcomes comparable to those of standard one-to-one RR while serving 3–4 times the number of students per instructor" (2000, p. 615). This conclusion rests on two unpublished studies with only four first-grade students in each condition (Acalin, 1995; Evans, 1996; for a detailed discussion, see Reading Recovery Council of North America, 2002). In a best evidence review, the studies included are vetted through the peer review process using the selection criteria for best evidence. The procedures and conclusions from a published meta-analysis are subject to peer review, but reviewers may not have access to or the resources necessary to check the original sources contributing to the meta-analysis. Recognition of the difficulties in Chapman et al.'s (2001) study requires a detailed examination of the experimental design.

Sustained Gains

Reynolds and Wheldall devoted considerable attention to the issue of sustained gains in their evaluation of RR. They argued that evidence of long-term change is an area that requires well-designed experiments. Center, Wheldall, Freeman, Outhred, and McNaught (1995) provided relevant experimental evidence on this issue. Reynolds and Wheldall (2007) characterised this study as one of the few true experimental studies of RR. We agree with Reynolds and Wheldall, and disagree with the WWC's exclusion of Center et al.'s study from their evidence analysis. The WWC intervention report listed Center et al.'s study as not meeting their evidence standards with the following note: "Incomparable groups: this study was a quasi-experimental design that used achievement pretests but it did

not establish that the comparison group was comparable to the treatment group prior to the start of the intervention" (WWC, 2007c, p. 11). In the subsequent discussion we examine evidence from Center et al.'s study that is directly related to several of the major concerns raised in Reynolds and Wheldall's review.

Center et al. (1995) included randomised experimental and control groups as well as a matched comparison group of at-risk students in schools that had not yet implemented RR. At the end of the intervention period the authors reported significant and large effect sizes favouring the RR group over the random control group on six of eight reading measures, five of which would be considered independent of the RR intervention (Reynolds & Wheldall, 2007, p. 215, point #8).

For examination of sustained gains, Center et al. (1995) provided relevant experimental evidence from the testing one year after the intervention, in the middle of second grade (called medium-term maintenance). At this point, analysis comparing the RR group with the control group is no longer appropriate since the lowest students from the control group entered RR after the initial intervention period. In Center et al.'s study, the at-risk students from the schools that had not implemented RR provided the needed comparison group against which to judge the progress of the RR group one year after intervention. Center et al. reported no statistical analysis of these data.

However, Center et al. (1995) did provide sufficient information to conduct this analysis and to confirm the initial equivalence of the groups on the pre-test measures. Center et al. reported individual data on all students in the RR group and the comparison group on seven outcome measures at medium-term maintenance (1995, p. 256, Table 10 and p. 258, Table 12). This allowed us to conduct an analysis of the differences on these measures. The means and standard deviations for each measure are presented in Table 1, along with the effect size. Six of the comparisons were statistically significant ($p < 0.01$), with only the phonemic awareness analysis yielding a marginally significant p level between 0.05 and 0.01.

From Reynolds and Wheldall's (2007) perspective this is an important analysis. The one-year maintenance scores provide an experimental demonstration of sustained effects of RR through the middle of second grade. Retention of gains resulting from early intervention is an important consideration for education systems implementing an intensive early intervention like RR. Reynolds and Wheldall returned to this issue throughout their review.

Retention of gains is a difficult issue to research. A strong design would require comparison of at-risk students who received RR in first grade with a random or well-matched control group, like that presented in our analysis of Center et al.'s (1995) data. If we looked at retention of gains in fifth grade what would we learn? Certainly the control group would receive service over this time. If the groups performed at equal levels what would this indicate? Would it suggest that four years of small-group support produced similar results to a 20-week early intervention? Which would be more cost-effective? What would be the educational benefits of greatly reducing the number of struggling readers in classrooms with an early intervention? What are the system responsibilities for maintaining the gains resulting from early intervention?

Because experimental studies of long-term effects are difficult to conduct and interpret, most of the evidence reviewed by Reynolds and Wheldall (2007) consisted of evaluation studies that compared the achievement of the students served in RR with the remainder of the grade-level cohort. The larger-scale, better-conducted evaluation studies suggested that the gains due to early intervention are retained. These findings were reflected in the reports of the New South Wales Department of Education and Training (2003, 2004); Rowe's

Table 1. One-year maintenance score means and standard deviations for Reading Recovery and comparison groups from Center et al. (1995).

	Reading Recovery Group (n = 23)	Comparison group (n = 32)	Effect size
Neale Accuracy Subtest			
Mean	28.57	14.38	1.12
Standard deviation	14.08	11.49	
Clay's Book-level Test			
Mean	17.43	6.44	1.55
Standard deviation	6.86	7.23	
Burt Word Reading Test			
Mean	36.57	22.88	1.07
Standard deviation	13.59	12.24	
Passage Reading Test			
Mean	76.04	41.78	1.23
Standard deviation	25.42	29.47	
Phoneme Awareness Test			
Mean	52.13	40.56	0.64
Standard deviation	18.65	17.47	
Word Attack Skills Test			
Mean	102.52	70.63	0.77
Standard deviation	42.62	39.74	
Woodcock–Passage Comprehension Subtest			
Mean	22.30	14.38	0.98
Standard deviation	7.19	8.66	

(1995) analysis in Victoria; Douëtil (2004) in the United Kingdom; and Brown, Denton, Kelly, and Neal (1999), and Lyons and Beaver (1995) in the United States.

Evaluation data on sustained gains are difficult to interpret since there is no appropriate control group to indicate how these students would have performed in higher grades without the benefit of early intervention. D'Agostino and Murphy (2004) used regression procedures and meta-analysis to determine whether RR gains were sustained into second grade. They concluded that:

> Compared to similar needy students, discontinued students actually widened the gap from post-test to second-grade follow-up on standardised achievement tests, and they closed the gap at follow-up with regular students. At follow-up, the not-discontinued students surpassed the low achievers on standardized tests. In sum, the results seem to indicate a lasting programme effect, at least by the end of second grade, on broad reading skills. (D'Agostino & Murphy, 2004, p. 35)

Programme Evaluation Data from a Response to Intervention Perspective

Only a few high-quality studies are required to demonstrate causal validity (Shadish et al., 2002; Whitehurst, 2004). There is more than enough experimental evidence in the WWC (2007c) report to conclude that RR can make a large and significant increase in the early

literacy learning of the most at-risk group of students. Still, this is only a first step in establishing the effectiveness of an evidence-based intervention. Whitehurst pointed out that:

> A school that adopted the reading practice that had been demonstrated to be efficacious in a scientific evaluation should collect data on how children are performing in the classrooms using that practice to identify whether the program is working as deployed and to address potential problems, such as weak implementation. Together, scientific research and performance data comprise empirical evidence. (2004, p. 3)

RR implementations collect entry and exit data on every student participating in the intervention. These data are aggregated and reported at local, regional, and national levels (see the National Data Evaluation Center website for examples of these reports in the US context: http://www.ndec.us/). As indicated in the sustained effects discussion, many school districts continue to track the progress of RR students into higher grades.

From its inception in 1976, RR has functioned as a Response to Intervention (RTI) approach to the identification of children with learning disabilities related to literacy (Clay, 1987; Lose et al., 2007; McEneaney, Lose, & Schwartz, 2006; O'Connor & Simic, 2002). RR evaluation data provides evidence of effective early intervening services and an RTI-based identification process. As Reynolds and Wheldall (2007) indicated, this intensive intervention returns 70–90% of struggling readers who received a full intervention to average performance levels. The 10–30% of students who do not reach grade-level criteria to discontinue their series of lessons have received a dynamic assessment that can guide recommendations for further literacy support. It would be unusual to include these students in an analysis of retention of gains, although it would be important to track their progress in response to various forms of subsequent support (Phillips & Smith, 1997). For some of these students this support might be in the form of special education service. This is exactly the procedure that Vellutino et al. (1996) used to identify candidates for special education in the research that has fuelled the current interest in RTI.

Reynolds and Wheldall were particularly concerned that some of the lowest-performing students who enter the RR programme do not make accelerated gains (2007, p. 213, point #1). Evaluation data from RR implementations around the world confirm this observation. Accelerated gains are necessary to return the lowest students to average levels of performance. An RTI approach has two positive outcomes, one of which is identifying students who need longer-term support to overcome their literacy learning difficulties (Jones, Johnson, Schwartz, & Zalud, 2005). Some of the very lowest-performing students do make accelerated gains and reach average performance levels. These students perform well on any number of post-intervention measures including phonemic awareness tasks (Center et al., 1995; Iversen & Tunmer, 1993; Schwartz, 2005). However, a small percentage of students do not make accelerated gains. These students continue to struggle on multiple measures of reading ability. Effective early intervention provides a means to differentiate between these two sets of students.

Center et al.'s (1995) study provides a good illustration of RR's effectiveness as an RTI approach. Rather than conducting the analysis of sustained effects that we present above, Center et al. ranked all of the students in the RR and comparison group according to performance on the Neale Accuracy Subtest in the middle of the second grade. Based on these rankings and the norms for this test, they categorised each student as "still below average", "average or near average", or "fully recovered". Since 28% of the comparison group students were classified as average or fully recovered, Center et al. argued that this percentage of the RR group would have made progress without the intervention, so RR is less effective than it appears. Schwartz (2005) conducted a similar efficiency analysis over a shorter period of time and found that 14% of the control group students identified as at-risk

in the fall of first grade appeared to be making reasonable progress by the middle of first grade, reading material above text Level Six at that point.

Reynolds and Wheldall discussed this aspect of intervention efficiency in their section on experimental research (2007, pp. 207–208). Providing an intensive intervention for students who might eventually progress without this type of support is inefficient. Any early intervention programme runs this risk. Certainly these children benefit from RR and make very rapid progress when given individual support, thus shortening their series of lessons and making room for additional students to enter the programme. An alternative to early intervention is to allow many children to struggle longer in the classroom context, increasing the gap between themselves and their average peers, practicing ineffective reading strategies and experiencing greater degrees of failure. This is what an RTI approach is meant to prevent. Delaying the selection process would increase selection efficiency, but at a cost to children that many administrators, teachers, and parents might consider too high.

Selection efficiency can be increased by providing greater opportunities for literacy learning in the first year of school. This would help to identify children who are unlikely to progress without an intensive intervention. The efficiency estimates in Center et al. (1995) and Schwartz (2005) are based on children selected for intervention in the beginning of Year Two (first grade). Selection efficiency is probably higher for children who enter the intervention later in this year since these students have had longer to respond to classroom instruction.

Reynolds and Wheldall (2007) did not discuss the other critical aspect of intervention efficiency reflected in Center et al.'s (1995) data—the effectiveness of the intervention in reducing the need for long-term support services such as special education (O'Connor & Simic, 2002; Schwartz, 2005). Yet this is a central question in Reynolds and Wheldall's review. Center et al.'s data indicated for those students classified as "still below average" that 21 of the 23 students in the comparison schools had text reading levels of four or lower (1995, p. 258, Table 12) in the middle of Year 3 (second grade)! This low performance makes all of these students potential candidates for special education services. For the RR group, one student was reading at text Level Three, one at Level Seven, and the remaining six students classified as "still below average" were reading at text levels of 10 or higher. Students reading at text Level 10 in the middle of second grade are well below grade-level norms, but they are at least under way in developing a processing system for literacy, leaving just two students for more intensive special education services. The efficiency of RR in reducing the proportion of children in the "still below average" category and improving the outcomes for these children provides strong evidence to warrant sufficient funding to support interventions for all at-risk first-grade children (Every Child A Reader, 2006; O'Connor & Simic, 2002).

Reducing the percentage of children who need long-term special education services is a primary characteristic of an effective RTI programme. Clay (1987) provided a foundation for this perspective on early intervention and RTI in her discussion of learning to be learning disabled. Based on Center et al.'s (1995) data, RR lowers this percentage from 66% of the at-risk population to somewhere between 9% and 26% of that population. As a percentage of the total first-grade cohort this would be between 2% and 5% versus 13% of the cohort appearing to need special education service without effective early intervention. This result is consistent with estimates from Schwartz (2005) and Vellutino et al. (1996).

From Reynolds and Wheldall's (2007) perspective, this analysis requires cautious interpretation since it is based on RR text levels (see 2007, p. 215, What RR has Not Done Well, #8). To counter this concern we calculated the correlations among the outcome measures

reported in Center et al. (1995) for the individual scores of all three groups in the mid-second grade. Clay's book-level measure correlated over 0.85 with the four other measures of reading ability (Neale Accuracy Subtest = 0.91, Burt Word Reading = 0.87, Passage Reading = 0.86, Woodcock Passage Comprehension = 0.87). For the two sub-skill measures, the correlations were slightly lower (Phonemic Awareness = 0.63, Word Attack Skill = 0.80). These high correlations suggest that an efficiency analysis based on other independent reading measures would yield similar results.

Reynolds and Wheldall devoted a large portion of their review to summarising evaluation data from the different countries implementing RR (2007, pp. 201–207). They discussed the outcome categories in these evaluation data as if these percentages reflect causal validity in a manner similar to the experimental evidence. We view these data from a very different perspective. Evaluation data indicate that a research-based early intervention can be successfully disseminated and brought to scale across multiple settings and international boundaries. The gains demonstrated in carefully controlled experimental studies are mirrored in these large-scale evaluation studies. Variations in outcomes across settings are to be expected. These variations reflect both local and national implementation issues. Systems like New Zealand's, which provide support for full implementation, with sufficient resources to serve all children judged to need an early intervention, are able to benefit fully from an RTI approach.

However, even a partial implementation can produce impressive results, as more up-to-date data from the United Kingdom demonstrates (Reading Recovery National Network, 2003–04, 2004–05, 2005–06). Reynolds and Wheldall (2007) cited data from England dating from 1994; more recent national reports (Douëtil, 2004) show that over 60% of all children served, and 85% of those who received a complete RR programme, met the criteria for reaching average performance levels. Furthermore, in national reading assessments one year on, 74% reached average reading levels, and four years later, aged 10 years, 60% were still achieving the expected level (Douëtil, 2004). These findings persuaded the government in England to collaborate with charitable trusts and the business sector in a three-year, £10 million (approximately AU$22 million) project aimed to show that, "with the right resources, it is possible to tackle the literacy difficulties that blight many children's lives—particularly those of children who live in poverty" (Every Child A Reader, 2006, p. 3).

Similarly impressive results have been obtained in other systems that support strong implementation, such as the New South Wales and Victorian (Auditor General Victoria, 2003) educational systems in Australia, and systems in the United States and Canada that provide support for early intervention. Where funding only allows for partial coverage, perhaps sufficient resources to serve only the bottom 5% of the first-grade cohort, evaluation data show lower percentages of students returned to average performance levels and more incomplete programmes at the end of first grade.

Local evaluation data provide the information needed to improve the site's implementation. This may suggest changes that need to be made to reduce the number of missed lessons, to support professional development, to improve classroom instruction in the first years of school, or to shift resource allocations to support higher levels of implementation.

Cost-effectiveness versus Cost-benefit

Providing sufficient resources for full implementation is certainly an issue of cost. Reynolds and Wheldall (2007) summarise a number of estimates of these costs that range between AU$2350 and AU$9088. The actual cost per student is likely to lie within this

broad range. (For a further discussion of programme costs, see Schmitt, Askew, Fountas, Lyons, & Pinnell, 2005.)

Cost-effectiveness analysis requires a comparison between different approaches to achieving similar outcomes with the same population of students (Hummel-Rossi & Ashdown, 2002). Center et al.'s (1995) data could be used to calculate the cost related to raising text-reading levels for at-risk students or reducing the percentage of students needing long-term literacy support. RR was much more effective at achieving these goals than the control conditions. Information on the costs associated with each condition could support a cost-effectiveness analysis.

Other studies of RR have attempted some form of cost-benefit analysis (Lyons & Beaver, 1995; Schmitt et al., 2005). Cost-benefit analysis does not require a comparison among alternative methods of achieving a particular goal, but rather a balance sheet approach to the cost and benefits associated with a particular intervention. In the United Kingdom the KPMG Foundation commissioned a project team to carry out a cost-benefit analysis of the long-term costs of literacy difficulties (Gross, Jones, Raby, & Tolfree, 2006). This team included the director of the Every Child A Reader initiative, the director of the KMPG Foundation Educational Advisory team, a qualified accountant, and an economist. They identified the consequences of reading failure, including special support in school across grades, truancy and exclusion from school, reduced employment opportunities, increased health risks, and greater risk of involvement in the criminal justice system. Costs related to these consequences of low literacy were then calculated over time to age 37 years.

Using the data on the effectiveness of RR in reducing reading difficulties in the Every Child a Reader initiative, the team estimated the savings possible from providing early intervention services at age six years to the 38,700 pupils per year in England who currently leave primary school with very low literacy skills. The savings to the society over the next 30 years was estimated at over £1.37 billion (more than AU$3.17 billion). This was a 15:1 return on the investment in RR services. Furthermore, they say:

> These estimates are conservative. They do not include savings that could not readily be quantified, such as the economic effects of reduced spending power, social housing costs, the costs of generally poorer health, the costs of substance abuse over the age of 18 and the costs of intergenerational effects on literacy skills. (Gross et al., 2006, p. 27)

Conclusion

Reynolds and Wheldall argued that "If RR is to maintain its reputation as an effective programme it needs to demonstrate to the general research community that it has a strong research base" (2007, p. 217). We do not agree that the research and evaluation data are equivocal on the effectiveness of the RR intervention. The WWC (2007a) intervention report supports this claim for the experimental evidence. The regional and national evaluation data show variations in outcomes, but this is to be expected. Implementing a comprehensive literacy programme that includes early intervention as one of its components is challenging work. It requires the coordinated effort of administrators, tutors/teacher leaders, classroom teachers, RR teachers, and special education teachers working together to support student learning.

References

Acalin, T. A. (1995). *A comparison of Reading Recovery to Project Read.* Unpublished master's thesis. California State University, Fullerton, CA.

Auditor General Victoria. (2003). *Improving literacy standards in government schools.* Retrieved August 27, 2008 from http://www.audit.vic.gov.au/reports_par/agp90dl.html

Brown, W., Denton, E., Kelly, P., & Neal, J. (1999). Reading Recovery effectiveness: A five-year success story in San Luis Costal Unified School District. *ERS Spectrum: Journal of School Research and Information, 17,* 3–12.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Center, Y., Wheldall, K., Freeman, L., Outhred, L., & McNaught, M. (1995). An evaluation of Reading Recovery. *Reading Research Quarterly, 30,* 240–263.

Chapman, J. W., Tunmer, W. C., & Prochnow, J. E. (2001). Does success in the Reading Recovery program depend on developing proficiency in phonological-processing skills? A longitudinal study in a whole language instructional context. *Scientific Studies of Reading, 5,* 141–176.

Clay, M. M. (1987). Learning to be learning disabled. *New Zealand Journal of Educational Studies, 22,* 155–173.

Coalition for Evidence-Based Policy. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide.* US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Washington, DC.

D'Agostino, J. V., & Murphy, J. A. (2004). A meta-analysis of Reading Recovery in United States schools. *Educational Evaluation and Policy Analysis, 26,* 23–38.

Douëtil, J. (2004). *Reading Recovery: Key stage 1 effectiveness.* Retrieved August 27, 2008 from http://ioewebserver.ioe.ac.uk/ioe/cms/get.asp?cid=9263&9263_0=9268

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure: A meta-analysis of the intervention research. *Journal of Educational Psychology, 92,* 605–619.

Evans, T. L. P. (1996). *I can read deze books: A qualitative comparison of the Reading Recovery program and small group intervention.* Unpublished doctoral dissertation, Auburn University, Auburn, Alabama.

Every Child A Reader (2006). *Every child a reader: The results of the first year.* London: Every Child A Reader. Retrieved August 27, 2008 from http://www.everychildareader.org/pub/index.cfm

Gross, J., Jones, D., Raby, M., & Tolfree, T (2006). *The long-term costs of literacy difficulties.* London: KPMG Foundation.

Hummel-Rossi, B., & Ashdown, J. (2002). The state of cost-benefit and cost-effectiveness analyses in education. *Review of Educational Research, 72,* 1–30.

Iversen, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology, 85,* 112–126.

Jones, N., Johnson, C., Schwartz, R., & Zalud, G. (2005). Two positive outcomes of Reading Recovery: Exploring the interface between Reading Recovery and special education. *The Journal of Reading Recovery, 4*(3), 19–34.

Lose, M. K., Schmitt, M. C., Gómez-Bellengé, F. X., Jones, N., Honchell, B. A., & Askew, B. J. (2007). Reading Recovery and IDEA legislation: Early intervention services (EIS) and response to intervention (RTI). *The Journal of Reading Recovery, 6*(2), 42–47.

Lyons, C. A., & Beaver, J. (1995). Reducing retention and learning disability placement through Reading Recovery: An educationally sound, cost-effective choice. In R. L. Allington & S. A. Walmsley (Eds.), *No quick fix: Rethinking literacy programs in America's elementary schools* (pp. 116–136). Portsmouth, NH: Heinemann.

McEneaney, J. E., Lose, M. K., & Schwartz, R. M. (2006). A transactional perspective on reading difficulties and response to intervention. *Reading Research Quarterly, 41,* 117–128.

New South Wales Department of Education and Training. (2003). *2003 annual report, statistical compendium.* Retrieved August 27, 2008 from https://www.det.nsw.edu.au/reports_stats/annual_reports/report2003.htm

New South Wales Department of Education and Training. (2004). *2004 annual report, statistical compendium.* Retrieved August 27, 2008 from https://www.det.nsw.edu.au/reports_stats/annual_reports/report2004.htm

O'Connor, E. A., & Simic, O. (2002). The effect of Reading Recovery on special education referrals and placements. *Psychology in the Schools, 39,* 635–646.

Phillips, G., & Smith, P. (1997). *A third chance to learn: The development and evaluation of specialised interventions for young children experiencing the greatest difficulty in learning to read.* Distributed by New Zealand Council for Educational Research, Wellington, New Zealand.

Reading Recovery Council of North America. (2002). *What evidence says about Reading Recovery.* Worthington, OH: Author. Retrieved August 27, 2008 from http://www.readingrecovery.org/research/analysis/What_Evidence_Says.asp

Reading Recovery National Network. (2003–04). *Annual monitoring report, 2003–4.* Retrieved August 27, 2008 from http://ioewebserver.ioe.ac.uk/ioe/cms/get.asp?cid=9263&9263_0=7104

Reading Recovery National Network. (2004–05). *Annual monitoring reports, 2004–5.* Retrieved August 27, 2008 from http://ioewebserver.ioe.ac.uk/ioe/cms/get.asp?cid=9263&9263_0=7104

Reading Recovery National Network. (2005–06). *Annual monitoring report 2005–6.* Retrieved August 27, 2008 from http://ioewebserver.ioe.ac.uk/ioe/cms/get.asp?cid=9263&9263_0=7104

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

Rowe, K. J. (1995). Factors affecting students' progress in reading: Key findings from a longitudinal study. *Literacy, Teaching and Learning: An International Journal of Early Literacy, 1*(2), 57–110.

Schmitt, M. C., Askew, B. J., Fountas, I. C., Lyons, C. A., & Pinnell, G. S. (2005). *Changing futures: The influence of Reading Recovery in the United States.* Worthington, OH: Reading Recovery Council of North America.

Schwartz, R. M. (2005). Literacy learning of at-risk first grade students in the Reading Recovery early intervention. *Journal of Educational Psychology, 97,* 257–267.

Schwartz, R. M., Askew, B. J., & Gómez-Bellengé, F. X. (2007). What works? Reading Recovery: An analysis of the What Works Clearinghouse intervention report issued March 19, 2007. *The Journal of Reading Recovery, 6*(2), 49–52.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analyic and traditional reviews. *Educational Researcher, 15*(9), 5–11.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluation. *Educational Researcher, 37*(1), 5–14.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly, 43,* 290–322.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., et al. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88,* 601–638.

What Works Clearinghouse. (2006). *What works clearinghouse study design classification.* Retrieved 27 August, 2008, from Institute of Education Science, US Department of Education web site: http://ies.ed.gov/ncee/wwc/twp.asp

What Works Clearinghouse. (2007a). *Beginning reading: WWC topic report.* Retrieved 27 August, 2008, from Institute of Education Science, US Department of Education web site: http://ies.ed.gov/ncee/wwc/reports/beginning_reading/topic/

What Works Clearinghouse. (2007b). *Computation of the Improvement Index. Technical details of WWC-conducted computations.* Retrieved 27 August, 2008, from Institute of Education Science, US Department of Education web site: http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=9&tocId=4

What Works Clearinghouse. (2007c). *Reading Recovery intervention report.* Retrieved 19 March, 2007, from Institute of Education Science, US Department of Education web site: http://ies.ed.gov/ncee/wwc/reports/beginning_reading/reading_recovery/

What Works Clearinghouse. (2008). *Who we are.* Retrieved 27 August, 2008, from Institute of Education Science, US Department of Education web site: http://ies.ed.gov/ncee/wwc/overview/

Whitehurst, G. J. R. (2004). *IPR distinguished public policy lecture series 2003–04—Making education evidence-based: Premises, principles, and politics.* Retrieved August 27, 2008 from Institute for Policy Research, Northwestern University web site: http://ies.ed.gov/director/pdf/2004_04_26.pdf