Phonics training for English-speaking poor readers (Review)

McArthur G, Eve PM, Jones K, Banales E, Kohnen S, Anandakumar T, Larsen L, Marinus E, Wang HC, Castles A

This is a reprint of a Cochrane review, prepared and maintained by The Cochrane Collaboration and published in *The Cochrane Library* 2012, Issue 12

http://www.thecochranelibrary.com

TABLE OF CONTENTS

[Intervention Review]

Phonics training for English-speaking poor readers

Genevieve McArthur[1], Philippa M Eve[1], Kristy Jones[1], Erin Banales[1], Saskia Kohnen[1], Thushara Anandakumar[1], Linda Larsen[1], Eva Marinus[1], Hua-Chen Wang[1], Anne Castles[1]

[1]ARC Centre of Excellence of Cognition and its Disorders, Department of Cognitive Science, Macquarie University, Sydney, Australia

Contact address: Genevieve McArthur, ARC Centre of Excellence of Cognition and its Disorders, Department of Cognitive Science, Macquarie University, Sydney, NSW, 2109, Australia. genevieve.mcarthur@mq.edu.au.

Editorial group: Cochrane Developmental, Psychosocial and Learning Problems Group.
Publication status and date: New, published in Issue 12, 2012.
Review content assessed as up-to-date: 3 November 2012.

Citation: McArthur G, Eve PM, Jones K, Banales E, Kohnen S, Anandakumar T, Larsen L, Marinus E, Wang HC, Castles A. Phonics training for English-speaking poor readers. *Cochrane Database of Systematic Reviews* 2012, Issue 12. Art. No.: CD009115. DOI: 10.1002/14651858.CD009115.pub2.

ABSTRACT

Background

Around 5% of English speakers have a significant problem with learning to read words. Poor word readers are often trained to use letter-sound rules to improve their reading skills. This training is commonly called phonics. Well over 100 studies have administered some form of phonics training to poor word readers. However, there are surprisingly few systematic reviews or meta-analyses of these studies. The most well-known review was done by the National Reading Panel (Ehri 2001) 12 years ago and needs updating. The most recent review (Suggate 2010) focused solely on children and did not include unpublished studies.

Objectives

The primary aim of this review was to measure the effect that phonics training has on the literacy skills of English-speaking children, adolescents, and adults whose reading was at least one standard deviation (SD), one year, or one grade below the expected level, despite no reported problems that could explain their impaired ability to learn to read. A secondary objective was to explore the impact of various factors, such as length of training or training group size, that might moderate the effect of phonics training on poor word reading skills.

Search methods

We searched the following databases in July 2012: CENTRAL 2012 (Issue 6), MEDLINE 1948 to June week 3 2012, EMBASE 1980 to 2012 week 26, DARE 2013 (Issue 6), ERIC (1966 to current), PsycINFO (1806 to current), CINAHL (1938 to current), Science Citation Index (1970 to 29 June 2012), Social Science Citation Index (1970 to 29 June 2012), Conference Proceedings Citation Index - Science (1990 to 29 June 2012), Conference Proceedings Citation Index - Social Science & Humanities (1990 to 29 June 2012), ZETOC, Index to Theses-UK and Ireland, ClinicalTrials.gov, ICTRP, the metaRegister of Controlled Trials, ProQuest Dissertations and Theses, DART Europe E-theses Portal, Australasian Digital Theses Program, Education Research Theses, Electronic Theses Online System, Networked Digital Library of Theses and Dissertations. Theses Canada portal, www.dissertation.com, and www.thesisabstracts.com. We also contacted experts and examined the reference lists of published studies.

Selection criteria

We included studies that use randomisation, quasi-randomisation, or minimisation to allocate participants to either a phonics intervention group (phonics alone, phonics and phoneme awareness training, or phonics and irregular word reading training) or a control group (no training or alternative training, such as maths). Participants were English-speaking children, adolescents, or adults whose

word reading was below the level expected for their age for no known reason (that is, they had adequate attention and no known physical, neurological, or psychological problems).

Data collection and analysis

Two review authors independently selected studies, assessed risk of bias, and extracted data.

Main results

We found 11 studies that met the criteria for this review. They involved 736 participants. We measured the effect of phonics training on eight outcomes. The amount of evidence for each outcome varied considerably, ranging from 10 studies for word reading accuracy to one study for nonword reading fluency. The effect sizes for the outcomes were: word reading accuracy standardised mean difference (SMD) 0.47 (95% confidence interval (CI) 0.06 to 0.88; 10 studies), nonword reading accuracy SMD 0.76 (95% CI 0.25 to 1.27; eight studies), word reading fluency SMD -0.51 (95% CI -1.14 to 0.13; two studies), reading comprehension SMD 0.14 (95% CI -0.46 to 0.74; three studies), spelling SMD 0.36 (95% CI -0.27 to 1.00; two studies), letter-sound knowledge SMD 0.35 (95% CI 0.04 to 0.65; three studies), and phonological output SMD 0.38 (95% -0.04 to 0.80; four studies). There was one result in a negative direction for nonword reading fluency SMD 0.38 (95% CI -0.55 to 1.32; one study), though this was not statistically significant.

We did five subgroup analyses on two outcomes that had sufficient data (word reading accuracy and nonword reading accuracy). The efficacy of phonics training was not moderated significantly by training type (phonics alone versus phonics and phoneme awareness versus phonics and irregular word training), training intensity (less than two hours per week versus at least two hours per week), training duration (less than three months versus at least three months), training group size (one-on-one versus small group training), or training administrator (human administration versus computer administration).

Authors' conclusions

Phonics training appears to be effective for improving some reading skills. Specifically, statistically significant effects were found for nonword reading accuracy (large effect), word reading accuracy (moderate effect), and letter-sound knowledge (small-to-moderate effect). For several other outcomes, there were small or moderate effect sizes that did not reach statistical significance but may be meaningful: word reading fluency, spelling, phonological output, and reading comprehension. The effect for nonword reading fluency, which was measured in only one study, was in a negative direction, but this was not statistically significant.

Future studies of phonics training need to improve the reporting of procedures used for random sequence generation, allocation concealment, and blinding of participants, personnel, and outcome assessment.

PLAIN LANGUAGE SUMMARY

Phonics training for English-speaking poor readers

Around 5% of English speakers have a significant problem with learning to read words. Poor word readers are often trained to use letter-sound rules to improve their reading skills. This training is commonly called phonics. The primary aim of this review was to determine the effectiveness of phonics training for improving eight literacy skills in English-speaking poor word readers. A secondary objective was to explore the impact of various factors, such as training duration and training group size, that might moderate the effect of phonics training on poor word reading skills.

We found 11 studies that met the criteria for this review. These studies involved a total of 736 people. The amount of evidence for each literacy skill varied considerably, ranging from around 10 studies for word reading accuracy to just one study for nonword reading fluency.

The outcomes suggests that phonics training may be effective for improving some reading skills. Specifically, it seems to have a large effect on nonword reading accuracy, a moderate effect on word reading accuracy, and a small-to-moderate effect on letter-sound knowledge. For some outcomes (word reading fluency, spelling, phonological output, and reading comprehension), phonics training may have a small or moderate effect but it is difficult to be sure as the results found could also be due to chance. Results for nonword reading fluency, which was measured in only one study, were in a negative direction but again, this may be a chance finding.

Future studies of phonics training need to improve how they report the procedure used to put participants into groups and how they try to ensure participants do not know whether they are part of the 'experimental' group or the 'control' group. Studies should also

report clearly how they ensure those measuring children's reading progress do not know if they have been part of the phonics training group or not.

SUMMARY OF FINDINGS FOR THE MAIN COMPARISON *[Explanation]*

Phonics training compared with control (no training or alternative training) for English-speaking poor readers

Patient or population: English-speaking poor readers
Setting: English-speaking countries
Intervention: phonics
Comparison: no training or alternative training

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	No of participants (studies)	Quality of the evidence (GRADE) **	Comments *
	Assumed risk	Corresponding risk				
	No training or alternative training	English-speaking poor readers				
Word reading accuracy Immediate follow-up	The mean score in the intervention groups was on average **0.47 SD better**[1] (95% CI 0.06 to 0.88)			683 (10 studies)	High	
Nonword reading accuracy Immediate follow-up	The mean score in the intervention groups was on average **0.76 SD better**[1] (95% CI 0.25 to 1.27)			512 (8 studies)	High	
Word reading fluency Immediate follow-up	The mean score in the intervention groups was on average **0.51 SD better**[1] (95% CI -1.14 to 0.13)			54 (2 studies)	Moderate	

Nonword reading fluency Immediate follow-up	The mean score in the intervention groups was on average **0.38 SD worse**[1] (95% CI -0.55 to 1.32)		18 (1 study)	Moderate
Reading comprehension Immediate follow-up	The mean score in the intervention groups was on average **0.14 SD better**[1] (95% CI -0.46 to 0.74)		173 (3 studies)	Moderate
Spelling Immediate follow-up	The mean score in the intervention groups was on average **0.36 SD better**[1] (95% CI -0.27 to 1.00)		140 (2 studies)	Moderate

CI: confidence interval

**GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate
Very low quality: We are very uncertain about the estimate

1. Different studies used different continuous measures. Thus, comparative risk is reflected by size of the phonics training effect which was indexed with standardised mean differences (SMDs). The results are expressed as SD (standard deviation) units. As a rule of thumb, 0.2 SD represents a small difference, 0.5 a moderate difference, and 0.8 a large difference

2. Downgraded for imprecision. The confidence intervals are compatible with no effect at all as well as important improvement or deterioration

BACKGROUND

Description of the condition

Around 5% of English speakers have a significant problem with learning to read words. According to the 'self teaching hypothesis' (Share 1995) and models of word reading (Coltheart 2001; Harm 1999), reading a word depends on the ability to (1) identify each letter in the word (for example, S H I P), (2) translate each letter (I and P) or letter cluster (SH) into its correct phoneme ('sh' 'i' 'P'), and (3) blend these phonemes into a phonological output ('ship'). Reading words using these processes is sometimes called 'letter-sound reading'.

Once a word has been read via letter-sound reading, a memory of that whole written word - that is, the combination and order of the letters in the word - begins to form (for example, SHIP). This memory activates the meaning of that word (a boat), plus the memory of the phonological version of that word ('ship'), and the phonological output of that word ('ship'). Reading words using these processes is sometimes called 'sight word reading'.

Sight word reading is particularly important for reading English for two reasons. First, it is quicker than letter-sound reading. Second, one-third of written words in English contain letters that do not follow the letter-sound rules (that is, they are 'irregular'). For example, ACH in YACHT sounds like 'o' and not 'a' 'ch'. Most irregular words can be partially read with letter-sound reading since some of the letters are regular (for example, Y and T in YACHT follow the letter-sound rules 'y' and 't'). However, to be read fully, irregular words must be recognised as a whole via sight word reading.

If a person has a problem with any of the processes involved in letter-sound reading or sight word reading, then this will impair their ability to read words. For example, if a person has poor letter-sound knowledge, then they should have difficulty reading new regular words (EXPELLIARMUS) or regular names (HEDWIG) or nonwords (that is, nonsense words that follow the letter-sound rules, such as GREMP). Alternatively, if a reader has a poor store of memories for whole written words, then they should find it difficult with regular words (such as SHIP) and irregular words (such as YACHT). If a person has poor letter identification or phonological output then they should find it difficult to read regular words, irregular words, and nonwords because these processes are involved in both letter-sound reading and sight word reading.

Description of the intervention

This review focuses on the most popular reading intervention for poor word readers: phonics. Phonics teaches people to learn to read via letter-sound rules. However, not all programmes that claim to be phonics programmes focus on training letter-sound rules alone. Most programmes train numerous skills in addition to phonics, such as letter identification, sight word reading, phonological output, and reading comprehension. The results of these 'complex' phonics programmes are difficult to interpret because reading gains could stem from phonics training, non-phonics training, or an interaction between the two. Thus, the best way to test the efficacy of phonics training is to use 'pure' phonics programmes that focus training on learning to read via the letter-sound rules alone.

How the intervention might work

Since phonics programmes train letter-sound reading, then phonics training should improve performance on tests of the individual processes that are involved in letter-sound reading (such as letter identification, letter-sound knowledge, and phonological output) and tests that tax these processes simultaneously (such as nonword reading and regular word reading). Further, since improvements in letter-sound reading should increase memories of whole written words, then phonics should also improve performance on tests of the individual processes involved in sight word reading (memories of written words, the meaning of words, and phonological memories of words) as well tests that tax all these processes simultaneously (regular and irregular word reading). These gains in word reading should in turn have knock-on effects on more complex literacy skills that depend on word reading, such as reading comprehension and spelling.

The effect of phonics training on these reading skills may be influenced (or moderated) by a number of factors. One factor is the type of training. For reasons outlined above, this review considered the effects of pure phonics programmes. The most extreme version of a pure phonics programme focuses training exclusively on the letter-sound rules. Such programmes are rare, and so this review also considered phonics programmes that trained letter-sound rules plus one other ability. The most common abilities trained alongside phonics are phoneme awareness (that is, the ability to perceive, identify, discriminate, and manipulate speech sounds; see Hatcher 1994 for example) and the ability to read irregular words. Therefore, this review used subgroup analyses to compare the effects of phonics training only, phonics training plus phoneme awareness training, and phonics training plus irregular word reading training on literacy outcomes.

A second factor that may moderate the effect of phonics training is training intensity. Previous research with average readers has shown that phonics programmes that include a greater number of training sessions per week are more likely to have a greater effect than programmes with fewer sessions (Bus 1999). Although logic would dictate the same would be true for poor word readers, this has yet to be established. We tested this possibility with subgroup analyses that compared the effect on literacy outcomes of phonics programmes done up to two hours per week versus those done for two hours or more per week.

A third moderating factor on phonics training may be duration. We would predict that longer periods of phonics training would

lead to greater reading gains than shorter programmes. We used subgroup analyses to compare the effects of phonics programmes shorter than three months to those that were least three months long on literacy outcomes.

A fourth factor that may moderate the effect of phonics is training group size. Previous research has shown that one-on-one phonics training is more effective than phonics training in a group for average readers (Ehri 2001). We would expect the same to be true for poor word readers. We tested this possibility using subgroup analyses that compared the effects of phonics in studies that did one-on-one training and studies that trained small groups of poor word readers on literacy outcomes.

A fifth moderating factor of phonics training may be the training administrator. One study has reported that a reading training programme administered by a teacher is more effective than a programme administered by a computer (Dawson 2000). However, another study has found that delivering a reading programme via a computer alone is just as effective as delivering the same programme via a teacher and a computer (Torgesen 2010). In this review, we used subgroup analyses to compare the effects of phonics training administered by a human versus a computer on literacy outcomes.

Why it is important to do this review

Well over 100 studies have tested the effect of phonics training in poor readers. Yet, there are surprisingly few systematic reviews or meta-analyses on the effect of phonics training in people with poor reading. A very early review by Chall 1967 supported the use of phonics training for reading instruction, particularly for children from low socioeconomic backgrounds. However, this review did not assess the effect of specific effect of phonics in poor readers per se. The same is true for later meta-analyses by Elbaum 2000, Therrien 2004, and Swanson 1999. In contrast, the National Reading Panel (Ehri 2001) and Suggate 2010 have used meta-analyses to measure the effect of phonics programmes specifically in poor readers. The National Reading Panel's review was conducted 12 years ago and needs updating; and Suggate's review excluded unpublished studies and focused solely on children.

We are not aware of any study that has looked at the effect of phonics training on each of the processes involved in letter-sound reading and sight word reading. It would be clinically and theoretically useful to look at the effects of phonics training on letter identification, letter-sound rules, phonological output, written-word memories, semantics, and phonological memories. It would also be informative to look at the efficacy of phonics training on reading skills that depend on these processes, such as word and nonword reading accuracy and fluency, reading comprehension, and spelling.

Finally, we currently have little knowledge about the impact of moderating factors on phonics training in poor word readers. For example, we do not know how intense or long training has to be; if it should be administered individually or in a small group; or if it is best conducted by a human or a computer. Again, this information will help teachers and therapists maximise the efficacy of their phonics training programmes.

OBJECTIVES

The primary aim of this review is to determine the effectiveness of phonics training for improving reading skills in English-speaking poor word readers.

A secondary objective is to explore the impact of five potential moderating factors on phonics interventions in poor word readers: training type, training intensity, training duration, training group size, and training administrator.

METHODS

Criteria for considering studies for this review

Types of studies

This review included studies that allocated participants using random allocation (that is, uses a random component in sequence generation), minimisation (that is, minimises differences between groups for one or more factors), or quasi-randomisation (that is, uses systematic component for sequence generation, such as alternation, date of birth, case number). In terms of control data, this review included studies that had a control group that was either untrained or did alternative training (for example, maths training). It is noteworthy that an untrained control group in a reading training study is the same as a treatment-as-usual control group since most untrained participants continue to practice reading on a daily basis either at school (children) or work (adults).

Types of participants

In this review, we included studies that tested children, adolescents, and adults whose reading was at least one standard deviation (SD), one year, or one grade below the expected level, despite no reported problems that could explain their impaired ability to learn to read. Thus, children with attention deficit hyperactivity disorder (ADHD) or medical problems relating to cognition were not included in the review. This review did not exclude samples of poor word readers with a low intelligence quotient (IQ) since a discrepancy between IQ and reading is not predictive of prognosis or response to intervention (Fletcher 2005). Nor did we exclude participants based on age, gender, or socioeconomic status (SES), since poor reading is not associated with a particular age, gender,

or SES. The review was restricted to English-speaking poor word readers (as first or second language) because most other languages do not comprise the same large proportion of irregular words that cannot be read with the letter-sound rules. This results in a different ratio of words that can and cannot be read using the letter-sound rules (that is, regular versus irregular words) compared to other languages.

Types of interventions

We included studies that used a phonics programme that trained reading via the letter-sound rules alone (phonics only) or with one other type of training (that is, a programme that combined phonics with phoneme awareness or irregular word reading). In addition, we included studies that tested a control group that was either (1) untrained, (2) did alternative training (for example, maths training), (3) or did the same training as the intervention group minus the phonics component (that is, phoneme awareness training or irregular word reading training). See Table 1 for additional phonics programmes that will be included in future updates if data are available.

Types of outcome measures

We measured the effect of phonics training on six primary outcomes and two secondary outcomes (see Table 1 for outcome measures that will be included in future updates if data are available).

Primary outcomes

1. Word reading accuracy.
2. Nonword reading accuracy.
3. Word reading fluency.
4. Nonword reading fluency.
5. Reading comprehension.
6. Spelling.

Secondary outcomes

7. Letter-sound knowledge.
8. Phonological output (as measured by phoneme awareness tasks such as nonword blending).

Timing of outcome assessment

In this review's protocol, we planned four points in time for outcome assessment:

1. immediately after training;
2. one to six months after training;
3. seven to 18 months after training;
4. more than 18 months after training.

All of the studies in this review reported data for outcomes immediately after training, and so all effects are based on immediate assessment. In future updates, we will index four periods of assessment if the data allows (see Table 1). We included the six primary outcomes in the Summary of findings for the main comparison.

Search methods for identification of studies

We ran the initial searches for this review in May 2011. We used the Cochrane highly sensitive search strategy for identifying randomised trials in Ovid MEDLINE (Lefebvre 2008) and adapted this for other databases where appropriate. No date or language limits were applied. When the searches were re-run in July 2012, they were restricted to the period following the first searches either by publication year or by the date on which they were added to the database. We also adapted the original search strategies for ERIC and PsycINFO because these had previously been searched on different platforms. Search strategies for each database are reported in Appendix 1.

Electronic searches

We searched the following databases and websites.

1. The Cochrane Central Register of Controlled Trials (CENTRAL), 2012 (Issue 6), part of *The Cochrane Library*, last searched 3 July 2012
2. Ovid MEDLINE (R), 1948 to June week 3 2012, last searched 3 July 2012
3. EMBASE (Ovid), 1980 to 2012 week 26, last searched 3 July 2012
4. Database of Reviews of Abstracts of Effects (DARE) part of *The Cochrane Library*, 2012 (Issue 2), last searched 3 July 2012
5. ERIC (Proquest), 1966 to current, last searched 4 July 2012
6. ERIC (Dialog Datastar), 1966 to current, searched 31 May 2011
7. PsycINFO (Ovid), 1806 to current, last searched 4 July 2012
8. PsycINFO (EBSCOhost)), 1887 to current, last searched 31 May 2011
9. CINAHL (EBSCOhost), 1938 to current, last searched 4 July 2012
10. Science Citation Index (Web of Science), 1970 to 29 June 2012, last searched 4 July 2012
11. Social Science Citation Index (Web of Science), 1970 to 29 June 2012, last searched 4 July 2012
12. Conference Proceedings Citation Index - Science (CPCI-S),1990 to 29 June 2012, last searched 4 July 2012
13. Conference Proceedings Citation Index - Social Sciences & Humanities (CPCI-SSH), 1990 to 29 June 2012, last searched 4 July 2012
14. Zetoc, last searched 4 July 2012
15. ClinicalTrials.gov, last searched 4 July 2012
16. International Clinical Trials Registry Platform (ICTRP), last searched 4 July 2012

17. metaRegister of Clinical Trials, last searched 4 July 2012
18. Index to Theses in the UK and Ireland (ProQuest), last searched 4 July 2012
19. ProQuest Dissertations and Theses, last searched July 2012
20. DART Europe E-theses Portal, last searched July 2012
21. Australasian Digital Theses Program, last searched July 2012
22. Education Research Theses, last searched July 2012
23. Electronic Theses Online System, last searched July 2012
24. Networked Digital Library of Theses and Dissertations, last searched July 2012
25. Theses Canada portal, last searched July 2012
26. www.dissertation.com, last searched July 2012
27. www.thesisabstracts.com, last searched July 2012

Searching other resources

We examined the reference lists of published studies to identify further relevant studies. We contacted experts in the field and asked them to forward any published or unpublished studies that we may have missed.

Data collection and analysis

Selection of studies

Two review authors independently assessed each potentially relevant paper against the inclusion criteria. Any disagreement was discussed between the two review authors to determine if there was an oversight by one review author. If this did not resolve the issue, a third review author from the team (typically the first author) made the decision.

Data extraction and management

Two review authors extracted the data from each included study using a data extraction form. Data were collected on sample characteristics (including sample size); intervention characteristics (training type, training intensity, training duration, training group size, training administrator); and primary and secondary outcome measures (means, SDs, N values, and statistics). We settled disagreements between review authors via reference to the study papers and study authors. We dealt with any data missing from a study using the procedures outlined in the Dealing with missing data section. Once all data had been extracted into the forms, we entered it into the Data and analyses section. An independent author double checked the accuracy of the data entered.

Assessment of risk of bias in included studies

A 'Risk of bias' table for each study was constructed as outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2008; see Characteristics of included studies). This table was used by two review authors to describe and judge the risk of bias in each study independently. This was done by describing the method and then answering six questions (see below) to judge whether there was a low risk of bias, a high risk of bias, or an unclear risk of bias. If the two review authors provided different judgements for a domain, then they discussed their judgements. There was no unresolved disagreement between authors.

1. Random sequence generation. Description: the method used to generate the allocation sequence is described in enough detail to determine if it should produce comparable groups. Question: was the allocation sequence adequately generated?
2. Allocation sequence concealment. Description: the method used to conceal the allocation sequence is described in sufficient detail to determine if intervention allocations could have been foreseen before or during enrolment. Question: was allocation adequately concealed?
3. Blinding of participants and personnel. Description: the method used to blind participants and personnel from knowledge about which intervention a participant received, and any information relating to its effectiveness. Question: was knowledge of the allocated intervention adequately prevented?
4. Incomplete outcome data. Description: state the completeness of outcome data for each of the main outcomes, including attrition and exclusions from the analysis; state if attrition and exclusions were reported, the number in each group, reported reasons for attrition or exclusions, and re-inclusions done by the review authors. Question: was incomplete outcome data adequately assessed?
5. Selective outcome reporting. Description: state how the study was examined for potential selective outcome reporting, and state what was found. Question: were reports of the study free of suggested selective outcome reporting?
6. Other sources of bias. Description: state any concerns about bias not previously addressed. Question: was the study free of other problems that could make it at high risk of bias?

Measures of treatment effect

Continuous data

The studies reported continuous data for behavioural and drop-out measures. We calculated effect sizes using the mean difference (MD) (with 95% confidence intervals (CIs)) between the post-training means and SDs of intervention and control groups. Each outcome was measured with different tests (see Table 2 for measures used in each study) and so we calculated standardised mean differences (SMDs). We considered SMDs of around 0.2, 0.5, and 0.8 to represent small, moderate, and large effects, respectively (Cohen 1988). We considered effects with $P \leq 0.05$ to be statistically significant.

Unit of analysis issues

As no cross-over trials or cluster-randomised designs were used by studies in this review, we were able to make direct calculations of effect size for all outcomes without imputation or adjustment (see Table 1 for how future updates will address cross-over trials or cluster-randomised studies).

For the four studies that included more than one intervention group that received phonics training, we combined the post-training means, SDs, and N values of the groups (Hurford 1994; Levy 1997; Levy 1999; Savage 2003). See Characteristics of included studies for more details of these studies. See Table 1 for how future updates will combine group data if required.

Three studies (Barker 1995; Lovett 1990; Lovett 2000) tested word reading accuracy with two tests (regular word reading and irregular word reading). One study (Lovett 1990) tested word reading fluency with two tests (a regular word test and an an irregular word test) and tested spelling with two tests (regular word spelling and irregular word spelling). For each study that used two tests to measure a single outcome, we (1) used RevMan meta-anlaysis to calculate the SMDs for each test separately, (2) calculated the mean SMDs for the two tests, (3) removed the data entries for the two tests, and (4) inserted a new entry that used the mean SMD for the experimental group, 0 for the control mean, 1 for the SDs of both groups, and the N of the study. In future updates of this review, we will estimate effect sizes for regular and irregular words separately, if we have enough data to allow us to do so (see Table 1).

Dealing with missing data

If a study had missing data (for example, means, SDs, amount of training, drop-out rates), we requested that data from the corresponding author (see Characteristics of included studies for details of communications). If this request failed, we contacted the co-authors for the missing data. If a study excluded data for participants who (1) failed to complete the training, or (2) failed to adhere to the treatment programme, we asked the study authors for information about these cases. If an appeal for missing data did not leave us with a full data set, we only included data for participants whose results were known. We addressed the potential impact of any missing data in each study's 'Risk of bias' table and the Risk of bias in included studies section.

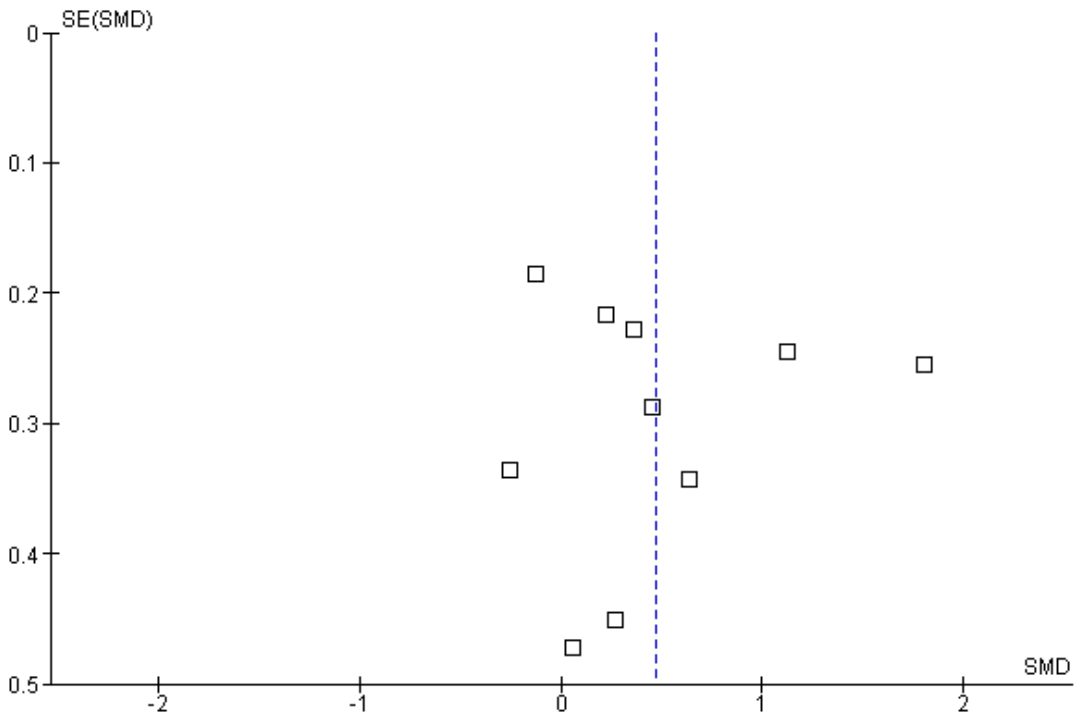
Assessment of heterogeneity

A Chi^2 test with a P value of 0.10 was used to examine the degree of consistency in the effect sizes found by the included studies (that is, heterogeneity). Further, the I^2 statistic (with a cut-off value of 70%) was used to estimate the percentage of variance in the effects owing to heterogeneity rather than chance. The I^2 value was greater than 70% for two outcomes: word reading accuracy and nonword reading accuracy (see Table 3 for a summary of the heterogeneity statistics for each outcome). For word reading accuracy, we wondered if the atypical negative effect found by Barker 1995 (SMD -0.35) and an unusually large effect found by Levy 1999 (SMD 1.80) might be the cause. For nonword reading accuracy, we wondered if it was due to an atypical negative effect found by Barker 1995 (SMD -0.50. To determine if the outlying effects should be removed from each analysis we: (1) double-checked the data, (2) reconsidered the validity and reliability of the measures, and (3) examined outlier studies to see if there was an obvious reason for the outlying result. These steps revealed no explanation for the outlying outcomes, and so we could not justify excluding these studies.

Assessment of reporting biases

Due to the small number of studies in this review, we were only able to use a funnel plot to explore reporting bias for one outcome (word reading accuracy), which had data from 10 studies (see Figure 1). The shape of the funnel plot did not suggest a bias (1) against publishing small studies with non-significant effects (in which case there would be a clear gap in the bottom left of the graph), or (2) towards publishing studies based on P values alone (in which case, the plot would have more studies at the left and right sides of the graph than in the middle; Sterne 2008). Thus, although this can be a difficult judgement with relatively few studies, publication bias did not appear to account for the heterogeneity for the word reading accuracy outcome at least.

Figure 1. Funnel plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.1 Word reading accuracy



Data synthesis

The primary aim of this review was to determine the effectiveness of phonics training for improving reading skills of English-speaking poor word readers. To this end, we merged together studies that reported outcomes for each primary outcome and secondary outcome:

1. word reading accuracy;
2. nonword reading accuracy;
3. word reading fluency;
4. nonword reading fluency;
5. reading comprehension;
6. spelling;
7. letter-sound knowledge;
8. phonological output.

As discussed above, we found heterogeneous effect sizes between studies for word reading accuracy and nonword reading acuracy. To test the impact of this heterogeneity, we calculated and compared (inverse variance) effect sizes using fixed-effect meta-analyses (which assumes the treatment effect is the same in each study) and random-effects meta-analyses (which assumes the treatment effect follows a distribution across studies; see Table 3). The results for all outcomes were similar, which suggested a degree of statistical reliability. In this review, we reported the effects from the random-effects analysis because studies did not use the same phonics training programme (and so it is likely that the treatment effect differed between studies) and because random-effects analyses adjusts estimates to incorporate heterogeneity (Deeks 2008). See Table 3 for a summary of the effect sizes for each outcome. See Table 1 for additional outcomes that we will use in future updates if data are available.

Subgroup analysis and investigation of heterogeneity

The secondary aim of this review was to explore potential moderators on the efficacy of phonics interventions. In the protocol for this review, we planned seven subgroup analyses:

1. training type (phonics alone, phonics and phoneme awareness, phonics and sight words);
2. training intensity (less than two hours per week, at least two hours per week);
3. training duration (less than three months, at least three months);
4. training group size (one-on-one, small group);

5. training administrator (human, computer);
6. poor-reading profile (phonological, surface, mixed, unknown);
7. spoken language (impaired; unimpaired; unknown).

The studies included in this review did not provide data for two planned subgroups ('poor reading profile' and 'spoken language'). These subgroups will be reinstated in future updates (see Table 1). Further, we were only able to conduct subgroup analyses in relation to two of the eight outcomes included in this review, namely word reading accuracy and nonword reading accuracy. For these outcomes we had data from 10 and eight studies, respectively. Unfortunately, lack of data meant that some analyses were not possible. First, for neither outcome were we able to perform the analysis "phonics and sight words subgroup" (subgroup 1: training type). Secondly, for nonword reading accuracy we were unable to perform the analysis "at least two hours per week" (subgroup 2: training intensity) or the analysis "at least three months" (subgroup 3: training duration).

We were unable to perform subgroup analyses for the remaining six outcomes, as data were available from fewer than five studies: word reading fluency, nonword reading fluency, reading comprehension, spelling, letter-sound knowledge, and phonological output.

Sensitivity analysis

To determine if the meta-analyses for each outcome were affected by risk of bias of included studies, we re-did the meta-analyses excluding the one study that had unclear random sequence generation (Hurford 1994). Hurford 1994 contributed data to just two outcomes: word reading accuracy and nonword reading accuracy. The SMDs for word reading accuracy with and without Hurford 1994 were 0.47 (95% CI 0.06 to 0.88; Z = 2.22; P = 0.03) and 0.47 (95% CI 0.01 to 0.93 Z = 2.00; P = 0.05), respectively. The SMDs for nonword reading accuracy with and without Hurford 1994 were 0.76 (95% CI 0.25 to 1.27 Z = 2.91; P < 0.01) and 0.38 (95% CI -0.55 to 1.32 Z = 2.71; P < 0.01), respectively. These similar outcomes suggest that the unclear random allocation for Hurford 1994 did not have undue influence on the overall outcomes.

RESULTS

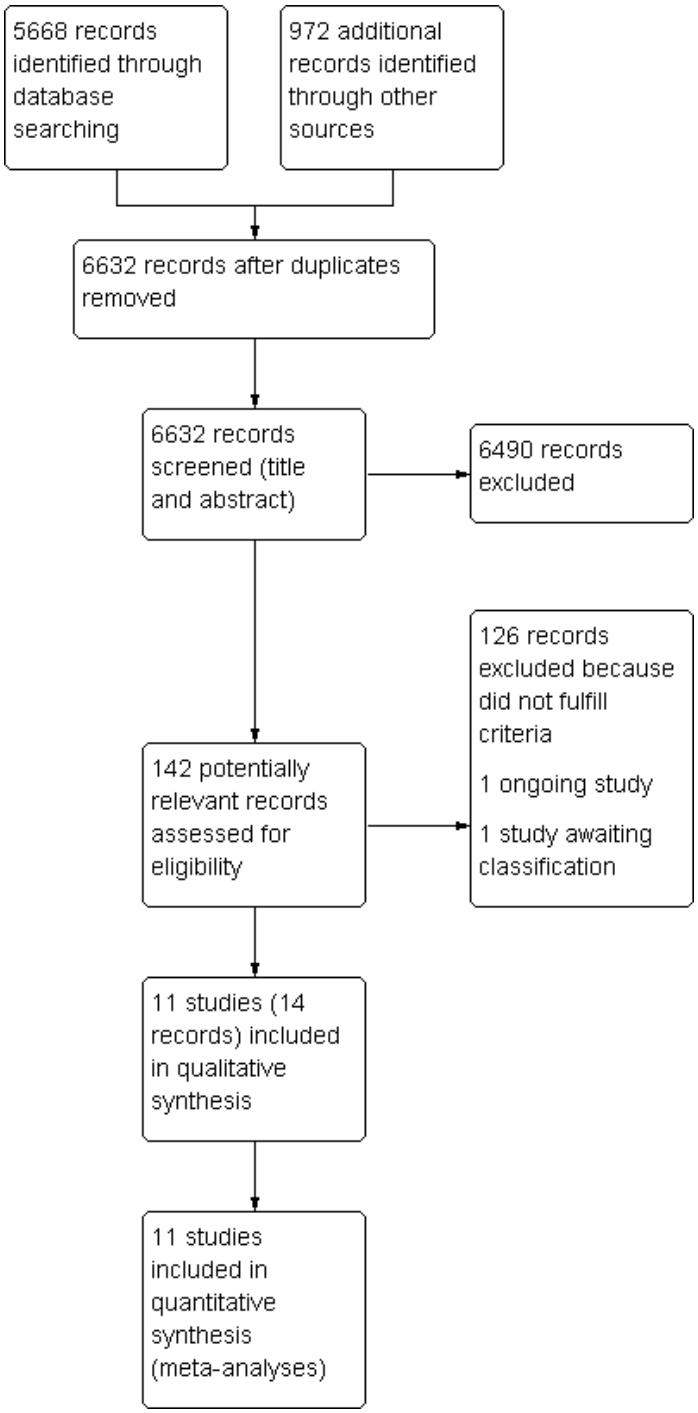
Description of studies

See: Characteristics of included studies; Characteristics of excluded studies; Characteristics of studies awaiting classification; Characteristics of ongoing studies.

Results of the search

The first search was conducted on 31 May 2011. This search was updated in July 2012 (for references published from May 2011 to July 2012). The search identified 6632 records. Examination of the titles and abstracts by two review authors independently identified 142 potential studies. One was our own ongoing study (ACTRN12608000454370) and one was an unpublished PhD thesis we were unable to obtain (see Studies awaiting classification). Examination of the full-text articles of the other articles by two review authors independently led to the rejection of 126 records. The remaining 14 records represented 11 studies. We therefore included 11 studies in the review. See Figure 2.

Figure 2. Study flow diagram



Included studies

Eleven studies, including 736 participants, met the criteria of this review (Lovett 1990; Hurford 1994; Barker 1995; Levy 1997; Levy 1999; Lovett 2000; Savage 2003; Savage 2005; Blythe 2006; Hurry 2007; Ford 2009). Subsamples from Lovett 2000 were also described in three other papers.

In terms of study design, all studies compared phonics training to a control group. In publications, all studies reported that they allocated participants using some form of randomisation, quasi-randomisation, or minimisation. However, only one study explicitly described their random sequence procedure (Ford 2009); and all studies were missing information about allocation concealment and blinding. Thus, we contacted all study authors for further information regarding risk of bias factors (see 'Risk of bias' tables for each study for more information, and 'Risk of bias in included studies below'). Combined, the information from publications and personal communication indicated that the 11 studies were controlled trials that used randomisation or minimisation. We were slightly unsure about Hurford 1994 due to the discrepancy between the published report and personal contact with the author, so we included the study but undertook sensitivity analyses to see the impact of removing it.

In terms of intervention, four of the 11 studies included more than one 'phonics only' or 'phonics and phoneme awareness' groups, in which case we merged the data of the two groups (Hurford 1994; Levy 1997; Levy 1999; Savage 2003). Six of the 11 studies included additional non-phonics training groups that were not included in the review (Lovett 1990; Barker 1995; Levy 1997; Levy 1999; Lovett 2000; Hurry 2007). See 'Characteristics of included studies' for more details.

In terms of comparison groups, all studies compared a phonics intervention group to a control group that either did no training (that is, treatment as usual; Hurford 1994; Levy 1997; Savage 2003; Savage 2005; Blythe 2006; Hurry 2007; Ford 2009) or did alternative training (Lovett 1990; Barker 1995; Levy 1999; Lovett 2000)

Location of studies

The studies were carried out in Canada (four studies), the UK (three studies), the US (three studies), and Australia (one study).

Participants

See Table 4 for details about the participants in the individual studies.

All studies reported participant details for participants who started the study rather than completed the study. However, it is noteworthy that all studies had very low or zero drop-out rates.

Reading ability

The criteria used to identify poor readers differed between studies. Most used some kind of 'cut-off' point on a reading measure (or measures), such as below the 40th, 20th, or 25th percentile (Barker 1995; Lovett 1990; Lovett 2000); a standard score less than 91 (Hurford 1994) or less than 90 (Levy 1999); or less than seven words read correctly in an experimental measure (Levy 1997). Some studies recruited the poorest readers from a large sample of screened children (Savage 2003; Savage 2005; Hurry 2007), while others recruited children if they were participating in remedial reading at school (Blythe 2006; Ford 2009; Note: data presented by these studies showed that the reading scores of these samples fell more than one SD below the level expected for their age and so the samples were known to meet the criteria for this review). Some studies also required participants to perform poorly on non-reading tests such as phoneme awareness tasks (Barker 1995; Savage 2003; Savage 2005). The large differences in inclusion criteria mean that there was a great deal of heterogeneity in the characteristics of poor readers both within and between studies.

Common exclusion criteria

Three of the 11 studies reported criteria for exclusion from the study. The most common exclusion criteria were low IQ scores (Lovett 1990; Blythe 2006); English as a second language (Lovett 1990; Lovett 2000); and history of perceptual, psychological, or neurological problems (Lovett 1990). The remaining studies did not state exclusion criteria. Thus, differences between studies in exclusionary criteria added to the heterogeneity of samples both within and between studies.

IQ

Two of the 11 studies excluded low IQ scorers from their samples (Lovett 1990; Blythe 2006). Eight studies reported the verbal, non-verbal IQ, or full IQ scores of their participants. The data suggest that most poor readers in these studies had IQ scores within or above the average range.

English speakers (first or second language)

Four of the 11 studies reported the ethnicity of their samples, which were either mixed (Levy 1999; Hurry 2007; Ford 2009), or predominantly white (Hurford 1994).

Age

Eight of the 11 studies tested children aged from five to eight years (Hurford 1994; Barker 1995; Levy 1997; Levy 1999; Savage 2003; Savage 2005; Blythe 2006; Hurry 2007). Two studies tested a slightly older and wider age group: seven to 13 years (Lovett 1990; Lovett 2000). One study tested adolescents (Ford 2009).

Gender

Seven of the 11 studies tested roughly equal numbers of girls and boys (Hurford 1994; Levy 1997; Levy 1999; Savage 2003; Savage 2005; Hurry 2007; Ford 2009). Three studies tested a larger pro-

portion of males (around 70% to 75%) than females (around 25% to 30%; Lovett 1990; Lovett 2000; Blythe 2006). One study did not report the numbers of girls and boys in the study (Barker 1995).

Socioeconomic status (SES)

Three of the 11 studies reported the SES of their sample, which was lower SES (Savage 2005; Ford 2009) or middle SES (Lovett 1990).

Interventions

Studies in this review used training programmes that differed in training type (phonics only, phonics and phoneme awareness, phonics and irregular word training); in training intensity (less than two hours per week, at least two hours per week), in training duration (less than three months, at least three months), in training group size (one-on-one or small group), and in training administrator (human, computer). These five categories correspond to the five subgroup analyses we were able to carry out. The studies that fall into each of the subgroups are summarised in Table 5 and are discussed in turn below.

Training type

Phonics only

Three of the 11 studies trained poor readers with a programme that focused on training children to read using letter-sound rules (Barker 1995; Levy 1997; Levy 1999). Barker 1995 used the Hint and Hunt programme that taught children to read with the letter-sound rules for short vowel sounds. Levy 1997 and Levy 1999 taught children to read using the letter-sound rules for rime segments in words (that is, the vowel and consonant that fall after the initial onset a word, such as r (onset) ime (rime)).

Phonics and phoneme awareness

Seven of the 11 studies trained poor readers with a programme that focused training on phoneme awareness as well as training reading with letter-sound rules (Hurford 1994; Lovett 2000; Savage 2003; Savage 2005; Blythe 2006; Hurry 2007; Ford 2009). Blythe 2006 trained phoneme awareness, letter-sound rules, speech blending, and letter blending. Ford 2009 trained phonemic awareness and decoding multi-syllabic words using letter-sound rules. Hurford 1994 trained various phoneme awareness skills (discrimination, segmentation, blending) with letters. Hurry 2007 trained various phoneme awareness skills (alliteration, rhyme, boundary sounds, vowels, digraphs (that is, two or more letters that make a single sound, such as TH), as well as using plastic letters to build words using letter-sound rules. Lovett 2000 trained various phoneme awareness skills (segmentation, blending, rhyming) and used a special orthography (highlighting salient features of some letters) to teach letter-sound rules. Savage 2003 and Savage 2005 trained children to read using the letter-sound rules for phonemes (for example, C S M) and rhymes (for example, AT (as in CAT SAT MAT), and trained phoneme awareness for phonemes and rhymes.

Training intensity

Less than two hours per week

Nine of the 11 studies trained poor readers for less than two hours per week. Most of these studies trained children between 60 and 90 minutes per week (Barker 1995; Levy 1997; Levy 1999; Savage 2003; Savage 2005; Blythe 2006). The remaining studies trained children for 15 to 45 minutes per week on average (Hurford 1994; Hurry 2007; Ford 2009).

At least two hours per week

Two of the 11 studies trained poor readers for four hours per week (Lovett 1990; Lovett 2000).

Training duration

Less than three months

Nine of the 11 studies conducted their training for less than three months (Lovett 1990; Barker 1995; Levy 1997; Levy 1999; Lovett 2000; Savage 2003; Savage 2005; Blythe 2006; Ford 2009).

At least three months

Only Hurford 1994 (five months) and Hurry 2007 (seven months) carried out training for over three months.

Training group size

One-on-one

Six of the 11 studies provided poor readers with one-on-one training by a reading professional (teachers, clinician, researcher) or computer (Hurford 1994; Levy 1997; Levy 1999; Blythe 2006; Hurry 2007; Ford 2009).

Small group

Five of the 11 studies trained poor readers in small groups comprising fewer than five trainees (Lovett 1990; Barker 1995; Lovett 2000; Savage 2003; Savage 2005).

Training administrator

Human

Seven of the 11 studies administered training primary via a human, that is, researcher, teacher, reading specialist (Lovett 1990; Levy 1997; Levy 1999; Lovett 2000; Savage 2003; Savage 2005; Hurry 2007).

Computer

Four of the 11 studies used computers as the primary training method (Hurford 1994; Barker 1995; Blythe 2006; Ford 2009).

Outcome measures

The tests used by each study to measure primary and secondary outcomes are outlined in Characteristics of included studies, summarised in Table 2, and discussed below.

Primary outcomes

Word reading accuracy

Ten of the 11 studies measured word reading accuracy. Six tests were experimental tasks designed specifically for the study that presented readers with regular or irregular words (Lovett 1990; Barker 1995; Levy 1997; Levy 1999; Lovett 2000; Savage 2003). Three tests were a version of the Word Identification from the Woodcock-Johnson Reading Mastery Test (Barker 1995; Ford 2009; Hurford 1994). One test was the Wechsler Individual Achievement Test (Blythe 2006). One test was the Word Reading test from the British Ability Scale (Hurry 2007).

Nonword reading accuracy

Eight of the 11 studies tested nonword reading accuracy. Four studies used a nonword reading test from a version of the Woodcock-Johnson Reading Mastery Test (Hurford 1994; Barker 1995; Lovett 2000; Ford 2009), three studies used experimental nonword reading tests that were developed for the study (Levy 1997; Levy 1999; Savage 2003), and one study used a nonword reading test from the Wechsler Individual Achievement Test 2nd Edition (Blythe 2006).

Word reading fluency

Two of the 11 studies measured word reading fluency. One study used the Sight Word test from the Test of Word Reading Efficiency (Ford 2009). A second study used two experimental tests of regular and irregular words that were designed specifically for the study (Lovett 1990). For the meta-anlaysis in this review, we averaged the effect sizes of these two outcomes for the same reasons, and using the same procedures, as outlined above under Word reading accuracy above.

Nonword reading fluency

One study tested nonword reading fluency used the Phonemic Decoding test from the Test of Word Reading Efficiency (Ford 2009).

Reading comprehension

Three of the 11 studies tested reading comprehension. One study used the Neale Analysis of Reading Ability (Hurry 2007), one used the Wechsler Individual Achievement Test 2nd Edition (Blythe 2006), and one used the Gates-MacGinitie Reading Test (Ford 2009).

Spelling

Two of the 11 studies tested regular or irregular word spelling (Lovett 1990; Savage 2003). Lovett 1990 tested spelling with separate regular and irregular spelling tests. For the meta-anlaysis in this review, we averaged the effect sizes of these two outcomes for the same reasons, and using the same procedures, as outlined above under Word reading accuracy above.

Secondary outcomes

Letter-sound knowledge

It is noteworthy that only three of the 11 studies tested letter-sound knowledge. This was unexpected since letter-sound knowledge is the focus of phonics training. The three studies tested letter-sound knowledge using experimental tasks designed specifically for the study (Lovett 1990; Savage 2003; Savage 2005).

Phonological output

Four of the 11 studies tested phonological output. Three tests were experimental tasks designed specifically for the study (Barker 1995; Savage 2003; Savage 2005). And one test was the Goldman Fristoe Woodcock Sound Analysis test (Lovett 2000).

Excluded studies

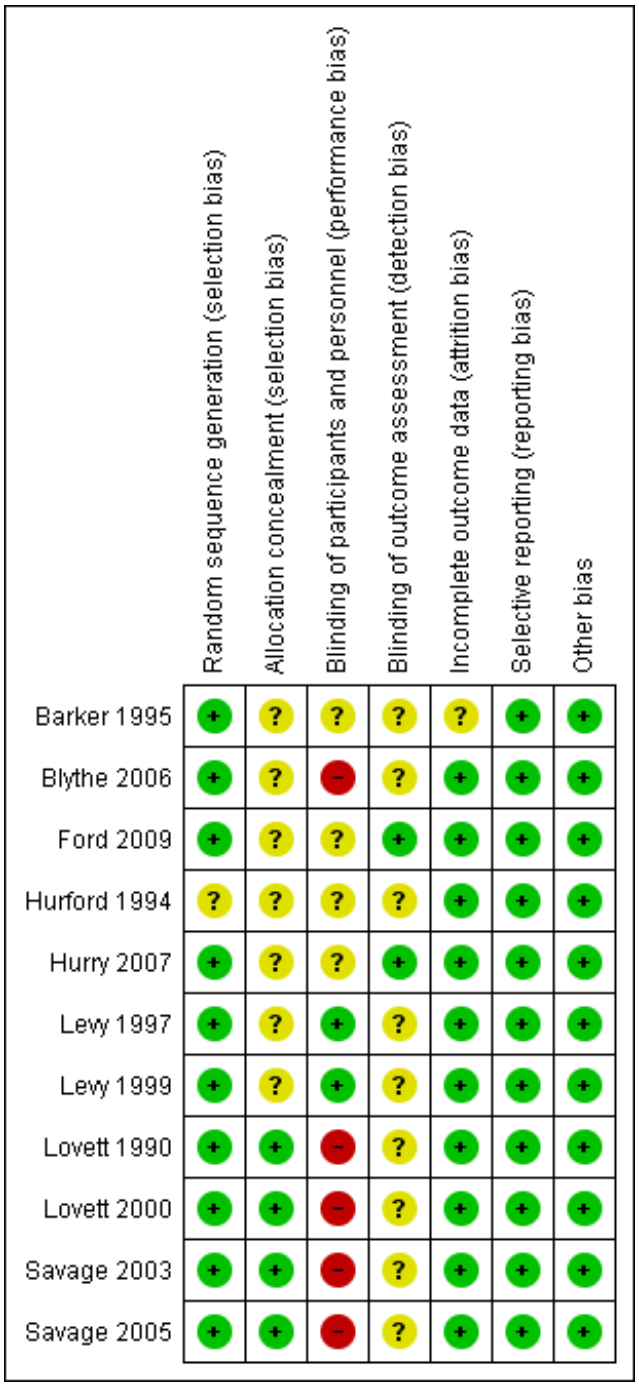
See Characteristics of excluded studies table. We listed studies that reading researchers might expect to be included in this review but were excluded because they failed to meet our review criteria. These were studies that trained the ability to read via the letter-sound rules (that is, phonics training) alongside two or more other skills such as text reading, phoneme awareness, and reading comprehension (for example, Lovett 2011; Vellutino 1986; Vellutino 1987; Lovett 1988; Lovett 1989; Lovett 1990; Hatcher 1994; Wise 1995; Vellutino 1996; Foorman 1997; Gillon 1997; Foorman Francis 1998; Wise 1997; Olson 1997; Wise 1999; Torgesen 1999; Wise 2000; Rashotte 2001; Torgesen 2001; Torgesen 2006; Hatcher 2006); studies that did not include a control group that was untrained or did non-phonics alternative training (Alexander 1991; Wise 1995; Wise 2000; Torgesen 2001; Hatcher 2006); studies that did not use randomisation, quasi-randomisation, or minimisation (Gillon 2000; Gillon 2002).

Risk of bias in included studies

Details about risk of bias for each study are shown in 'Risk of bias' tables for each study (see 'Characteristics of included studies') and are summarised in Figure 3. Information in publications initially indicated that risk of bias was unclear for almost all studies for random sequence generation, allocation concealment, and blinding of participants, personnel, and outcome assessment. Thus, we

requested information from all study authors. This information suggested low risk of bias for random sequence allocation for all studies except for Hurford 1994. There was unclear risk of bias for most studies for allocation concealment and blinding of outcome assessment. There was unclear or high risk of bias for most studies for blinding of personnel and participants. However, there was low risk of bias for almost all studies for incompleted outcomes data and selective reporting.

Figure 3. Risk of bias summary: review authors' judgements about each risk of bias item for each included study



Allocation

Random sequence generation

Regarding sequence generation, information provided in publications and from personal communication clearly indicated that all bar one study allocated participants to groups using randomisation. The exception was Hurford 1994 for which there was inconsistent information between the publication and personal communication. We therefore rated the risk of bias of this study as high.

Allocation concealment

Regarding allocation concealment, four studies used a procedure that clearly minimised minimised this risk of bias (Lovett 1990; Lovett 2000; Savage 2003; Savage 2005). It could not be ascertained from information provided by the remaining studies if allocation concealment was adequate or not (Barker 1995; Blythe 2006; Ford 2009; Hurford 1994; Hurry 2007; Levy 1997; Levy 1999).

Blinding

Participants and personnel

In terms of performance bias, it is difficult to ensure blinding of personel in cognitive treatment trials where a human administers the training because it is practically impossible to blind personnel to the treatment that they are administering. However, blinding of partipants is easier since participants in reading treatment studies (typically children) seldom have the expertise to discern which treatment or control group they have been allocated to. Thus, degree of performance bias in the current review was primarily driven by how a study tackled the blinding of personel. Four studies provided no information on this issue, and so were deemed unclear (Barker 1995; Ford 2009; Hurford 1994; Hurry 2007). Five studies stated that they did not blind personnel to the experimental and treatment groups, and so were deemed high risk (Blythe 2006; Lovett 1990; Lovett 2000; Savage 2003; Savage 2005). Two studies stated that they blinded personnel to expectations about the efficacy of the training that they were administering to minimise bias (Levy 1997; Levy 1999). These were deemed a low risk of bias.

Outcome assessment

In terms of blinding outcome assessment, two studies employed blind assessment of treatment outcomes (Hurry 2007; Ford 2009); seven studies used methods that made it unclear if outcome assessment was blind or not (Lovett 1990; Levy 1997; Levy 1999; Lovett 2000; Savage 2003; Savage 2005; Blythe 2006), and two studies did not report information about blinding of assessment bias, which again made the risk of this bias unclear (Barker 1995; Hurford 1994).

Incomplete outcome data

Four of the 11 studies indicated that there was no attrition across the study (Levy 1997; Levy 1999; Lovett 2000; Blythe 2006). Four other studies reported a minor attrition across the study (one to four data points; Lovett 1990; Savage 2003; Savage 2005; Ford 2009). Two studies reported moderate losses that appeared to be random in nature (Hurford 1994 - 13.3%; Hurry 2007 - N = 23). And one study did not provide any information about incomplete outcome data (Barker 1995). Thus, all bar one study had a low risk of bias for incomplete outcome data.

Selective reporting

There were no missing literacy tests in any studies, and so there was no evidence for selective reporting in any study.

Other potential sources of bias

None known.

Effects of interventions

See: **Summary of findings for the main comparison** Phonics training compared with control (no training or alternative training) for English-speaking poor readers

The primary aim of this review was to determine the effectiveness of phonics training for improving reading skills of English-speaking poor word readers. To this end, we calculated the effects of phonics training on six primary and two secondary outcomes. A summary of the statistics can be found in Data and analyses, Table 3, and Summary of findings for the main comparison.

A secondary objective was to explore the impact of various moderating factors on the efficacy of phonics interventions in poor word readers. Thus, for outcomes that had data from more than five studies (word reading accuracy and nonword reading accuracy) we carried out five subgroup analyses for training type (phonics alone, phonics versus phonics plus phoneme awareness versus phonics plus irregular word training), training intensity (less than

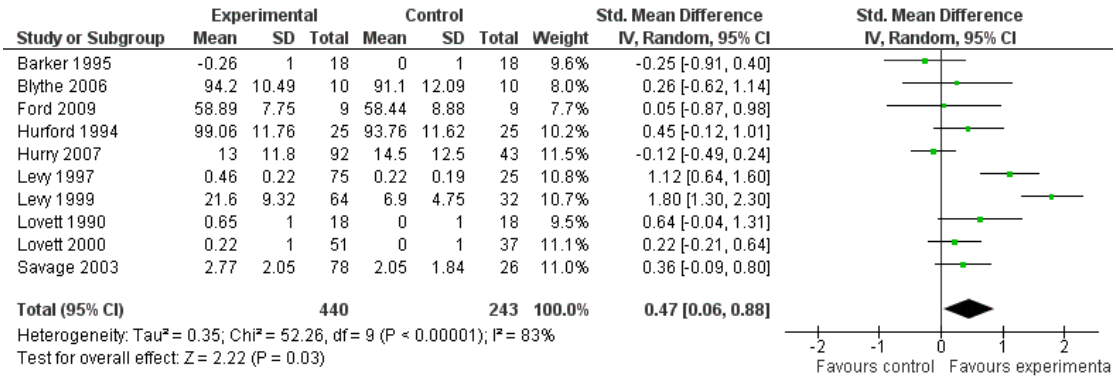
two hours per week versus at least two hours per week), training duration (less than three months versus at least three months), training group size (one-on-one versus small groups), and training administrator (human versus computer). See Table 6 and Data and analyses for a summary of the results for these subgroup analyses. The heterogeneity for two of the outcomes exceeded 70% which we addressed using the approach outlined in the Assessment of heterogeneity section above (word reading accuracy: $Chi^2 = 52.26$; DF = 9; P < 0.01; I^2 = 79%; nonword reading accuracy: Chi^2 = 44.04; DF = 7; P < 0.01; I^2 = 84%). This approach did not alter the effect calculations for the eight outcomes, which are outlined in turn below.

Primary outcomes

Word reading accuracy

Ten of the 11 studies tested the effect of phonics on word reading accuracy (see Figure 4) (Barker 1995; Blythe 2006; Ford 2009; Hurford 1994; Hurry 2007; Levy 1997; Levy 1999; Lovett 1990; Lovett 2000; Savage 2003;). Three studies (Barker 1995; Lovett 1990; Lovett 2000) used two separate regular and irregular word reading tests. We dealt with repeated measures of the same outcome using the procedure outlined above under Unit of analysis issues.

Figure 4. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.1 Word reading accuracy



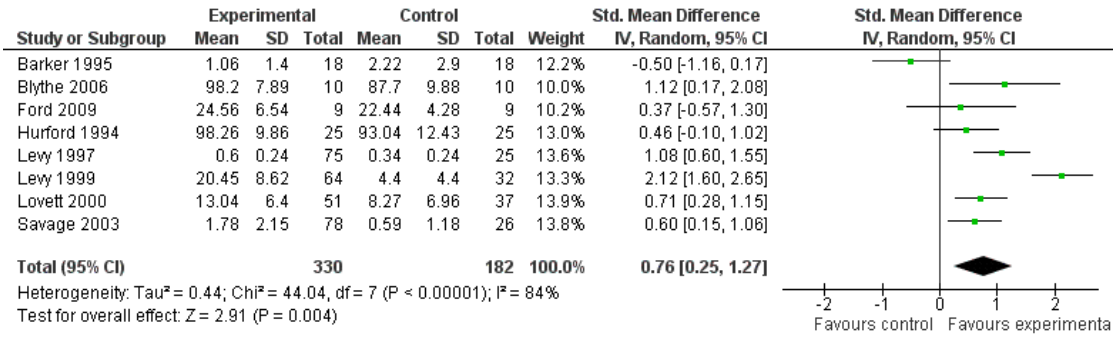
The SMD was 0.47, which was statistically significant (95% CI 0.06 to 0.88; Z = 2.22; P =0.03) (Analysis 1.1). This suggests a moderate effect of phonics training on word reading accuracy in poor word readers.

The subgroup analysis (Analysis 2.1 and Table 6) revealed no statistically significant difference between training type (Chi^2 = 1.23; DF = 1; P = 0.27; I^2 = 18.8%), training intensity (Chi^2= 0.17; DF = 1; P = 0.67; I^2 = 0%), training duration (Chi^2 = 1.36; DF = 1; P = 0.24; I^2 = 26.3%), training group size (Chi^2 = 0.94; DF = 1; P = 0.33; I^2 = 0%), or training administrator (Chi^2 = 2.13; DF = 1; P = 0.14; I^2 = 53%).

Nonword reading accuracy

Eight of the 11 studies used eight measures to test the effect of phonics on nonword reading accuracy (see Figure 5) (Hurford 1994; Barker 1995; Levy 1997; Levy 1999; Lovett 2000; Savage 2003; Blythe 2006; Ford 2009). The SMD was 0.76, which was statistically significant (95% CI 0.25 to 1.27; Z = 2.91; P < 0.01) (Analysis 1.2). This suggests a large effect of phonics training on nonword reading accuracy in poor readers.

Figure 5. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.2 Nonword reading accuracy

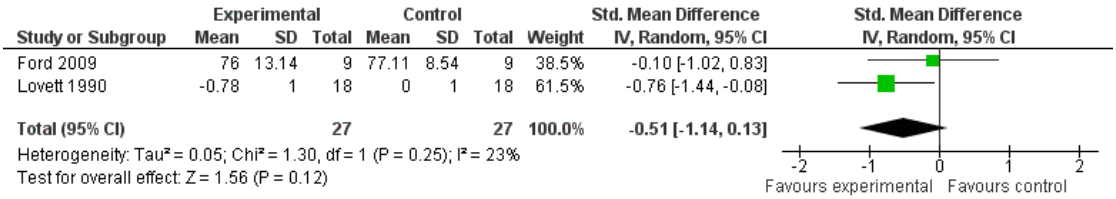


The subgroup analysis (Analysis 2.2 and Table 6) revealed no statistically significant difference between training type (Chi^2 = 0.16; DF = 1; P = 0.69; I^2 = 0%), training group size (Chi^2 = 2.43; DF = 1; P = 0.12; I^2 = 58.8%), or training administrator (Chi^2 = 3.02; DF = 1; P = 0.08; I^2 = 66.8%).

Word reading fluency

Two of the 11 studies tested the effect of phonics on word reading fluency (see Figure 6) (Lovett 1990; Ford 2009). We dealt with repeated measures of the same outcome in Lovett 1990 using the procedure outlined above under Unit of analysis issues. Since improvements in fluency are reflected by a reduction in scores (that is, less time taken to read a set number of words or text) then improvement is reflected by negative SMDs rather than positive SMDs. The SMD was -0.51, which was not statistically significant, probably due to lack of data (95% CI -1.14 to -0.13; Z = 1.26; P < 0.01) (Analysis 1.3). This suggests a moderate effect of phonics on word reading fluency in poor readers.

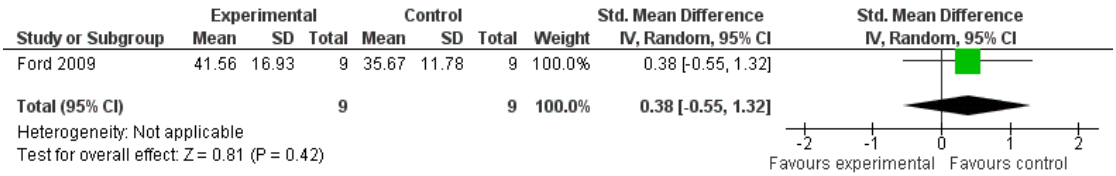
Figure 6. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.3 Word reading fluency



Nonword reading fluency

One of the 11 studies tested the effect of phonics on nonword reading fluency (see Figure 7) (Ford 2009). The SMD was 0.38 (a negative effect since an increase in score represents an increase in reading time), which was not statistically significant (95% CI -0.55 to 1.32; Z = 0.81; P = 0.42) (Analysis 1.4). This suggests a small-to-moderate negative effect of phonics on nonword reading fluency in poor readers.

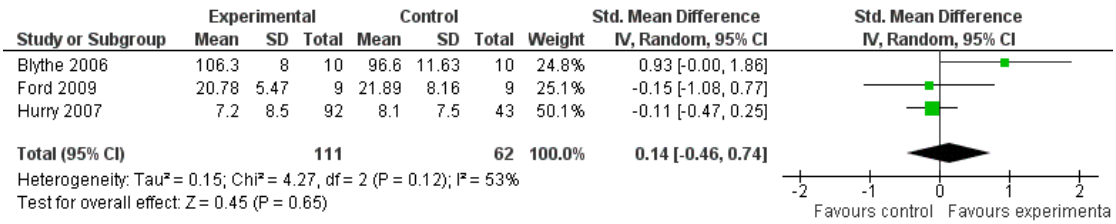
Figure 7. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.4 Nonword reading fluency



Reading comprehension

Three of the 11 studies tested the effect of phonics on reading comprehension (see Figure 8) (Blythe 2006; Hurry 2007; Ford 2009). The SMD was 0.14, which was not statistically significant (95% CI -0.46 to 0.74; Z = 0.45; P = 0.65) (Analysis 1.5). This suggests a small effect of phonics on reading comprehension in poor readers.

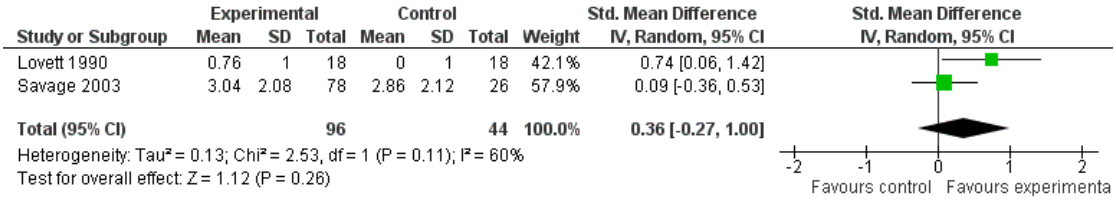
Figure 8. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.5 Reading comprehension



Spelling

Two of the 11 studies tested the effect of phonics on spelling words (see Figure 9) (Lovett 1990; Savage 2005). We dealt with repeated measures of the same outcome in Lovett 1990 using the procedure outlined above under Unit of analysis issues. The SMD was 0.36, which was not statistically significant (95% CI -0.27 to 1.00; Z = 1.12; P = 0.26) (Analysis 1.6). This suggests a small-to-moderate effect of phonics on spelling words in poor readers.

Figure 9. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.6 Spelling

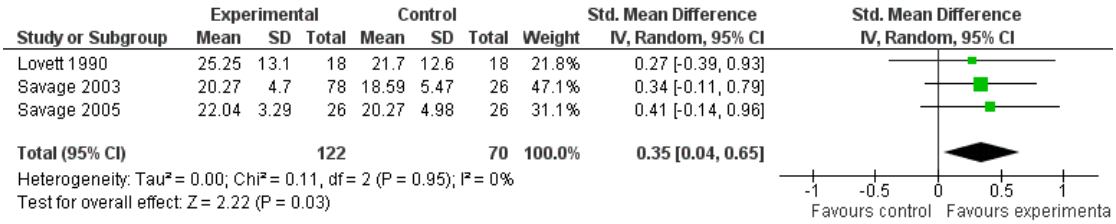


Secondary outcomes

Letter-sound knowledge

Three of the 11 studies tested the effect of phonics on letter-sound knowledge (see Figure 10) (Lovett 1990; Savage 2003; Savage 2005). The SMD was 0.35, which was statistically significant (95% CI 0.04 to 0.65; Z = 2.22; P = 0.03) (Analysis 1.7). This suggests a small-to-moderate effect of phonics on letter-sound knowledge in poor readers.

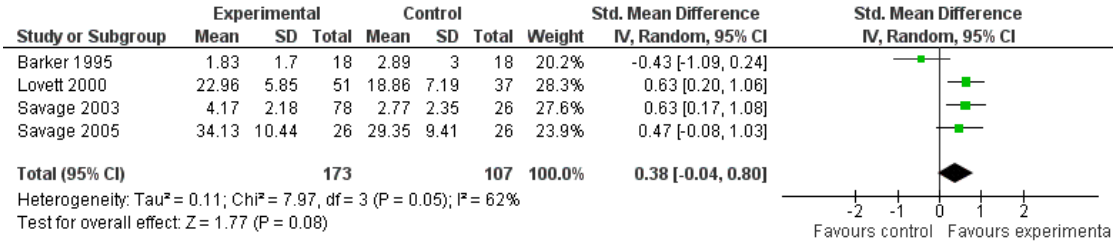
Figure 10. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.7 Letter-sound knowledge



Phonological output

Four of the 11 studies tested the effect of phonics on phonological output (see Figure 11) (Barker 1995; Lovett 2000; Savage 2003; Savage 2005). The SMD was 0.38, which was not statistically significant (95% CI -0.04 to 0.80; Z = 1.77; P = 0.08) (Analysis 1.8). This suggests a small-to-moderate effect of phonics training on phonological output in poor readers.

Figure 11. Forest plot of comparison: 1 Treatment versus control random-effects model, outcome: 1.8 Phonological output



DISCUSSION

Summary of main results

Out of 6632 records, we found 11 studies that met the criteria for this review. A meta-analysis of the data revealed that the effect of phonics training was large for nonword reading accuracy (SMD 0.76), moderate for word reading fluency (SMD -0.51) and word reading accuracy (SMD 0.47), small-to-moderate for spelling (SMD 0.36), letter-sound knowledge (SMD 0.35) and phonological output (SMD 0.38), small for reading comprehension (SMD 0.14), and slightly negative for nonword reading fluency (SMD 0.38). The effect sizes were only statistically significant for nonword reading accuracy, word reading accuracy, and letter-sound knowledge. A subgroup analysis revealed that the efficacy of phonics training was not significantly affected by training type, training intesnity, training duration, training group size, or training administrator.

Overall completeness and applicability of evidence

The outcomes of the 11 studies in this review appear applicable to English-speaking poor word readers in the general population for at least seven reasons. First, the 11 studies were published in 1990 to 2009, and so are applicable to poor word readers in modern times. Second, the proportion of studies done in each location was representative of the proportion of populations of English-speaking countries. Specifically, most studies were done in Canada (four studies), the US (three studies), and the UK (three studies), with fewer done in Australia (one study). Third, one of the debates in reading research is how poor reading (that is, 'dyslexia') should be defined. The studies in this review used a variety of inclusion and exclusion criteria to recruit poor readers, which resulted in heterogeneous samples both within and between studies. This is representative of samples in reading and dyslexia research studies. Fourth, research has established that poor reading is not restricted to a particular culture or SES. The studies in this review recruited samples with a variety of ethnicity and SES, which is representative of English-speaking poor word readers in the general population. A fifth reason why the current studies appear applicable is because they included similar numbers of males and females. There is a popular perception that more males than females are poor readers. This view has arisen from recruitment bias: people are more likely to notice poor reading in boys than girls, possibly because boys are more likely to misbehave when they are frustrated or bored. Studies minimising recruitment bias have found roughly equal proportions of male and female poor readers (Shaywitz 2001). Thus, by recruiting similar numbers of males and females, the studies included in this review represent the proportion of males and females with poor reading in the general population.

A sixth reason why the current studies are applicable to poor word readers in the general population is because of their IQ. Most poor readers in the studies in this review had IQ scores within or above the average range. This reflects the type of poor reader who gains the most attention in society (that is, those with poor reading despite average intelligence). As mentioned in the Background section, there is growing evidence that IQ is not predictive of poor reading or response to intervention. Thus, the outcomes of this review are application to poor word readers with low IQ.

And seventh, it is noteworthy that all bar one study tested children, and so the results of this review are more directly applicable to children than adults. While it is unlikely that adults with poor reading will respond differently to phonics training than children with poor reading, this issue requires much more investigation. It is also noteworthy that only three studies measured letter-sound knowledge. This is somewhat surprising given that phonics training focuses on letter-sound knowledge. Future studies should include letter-sound knowledge measures to ensure a more complete understanding of the effects of phonics on poor word readers.

Quality of the evidence

There are at least four factors that have the potential to affect the quality of evidence in this review. First, there is risk of bias. As illustrated by the 'Risk of bias' tables, nine of the 11 studies had a "low-risk" judgement for most of the seven biases assessed in this review (see Figure 2). The two remaining studies had an "unclear" judgement for most of the seven areas of potential bias considered in this review. According to GRADE criteria, quality of evidence should be judged as 'high' if "most information is from studies at low risk of bias". Quality of evidence should be downgraded to 'moderate' if "most information is from studies at low/unclear risk of bias" and if "potential limitations are likely to lower confidence in the estimate of effect" (Schunemann 2008; page 364). According to these criteria, the quality of evidence for all the outcomes measured in this review was moderate or high.

Second, there is the amount of data used to calculate effects for each outcome. While the effects for word reading accuracy and nonword reading accuracy were calculated from eight to 10 studies, the effect for phonological output was based on four studies, and the effects for word reading fluency, nonword reading fluency, reading comprehension, spelling, and letter-sound knowledge were based on data from only one to three studies.

Third, there is a possible violation of statistical independence at the student level in some studies. Specifically, five of the 11 studies in this review (Barker 1995; Lovett 1990; Lovett 2000; Savage 2003; Savage 2005) carried out treatment in small groups, rather than on an individual basis. Poor readers receiving treatment in the same group (for example, with the same instructor) may respond more similarly to each other than to poor readers in other groups. This will produce standard errors that are spuriously small and hence increase the likelihood of a Type 1 error. It is possible to adjust for this problem using intraclass correlations, if it is possible to make a reasonable estimate of the extent to which statistical independence has been violated (Hedges 2007). In future updates, when more studies are available, we will attempt to make such an estimate, and hence adjust for any potential violations of statistical independence.

Fourth, there is the chance that some training studies expose participants in a treatment group - but not a control group - to content that is included in the outcomes. While it is possible that some phonics training programmes may expose children to words, or parts of words, that may be included in the post-tests, phonics training programmes typically use a wide range of constantly changing stimuli to teach children the letter-sound "rules", rather than repeatedly using the same content (that is, specific words or nonwords). Since phonics training typically focuses on repeatedly training rules, rather than specific content, the effect of content exposure during training should be minimal in typical phonics training studies.

In sum, risk of bias, amount of data, statistical independence, and content exposure have a minimal impact on the quality of evidence in this review - particularly for outcomes based on data from a larger number of studies, that is, word reading accuracy, nonword reading accuracy, and perhaps phonological output. Future studies may result in changes to the strength of evidence in relation to these outcomes for which there are currently limited data, such as word and nonword reading fluency, reading comprehension, spelling, and letter-sound knowledge.

Potential biases in the review process

The various anayses conducted in this review suggest that potential biases in this review are minimal for six reasons. First, almost all studies had low risk of bias for random sequence generation, incomplete outcome data, and selective reporting. The majority also had low or unclear risk of bias for allocation concealment, blinding of outcome assessment, and blinding of personnel and participants. Second, excessive heterogeneity only applied to two outcomes, and an analysis of this heterogeneity reveal no systematic explanation for the variance. Third, a funnel plot of one outcome suggested no evidence of publication bias, bias introduced by using P values, or bias owing to outliers. Fourth, a comparison of effects using fixed- and random-effects analyses revealed very similar outcomes, suggesting a degree of statistical reliability. Fifth, a sensitivity analysis that only included studies with low risk of bias produced very similar results to the primary analysis. And sixth, the quality of evidence was moderate or high for all outcomes.

Agreements and disagreements with other studies or reviews

There are two previous meta-analyses that are highly relevant to this review. The National Reading Panel (Ehri 2001) found small-to-moderate effects of phonics on the reading skills of poor readers. In line with this, the current review found small-to-moderate effects on spelling, letter-sound knowledge, and phonological output. However, in addition, the current review found moderate effects of phonics training on word reading accuracy and word reading fluency, and a large effect on nonword reading accuracy.

A likely explanation for the slightly discordant results between the two studies is the different criteria used for study inclusion. In the current study, we were interested in the specific effect of phonics training. Ideally, we would have only included studies that used 'pure' phonics training programmes (that is, programmes that only taught reading via the letter-sound rules). However, prior to doing this review, we suspected pure phonics training studies might be rare. Thus, our criteria for phonics training included programmes that trained phonics alone, or trained phonics plus one other reading-related skill (irregular word, phonological output). The National Reading Panel (Ehri 2001) did not use such strict criteria, and so included many more studies that used programmes that trained at least two other reading skills in addition to phonics. As discussed above, the outcomes of such complex phonics programmes are difficult to interpret because reading gains could stem from phonics training, non-phonics training, or an interaction between the two. The fact that the current review found moder-

ate and larger effect on for some outcomes suggests that the inclusion of non-phonics training in complex phonics programmes may weaken training effects on some reading-related outcomes - perhaps because less time is dedicated to phonics training per se.

The second previous meta-analysis was conducted by Suggate 2010, who found a moderate effect size of phonics training on reading skills, pre-reading skills, and comprehension skills in children who were struggling readers. Fortunately, Suggate's criteria for phonics training were quite similar to the current study, and Suggate's criteria for struggling readers were similar to our criteria for poor word readers. This may explain why our moderate effects reflect those of Suggate, and why Suggate identified a similar number of relevant phonics training studies (13) in struggling readers in Grade 1 to 7. However, unlike the current review, Suggate 2010 focused on children and did not include unpublished studies. Thus, the slightly different outcomes of the two studies could be explained by different study sets.

AUTHORS' CONCLUSIONS

Implications for practice

The results of this review suggest that phonics training had a large effect on nonword reading accuracy, a moderate effect on word reading accuracy, word reading fluency, spelling, letter-sound knowledge, and phonological output. Preliminary evidence from just three studies suggests that phonics training may only have a small effect on reading comprehension. A small-to-moderate negative effect was found for nonword reading fluency. Only three of the results were statistically significant (for nonword reading accuracy, word reading accuracy, and letter-sound knowledge). Whether results for other outcomes were statistically significant or not may have depended on the amount of data from which they were calculated. Overall, the findings suggest that teachers and reading professionals should test poor word readers for a wide range of reading skills to determine if they have the type of poor reading that responds to phonics.

Implications for research

The outcomes of this review have at least eight implications for research. First, there is a widely held belief that phonics training is the best way to treat poor reading. Given this belief, we were surprised to find that of 6632 records, we found only 11 studies that examined the effect of a relatively pure phonics training programme in poor readers. While the outcomes of these studies generally support the belief in phonics, many more randomised controlled trials (RCTs) are needed before we can be confident about the strength and extent of the effects of phonics training per se in English-speaking poor word readers.

Second, more studies are needed to look at the effects of combining phonics training with other reading skills. At this early stage of research, it would be best to look at the effects of training phonics with just one other reading skill. As our understanding of these simple effects increases, we can start to look at the effects of training phonics with two other reading skills, and so on.

Third, as mentioned above, this review revealed that phonics training has different effects on different types of reading skills. Most of the studies in this review included measures of word reading accuracy. Only one study tested nonword reading fluency and no study tested letter identification. Further, only three studies measured letter-sound knowledge, which is surprising given that phonics training focuses on letter-sound knowledge. Future RCTs of phonics training would do well to include a more comprehensive range of reading outcomes to understand the true effects of phonics training on poor word readers.

Fourth, more research is needed to understand the effect that non-reading moderator variables - such as training type, training intensity, training duration, training group size, training administrator - have on the effectiveness of phonics training on poor reading. In this review, we attempted to address these issues via the subgroup analyses for each outcome. However, only two outcomes had enough studies to conduct these subgroup analyses. Thus, more research is needed on the effects of moderator variable on the efficacy of phonics training

Fifth, the small-to-moderate effect of phonics on phonological output, which we indexed with phoneme awareness outcome measures, was interesting because it addressed a controversial issue regarding the strong relationship between reading and phoneme awareness. There is a widespread assumption by many researchers and clinicians that poor readers have poor phoneme awareness because phoneme awareness causes poor reading. However, there is good evidence that reading ability affects phoneme awareness (Bishop 2004; Castles 2004). The current review suggests that the effect of reading ability on phoneme awareness is small-to-moderate in size.

Sixth, the 'Risk of bias' analyses in this review revealed that studies of phonics training on poor readers need to improve the reporting of their methods. While most studies in this review stated that they used randomised allocation of participants to groups, few actually described how they generated the allocation sequence or concealment in their publications, and so we had to ask for this information personally. While double-blinding is difficult to guarantee in cognitive treatment trials, few studies explained how they at least attempted to instigate double-blinding. Thus, future RCTs of phonics programmes need to explain the methods of their RCTs in more detail. The CONSORT (Consolidated Standards of Reporting Trials) 2010 guidelines may prove useful in this respect (Shultz 2010).

ACKNOWLEDGEMENTS

We are extremely and eternally grateful for the help provided by the Cochrane Developmental, Psychosocial and Learning Problems Group (CDPLPG) editors on this review. In particular, we would like to thank Laura MacDonald, Geraldine MacDonald, Margaret Anderson, and Nuala Livingstone. We would also like to thank the statistician and external reviewers of the protocol and review. And we would like to thank the authors of the studies included in this review for responding so willingly to repeated requests for information.

REFERENCES

References to studies included in this review

Barker 1995 ***{published data only (unpublished sought but not used)}***
Barker T, Torgese J. An evaluation of computer-assisted instruction in phonological awareness with below average readers. *Journal of Educational Computing Research* 1995;**13**(1):89–103.

Blythe 2006 ***{published data only (unpublished sought but not used)}***
Blythe J. Computer-based phonological skills training for primary students with mild to moderate dyslexia- a pilot study. *Australian Journal of Educational and Developmental Psychology* 2006;**6**:39–49.

Ford 2009 ***{unpublished data only}***
Ford C. *The effect of the backward-chaining method of decoding with computer-assisted instruction on the reading skills of struggling adolescent readers [thesis]*. DeKalb, Illinois: Northern Illinois University, 2009.

Hurford 1994 ***{published data only (unpublished sought but not used)}***
Hurford D, Johnston M, Nepote P, Hampton S, Moore S, Neal J, et al.Early identification and remediation of phonological processing deficits in first-grade children at risk for reading disabilities. *Journal of Learning Disabilities* 1994;**27**(10):647–59.

Hurry 2007 ***{published data only (unpublished sought but not used)}***
Hurry J, Sylva K. Long-term outcomes of early reading intervention. *Journal of Research in Reading* 2007;**30**(3):227–48.

Levy 1997 ***{published data only (unpublished sought but not used)}***
Levy B, Lysynchuk L. Beginning word recognition: benefits of training by segmentation and whole word methods. *Scientific Studies of Reading* 1997;**1**(4):359–87.

Levy 1999 ***{published data only (unpublished sought but not used)}***
Levy B, Bourassa D, Horn C. Fast and slow namers: benefits of segmentation and whole word training. *Journal of Experimental Child Psychology* 1999;**73**(2):115–38.

Lovett 1990 ***{published data only (unpublished sought but not used)}***
Lovett M, Warren-Chaplin P, Ransby M, Borden S. Training the word recognition skills of reading disabled children: treatment and transfer effect. *Journal of Educational Psychology* 1990;**82**:769–80.

Lovett 2000 ***{published data only (unpublished sought but not used)}***
*Lovett M, Steinbach K, Frijters J. Remediating the core deficits of developmental reading disability: a double-deficit perspective. *Journal of Learning Disabilities* 2000;**33**(4):334–58.
Lovett M, Borden S, DeLuca T, Lacerenza L, Benson N, Brackstone D. Treating the core deficits of developmental dyslexia: evidence of transfer of learning after phonologically- and strategy-based reading training programs. *Developmental Psychology* 1994;**30**:805–22.
Lovett M, Lacerenza L, Borden S, Frijters J, Steinbach K, De Palma M. Components of effective remediation for developmental reading disabilities: Combining phonological and strategy-based instruction to improve outcomes. *Journal of Educational Psychology* 2000;**92**:263–83.
Lovett M, Steinbach K. The effectiveness of remedial programs for reading disabled children of different ages: does the benefit decrease for older children?. *Learning Disability Quarterly* 1997;**20**:189–210.

Savage 2003 ***{published data only (unpublished sought but not used)}***
Savage R, Carless S, Stuart M. The effects of rime- and phoneme- based teaching delivered by Learning Support Assistants. *Journal of Research in Reading* 2003;**26**(3):211–33.

Savage 2005 ***{published data only (unpublished sought but not used)}***
Savage R, Carless S. Learning support assistants can deliver effective reading interventions for 'at-risk' children. *Educational Research* 2005;**47**(1):45–61.

References to studies excluded from this review

Alexander 1991 ***{published data only}***
Alexander A, Andersen H, Heilman P, Voeller K, Torgesen J. Phonological awareness training and remediation of analytic decoding deficits in a group of severe dyslexics. *Annals of Dyslexia* 1991;**41**:193–206.

Foorman 1997 ***{published data only}***
Foorman B, Francis D, Winikates D, Mehta P, Schatschneider C, Fletcher J. Early interventions for children with reading disabilities. *Scientific Studies of Reading* 1997;**1**(3):255–76.

Foorman Francis 1998 *{published data only}*
Foorman B, Francis D, Fletcher J, Schatschneider C, Mehta P. The role of instruction in learning to read: preventing reading failure in at-risk children. *Journal of Educational Psychology* 1998;**90**(1):37–55.

Gillon 1997 *{published data only}*
Gillon G, Dodd B. Enhancing the phonological processing skills of children with specific reading disability. *European Journal of Disorders of Communication* 1997;**32**(2):67–90.

Gillon 2000 *{published data only}*
Gillon, G. The efficacy of phonological awareness intervention for children with spoken language impairment. *Language, Speech, and Hearing Services in Schools* 2000;**31**: 126–41.

Gillon 2002 *{published data only}*
Gillon G. Follow-up study investigating with benefits of phonological awareness intervention for children with spoken language impairment. *International Journal of Language & Communication Disorders* 2002;**37**(4):381–400.

Hatcher 1994 *{published data only}*
Hatcher P, Hulme C, Ellis A. Ameliorating early reading failure by integrating the teaching of reading and phonological skills: the phonological linkage hypothesis. *Child Development* 1994;**65**(1):41–57.

Hatcher 2006 *{published data only}*
Hatcher P, Goetz K, Snowling M, Hulme C, Gibbs S, Smith G. Evidence for the effectiveness of the Early Literacy Support programme. *British Journal of Educational Psychology* 2006;**76**:351–67.

Lovett 1988 *{published data only}*
Lovett M, Ransby M, Barron R. Treatment subtype and word type effects in dyslexic children's response to remediation. *Brain and Language* 1988;**34**(2):328–49.

Lovett 1989 *{published data only}*
Lovett M, Ransby M, Hardwick N, Johns M, Donaldson S. Can dyslexia be treated? Treatment-specific and generalized treatment effects in dyslexic children's response to remediation. *Brain and Language* 1989;**37**(1):90–121.

Lovett 1994 *{published data only}*
Lovett M, Barron R, Forbes J, Cuksts B. Computer speech-based training of literacy skills in neurologically impaired children: a controlled evaluation. *Brain and Language* 1994; **47**(1):117–54.

Lovett 2011 *{published data only}*
Lovett M, Lacerenza L, De Palma M, Frijters J. Evaluating the efficacy of remediation for struggling readers in high school. *Journal of Learning Disabilities* 2011 Dec 29. [Epub ahead of print].

Olson 1992 *{published data only}*
Olson R, Wise B. Reading on the computer with orthographic and speech feedback - an overview of the Colorado Remediation Project. *Reading and Writing* 1992; **4**(2):107–44.

Olson 1997 *{published data only}*
Olson R, Wise B, Ring J, Johnson M. Computer-based remedial training in phoneme awareness and phonological decoding: effects on the posttraining development of word recognition. *Scientific Studies of Reading* 1997;**1**(3):235–53.

Rashotte 2001 *{published data only}*
Rashotte C, MacPhee K, Torgesen J. The effectiveness of a group reading instruction program with poor readers in multiple grades. *Learning Disability Quarterly* 2001;**24**(2): 119–34.

Torgesen 1997 *{published data only}*
Torgesen J, Wagner R, Rashotte C. Prevention and remediation of severe reading disabilities: keeping the end in mind. *Scientific Studies of Reading* 1997;**1**(3):217–34.

Torgesen 1999 *{published data only}*
Torgesen J, Wagner R, Rashotte C, Rose E, Lindamood O, Conway T, et al.Preventing reading failure in young children with phonological processing disabilities: group and individual responses to instruction. *Journal of Educational Psychology* 1999;**91**(4):579–93.

Torgesen 2001 *{published data only}*
Torgesen J, Alexander A, Wagner R, Rashotte C, Voeller K, Conway T. Intensive remedial instruction for children with severe reading disabilities: immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities* 2001;**34**(1):33–58.

Torgesen 2006 *{published data only}*
Torgesen J, Myers D, Schirm A, Stuart E, Vartivarian S, Mansfield W, et al.Closing the Reading Gap: First Year Findings from a Randomized Trial of Four Reading Interventions for Striving Readers. National Assessment of Title I: Interim Report. National Center for Education Evaluation and Regional Assistance 2006; Vol. 2:1–136.

Vellutino 1986 *{published data only}*
Vellutino F, Scanlon D. Experimental evidence for the effects of instructional bias on word identification. *Exceptional Children* 1986;**53**(2):145–55.

Vellutino 1987 *{published data only}*
Vellutino F, Scanlon D. Phonological coding, phonological awareness, and reading ability: evidence from a longitudinal and experimental study. *Merrill-Palmer Quarterly* 1987;**33** (3):321–63.

Vellutino 1996 *{published data only}*
Vellutino F, Scanlon DM, Sipay ER, Small SG, Pratt A, Chen R, et al.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* 1996;**88**(4):601–38.

Wise 1995 *{published data only}*
Wise B, Olson R. Computer-based phonological awareness and reading-instruction. *Annals of Dyslexia* 1995;**45**: 99–122.

Wise 1997 *{published data only}*
Wise B, Ring J, Sessions L, Olson R. Phonological awareness with and without articulation: a preliminary study. *Learning Disability Quarterly* 1997;**20**(3):211–25.

Wise 1999 *{published data only}*
Wise B, Ring J, Olson R. Training phonological awareness with and without explicit attention to articulation. *Journal of Experimental Child Psychology* 1999;**72**(4):271–304.

Wise 2000 *{published data only}*
Wise B, Ring J, Olson R. Individual differences in gains from computer-assisted remedial reading. *Journal of Experimental Child Psychology* 2000;**77**(3):197–235.

References to studies awaiting assessment

Pye 2008 *{unpublished data only}*
Pye RE. Predicting improvement in dyslexia: A longitudinal treatment study. Unpublished PhD thesis. University of Reading, UK. 2008.

References to ongoing studies

ACTRN12608000454370 *{unpublished data only}*
G McArthur, A Castles, S Kohnen, L Larsen, K Jones, T Anandakumar, E Banales. The effect of phonics and sight-word reading treatment on the reading skills of children with different types of reading impairment.

Additional references

Bishop 2004
Bishop D, Snowling M. Developmental dyslexia and specific language impairment: same or different?. *Psychological Bulletin* 2004;**130**:858–86.

Bus 1999
Bus A, Van IJzendoorn M. Phonological awareness and early reading: a meta-analysis of experimental training studies. *Journal of Educational Psychology* 1999;**91**(3):403–14.

Castles 2004
Castles A, Coltheart M. Is there a causal link from phonological awareness to success in learning to read?. *Cognition* 2004;**91**:77–111.

Chall 1967
Chall J S. *Learning to Read: The Great Debate*. New York: McGraw-Hill, 1967.

Cohen 1988
Cohen J. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New Jersey: Lawrence Erlbaum Associates,. *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edition. New Jersey: Lawrence Erlbaum Associates, 1988.

Coltheart 2001
Coltheart M, Rastle K, Perry C, Langdon R, Ziegler J. A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* 2001;**108**:204–56.

Dawson 2000
Dawson L, Venn M, Gunter P. The effects of teacher versus computer reading models. *Behavioral Disorders* 2000;**25**(2): 105–13.

Deeks 2008
Deeks J, Higgins J, Altman D. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins J, Green S editor(s). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, 2008: 243–96.

Ehri 2001
Ehri L, Nunes S, Stahl S, Willows D. Systematic phonics instruction helps students learn to read: evidence from the national reading panel's meta-analysis. *Review of Educational Research* 2001;**71**:393–447.

Elbaum 2000
Elbaum B, Vaughn S, Hughs M, Moody S. How effective are one-on-one tutoring programs in reading for elementary students at risk for reading failure?. *Journal of Educational Research* 2000;**92**:605–619.

Fletcher 2005
Fletcher J, Denton C, Francis D. Validity of alternative approaches for the identification of learning disabilities: operationalizing unexpected underachievement. *Journal of Learning Disabilities* 2005;**38**(6):545–52.

Harm 1999
Harm M, Seidenberg M. Reading acquisition, phonology, and dyslexia: insights from a connectionist model. *Psychological Review* 1999;**106**:491–528.

Hedges 2007
Hedges L V, Hedberg E C. Intraclass correlation values for planning group-randomized trials in education.. *Educational Evaluation and Policy Analysis* 2007;**29**:60–87.

Higgins 2008
Higgins J, Altman D. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S editor(s). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, 2008:187–242.

Higgins 2008b
Higgins J, Deeks J, Altman D. Chapter 16: Special topics in statistics. In: Higgins JPT, Green S editor(s). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, 2008:481–529.

Lefebvre 2008
Lefebvre C, Manheinmer E, Glanville J. Chapter 6: Searching for studies. *Higgins JPT, Green S (editors) Cochrane Handbook for Systematic Reviews of Interventions*. Wiley, 2008.

Schunemann 2008
Schünemann H, Oxman A, Higgins J, Vist G, Glasziou P, Guyatt G. Chapter 11: Presenting results and 'summary of findings' tables. In: Higgins J, Green S editor(s). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, 2008.

Share 1995
Share D. Phonological recoding and self-teaching: sine qua non of reading acquisition. *Cognition* 1995;**55**:151–218.

Shaywitz 2001
Shaywitz S, Shaywitz B. The neurobiology of reading and dyslexia. *Focus on Basics* 2001;**5**:11–5.

Shultz 2010
Schulz K, Altman D, Moher D. Statement: updated guidelines for reporting parallel group randomised trials. *Annals of Internal Medicine* 2010;**152**:1–7.

Sterne 2008
Sterne J, Egger M, Moher D. Chapter 10: Addressing reporting biases. In: Higgins JPT, Green S editor(s). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, 2008:298–333.

Suggate 2010
Suggate S. Why what we teach depends on when: grade and reading intervention modality moderate effect size. *Developmental Psychology* 2010;**46**:1556–79.

Swanson 1999
Swanson H, Hoskyn M, Lee C. Interventions for students with learning disabilities. *Interventions for students with learning disabilities*. New York: Guildford Press, 1999.

Therrien 2004
Therrien W. Fleuncy and comprehension gains as a result of repeated reading: a meta-analysis. *Remedial and Special Education* 2004;**25**:252–61.

Torgesen 2010
Torgeson J, Wagner R, Rachotte C, Herron J, Lindamood P. Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: outcomes from two instructional approaches. *Annals of Dyslexia* 2010;**60** (1):40–56.

* *Indicates the major publication for the study*

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Barker 1995

Methods	Randomised controlled trial Two intervention groups (1 relevant: phonics training) and 1 control group (alterative training)
Participants	Criteria: a score at or below the 40th percentile on the WJRMT Word Identification test; a score below the 50th percentile on the Sound Categorization subtest Recruits: 54 English-speaking children (ranging from 6 years, 2 months to 7 years, 8 months) from 2 elementary schools. Scored slightly below average range on Vocabulary subtest form Stanford Binet IV-Revised (mean 16.5, SD = 2.36; range = 11 to 22) Allocation: "Children were randomly assigned to one of three conditions" (p. 95). This review used the phonological decoding training group as the intervention group and the maths training group as the control group. There was also a phonological awareness control group (see notes), which was not used by this review Intervention group: N = 18 Control group: N = 18
Interventions	Intervention: phonological decoding training: Hint and Hunt I programme: "Designed to acquaint children with the basic short vowel sounds and provide practice in identifying words containing those sounds" (p. 94; phonics) Control: attentional control group: maths-oriented software programmes (Alien Addition, Math Rabbit, Math Blaster) Procedure: training took place in school psychologist's office. Groups of 3 and 4 throughout the school day. 25-minute sessions, 4 times a week (Monday to Thursday) for 8 weeks. Friday used as make-up sessions. One experimenter at each site who set up each station with appropriate programme for each student. Training done via computer. Experimenter helped with technical issues but no conceptual issues. Students rewarded with 1 sticker at end of session
Outcomes	Time of post-test: immediately after training completed Relevant measures: nonword reading accuracy (Word Analysis subtest from WJRMT), regular word reading accuracy (experimental: analogue reading: matching a spoken word to 1 of 2 printed words), regular and irregular word reading accuracy (Word Identification subtest from WJRMT) and phonological awareness (experimental: phoneme elision)
Notes	1. The phonological awareness training group used *Daisy Quest* and *Daisy's Castle. Daisy Quest* trains recognising words that rhyme; recognising words that have the same beginning, middle, and ending sounds. *Daisy's Castle* teaches these additional skills: recognising words formed from a series of phonemes presented as onset and rhyme; recognising words that can be formed from a series of separately presented phonemes; counting the number of sounds in words. These programmes do not include phonics and so were not included as an intervention in this review 2. Two measures were used to test Reading Accuracy: nonwords. We only included the Word Analysis subtest from the WJRMT as it is a published test with known reliability
Risk of bias	

Barker 1995 (*Continued*)

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Quote from publication: "children were randomly assigned to one of three conditions" Comment: No other information provided
Allocation concealment (selection bias)	Unclear risk	No information provided
Blinding of participants and personnel (performance bias) All outcomes	Unclear risk	No information provided participants or personel, however, participants unlikely to know allocation
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	No information provided
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	No information provided; allocated group sizes not reported in publication, and no response to request for information
Selective reporting (reporting bias)	Low risk	Data reported for all phonological and reading tests listed in methods; adequate detail for data to be included in analysis
Other bias	Low risk	None apparent

Blythe 2006

Methods	Randomised controlled trial One intervention group (phonics + phonological awareness) and 1 control group (untrained)
Participants	Criteria: received weekly group-based remedial reading instruction at the school and referred to the study by a support teacher Recruits: 20 English-speaking dyslexic primary students (15 male, 5 female; mean age 101.35 months, SD 17.58 months; mean FSIQ-2 100.15, SD 9.38) from a medium-sized private primary school in Western Sydney (Australia). The participants had no other co-morbid specific learning disorders. They had an average delay of 13 months on a word reading task (subtest on WIAT-II); 11 months on a reading comprehension task (subtest on WIAT-II), and 25 months on a pseudoword decoding task (subtest on WIAT-II) Allocation: random allocation Intervention group: N = 10 (mean age: 99.8 months; SD 18.94) Control group: N = 10 (mean age: 102.9 months; SD 16.98)

Blythe 2006 (*Continued*)

Interventions	Intervention: *Phonics Alive! 2: The Sound Blender* (version 1.2): 10-week training programme. "Program consists of 12 modules which systematically build skills in phoneme awareness, phoneme-grapheme correspondences, sound and letter blending and speed of processing" (P. 41) Control: students continued to receive their school-based reading instruction (both in-class and at a weekly remedial group with the support teacher) Procedure: children in the intervention group continued their school-based instruction while they did their training at home and at school on a computer. At home, each training module took approximately 15 minutes to complete. Students were instructed to repeat each module until they reached a mastery level of 90% correct. Upon mastery of a module, students had to complete review worksheets. According to parents, an average of 3.6 computer modules were attempted per child per week. "Thus, over the 10-week training period, students completed an average of 46 module attempts which represented approximately 11.5 hours of on-computer time" (in addition to 30 minutes per week with researcher: 5 hours). At school, children did "a weekly, 30 minute, one-on-one session with the researcher where the student's progress was assessed by reviewing their progress chart and completed worksheets (5 minutes) and completing the current module on a computer (to verify mastery)." Any remaining time was spent playing a "nonsense word game"
Outcomes	Time of post-test: immediately after training completed Relevant measures: nonword reading accuracy (WIAT-II: Pseudoword Decoding subtest) , regular and irregular word reading accuracy (WIAT-II: Word Reading subtest) and reading comprehension (WIAT-II subtest)
Notes	Contacted author for post-test standard deviations 14/09/2011 (supplied)

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote from publication: "participants were randomly assigned to either a control or treatment condition" Quote from personal communication: "participants who met selection criteria were randomly assigned to either the Tx or Ct condition by drawing eligible names from a hat and placing sequentially into Tx/Ct until all were assigned"
Allocation concealment (selection bias)	Unclear risk	No information provided
Blinding of participants and personnel (performance bias) All outcomes	High risk	Quote from personal communication: "given this was a simple Tx/Ct design there was no way to blind study participants or personnel from knowledge of who was in the treatment group. However, the trainer

Blythe 2006 (*Continued*)

		had no previous knowledge of or awareness of the participants and was not involved in the referral process (they were referred by the school counsellor)" Comment: it is difficult to ensure double blinding in cognitive treatment trials where a human administers the training. It is practically impossible to blind expert personnel to the treatment that they are administering. However, it is unlikely that the participants have the expertise to discern which treatment of control group they have been allocated to
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Quote from personal communication: "Initial assessment of IQ and Reading was conducted by the investigator on all participants PRIOR to their random assignment to Tx or Ct conditions, and thus the assessor was unaware of their future status in the study. Given this was a pilot study, post-treatment assessment was conducted by the same assessor on all students and this precluded the assessor conducting blind post-treatment assessments"
Incomplete outcome data (attrition bias) All outcomes	Low risk	Comment: degrees-of-freedom values indicate that data for all randomised participants were included in the analyses. Author sent post-test SDs
Selective reporting (reporting bias)	Low risk	Data reported for all reading tests listed in methods; adequate detail for data to be included in analysis
Other bias	Low risk	None apparent

Ford 2009

Methods	Randomised controlled trial One intervention group (phonics + phoneme awareness) and 1 control group (untrained)
Participants	Criteria: enrolled in a remedial reading programme Recruits: 20 English-speaking participants from American alternative high school in Illinois. Most participants were bilingual. Mostly Title 1 (lower SES) Allocation: "students were randomly assigned to an experimental or control group by drawing names" (p. 49) Intervention group: N = 9 (5 females, 4 males); mean age 16 years, 2 months; 3 African-American, 5 Hispanic, 1 white); average standard score on TOWRE sight-words: 85

	(within average) and TOWRE phonemic decoding: 83 (below average) Control group: N = 9 (5 females, 4 males); mean age 16 years, 1.5 months; 1 African-American, 7 Hispanic, 1 white; average standard score on TOWRE sight-words: 85 (within average) and TOWRE phonemic decoding: 81 (below average)
Interventions	Intervention: practice in phonemic awareness and decoding multi-syllable words using backwards chaining, followed by practice on Word Workout computer program (practice skills learned in teacher-instructed sessions; phonics + phonological awareness) Control: not explicitly stated; however, probably treatment and schooling as usual throughout the training period (since all participants came from a remedial reading programme and participants were divided into experimental and control groups via random drawing (p. 48)) Procedure: training was conducted by the researcher with small groups or one-on-one. For 7 weeks, children did three 15-minute sessions per week
Outcomes	Time of post-test: immediately after training completed Relevant measures: nonword reading accuracy (WJTA-III: Word Attack subtest; forms A and B), regular and irregular word reading accuracy (WJTA-III: Letter Word Identification subtest; forms A and B), nonword reading fluency (TOWRE: Phonemic Decoding Efficiency subtest (forms A and B)), regular and irregular word reading fluency (TOWRE: Sight Word Efficiency subtest (forms A and B)) and reading comprehension (Gates-MacGinitie Reading Comprehension Subtest)
Notes	1. TOWRE sight-words and TOWRE phonemic decoding standard scores calculated using raw scores given (see pp. 771-772) 2. Two participants dropped out (1 from each group) and thus their pre-test scores were removed. The thesis only provides information on the 18 participants who completed the training 3. Qualitative data (survey and focus groups) also collected

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote from publication: "students were randomly assigned to an experimental or control group by drawing names"
Allocation concealment (selection bias)	Unclear risk	No information provided
Blinding of participants and personnel (performance bias) All outcomes	Unclear risk	No information provided for participants or personel, however, it is unlikely participants were aware of allocation
Blinding of outcome assessment (detection bias) All outcomes	Low risk	Quote from publication: "Tests were administered by an experienced teacher who was not otherwise involved with the study ... to reduce tester bias"

Incomplete outcome data (attrition bias) All outcomes	Low risk	Quote from publication: "One student from each group dropped out before the conflusion" Comment: Both groups experienced the same (low) drop-out rate
Selective reporting (reporting bias)	Low risk	Data reported for all reading tests listed in methods; adequate detail for data to be included in analysis
Other bias	Low risk	No other apparent bias.

Hurford 1994

Methods	Randomised controlled trial (most likely) Two intervention groups (2 trained groups; phonics + phoneme awareness) and 2 control groups (untrained)
Participants	Criteria: a standard score less than 91 on the Word Attack test of the WRMT-R; a standard score less than 91 on the Word Identification subtests of the WRMT-R Recruits: 99 children identified as at risk for reading difficulties with normal IQ (RD) or at risk for becoming poor readers with low IQ ('garden variety' (GV) poor readers). Children came from US schools, mostly middle-class elementary schools. 92.8% were white, 6% African-American, 5% Hispanic, and 7% Asian-American. 45.9% females, 54.1% males Allocation: 'Half of the RDs and half of the GVs were included in the training group and the other half comprised the control groups. So, there were four groups (RD Trained, GV Trained, RD Control and GV Control). Group membership was determined by matching the students at risk for RD on the variables outlined in the method section and then RANDOMLY assigning them to either the T or C group. Statistical analysis was performed to determine that the T and C groups were equivalent' (personal communication in email). Since this review did not use IQ as an exclusionary criteria, we merged the 2 trained groups (RD and GC) to form the Intervention group, and merged the two untrained groups (RD and GV) to form the Control group Intervention group: N = 49; 25 females and 24 males; mean age 79.8 months Control group: N = 50; 26 females and 24 males; mean age 80.9 months
Interventions	Intervention: intrasyllable discrimination training, phonemic blending and phonemic segmentation with letters (phonics + phonological awareness). The training sequence was the same for each participant Control: no training Procedure: intervention was one-on-one, 15 to 20 minutes/session. Approximately 40 sessions - twice a week for approximately 20 weeks by computer and trainer
Outcomes	Time of post-test: less than 1 month after training completed Relevant measures: nonword reading accuracy (WJRMT-R: Word Attack subtest) and regular and irregular word reading accuracy (WJRMT-R: Word Identification subtest)

Notes	1. The study also included 332 non-disabled children; however, we have excluded them from our review 2. Drop-outs for 486 participants initially screened: 55 (13.3%), "this loss in the participant pool due to attrition (13.3%) is similar to the attrition rate these school systems typically experience" (p. 649) 3. We used the word attack and word identification measures. Since we are including all poor readers regardless of IQ, we took the mean of the 2 untrained groups (RD and GV) for control data and the 2 trained groups (RD and GV) for experimental data 4. Contacted Hurford (20/09/2011) for means and SDs for primary outcomes (discrimination, segmentation, word identification and word attack measures) at pre and post test (supplied)

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Unclear risk	Quote from publication: 'at-risk training and control groups were then matched as closely as possible on age, gender, race, reading ability, phonological processing ability, and intelligence' Quote from personal communication: 'Half of the RDs and half of the GVs were included in the training group and the other half comprised the control groups. So, there were four groups (RD Trained, GV Trained, RD Control and GV Control). Group membership was determined by matching the students at risk for RD on the variables outlined in the method section and then randomly assigning them to either the T or C group. Statistical analysis was performed to determine that the T and C groups were equivalent' Comment: Given the different information in the publication (groups were matched) and personal communication (groups were randomised to ensure matching), random sequence generation was deemed unclear
Allocation concealment (selection bias)	Unclear risk	No information provided
Blinding of participants and personnel (performance bias) All outcomes	Unclear risk	No information provided
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Personal communication: Testing was done by someone who did not know the students

Incomplete outcome data (attrition bias) All outcomes	Low risk	Quote from publication: "three groups lost approximately same percentage [13.3%] of participants" Comment: All groups experienced the same (relatively low) drop-out rate
Selective reporting (reporting bias)	Low risk	Comment: Data reported for all reading tests listed in methods; adequate detail for data to be included in analysis
Other bias	Low risk	Comment: None apparent.

Hurry 2007

Methods	Randomised controlled trial Two intervention groups (1 relevant: phonics + phoneme awareness) and 2 control groups (1 relevant; untrained)
Participants	Criteria: 1 of 7 poorest Year 2 scorers in 18 schools on the Diagnostic Survey (Clay, 1985) Recruits: 142 children in schools where 61% were boys, 42% received free school meals, and 16% spoke English as a second language. One child was excluded from the study because of missing baseline data. All children had IQ in the average range (92 to 96) Allocation: random allocation (within schools) of poor readers to intervention and control groups Intervention group: N = 96 Control group: N = 46
Interventions	Intervention: Phonological Training: "Following Bradley and Bryant (1985), this involved sound awareness training plus word building with plastic letters. The training initially focused on alliteration and rhyme but also included work on boundary sounds and vowels and digraphs in response to the child's progress. Children also matched sounds with plastic letters and constructed words" (p. 234; phonics + phonological awareness) Control: children in within-school control groups received standard provision available in school. Since these children were poor readers, they received around 21 minutes extra help per week with reading Procedure: intervention was 40 sessions (10 minutes each, one-on-one with tutor, spread over 7 months). Five tutors who delivered phonological training. Did not share details of intervention with classroom teachers
Outcomes	Time of post-test: immediately after training completed Relevant measures: regular and irregular word reading accuracy (BAS word reading) and reading comprehension (Neale Prose Reading)
Notes	1. For the phonological training group, the article reported that the 6 poorest readers from 23 schools were allocated to either the phonological training (N = 4) or the within-school control (N = 2). This would equate to 92 participants in the phonological training group.

	However Table 1 (p. 232) reports that there were 96 participants in the phonological training group. We contacted Jane Hurry to explain this. We received a reply on the 16/01/2012: "I have now looked at the file and find that of the 23 Phon schools we actually selected the bottom 7 children from 5 of the schools. Of those 5 extra children, there was missing baseline data for 1, so that child never made it into the study. The other extra 4 were assigned to the intervention, hence the 96" 2. We have excluded the 22 Reading Recovery schools (and controls) from our analysis since it involved text reading (an exclusion criterion of our review) 3. We have excluded the 18 untrained control schools since the within-school controls are superior controls for the trained children because they were better matched for SES and learning environment 4. Contacted Hurry on which subtests were used from the Neale Prose Reading. Replied that they used the accuracy and comprehension subtests to make up their Neale Prose Reading measure (see Table 2). We have used this as a measure of reading comprehension 5. There were 3 post-tests: post-test 1 (after completion), post-test 2 (1 year later), post-test 3 (3.5 years later). We included the first post-test results in this review since all other studies in this review reported immediate post-test data 6. Contacted Hurry for clarification on: (1) participant numbers (Hurry, 16/01/2012; see above); (2) attrition (Hurry, 17/01/2012; see response in 'Risk of bias' table below), (3) which subtest of the Neale (Prose) Reading was used: Neale accuracy and comprehension scores (Hurry, 03/02/2012), (4) approximately how many minutes/hours the participants spent on phonological training per week (Hurry's response 16/02/2012: "I confirm that each child was given 40 x 10 min individual sessions = 400 minutes")

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Quote from publication for the relevant groups (phonological training and within-school controls): the "six poorest readers randomly assigned to Phonological Training (N=4) or to within-school control condition (N=2)"
Allocation concealment (selection bias)	Unclear risk	No information provided
Blinding of participants and personnel (performance bias) All outcomes	Unclear risk	No information provided for participants or personel, however, it is unlikely participants were aware of allocation
Blinding of outcome assessment (detection bias) All outcomes	Low risk	Quote from publication: "At each of the three post-tests, members of the research team tested the children 'blind', that is without knowing to which group children belonged"

Hurry 2007 (*Continued*)

Incomplete outcome data (attrition bias) All outcomes	Low risk	For the relevant groups (phonological training and within-school controls), 4 and 3 (respectively) children dropped out between pre- and post-test 1. We requested more information from J. Hurry. Response (received 17/01/2012) that some of children "had failed to receive a sufficient amount of the intervention, usually as a result of moving school" while others "could not be tested because they had moved too far or were not traced". Thus both groups experienced the same (relatively low) drop-out rate for reasons extraneous to the study
Selective reporting (reporting bias)	Low risk	Data reported for all reading tests listed in methods; adequate detail for data to be included in analysis
Other bias	Low risk	None apparent

Levy 1997

Methods	Randomised controlled trial Four intervention groups (3 relevant: phonics) and one1 control group (untrained)
Participants	Criteria: less than 7 words read correctly on the WRMT Word Identification test; or less than 7 words read correctly on the WRAT-R Word Identification test; or less than 7 training words reading correctly Recruits: 125 English-speaking Grade 2 children from 16 schools in Canada. Mean performance on WRMT at Grade 1.2 level and word identification subtest of WRAT-R scores in pre-school range. On average only read 3 or 4 words from the set of 32 words to be trained Allocation: children were randomly allocated to 5 groups: 4 intervention groups and 1 control group. Three intervention groups did phonics training and so all these children were grouped together for the Intervention group. The fourth group did whole word training (not relevant). The fifth (untrained) group was used as the control group Intervention group: N = 75 Control group: N = 25
Interventions	"The four training groups all learned to read the same set of 32 words, as well as participating in the classroom program... On each day of training, children in all groups read once only the entire set of 32 words printed on individual index cards. The groups differed in how the words were grouped during learning, and in the method of instruction" (p. 366) Intervention 1 (rime training): "32 words arranged so that the 4 written words of a rime family were shown together. First 15 days or until all 32 words pronounced correctly on 2 successive days: 'common rime segment for each family block was written in red to highlight the shared orthographic segment" (p. 366). Following 15 days or when

	criterion was met: "10 black and white trials where the child pronounced the 32 words printed in black ink once a day" (p. 368) Intervention 2 (onset training): "four written words per family block shared the initial consonant(s)-vowel segment". 15 colour trial days (or 2 successive correct readings): initial consonant(s)-vowel segment written in red. Following 15 days or when criterion met: maximum of 10 black and white trials Intervention 3 (phoneme training): "four written words for each block were randomly selected from the 32 words, with the restriction that no two onset or rime family members could be in the same block. The same eight random blocks were used on each day of training. There was no consistent relation among phonemic units in the four words, but for each word the letters of each phoneme were printed in a different colour.. maximum of 15 colour trials and 10 black and white trials" (p. 368) Control: received regular classroom regimen during the training phase Procedure: pre-test phase, training phase, post-training phase. One-on-one training
Outcomes	Time of post-test: immediately after training completed Relevant measures: nonword reading accuracy (experimental: 48 new nonwords) and regular word reading accuracy (experimental: 48 new regular words)
Notes	1. Paper presents 2 experiments. Experiment 1 focused on non-readers while experiment 2 focused on poor readers. Therefore we only included Experiment 2 in our review Intervention 4 (of experiment 2) trained irregular words and therefore we did not include this in our review/analysis 2. Contacted author (B. Levy) 26/09/2011 for: (1) mean age (and SDs) of participants: did not know; (2) number of males/females: did not know; (2) inclusion criteria: did not know; (4) details on the control group: same as the control group in experiment; (5) length of training: depended on child's progress and speed of responding; (6) training group size: one-on-one 3. Since the rime, onset and phoneme training groups all trained phonics, we merged their results for the experimental data 4. There were 2 measures that tested Reading Accuracy: nonwords (onset nonwords and rime nonwords). We merged these 2 tests for a measure of Reading Accuracy: nonwords. Similarly, there were 2 measures testing Reading Accuracy: regular words (onset words and rime words). We merged these 2 tests for 1 measure of Reading Accuracy: regular Words 5. There were 2 immediate post-tests: 1 the day after completion, and 1 a week after completion. We used the first post-test in this review

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote from publication: "twenty-five children were randomly assigned to each of the five training conditions" Quote from personal communication: "children were randomly assigned to conditions as they arrived for the study, with the intention to keep numbers per condi-

Levy 1997 (*Continued*)

		tion as equal as possible in each school at all times. The idea was to balance for time of year effects and conditions in schools. Otherwise assignment per condition was random and controlled by the tester“
Allocation concealment (selection bias)	Unclear risk	Quote from personal communication: ”the testers knew what condition the child was in, since that defined the experimental procedures“
Blinding of participants and personnel (performance bias) All outcomes	Low risk	Quote from publication: ”all phases of the experiment were conducted in a small room in the school where the experimenter worked with each child individually“ Quote from personal communication: ”the teachers and parents knew the general purpose of the study but no details of manipulations or child assignments or individual child outcomes“ Comment: unlikely participants were aware of allocation
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Quote from personal communication: ”the same testers scored all tests for both pre and post tests. No blinding of testers was attempted since the experimenters were largely the testers“
Incomplete outcome data (attrition bias) All outcomes	Low risk	No explicit information about attrition, but degrees-of-freedom suggest all randomised participants were included in the analysis
Selective reporting (reporting bias)	Low risk	Data reported for all outcome measures outlined in methods; adequate detail for data to be included in analysis
Other bias	Low risk	None apparent

Levy 1999

Methods	Randomised controlled trial (stratified randomisation) 6 intervention groups (4 relevant: phonics training) and 2 control groups (alternative training)
Participants	Criteria: in Grade 2; English speaker; a score below 90 on the WRAT-3 Word Identification test; a score more than half a grade below appropriate grade on the WRMT Word

	Identification test; less than 15 training words read correctly Recruits: 128 children (56 female, 72 male; mean age 7 years, 7 months) Allocation: poor readers were allocated to 4 fast RAN groups (rime training, phoneme training, whole word training, untrained controls) and 4 slow RAN groups (rime training, phoneme training, whole word training, untrained controls). Since (1) the rime and phoneme training trained phonics, and (2) this study did not discriminate poor readers based on their RAN, we grouped the 4 fast and slow RAN groups who did rime or phoneme trainees into the intervention group, and grouped the fast and slow RAN untrained controls into the control group Intervention group: N = 64 Control group: N = 32
Interventions	"On each day of training, children in all groups read through the set of 48 words once only. Each word was printed on a separate index card. On the 1st day only, the experimenter first read through the set only once, in a manner appropriate to modelling that training condition, and then the child read through the set in the same manner. On all subsequent days, the child read the words and the experimenter provided only corrective feedback. The critical differences among the three training conditions for the fast and the slow RAN groups were how the 48 words were grouped together during the presentation and how the words were segmented" (pp. 123-124) Intervention 1 (rime training): "48 words were presented 4 at a time, where the word on each of the four cards presented together was from the same rime family and each was segmented by colouring the rime unit in red and the onset unit in black" (p. 124). Colour trials: 15 days or until criterion of entire 48 words read correctly on 2 successive days was met. Following the colour trials the words were printed in black ink only Intervention 2 (phoneme training): "Each phonemic unit was printed in a different colour for the 1st 15 days of training or until the criterion of two successive perfect readings was met." Following the colour trials the words were printed in black and white Control (arithmetic training): "Help with addition and subtraction in one-on-one sessions" (p. 125) Procedure: all one-on-one training, outside of the classroom, for 15 minutes per day for 4 weeks
Outcomes	Time of test: day after completion of training Relevant measures: nonword reading accuracy (experimental: 48 new nonwords) and regular word reading accuracy (experimental: 48 new regular words)
Notes	1. While there were 6 intervention groups (fast and slow RAN rime, phoneme and whole word) our review focuses on the rime and phoneme conditions since they were phonics training 2. Since both the rime and phoneme intervention groups trained phonics, the experimental data used in this review is an average of the fast and slow RAN rime and phoneme training groups (that is, 4 groups). The control data is an average of the fast and slow RAN control groups 3. There were 2 immediate post-tests: 1 the day after completion, and 1 a week after completion. We only used the first post-test in this review
Risk of bias	

Levy 1999 (*Continued*)

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Quote: "the fastest RAN children were assigned to the four fast RAN training groups and the slowest RAN children were assigned to the four slow RAN training groups". And then 1 fast RAN group and 1 slow RAN group were allocated to each type of training and a control group Quote from personal communication: "Children were randomly assigned to conditions as they arrived for the study, with the intention to keep numbers per condition as equal as possible in each school at all times. The idea was to balance for time of year effects and conditions in schools. Otherwise assignment per condition was random and controlled by the tester"
Allocation concealment (selection bias)	Unclear risk	Quote from personal communication: "the testers knew what condition the child was in, since that defined the experimental procedures"
Blinding of participants and personnel (performance bias) All outcomes	Low risk	Quote from personal communication: "the teachers and parents knew the general purpose of the study but no details of manipulations or child assignments or individual child outcomes" Comment: unlikely participants were aware of allocation
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Quote from personal communication: "the same testers scored all tests for both pre and post tests. No blinding of testers was attempted since the experimenters were largely the testers"
Incomplete outcome data (attrition bias) All outcomes	Low risk	Comment: no explicit information about attrition, but analysis of number of children who met criterion after training suggests that all randomised participants were included in the analysis (that is, 16 in each group)
Selective reporting (reporting bias)	Low risk	Data reported for all outcome measures outlined in methods; adequate detail for data to be included in analysis

Levy 1999 (*Continued*)

Other bias	Low risk	None apparent

Lovett 1990

Methods	Randomised controlled trial Two intervention groups (1 relevant: phonics + sight words) and 1 control group (alternative training)
Participants	Criteria: a score below the 25th percentile on 4 out of 5 reading tests (WRAT-3: Reading; WRMT-R: Word Identification; WRMT-R: Word Attack; Peabody Individual Achievement Test - Revised: Reading Recognition; GFW Sound-symbol Tests: Reading of Symbols); WISC-R Verbal and Performance IQ of at least 85; no English as second language, extreme hyperactivity, hearing impairment, brain damage, a chronic medical condition or serious emotional disturbance, attention deficits; age between 7 and 13 years Recruits: 54 disabled readers (38 boys, 16 girls), aged 7 to 13 years (mean age 8.4 years, SD 1.6). WISC-R Mean Verbal IQ = 98.4, SD 10.6; Mean Performance IQ = 106.2, SD 12.6. Majority of participants were from families in the middle socioeconomic ranges according to the Blishen scales (Index *M* = 43.6 m, SD = 11.5, range 28.9 to 71.7) Allocation: randomly assigned to 3 groups: REGEXC, REG=EXC, and control (CSS). This review used the REGEXC group as the intervention group and the control group was the control group (see notes for remaining group) Intervention group: N = 18 (Jan Frijters contacted for this) Control group: N = 18 (Jan Frijters contacted for this)
Interventions	Intervention: REGEXC: "Regular words were taught by training the constituent letter-sound mappings. Exception words were introduced and rehearsed by whole-word methods alone... spelling training for regular words emphasized segmentation of the word into its individual *sounds*, with attention paid to the sequence of sounds, the sequence of individual letters, and any letter-sound patterns illustrated by the word" (pp. 770-771) Control: CSS programme: problem solving and study skills training Procedure: 35 one-hour sessions for each programme (4 per week). Children instructed in pairs in special laboratory classrooms at a paediatric teaching hospital by special education teachers. "There was no attempt to explicitly control for other educational experiences of the children enrolled in these programs. Some were in special education placements in their community schools; some were not and had never been. For those subjects receiving any other individualized remedial instruction, their teacher was asked to refrain from training, rehearsing, or elaborating in any way the instructional content the child was receiving as part of his or her experimental treatment program" (p. 771)
Outcomes	Time of post-test: not stated explicitly but appears to be immediate Relevant measures: regular word reading accuracy (experimental: trained and untrained words), irregular word reading accuracy (experimental: trained and untrained words), regular word reading fluency (experimental: trained and untrained words), irregular word reading fluency (experimental: trained and untrained words), regular word spelling (experimental: trained and untrained words), irregular word spelling (experimental: trained and untrained words) and letter-sound knowledge (experimental: trained and untrained letter-sound rules)

Lovett 1990 (*Continued*)

Notes	1. Contacted Frijters (4/10/2011) about number of people per training condition (replied N = 18) 2. A second intervention group: REG=EXC was excluded from the review since it trained orthographically regular words by whole word method alone 3. Pre- and post-test means and SDs were provided along with post-test means adjusted for pre-test performance differences. We have not included the post-test means adjusted for pre-test performance difference because there was very little difference between these adjusted post-test means and the unadjusted post-test means, and because the adjusted post-test means provided standard errors rather than SDs 4. For each of the outcomes there were both trained and untrained measures. We averaged the trained and untrained data for each outcome

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote from publication: "children were randomly assigned to a treatment condition and to a particular teacher" Quote from personal communication: "children were matched on decoding ability and then random number tables were used to random assign treatment to pair and to assign teacher to pair" Communication: Best described as matching with randomisation
Allocation concealment (selection bias)	Low risk	Quote from personal communication: "the PI assigned treatments and teachers to child pair based on participant identity alone. Neither children nor teachers would have had contact with the person doing the assignment, as all contact prior to this point was with study psychometrists"
Blinding of participants and personnel (performance bias) All outcomes	High risk	Quote from personal communication: "since this is a verbally-administered intervention with quite explicit and structured content, and teachers were trained on the materials used, teachers could not be blind to the particular treatment they were teaching. Participants were not told what their assignments were, but on consent forms were told that they would participate in one of three conditions, with all conditions described. Teachers did not reveal condition to participants"

Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Quote from personal communication: "All standardized/norm referenced assessments were administered by trained psychometrists who were blind to assignment; however, some content-related and experimental measures (e.g., the four word lists) were administered by teachers themselves at the pre-specified testing intervals. In the former case, psychometrists would have had the participants name and testing folder alone, not the master subject-list" Comment: We used data from the experimental measures that were administered by teachers, and so risk is unclear
Incomplete outcome data (attrition bias) All outcomes	Low risk	Comment: in the publication, there was no explicit information about attrition; the fact that degrees-of-freedom varied between tests suggests missing data for some children for some tests Quote from personal communication: "this one has puzzled us. We would typically report dropouts and/or discontinuations. Given the design, we would have expected a df of 50, which is what is reported for most measures. The lower df would likely indicate not dropped-out participants, but equipment errors, basal/ceiling problems, etc. that may have invalidated particular tests, or in the case of speed specifically (reported as 41 df) a failure of the voice onset recording device" Comment: given that equipment errors etc occur on a random basis, the lower dfs were unlikely to relate to bias
Selective reporting (reporting bias)	Low risk	Data reported for all outcome measures outlined in methods; adequate detail for data to be included in analysis
Other bias	Low risk	None apparent

Lovett 2000

Methods	Randomised controlled trial Two intervention programmes (1 relevant: phonics + phoneme awareness) and a control (alternative training)
Participants	Criteria: a score below the 25th percentile on 4 out of 5 reading tests (WRAT-3: Reading; WRMT-R: Word Identification; WRMT-R: Word Attack; Peabody Individual Achievement Test - Revised: Reading Recognition; GFW Sound-symbol Tests: Reading of Symbols); WISC-R Verbal and Performance IQ of at least 85; no English as second language, extreme hyperactivity, hearing impairment, brain damage, a chronic medical condition or serious emotional disturbance, attention deficits; age between 7 and 13 years Recruits: 166 reading disabled children (113 boys, 53 girls). Mean IQ on WISC-3 or WISC-R: Verbal IQ M = 92, SD = 13.7, Performance IQ M = 98.7, SD = 14.3. Mean age 9.9 (SD 1.6 years). On average, sample was more than 2 SD below age-norm expectations at referral, with half of the children consistently below the first percentile for age on standardised achievement measures. Of these 166, 84.3% of the sample (140 participants) could be classified into 1 of 3 subgroups: 54.3% double deficit, 22.1% phonological deficit, 23.6% visual naming-speed deficit Allocation: the 140 children randomly assigned to 1 of 3 treatments: PhAB training; WIST Program (not relevant to this review); and CSS (controls). In this review, the PhAB trainees are the intervention group and the controls are the control group Intervention group: N = 51 Control group: N = 37
Interventions	Intervention: PhAB skills were trained with oral and written presentations of letter-sound and letter-cluster-sound correspondences. Word segmenting and blending, sound segmentation and blending, rhyming. Special orthography used (based on Engelmann and colleagues) to teach letter sounds: "the special orthography is a temporary convention used to highlight salient features of some letters; it provides visual cues to the child with RD such as symbols over long vowels (macrons), letter size variation, and connected letters to facilitate initial learning" (p. 337) Control: the CSS Program taught organisational strategies, academic problem solving, study and self-help techniques. Children in the CSS programme received the same amount of individualised teacher attention as did children in the remedial reading programmes Procedure: children received 35 hours of instruction (1-hour sessions, 4 times per week) on a two-to-one or three-to-one ratio in special laboratory classrooms at a paediatric teaching hospital or in affiliated schools in the Toronto metropolitan area
Outcomes	Time of post-test: immediately after training completed Relevant measures: nonword reading accuracy (WJRMT: Word Attack subtest), regular word reading accuracy (experimental: 149 untrained regular words), irregular word reading accuracy (experimental: 149 untrained exception words), and phoneme awareness (GFW Sound Symbol Tests: Sound Analysis subtest)
Notes	1. Contacted Frijters (4/10/2011) about means and SDs for reading measures from each of the 3 training conditions. We received an Excel file with means and SDs. 2. Asked whether there was an overlap in participants across 1994, 1997, and 2000 papers published by their laboratory (N = 62 in 1994 paper, N = 122 in 1997 paper, and N = 166 in 2000 paper). It was confirmed that there was an overlap in participants between the papers. We therefore decided to only include the 2000 paper for this Cochrane review

	to limit any over representation of the data in the final meta-analysis 3. The second intervention group did the WIST Program. The WIST contained more than 2 training components (word identification by analogy, seeking the part of the word that you know, attempting variable vowel pronunciations, 'peeling off' prefixes and suffixes in an multi-syllabic word) and so was not included in the review 4. Two measures tested Reading Accuracy: nonwords (GFW: Reading of Symbols and WJRMT-R: Word Attack). We included the WJRMT-R as it is a very widely used test with known reliability 5. There were multiple measures of phoneme awareness. We selected GFW sound analysis because it was well matched between groups before training

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote from publication: "the experimental design in which the original 166 children participated involved random assignment to one of three active treatment programs" Quote from personal communication: "children were matched on decoding ability and then random number tables were used to random assign treatment to pair and to assign teacher to pair" Communication: Best described as matching with randomisation
Allocation concealment (selection bias)	Low risk	Quote from personal communication: "the PI assigned treatments and teachers to child pair based on participant identity alone. Neither children nor teachers would have had contact with the person doing the assignment, as all contact prior to this point was with study psychometrists"
Blinding of participants and personnel (performance bias) All outcomes	High risk	Quote from publication: "sessions were taught by trained special education teachers who taught in all programs to minimize the potential for teacher by program delivery bias" Quote from personal communication: "since this is a verbally-administered intervention with quite explicit and structured content, and teachers were trained on the materials used, teachers could not be blind to the particular treatment they were teaching. Participants were not told what their assignments were, but on consent forms were told that they would participate in one

Lovett 2000 (*Continued*)

		of three conditions, with all conditions described. Teachers did not reveal condition to participants“
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Quote from personal communication: ”all standardized/norm referenced assessments were administered by trained psychometrists who were blind to assignment; however, some content-related and experimental measures (e.g., the four word lists) were administered by teachers themselves at the pre-specified testing intervals. In the former case, psychometrists would have had the participants name and testing folder alone, not the master subject-list“ Comment: We used data from both experimental and normed tests and so risk is unclear
Incomplete outcome data (attrition bias) All outcomes	Low risk	Extra data provided by author revealed that the data of all randomised participants was included in the analyses
Selective reporting (reporting bias)	Low risk	Data reported for all outcome measures outlined in methods; adequate detail for data to be included in analysis
Other bias	Low risk	None apparent

Savage 2003

Methods	Randomised controlled trial Three intervention groups (all phonics + phonological awareness) and 1 control group (untrained)
Participants	Criteria: the 108 Year 1 children across 9 schools with the lowest scores on screening tests for phonological awareness (nursery rhymes, rhyme matching, rhyme generation, blending, segmentation) and reading (nonsense word reading, word reading and spelling, letter-sound knowledge); English speaking Recruits: 108 English-speaking readers in Year 1 were selected (64 boys and 44 girls) Allocation: ”within each school, children were allocated to an intervention condition (usually nine children) or to a control condition (usually three children)“ (p. 219) Personal communication: ”this was done using an (online) random number generator set with parameters 1-4, for each school allowing placing into each of the interventions...Child-level allocation to intervention versus control within each school was again undertaken using random number generator” Intervention group: N = 81 Control group: N = 27

Savage 2003 (*Continued*)

Interventions	Intervention: "in each session, all children started with letter-sound learning activities using a range of multi-sensory approaches (for example, saying, looking, tracing) to learn letter sounds supported by the Jolly Phonics stories and actions." (p. 53); and "principles of segmenting and blending with a limited number of sounds." (p. 53). This was followed by 10-minutes of training on phonemes (for the phoneme training group) , for rhymes (for the rhyme training group) or for both (for the mixed training group) . This in turn was followed by 5 minutes of phonological awareness training: "games tailored to phonemes or rhymes respectively" (p. 53). From this point in each session, the training varied between intervention groups. The phoneme training group trained with S*oundWorks*: (1) an 'a-board', (2) writing on lines (with 'slips' and 'foldovers': cards with vowel markers or spaces to write vowels), (3) 'spelling from your head', (4) 'read the word', and (5) 'sound it out' with an adult. The rhyme training group practiced rhymes with plastic letters along with writing words, simple word searches, using onset rhyme 'word fans', sorting words into '-an' and '-at' groups and using onset sound frames (depicted as elements in a picture of a caterpillar's body). The mixed training group did a mixture of the two interventions above along with analysing words using their phonemic elements (for example, 'at' made up of 'a' and 't') and using phonemes and rhymes in word building Control: "children remained in class and undertook the word-level work appropriate to the second term of Year 1 of the National Literacy Strategy in their normal fashion" (p. 55) Procedure: learning support assistants conducted training in small groups (typically 4 children per group-as per email from Savage). 20-minute sessions, 4 times a week, for a period of 9 weeks at school
Outcomes	Time of post-test: not stated explicitly but appears to be immediate Relevant measures: nonword reading accuracy (experimental: high rhyme nonwords and low rhyme nonwords), regular word reading accuracy (experimental: 6 regular words), regular word spelling (experimental: 6 regular words), letter-sound knowledge (experimental: "two sets of cards each containing 13 of the 26 letters of the alphabet presented one letter per card" (p. 218) and phoneme awareness (experimental: onset-rhyme segmentation)
Notes	1. Similar design to Savage, Carless and Stuart (2003) but done on a new sample of the same size (personal communication from Robert Savage 30/11/2011) 2. Contacted Savage about 1. drop-outs (24/01/2012): 4 drop-outs, 1 from each group; 2. training group size (11/02/2012): typically 4 in each training group 3. Since the 3 intervention groups all consisted of phonics and phonological awareness training, we have used the combined mean scores (and SDs) at pre- and post-tests (see Table 3, p. 222) 4. There were 2 tests used to measure Reading Accuracy: nonwords (high rime nonwords and low rime nonwords). These 2 tests were averaged 5. Three tests were used to measure phoneme awareness (1. rime matching, 2. onset-rime segmentation, 3. phoneme segmentation). We included the onset-rime segmentation as its intervention and control pre-test scores had the best match
Risk of bias	

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Quote from publications: "within each school, children were allocated to an intervention condition (usually nine children) or to a control condition (usually three children). Schools themselves decided on the precise composition of each of the subgroups of three to four children who went together with an LSA for each intervention session based upon their knowledge of the children's social networks, so intervention groups varied slightly in size across schools" Quote from personal communication: "this was done using an (online) random number generator set with parameters 1-4, for each school allowing placing into each of the interventions. Schools decided on suitability of children for intervention (as we note on page 219), though only 1 child was removed on teacher request. Child-level allocation to intervention versus control within each school was again undertaken using random number generator. However schools decided the precise composition of (the already selected) intervention child groups to create groups of children who got on well" Quote from further personal communication: "the allocation was random at school and student-level. The composition of small groups of children WITHIN the allocated random conditions was (and I recall, was very occasionally) adjusted only on the suggestion of classroom teachers to make the groups more functional at the social level (an e.g. I recall is a particular group of 4 randomly-allocated kids which included 3 'noisy' boys and a very shy girl) , thus we might move the groups a bit for the delivery of the intervention. The initial randomisation was always respected. It was to avoid major problems that we would do this rather than to find groups who particularly got on, hence it was rare this happened. The key point is that the initial randomizations of condition was always intact, the grouping for the purpose of intervention delivery was occasionally adjusted"

Allocation concealment (selection bias)	Low risk	Quote from personal communication: "I did this allocation independent of those running the study and of co-author(s) Carless and Stuart. Carless led the Teacher Assistant (TA) training, so I judge allocation to be concealed, and not possible to predict"
Blinding of participants and personnel (performance bias) All outcomes	High risk	Quote from publication: "teachers were told who the control children and intervention children were, and were also reinforced at training and during the intervention to treat the control children in the same way as they would if no intervention was taking place for other children" Quote from personal communication: "the TAs delivered [the training] based on sub-lexical phonological unit taught (rimes or phonemes) and this content is quite visible in the 'treatment' (no equivalent to a pill or placebo an option here). The one aspect that was blind was that we emphasized to TAs and all other school staff that each of the interventions (rime phoneme or mixed) was a proven evidence-based intervention, so we cast it as 3-horse race between them (with no favoured intervention) at all times, and emphasized the need for a 'fair-test' of each. TAs understood this. At the participant end, these are 6 years olds in both studies. They simply knew they were in an intervention (intervention condition children only of course) or receiving regular classroom teaching (control group children)"
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Quote from personal communication: 'Pre-testing was undertaken as a screen of all children in schools before we identified and allocated the 'at-risk readers', (se Consort flow diagrams in both papers) so in this sense it is entirely blind'. 'There was no blinding of post-testing in relation to the intervention condition as TAs did both (though see comments above on the 3 horse race). However classroom assistants also did not know of the theoretical contrasts (and they were definitely blind to the status of the high-rime and low-rime non-

		words in the 2003 study as these were randomised as a set of 12 items for pre-testing and post-testing). TAs were not told at any point of any research predictions regarding the relationship between intervention and outcome (e.g. hypothesis of possible link between phoneme-based intervention and raise phoneme awareness at post-test, and similar for rimes etc)' Comment: Although testers were blind to intervention that was supposed to have superior effect, and were educated about the importance of bias, they were not blind to whether the child had done training or not at post-test. Thus, it is not clear how much bias was related to the outcomes assessment
Incomplete outcome data (attrition bias) All outcomes	Low risk	Comment: Across all intervention groups, only four participants dropped out. We confirmed with the author that there were equal dropouts in each group (N=1). Thus both groups experienced the same (relatively low) drop-out rate
Selective reporting (reporting bias)	Low risk	Comment: Data reported for all outcome measures outlined in methods; adequate detail for data to be included in analysis
Other bias	Low risk	Comment: None apparent.

Savage 2005

Methods	Randomised controlled trial Three intervention groups (all phonics + phonological awareness) and 1 control group (untrained)
Participants	Criteria: the 108 Year 1 children across 9 schools with the lowest scores on screening tests for phonological awareness (nursery rhymes, rhyme matching, rhyme generation, blending, segmentation) and reading (nonsense word reading, word reading and spelling, letter-sound knowledge); English speaking Recruits: 108 English-speaking readers in Year 1 were selected (54 boys and 54 girls) Allocation: the same as Savage 2003. That is random allocation of schools to 1 of 4 groups: 3 intervention groups (1 doing phoneme training, 1 doing rhyme training, and 1 doing a mix of both) and 1 control group (untrained). And then random allocation of children to treatment and control groups within schools. Since the 3 interventions trained phonics and phonological awareness, their data were merged for the Intervention group Intervention group: N = 81

	Control group: N = 27
Interventions	Intervention: "In each session, all children started with letter-sound learning activities using a range of multi-sensory approaches (for example, saying, looking, tracing) to learn letter sounds supported by the Jolly Phonics stories and actions." (p. 53); and "principles of segmenting and blending with a limited number of sounds" (p. 53). This was followed by 10-minutes of training on phonemes (for the phoneme training group) , for rhymes (for the rhyme training group) or for both (for the mixed training group) . This in turn was followed by 5 minutes of phonological awareness training: "games tailored to phonemes or rhymes respectively" (p. 53). From this point in each session, the training varied between intervention groups. The phoneme training group trained with S*oundWorks*: (1) an 'a-board', (2) writing on lines (with 'slips' and 'foldovers': cards with vowel markers or spaces to write vowels), (3) 'spelling from your head', (4) 'read the word', and (5) 'sound it out' with an adult. The rhyme training group practiced rhymes with plastic letters along with writing words, simple word searches, using onset rhyme 'word fans', sorting words into '-an' and '-at' groups and using onset sound frames (depicted as elements in a picture of a caterpillar's body). The mixed training group did a mixture of the 2 interventions above along with analysing words using their phonemic elements (for example, 'at' made up of 'a' and 't') and using phonemes and rhymes in word building Control: "children remained in class and undertook the word-level work appropriate to the second term of Year 1 of the National Literacy Strategy in their normal fashion" (p. 55) Procedure: learning support assistants conducted training in small groups (typically 4 children per group - as per email from Savage). 20-minute sessions, 4 times a week, for a period of 9 weeks at school
Outcomes	Time of post-test: the week after training was completed. Relevant measures: letter-sound knowledge (experimental: "cards with 26 individual letters on them" (p. 51) and phoneme awareness (experimental: nursery rhymes, rhyme matching, rhyme generation, blending and segmentation; see note 2 below)
Notes	1. Contacted Savage (24/01/2012) about what measured 1) phonological awareness, 2) letter sounds, 3) decoding and training group sizes (11/02/2012). Replied that phonological awareness was measured by nursery rhymes, rhyme matching, rhyme generation, blending and segmentation; letter sounds was measured by 1 experimental test; and decoding skills was measured by nonsense word reading, word reading and spelling, and letter-sound knowledge. We asked for the individual scores for each of these tests however he only had combined scores. Finally, training groups typically had 4 children each 2. The combined score for phonological awareness was used in our analysis 3. We did not use the decoding skills measure

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote from publication: "a quasi-random allocation of schools to programs was undertaken: four schools whose catchment ar-

		eas were known to draw primarily from lower SES backgrounds were each allocated to separate intervention groups. After that, for the other schools the allocation was entirely arbitrary". "Children were, however, entirely arbitrarily allocated to an intervention condition (nine children) or to a control condition (three children)". "As the allocation of children to intervention condition was not entirely arbitrary, but contained a systematic element..." Quote from personal communication: "The same [as the 2003 study] except that 4 schools of known low socio-economic status were each first randomly allocated to one of the 4 groups first, using a random number generator. Then the process was repeated as above for all remaining schools. Child-level allocation was again undertaken using random number generator"
Allocation concealment (selection bias)	Low risk	Quote from personal communication: "I did this allocation independent of those running the study and of co-author(s) Carless and Stuart. Carless led the Teacher Assistant (TA) training, so I judge allocation to be concealed, and not possible to predict"
Blinding of participants and personnel (performance bias) All outcomes	High risk	Quote from publication: "teachers were told who the control children and intervention children were, and were also reinforced at training and during the intervention to treat the control children in the same way as they would if no intervention was taking place for other children" Quote from personal communication: "The TAs delivered [the training] based on sub-lexical phonological unit taught (rimes or phonemes) and this content is quite visible in the 'treatment' (no equivalent to a pill or placebo an option here). The one aspect that was blind was that we emphasized to TAs and all other school staff that each of the interventions (rime phoneme or mixed) was a proven evidence-based intervention, so we cast it as 3-horse race between them (with no favoured interven-

		tion) at all times, and emphasized the need for a 'fair-test' of each. TAs understood this. At the participant end, these are 6 years olds in both studies. They simply knew they were in an intervention (intervention condition children only of course) or receiving regular classroom teaching (control group children)"
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Quote from personal communication: 'Pre-testing was undertaken as a screen of all children in schools before we identified and allocated the 'at-risk readers', (se Consort flow diagrams in both papers) so in this sense it is entirely blind'. 'There was no blinding of post-testing in relation to the intervention condition as TAs did both (though see comments above on the 3 horse race). However classroom assistants also did not know of the theoretical contrasts .... TAs were not told at any point of any research predictions regarding the relationship between intervention and outcome (e.g. hypothesis of possible link between phoneme-based intervention and raise phoneme awareness at post-test, and similar for rimes etc)' Comment: Although testers were blind to intervention that was supposed to have superior effect, and were educated about the importance of bias, they were not blind to whether the child had done training or not. Thus, it is not clear how much bias was related to the outcomes assessment, which was not blind
Incomplete outcome data (attrition bias) All outcomes	Low risk	Quote: "One child per intervention group was unavailable, having moved away from the LEA in the interim between pre- and post-test" Comment: Both groups experienced the same (relatively low) drop-out rate
Selective reporting (reporting bias)	Low risk	Comment: Data reported for all outcome measures outlined in methods; adequate detail for data to be included in analysis
Other bias	Low risk	Comment: None apparent.

BAS: British Ability Scales; CSS: Classroom Survival Skills; FSIQ: Full Scale IQ; GFW: Goldman-Fristoe-Woodcock; IQ: intelligence quotient; PhAB: phonological analysis and blending; PI: principal investigator; RAN: rapid automatised naming; SD; standard deviation; SES: socioeconomic status; TOWRE: Test of Word Reading Efficiency; WIAT-II: Wechsler Individual Achievement Test Second Edition; WIST: Word Identification Strategy Training; WJRMT: Woodcock-Johnson Reading Mastery Test; WJTA-III: Woodcock-Johnson Test of Achievement III; WRAT-R: Wide Range Achievement Test; WRMT-R: Woodcock Reading Mastery Test-Revised.

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Alexander 1991	Did not include control group
Foorman 1997	Training did not match this review's criteria for phonics training
Foorman Francis 1998	Training did not match this review's criteria for phonics training
Gillon 1997	Training did not match this review's criteria for phonics training
Gillon 2000	Group allocation did not use randomisation, quasi-randomisation, or minimisation
Gillon 2002	A follow-up study of children in Gillon 2000, which did not use randomisation, quasi-randomisation, or minimisation for group allocation
Hatcher 1994	Training did not match this review's criteria for phonics training
Hatcher 2006	Training and control group did not match this review's criteria for phonics training
Lovett 1988	Training did not match this review's criteria for phonics training
Lovett 1989	Training did not match this review's criteria for phonics training
Lovett 1994	Participants with dyslexia had neurological impairments (that is, did not meet this review's criteria for participants)
Lovett 2011	Training did not match this review's criteria for phonics training
Olson 1992	A review paper
Olson 1997	Training did not match this review's criteria for phonics training
Rashotte 2001	Training did not match this review's criteria for phonics training
Torgesen 1997	Missing pre-test data for reading outcomes (pre-tests in Table 1 are non-reading measures). We requested data from first author twice, but did not receive a reply. The second author said only the first author had the data

(*Continued*)

Torgesen 1999	Training did not match this review's criteria for phonics training
Torgesen 2001	No control group and training did not match this review's criteria for phonics training
Torgesen 2006	Training did not match this review's criteria for phonics training
Vellutino 1986	Training did not match this review's criteria for phonics training
Vellutino 1987	Training did not match this review's criteria for phonics training
Vellutino 1996	Training did not match this review's criteria for phonics training
Wise 1995	No control group and training did not match this review's criteria for phonics training
Wise 1997	Training did not match this review's criteria for phonics training
Wise 1999	Training did not match this review's criteria for phonics training
Wise 2000	No control group and training did not match this review's criteria for phonics training

Characteristics of studies awaiting assessment *[ordered by study ID]*

Pye 2008

Methods	
Participants	
Interventions	
Outcomes	
Notes	Contacted for unpublished thesis, but received an automated reply that author was on maternity leave for 6 months. We will consider for update of review

SLI: spoken language impairment

Characteristics of ongoing studies *[ordered by study ID]*

ACTRN12608000454370

Trial name or title	Phonics and sight word training in children with dyslexia
Methods	Quasi-randomised trial There are 3 treatment groups. The first group is given 8 weeks of phonics training and then 8 weeks of sight-word training (phonics + sight-word group). The second group is given the same training but in reverse order (sight-word + phonics group). The third group is given phonics and sight-word training on alternate days (mixed group). Outcomes are tested after (1) a test-re-test period, (2) after the first 8 weeks of training, and (3) after 16 weeks of training
Participants	Children with dyslexia, aged 7 to 12 years
Interventions	Computerised phonics training (phonics) and computerised and human irregular word training (sight word training)
Outcomes	Tests of (1) trained and untrained irregular word reading accuracy and spelling; (2) nonword reading accuracy, fluency, and spelling; and (3) reading fluency and reading comprehension
Starting date	2008
Contact information	Professor Genevieve McArthur Department of Cognitive Science ARC Centre of Cognition and its Disorders Macquarie University NSW 2109 Australia genevieve.mcarthur@mq.edu.au
Notes	Outcomes should be peer reviewed and published in mid- to late 2012

DATA AND ANALYSES

Comparison 1. Phonics training versus control (random-effects)

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Word reading accuracy	10	683	Std. Mean Difference (IV, Random, 95% CI)	0.47 [0.06, 0.88]
2 Nonword reading accuracy	8	512	Std. Mean Difference (IV, Random, 95% CI)	0.76 [0.25, 1.27]
3 Word reading fluency	2	54	Std. Mean Difference (IV, Random, 95% CI)	-0.51 [-1.14, 0.13]
4 Nonword reading fluency	1	18	Std. Mean Difference (IV, Random, 95% CI)	0.38 [-0.55, 1.32]
5 Reading comprehension	3	173	Std. Mean Difference (IV, Random, 95% CI)	0.14 [-0.46, 0.74]
6 Spelling	2	140	Std. Mean Difference (IV, Random, 95% CI)	0.36 [-0.27, 1.00]
7 Letter-sound knowledge	3	192	Std. Mean Difference (IV, Random, 95% CI)	0.35 [0.04, 0.65]
8 Phonological output	4	280	Std. Mean Difference (IV, Random, 95% CI)	0.38 [-0.04, 0.80]

Comparison 2. Phonics training versus control - subgroups (random-effects)

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Word reading accuracy	10		Std. Mean Difference (IV, Random, 95% CI)	Subtotals only
1.1 Training type: phonics only	3	232	Std. Mean Difference (IV, Random, 95% CI)	0.91 [-0.17, 1.98]
1.2 Training type: phonics + phoneme awareness	6	415	Std. Mean Difference (IV, Random, 95% CI)	0.28 [2.06, 0.56]
1.3 Training intensity: < 2 hours/week	8	559	Std. Mean Difference (IV, Random, 95% CI)	0.48 [-0.04, 1.00]
1.4 Training intensity: ≥ 2 hours/week	2	124	Std. Mean Difference (IV, Random, 95% CI)	0.34 [-0.03, 0.72]
1.5 Training duration: < 3 months	8	498	Std. Mean Difference (IV, Random, 95% CI)	0.56 [0.07, 1.04]
1.6 Training duration: ≥ 3 months	2	185	Std. Mean Difference (IV, Random, 95% CI)	0.12 [-0.43, 0.67]
1.7 Training group size: 1	6	419	Std. Mean Difference (IV, Random, 95% CI)	0.62 [-0.06, 1.29]
1.8 Training group size: ≤ 5	4	264	Std. Mean Difference (IV, Random, 95% CI)	0.25 [-0.04, 0.54]
1.9 Training administrator: human	6	559	Std. Mean Difference (IV, Random, 95% CI)	0.66 [0.08, 1.23]
1.10 Training administrator: computer	4	124	Std. Mean Difference (IV, Random, 95% CI)	0.15 [-0.20, 0.51]
2 Nonword reading accuracy	8	1536	Std. Mean Difference (IV, Random, 95% CI)	0.76 [0.48, 1.04]
2.1 Training type: phonics only	3	232	Std. Mean Difference (IV, Random, 95% CI)	0.91 [-0.45, 2.28]
2.2 Training type: phonics + phoneme awareness	5	280	Std. Mean Difference (IV, Random, 95% CI)	0.63 [0.38, 0.88]
2.3 Training group size: 1	5	284	Std. Mean Difference (IV, Random, 95% CI)	1.06 [0.39, 1.73]
2.4 Training group size: ≤ 5	3	228	Std. Mean Difference (IV, Random, 95% CI)	0.32 [-0.32, 0.96]

2.5 Training administrator: human	4	388	Std. Mean Difference (IV, Random, 95% CI)	1.12 [0.48, 1.76]
2.6 Training administrator: computer	4	124	Std. Mean Difference (IV, Random, 95% CI)	0.31 [-0.33, 0.96]

Comparison 3. Phonics training versus control (fixed-effect)

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Word reading accuracy	10	683	Std. Mean Difference (IV, Fixed, 95% CI)	0.46 [0.29, 0.62]
2 Nonword reading accuracy	8	512	Std. Mean Difference (IV, Fixed, 95% CI)	0.82 [0.62, 1.01]
3 Word reading fluency	2	54	Std. Mean Difference (IV, Fixed, 95% CI)	-0.53 [-1.08, 0.02]
4 Nonword reading fluency	1	18	Std. Mean Difference (IV, Fixed, 95% CI)	0.38 [-0.55, 1.32]
5 Reading comprehension	3	173	Std. Mean Difference (IV, Fixed, 95% CI)	0.01 [-0.31, 0.32]
6 Spelling	2	140	Std. Mean Difference (IV, Fixed, 95% CI)	0.28 [-0.09, 0.65]
7 Letter-sound knowledge	3	192	Std. Mean Difference (IV, Fixed, 95% CI)	0.35 [0.04, 0.65]
8 Phonological output	4	280	Std. Mean Difference (IV, Fixed, 95% CI)	0.44 [0.19, 0.70]

Comparison 4. Phonics training versus control sensitivity analysis with Hurford 1994 removed (random-effects)

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Word reading accuracy	9	633	Std. Mean Difference (IV, Random, 95% CI)	0.47 [0.01, 0.93]
2 Nonword reading accuracy	7	462	Std. Mean Difference (IV, Random, 95% CI)	0.80 [0.22, 1.38]

Analysis 1.1. Comparison 1 Phonics training versus control (random-effects), Outcome 1 Word reading accuracy.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

Outcome: 1 Word reading accuracy

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Barker 1995	18	-0.26 (1)	18	0 (1)		9.6 %	-0.25 [-0.91, 0.40]
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		8.0 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		7.7 %	0.05 [-0.87, 0.98]
Hurford 1994	25	99.06 (11.76)	25	93.76 (11.62)		10.2 %	0.45 [-0.12, 1.01]
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		11.5 %	-0.12 [-0.49, 0.24]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		10.8 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		10.7 %	1.80 [1.30, 2.30]
Lovett 1990	18	0.65 (1)	18	0 (1)		9.5 %	0.64 [-0.04, 1.31]
Lovett 2000	51	0.22 (1)	37	0 (1)		11.1 %	0.22 [-0.21, 0.64]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		11.0 %	0.36 [-0.09, 0.80]
Total (95% CI)	**440**		**243**			**100.0 %**	**0.47 [0.06, 0.88]**

Heterogeneity: $Tau^2 = 0.35$; $Chi^2 = 52.26$, df = 9 (P<0.00001); $I^2 = 83\%$

Test for overall effect: Z = 2.22 (P = 0.026)

Test for subgroup differences: Not applicable

Analysis 1.2. Comparison 1 Phonics training versus control (random-effects), Outcome 2 Nonword reading accuracy.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

Outcome: 2 Nonword reading accuracy

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Barker 1995	18	1.06 (1.4)	18	2.22 (2.9)		12.2 %	-0.50 [-1.16, 0.17]
Blythe 2006	10	98.2 (7.89)	10	87.7 (9.88)		10.0 %	1.12 [0.17, 2.08]
Ford 2009	9	24.56 (6.54)	9	22.44 (4.28)		10.2 %	0.37 [-0.57, 1.30]
Hurford 1994	25	98.26 (9.86)	25	93.04 (12.43)		13.0 %	0.46 [-0.10, 1.02]
Levy 1997	75	0.6 (0.24)	25	0.34 (0.24)		13.6 %	1.08 [0.60, 1.55]
Levy 1999	64	20.45 (8.62)	32	4.4 (4.4)		13.3 %	2.12 [1.60, 2.65]
Lovett 2000	51	13.04 (6.4)	37	8.27 (6.96)		13.9 %	0.71 [0.28, 1.15]
Savage 2003	78	1.78 (2.15)	26	0.59 (1.18)		13.8 %	0.60 [0.15, 1.06]
Total (95% CI)	**330**		**182**			**100.0 %**	**0.76 [0.25, 1.27]**

Heterogeneity: $Tau^2 = 0.44$; $Chi^2 = 44.04$, df = 7 ($P<0.00001$); $I^2 = 84\%$

Test for overall effect: Z = 2.91 (P = 0.0036)

Test for subgroup differences: Not applicable

Analysis 1.3. Comparison 1 Phonics training versus control (random-effects), Outcome 3 Word reading fluency.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

Outcome: 3 Word reading fluency

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Ford 2009	9	76 (13.14)	9	77.11 (8.54)		38.5 %	-0.10 [-1.02, 0.83]
Lovett 1990	18	-0.78 (1)	18	0 (1)		61.5 %	-0.76 [-1.44, -0.08]
Total (95% CI)	**27**		**27**			**100.0 %**	**-0.51 [-1.14, 0.13]**

Heterogeneity: $Tau^2 = 0.05$; $Chi^2 = 1.30$, df = 1 (P = 0.25); $I^2 = 23\%$

Test for overall effect: Z = 1.56 (P = 0.12)

Test for subgroup differences: Not applicable

Analysis 1.4. Comparison 1 Phonics training versus control (random-effects), Outcome 4 Nonword reading fluency.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

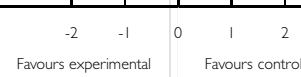
Outcome: 4 Nonword reading fluency

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Ford 2009	9	41.56 (16.93)	9	35.67 (11.78)		100.0 %	0.38 [-0.55, 1.32]
Total (95% CI)	**9**		**9**			**100.0 %**	**0.38 [-0.55, 1.32]**

Heterogeneity: not applicable

Test for overall effect: Z = 0.81 (P = 0.42)

Test for subgroup differences: Not applicable



Analysis 1.5. Comparison 1 Phonics training versus control (random-effects), Outcome 5 Reading comprehension.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

Outcome: 5 Reading comprehension

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Blythe 2006	10	106.3 (8)	10	96.6 (11.63)		24.8 %	0.93 [0.00, 1.86]
Ford 2009	9	20.78 (5.47)	9	21.89 (8.16)		25.1 %	-0.15 [-1.08, 0.77]
Hurry 2007	92	7.2 (8.5)	43	8.1 (7.5)		50.1 %	-0.11 [-0.47, 0.25]
Total (95% CI)	**111**		**62**			**100.0 %**	**0.14 [-0.46, 0.74]**

Heterogeneity: $Tau^2 = 0.15$; $Chi^2 = 4.27$, df = 2 (P = 0.12); $I^2 = 53\%$

Test for overall effect: Z = 0.45 (P = 0.65)

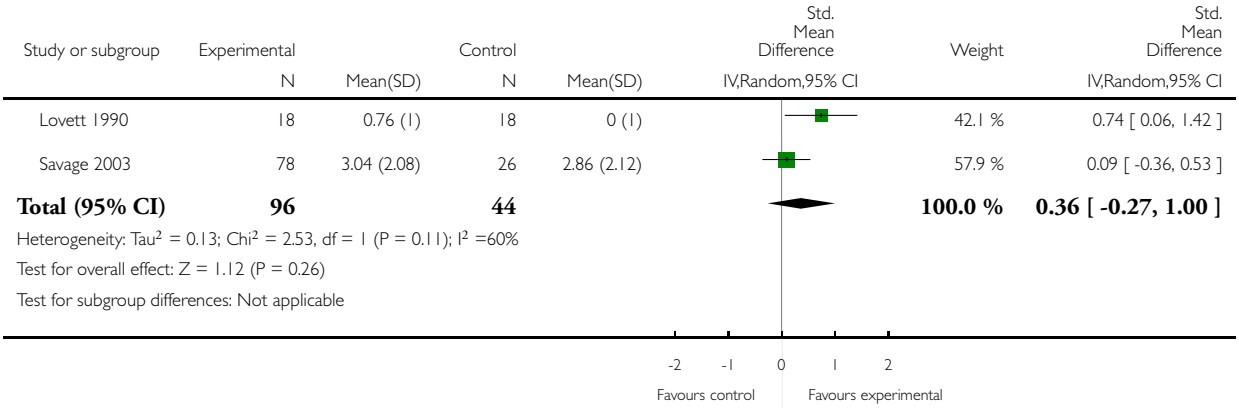
Test for subgroup differences: Not applicable

Analysis 1.6. Comparison 1 Phonics training versus control (random-effects), Outcome 6 Spelling.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

Outcome: 6 Spelling

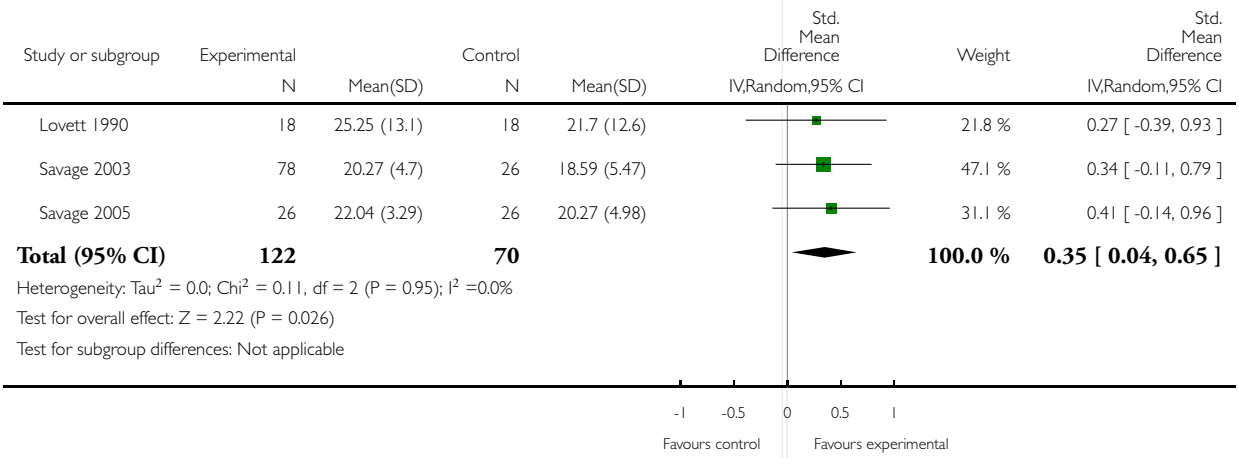


Analysis 1.7. Comparison 1 Phonics training versus control (random-effects), Outcome 7 Letter-sound knowledge.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

Outcome: 7 Letter-sound knowledge



Analysis 1.8. Comparison 1 Phonics training versus control (random-effects), Outcome 8 Phonological output.

Review: Phonics training for English-speaking poor readers

Comparison: 1 Phonics training versus control (random-effects)

Outcome: 8 Phonological output

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Barker 1995	18	1.83 (1.7)	18	2.89 (3)		20.2 %	-0.43 [-1.09, 0.24]
Lovett 2000	51	22.96 (5.85)	37	18.86 (7.19)		28.3 %	0.63 [0.20, 1.06]
Savage 2003	78	4.17 (2.18)	26	2.77 (2.35)		27.6 %	0.63 [0.17, 1.08]
Savage 2005	26	34.13 (10.44)	26	29.35 (9.41)		23.9 %	0.47 [-0.08, 1.03]
Total (95% CI)	**173**		**107**			**100.0 %**	**0.38 [-0.04, 0.80]**

Heterogeneity: $Tau^2 = 0.11$; $Chi^2 = 7.97$, df = 3 (P = 0.05); $I^2 = 62\%$

Test for overall effect: Z = 1.77 (P = 0.077)

Test for subgroup differences: Not applicable

Analysis 2.1. Comparison 2 Phonics training versus control - subgroups (random-effects), Outcome 1 Word reading accuracy.

Review: Phonics training for English-speaking poor readers

Comparison: 2 Phonics training versus control - subgroups (random-effects)

Outcome: 1 Word reading accuracy

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
1 Training type: phonics only							
Barker 1995	18	-0.26 (1)	18	0 (1)		32.1 %	-0.25 [-0.91, 0.40]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		34.0 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		33.8 %	1.80 [1.30, 2.30]
Subtotal (95% CI)	**157**		**75**			**100.0 %**	**0.91 [-0.17, 1.98]**
Heterogeneity: $Tau^2 = 0.83$; $Chi^2 = 23.93$, df = 2 (P<0.00001); $I^2 = 92\%$							
Test for overall effect: Z = 1.65 (P = 0.098)							
2 Training type: phonics + phoneme awareness							
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		8.2 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		7.6 %	0.05 [-0.87, 0.98]
Hurford 1994	25	99.06 (11.76)	25	93.76 (11.62)		16.0 %	0.45 [-0.12, 1.01]
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		25.7 %	-0.12 [-0.49, 0.24]
Lovett 2000	51	0.65 (1)	37	0 (1)		21.6 %	0.64 [0.21, 1.08]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		21.0 %	0.36 [-0.09, 0.80]
Subtotal (95% CI)	**265**		**150**			**100.0 %**	**0.28 [0.00, 0.56]**
Heterogeneity: $Tau^2 = 0.04$; $Chi^2 = 8.12$, df = 5 (P = 0.15); $I^2 = 38\%$							
Test for overall effect: Z = 1.96 (P = 0.050)							
3 Training intensity: < 2 hours/week							
Barker 1995	18	-0.26 (1)	18	0 (1)		12.2 %	-0.25 [-0.91, 0.40]
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		10.5 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		10.2 %	0.05 [-0.87, 0.98]
Hurford 1994	25	99.06 (11.76)	25	93.76 (11.62)		12.8 %	0.45 [-0.12, 1.01]
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		14.1 %	-0.12 [-0.49, 0.24]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		13.4 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		13.3 %	1.80 [1.30, 2.30]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		13.6 %	0.36 [-0.09, 0.80]
Subtotal (95% CI)	**371**		**188**			**100.0 %**	**0.48 [-0.04, 1.00]**

(*Continued* . . .)

(. . . *Continued*)

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Heterogeneity: Tau² = 0.47; Chi² = 50.65, df = 7 (P<0.00001); I² =86%							
Test for overall effect: Z = 1.80 (P = 0.072)							
4 Training intensity: ≥ 2 hours/week							
Lovett 1990	18	0.65 (1)	18	0 (1)		29.8 %	0.64 [-0.04, 1.31]
Lovett 2000	51	0.22 (1)	37	0 (1)		70.2 %	0.22 [-0.21, 0.64]
Subtotal (95% CI)	**69**		**55**			**100.0 %**	**0.34 [-0.03, 0.72]**
Heterogeneity: Tau² = 0.00; Chi² = 1.06, df = 1 (P = 0.30); I² =6%							
Test for overall effect: Z = 1.79 (P = 0.073)							
5 Training duration: < 3 months							
Barker 1995	18	-0.26 (1)	18	0 (1)		12.2 %	-0.25 [-0.91, 0.40]
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		10.4 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		10.0 %	0.05 [-0.87, 0.98]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		13.7 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		13.5 %	1.80 [1.30, 2.30]
Lovett 1990	18	0.65 (1)	18	0 (1)		12.1 %	0.64 [-0.04, 1.31]
Lovett 2000	51	0.22 (1)	37	0 (1)		14.1 %	0.22 [-0.21, 0.64]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		13.9 %	0.36 [-0.09, 0.80]
Subtotal (95% CI)	**323**		**175**			**100.0 %**	**0.56 [0.07, 1.04]**
Heterogeneity: Tau² = 0.38; Chi² = 39.40, df = 7 (P<0.00001); I² =82%							
Test for overall effect: Z = 2.25 (P = 0.024)							
6 Training duration: ≥ 3 months							
Hurford 1994	25	99.06 (11.76)	25	93.76 (11.62)		42.6 %	0.45 [-0.12, 1.01]
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		57.4 %	-0.12 [-0.49, 0.24]
Subtotal (95% CI)	**117**		**68**			**100.0 %**	**0.12 [-0.43, 0.67]**
Heterogeneity: Tau² = 0.10; Chi² = 2.80, df = 1 (P = 0.09); I² =64%							
Test for overall effect: Z = 0.42 (P = 0.67)							
7 Training group size: 1							
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		14.7 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		14.3 %	0.05 [-0.87, 0.98]
Hurford 1994	25	99.06 (11.76)	25	93.76 (11.62)		17.2 %	0.45 [-0.12, 1.01]
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		18.5 %	-0.12 [-0.49, 0.24]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		17.8 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		17.6 %	1.80 [1.30, 2.30]
Subtotal (95% CI)	**275**		**144**			**100.0 %**	**0.62 [-0.06, 1.29]**
Heterogeneity: Tau² = 0.61; Chi² = 44.35, df = 5 (P<0.00001); I² =89%							
Test for overall effect: Z = 1.78 (P = 0.075)							

(*Continued* . . .)

(. . . *Continued*)

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
8 Training group size: ≤ 5							
Barker 1995	18	-0.26 (1)	18	0 (1)		17.2 %	-0.25 [-0.91, 0.40]
Lovett 1990	18	0.65 (1)	18	0 (1)		16.6 %	0.64 [-0.04, 1.31]
Lovett 2000	51	0.22 (1)	37	0 (1)		34.3 %	0.22 [-0.21, 0.64]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		31.9 %	0.36 [-0.09, 0.80]
Subtotal (95% CI)	**165**		**99**			**100.0 %**	**0.25 [-0.04, 0.54]**
Heterogeneity: Tau² = 0.02; Chi² = 3.78, df = 3 (P = 0.29); I² =21%							
Test for overall effect: Z = 1.67 (P = 0.096)							
9 Training administrator: human							
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		17.6 %	-0.12 [-0.49, 0.24]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		16.7 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		16.6 %	1.80 [1.30, 2.30]
Lovett 1990	18	0.65 (1)	18	0 (1)		15.0 %	0.64 [-0.04, 1.31]
Lovett 2000	51	0.22 (1)	37	0 (1)		17.1 %	0.22 [-0.21, 0.64]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		17.0 %	0.36 [-0.09, 0.80]
Subtotal (95% CI)	**378**		**181**			**100.0 %**	**0.66 [0.08, 1.23]**
Heterogeneity: Tau² = 0.46; Chi² = 46.02, df = 5 (P<0.00001); I² =89%							
Test for overall effect: Z = 2.24 (P = 0.025)							
10 Training administrator: computer							
Barker 1995	18	-0.26 (1)	18	0 (1)		29.2 %	-0.25 [-0.91, 0.40]
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		16.2 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		14.7 %	0.05 [-0.87, 0.98]
Hurford 1994	25	99.06 (11.76)	25	93.76 (11.62)		39.9 %	0.45 [-0.12, 1.01]
Subtotal (95% CI)	**62**		**62**			**100.0 %**	**0.15 [-0.20, 0.51]**
Heterogeneity: Tau² = 0.0; Chi² = 2.63, df = 3 (P = 0.45); I² =0.0%							
Test for overall effect: Z = 0.85 (P = 0.40)							
Test for subgroup differences: Chi² = 6.10, df = 9 (P = 0.73), I² =0.0%							

Analysis 2.2. Comparison 2 Phonics training versus control - subgroups (random-effects), Outcome 2 Nonword reading accuracy.

Review: Phonics training for English-speaking poor readers

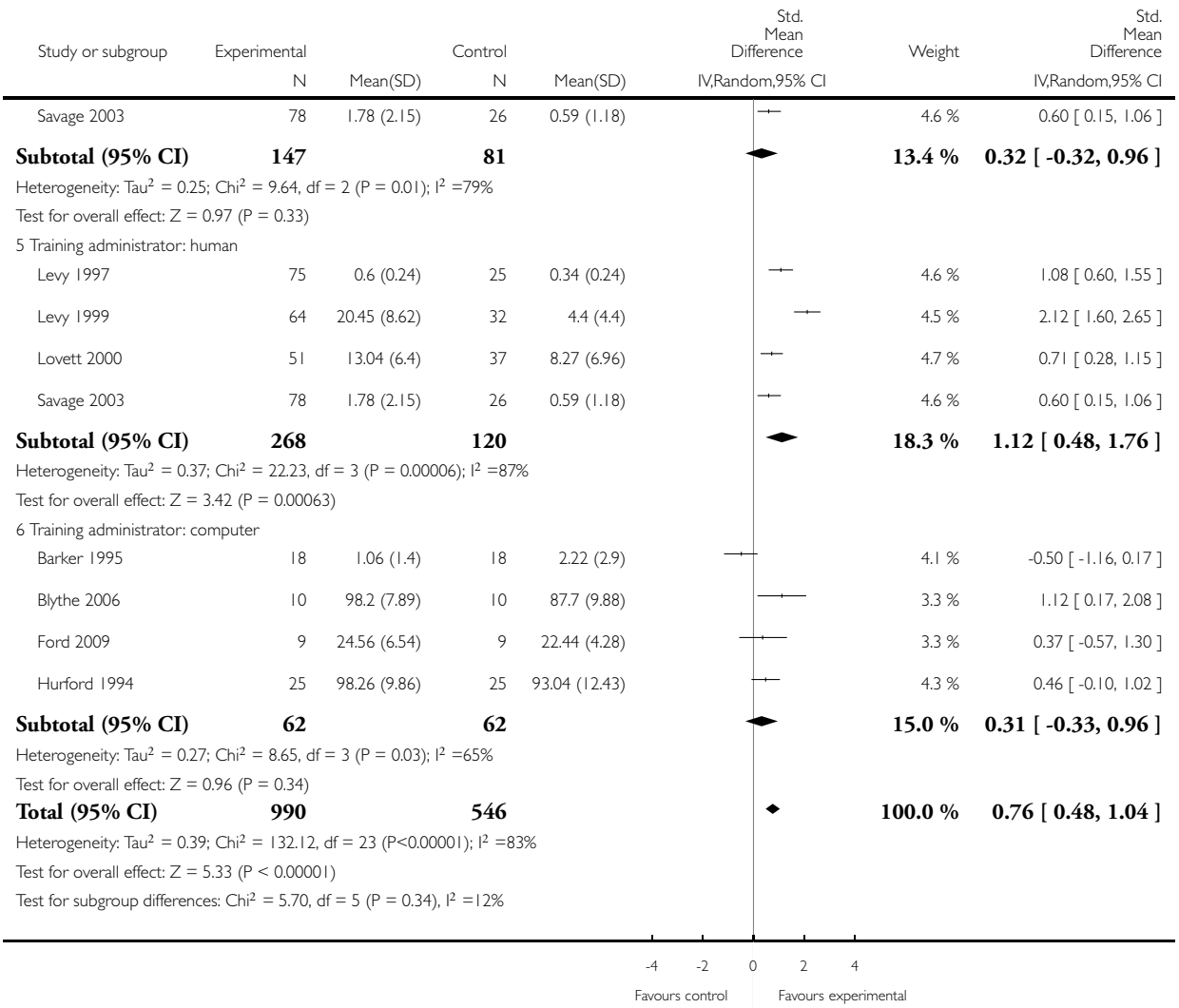
Comparison: 2 Phonics training versus control - subgroups (random-effects)

Outcome: 2 Nonword reading accuracy

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
1 Training type: phonics only							
Barker 1995	18	1.06 (1.4)	18	2.22 (2.9)		4.1 %	-0.50 [-1.16, 0.17]
Levy 1997	75	0.6 (0.24)	25	0.34 (0.24)		4.6 %	1.08 [0.60, 1.55]
Levy 1999	64	20.45 (8.62)	32	4.4 (4.4)		4.5 %	2.12 [1.60, 2.65]
Subtotal (95% CI)	**157**		**75**			**13.1 %**	**0.91 [-0.45, 2.28]**
Heterogeneity: Tau^2 = 1.37; Chi^2 = 36.92, df = 2 (P<0.00001); I^2 =95%							
Test for overall effect: Z = 1.32 (P = 0.19)							
2 Training type: phonics + phoneme awareness							
Blythe 2006	10	98.2 (7.89)	10	87.7 (9.88)		3.3 %	1.12 [0.17, 2.08]
Ford 2009	9	24.56 (6.54)	9	22.44 (4.28)		3.3 %	0.37 [-0.57, 1.30]
Hurford 1994	25	98.26 (9.86)	25	93.04 (12.43)		4.3 %	0.46 [-0.10, 1.02]
Lovett 2000	51	13.04 (6.4)	37	8.27 (6.96)		4.7 %	0.71 [0.28, 1.15]
Savage 2003	78	1.78 (2.15)	26	0.59 (1.18)		4.6 %	0.60 [0.15, 1.06]
Subtotal (95% CI)	**173**		**107**			**20.2 %**	**0.63 [0.38, 0.88]**
Heterogeneity: Tau^2 = 0.0; Chi^2 = 1.84, df = 4 (P = 0.77); I^2 =0.0%							
Test for overall effect: Z = 4.86 (P < 0.00001)							
3 Training group size: 1							
Blythe 2006	10	98.2 (7.89)	10	87.7 (9.88)		3.3 %	1.12 [0.17, 2.08]
Ford 2009	9	24.56 (6.54)	9	22.44 (4.28)		3.3 %	0.37 [-0.57, 1.30]
Hurford 1994	25	98.26 (9.86)	25	93.04 (12.43)		4.3 %	0.46 [-0.10, 1.02]
Levy 1997	75	0.6 (0.24)	25	0.34 (0.24)		4.6 %	1.08 [0.60, 1.55]
Levy 1999	64	20.45 (8.62)	32	4.4 (4.4)		4.5 %	2.12 [1.60, 2.65]
Subtotal (95% CI)	**183**		**101**			**20.0 %**	**1.06 [0.39, 1.73]**
Heterogeneity: Tau^2 = 0.46; Chi^2 = 21.92, df = 4 (P = 0.00021); I^2 =82%							
Test for overall effect: Z = 3.09 (P = 0.0020)							
4 Training group size: ≤ 5							
Barker 1995	18	1.06 (1.4)	18	2.22 (2.9)		4.1 %	-0.50 [-1.16, 0.17]
Lovett 2000	51	13.04 (6.4)	37	8.27 (6.96)		4.7 %	0.71 [0.28, 1.15]

(*Continued* . . .)

(. . . *Continued*)



Analysis 3.1. Comparison 3 Phonics training versus control (fixed-effect), Outcome 1 Word reading accuracy.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

Outcome: 1 Word reading accuracy

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Barker 1995	18	-0.26 (1)	18	0 (1)		6.3 %	-0.25 [-0.91, 0.40]
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		3.5 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		3.2 %	0.05 [-0.87, 0.98]
Hurford 1994	25	99.06 (11.76)	25	93.76 (11.62)		8.6 %	0.45 [-0.12, 1.01]
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		20.7 %	-0.12 [-0.49, 0.24]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		11.8 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		11.0 %	1.80 [1.30, 2.30]
Lovett 1990	18	0.65 (1)	18	0 (1)		6.0 %	0.64 [-0.04, 1.31]
Lovett 2000	51	0.22 (1)	37	0 (1)		15.1 %	0.22 [-0.21, 0.64]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		13.6 %	0.36 [-0.09, 0.80]
Total (95% CI)	**440**		**243**			**100.0 %**	**0.46 [0.29, 0.62]**

Heterogeneity: Chi^2 = 52.26, df = 9 (P<0.00001); I^2 =83%

Test for overall effect: Z = 5.44 (P < 0.00001)

Test for subgroup differences: Not applicable

Analysis 3.2. Comparison 3 Phonics training versus control (fixed-effect), Outcome 2 Nonword reading accuracy.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

Outcome: 2 Nonword reading accuracy

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Barker 1995	18	1.06 (1.4)	18	2.22 (2.9)		8.8 %	-0.50 [-1.16, 0.17]
Blythe 2006	10	98.2 (7.89)	10	87.7 (9.88)		4.2 %	1.12 [0.17, 2.08]
Ford 2009	9	24.56 (6.54)	9	22.44 (4.28)		4.4 %	0.37 [-0.57, 1.30]
Hurford 1994	25	98.26 (9.86)	25	93.04 (12.43)		12.2 %	0.46 [-0.10, 1.02]
Levy 1997	75	0.6 (0.24)	25	0.34 (0.24)		17.0 %	1.08 [0.60, 1.55]
Levy 1999	64	20.45 (8.62)	32	4.4 (4.4)		14.1 %	2.12 [1.60, 2.65]
Lovett 2000	51	13.04 (6.4)	37	8.27 (6.96)		20.3 %	0.71 [0.28, 1.15]
Savage 2003	78	1.78 (2.15)	26	0.59 (1.18)		19.0 %	0.60 [0.15, 1.06]
Total (95% CI)	**330**		**182**			**100.0 %**	**0.82 [0.62, 1.01]**

Heterogeneity: Chi^2 = 44.04, df = 7 (P<0.00001); I^2 =84%

Test for overall effect: Z = 8.14 (P < 0.00001)

Test for subgroup differences: Not applicable

Analysis 3.3. Comparison 3 Phonics training versus control (fixed-effect), Outcome 3 Word reading fluency.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

Outcome: 3 Word reading fluency

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Ford 2009	9	76 (13.14)	9	77.11 (8.54)		35.1 %	-0.10 [-1.02, 0.83]
Lovett 1990	18	-0.78 (1)	18	0 (1)		64.9 %	-0.76 [-1.44, -0.08]
Total (95% CI)	**27**		**27**			**100.0 %**	**-0.53 [-1.08, 0.02]**

Heterogeneity: Chi² = 1.30, df = 1 (P = 0.25); I² =23%

Test for overall effect: Z = 1.89 (P = 0.058)

Test for subgroup differences: Not applicable

Analysis 3.4. Comparison 3 Phonics training versus control (fixed-effect), Outcome 4 Nonword reading fluency.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

Outcome: 4 Nonword reading fluency

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Ford 2009	9	41.56 (16.93)	9	35.67 (11.78)		100.0 %	0.38 [-0.55, 1.32]
Total (95% CI)	**9**		**9**			**100.0 %**	**0.38 [-0.55, 1.32]**

Heterogeneity: not applicable

Test for overall effect: Z = 0.81 (P = 0.42)

Test for subgroup differences: Not applicable

Analysis 3.5. Comparison 3 Phonics training versus control (fixed-effect), Outcome 5 Reading comprehension.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

Outcome: 5 Reading comprehension

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Blythe 2006	10	106.3 (8)	10	96.6 (11.63)		11.5 %	0.93 [0.00, 1.86]
Ford 2009	9	20.78 (5.47)	9	21.89 (8.16)		11.8 %	-0.15 [-1.08, 0.77]
Hurry 2007	92	7.2 (8.5)	43	8.1 (7.5)		76.7 %	-0.11 [-0.47, 0.25]
Total (95% CI)	**111**		**62**			**100.0 %**	**0.01 [-0.31, 0.32]**

Heterogeneity: Chi^2 = 4.27, df = 2 (P = 0.12); I^2 =53%

Test for overall effect: Z = 0.04 (P = 0.97)

Test for subgroup differences: Not applicable

Analysis 3.6. Comparison 3 Phonics training versus control (fixed-effect), Outcome 6 Spelling.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

Outcome: 6 Spelling

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Lovett 1990	18	0.76 (1)	18	0 (1)		30.0 %	0.74 [0.06, 1.42]
Savage 2003	78	3.04 (2.08)	26	2.86 (2.12)		70.0 %	0.09 [-0.36, 0.53]
Total (95% CI)	**96**		**44**			**100.0 %**	**0.28 [-0.09, 0.65]**

Heterogeneity: Chi^2 = 2.53, df = 1 (P = 0.11); I^2 =60%

Test for overall effect: Z = 1.49 (P = 0.14)

Test for subgroup differences: Not applicable

Analysis 3.7. Comparison 3 Phonics training versus control (fixed-effect), Outcome 7 Letter-sound knowledge.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

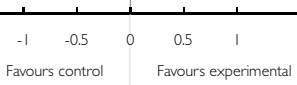
Outcome: 7 Letter-sound knowledge

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Lovett 1990	18	25.25 (13.1)	18	21.7 (12.6)		21.8 %	0.27 [-0.39, 0.93]
Savage 2003	78	20.27 (4.7)	26	18.59 (5.47)		47.1 %	0.34 [-0.11, 0.79]
Savage 2005	26	22.04 (3.29)	26	20.27 (4.98)		31.1 %	0.41 [-0.14, 0.96]
Total (95% CI)	**122**		**70**			**100.0 %**	**0.35 [0.04, 0.65]**

Heterogeneity: $Chi^2 = 0.11$, df = 2 (P = 0.95); $I^2 = 0.0\%$

Test for overall effect: Z = 2.22 (P = 0.026)

Test for subgroup differences: Not applicable



Analysis 3.8. Comparison 3 Phonics training versus control (fixed-effect), Outcome 8 Phonological output.

Review: Phonics training for English-speaking poor readers

Comparison: 3 Phonics training versus control (fixed-effect)

Outcome: 8 Phonological output

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Fixed,95% CI	Weight	Std. Mean Difference IV,Fixed,95% CI
Barker 1995	18	1.83 (1.7)	18	2.89 (3)		14.5 %	-0.43 [-1.09, 0.24]
Lovett 2000	51	22.96 (5.85)	37	18.86 (7.19)		33.7 %	0.63 [0.20, 1.06]
Savage 2003	78	4.17 (2.18)	26	2.77 (2.35)		31.0 %	0.63 [0.17, 1.08]
Savage 2005	26	34.13 (10.44)	26	29.35 (9.41)		20.8 %	0.47 [-0.08, 1.03]
Total (95% CI)	**173**		**107**			**100.0 %**	**0.44 [0.19, 0.70]**

Heterogeneity: Chi^2 = 7.97, df = 3 (P = 0.05); I^2 =62%

Test for overall effect: Z = 3.45 (P = 0.00056)

Test for subgroup differences: Not applicable

Analysis 4.1. Comparison 4 Phonics training versus control sensitivity analysis with Hurford 1994 removed (random-effects), Outcome 1 Word reading accuracy.

Review: Phonics training for English-speaking poor readers

Comparison: 4 Phonics training versus control sensitivity analysis with Hurford 1994 removed (random-effects)

Outcome: 1 Word reading accuracy

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Barker 1995	18	-0.26 (1)	18	0 (1)		10.7 %	-0.25 [-0.91, 0.40]
Blythe 2006	10	94.2 (10.49)	10	91.1 (12.09)		9.1 %	0.26 [-0.62, 1.14]
Ford 2009	9	58.89 (7.75)	9	58.44 (8.88)		8.8 %	0.05 [-0.87, 0.98]
Hurry 2007	92	13 (11.8)	43	14.5 (12.5)		12.6 %	-0.12 [-0.49, 0.24]
Levy 1997	75	0.46 (0.22)	25	0.22 (0.19)		11.9 %	1.12 [0.64, 1.60]
Levy 1999	64	21.6 (9.32)	32	6.9 (4.75)		11.8 %	1.80 [1.30, 2.30]
Lovett 1990	18	0.65 (1)	18	0 (1)		10.6 %	0.64 [-0.04, 1.31]
Lovett 2000	51	0.22 (1)	37	0 (1)		12.3 %	0.22 [-0.21, 0.64]
Savage 2003	78	2.77 (2.05)	26	2.05 (1.84)		12.1 %	0.36 [-0.09, 0.80]
Total (95% CI)	**415**		**218**			**100.0 %**	**0.47 [0.01, 0.93]**

Heterogeneity: $Tau^2 = 0.40$; $Chi^2 = 52.26$, df = 8 (P<0.00001); $I^2 = 85\%$

Test for overall effect: Z = 2.00 (P = 0.046)

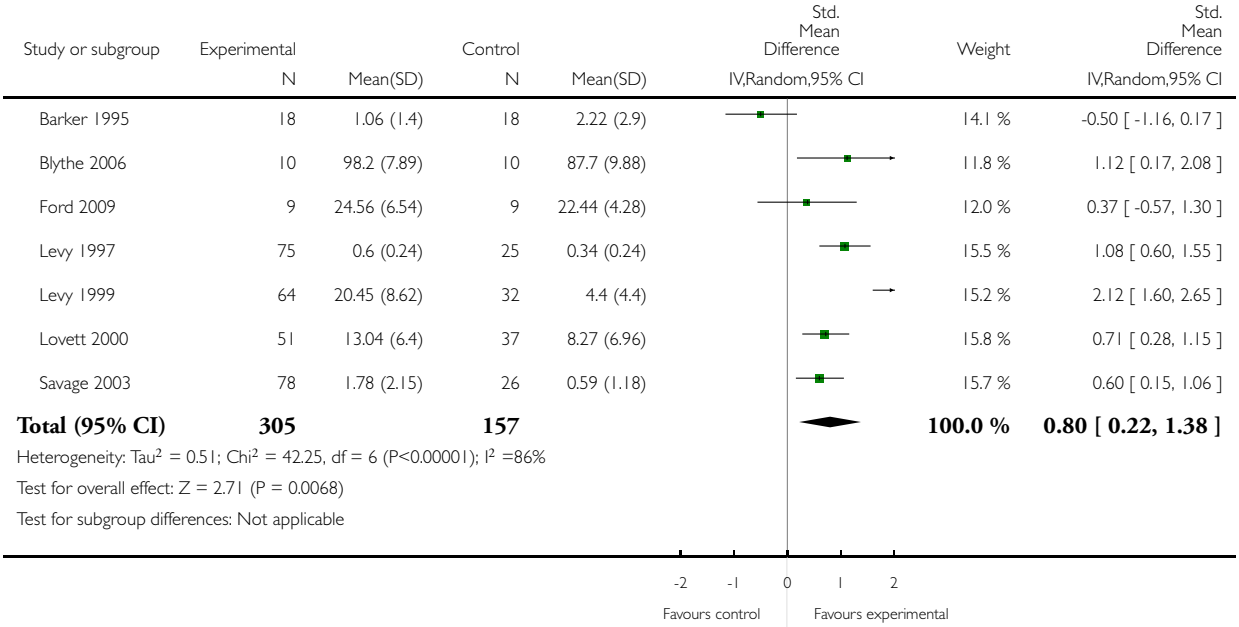
Test for subgroup differences: Not applicable

Analysis 4.2. Comparison 4 Phonics training versus control sensitivity analysis with Hurford 1994 removed (random-effects), Outcome 2 Nonword reading accuracy.

Review: Phonics training for English-speaking poor readers

Comparison: 4 Phonics training versus control sensitivity analysis with Hurford 1994 removed (random-effects)

Outcome: 2 Nonword reading accuracy



ADDITIONAL TABLES

Table 1. Additional methods for future updates

Issue	Method
Primary outcome measures	In the current review, we have combined measures for regular and irregular word reading (that is, word reading) to improve the power of the meta-analyses of the reading measures. However, from a theoretical point of view, it would be preferable to have separate estimates for these 2 types of reading. If relevant data are available, future updates will index the effect of phonics on (1) irregular and regular word reading separately, and (2) irregular and regular spelling separately
Secondary outcome measures	If relevant data are available, future updates will include letter identification, parsing, and blending as outcomes in this review
Training type	In future updates, we will include additional phonics training if the data are available (for example, phonics and letter identification training, phonics, and spoken vocabulary training)

Table 1. Additional methods for future updates (*Continued*)

Subgroups	If relevant data are available, future updates will include 4 additional subgroups to the analyses that were not possible in the current review owing to lack of data: (1) Age (children (below 12 years); adolescents (13 to 17 years); adults (18 years and above); (2) poor reading type (poor letter-sound reading; poor sight word reading; a combination of both); (3) spoken language ability (impaired unimpaired); and (4) timing of outcome assessment (immediately after training, 1 to 6 months after training, 7 to 18 months after training, more than 18 months). In addition, data allowing, we will reinstated a third categeory in the training type subgroup: phonics and sight words
Timing of outcome assessment	If relevant data are available, future updates will index 4 periods of assessment: (1) immediately after training, (2) 1 to 6 months after training, (3) 7 to 18 months after training, and (4) more than 18 months after training
Multiple measures	If a study includes multiple measures of a single outcome (for example, word reading accuracy) , and those measures are directly comparable in type and scale, an average of the 2 scores will be taken. If a study includes multiple measures of a single outcome that are not directly comparable, both measures will be used in the analysis
Multiple arms	If a study includes 2 or more comparable invention groups (for example, both 'phonics only' or both 'phonics and phoneme awareness'), the data of the 2 groups will be combined. If a study includes a 'phonics only' and 'phonics plus phoneme awareness' group (for example), the phonics only group will be used since this is a purer measure of phonics training. If a study includes more than 1 control group, the control group that does the least training of any type will be included in the review. For example, a control group that does 'school as usual' will be used over a control groups who does 'maths training'
Cross-over and cluster trials	If a cluster-randomised trial analyses the data as if individuals were randomised, we will adjust the calculations using 1 + (M-1) ICC, where M is the average cluster size, and ICC is an estimate of the relative variability between and within clusters (Higgins 2008b). If a cross-over trial does not appear to suffer carry-over or period effects, then a paired t-test will be used to measure the effect, or we will approximate a paired analysis by imputing standard deviations. If carry-over effects are a problem, we will use data from the first period (Higgins 2008b).

Table 2. Tests used by studies to measure outcomes

Outcomes	Tests	References	Studies
Word reading accuracy	1 experimental test		Barker 1995
	Woodcock Johnson Reading Mastery Test Third Edition: Word Identification	Woodcock 1987	Barker 1995
	Wechsler Individual Achievement Test 2nd Edition	Wechsler 2002	Blythe 2006

Table 2. Tests used by studies to measure outcomes (*Continued*)

	Woodcock Johnson Reading Mastery Test Third Edition: Word Identification	Woodcock 2001	Ford 2009
	Woodcock Johnson Reading Mastery Test Revised: Word Identification	Woodcock 1987	Hurford 1994
	British Ability Scale: Word Reading	Elliot, Murray & Pearson, 1984	Hurry 2007
	1 experimental test		Levy 1997
	1 experimental test		Levy 1999
	2 experimental tests		Lovett 2000
	2 experimental tests		Lovett 1990
	1 experimental test		Savage 2003
Nonword reading accuracy	Woodcock Johnson Reading Mastery Test: Word Analysis	Woodcock 1987	Barker 1995
	Wechsler Individual Achievement Test 2nd Edition	Wechsler 2002	Blythe 2006
	Woodcock-Johnson Tests of Achievement Third Edition: Word Attack	Woodcock 2001	Ford 2009
	Woodcock Johnson Reading Mastery Test Revised: Word Attack	Woodcock 1987	Hurford 1994
	1 experimental test		Levy 1997
	1 experimental test		Levy 1999
	Woodcock Johnson Reading Mastery Test Revised: Word Attack	Woodcock 1987	Lovett 2000
	1 experimental test		Savage 2003
Word reading fluency	Test of Word Reading Efficiency: Phonemic Decoding Efficiency	Torgesen 1999	Ford 2009

Table 2. Tests used by studies to measure outcomes (*Continued*)

	2 experimental tests		Lovett 1990
Nonword reading fluency	Test of Word Reading Efficiency: Sight Word Efficiency	Torgesen 1999	Ford 2009
Reading comprehension	Wechsler Individual Achievement Test 2nd Edition	Wechsler 2002	Blythe 2006
	Gates-MacGinitie Reading Test: Comprehension	MacGinitie 2002	Ford 2009
	Neale Analysis of Reading Ability	Neale 1988	Hurry 2007
Spelling	2 experimental tests		Lovett 1990
	1 experimental test		Savage 2003
Letter-sound knowledge	1 experimental test		Lovett 1990
	1 experimental test		Savage 2003
	1 experimental test		Savage 2005
Phonological output (measured with phoneme awareness tasks)	1 experimental test		Barker 1995
	Goldman Fristoe Woodcock Test of Auditory Discrimination: Sound analysis	Goldman 1974	Lovett 2000
	1 experimental test		Savage 2003
	1 experimental test		Savage 2005

Table 3. Effect sizes for random- and fixed-effect model analyses, and heterogeneity for random-effects analyses

			Random-effects model			**Heterogeneity**			**Fixed-effect model**		
Outcome measure	**N studies**	**N Participants**	**SMD [95% CI]**	**Z**	**P**	**Chi^2**	**P**	**I^2%**	**SMD [95% CI]**	**Z**	**P**
Word reading accuracy	10	683	0.47 [0.06, 0.88]	2.22	0.03	52.26	< 0.01	83	0.46 [0.29, 0.62]	5.44	< 0.01

Table 3. Effect sizes for random- and fixed-effect model analyses, and heterogeneity for random-effects analyses (*Continued*)

Non-word reading accuracy	8	512	0.76 [0.25, 1.27]	2.91	< 0.01	44.04	< 0.01	84	0.82 [0.62, 1.01]	8.14	< 0.01
Word reading fluency	2	54	-0.51 [-1.14, 0.13]	1.56	0.12	1.30	0.25	23	-0.53 [-1.08, 0.02]	1.89	0.06
Non-word reading fluency	1	18	0.38 [-0.55, 1.32]	0.81	0.42	NA	NA	NA	0.38 [-0.55, 1.32]	0.81	0.42
Reading comprehension	3	173	0.14 [-0.46, 0.74]	0.45	0.65	4.27	0.12	53	0.01 [-0.31, 0.32]	0.04	097
Spelling	2	140	0.36 [-0.27, 1.00]	1.12	0.26	2.53	0.11	60	0.28 [-0.09, 0.65]	1.49	0.14
Letter-sound knowledge	3	192	0.35 [0.04, 0.65]	2.22	0.03	0.11	0.95	0	0.35 [0.04, 0.65]	2.22	0.03
Phonological output	4	280	0.38 [-0.04, 0.80]	1.77	0.08	7.97	0.05	62	0.44 [0.19, 0.70]	3.45	< 0.01

CI: confidence interval; SMD: standardised mean difference

Table 4. Characteristics of participants in each study

Study	Location	Age	Gender	IQ	Ethnicity	SES	Inclusion criteria	Exclusion Criteria	Population
Barker 1995	USA	Range 6.2 to 7.8 years	Not reported	Verbal Mean = 16.5 SD = 2.36	Not reported	Not reported	Students nominated by teachers from 2 elementary schools who were	None stated	First-grade students

Table 4. Characteristics of participants in each study (*Continued*)

							given a short series of pre-tests assessing phonological awareness skills and basic word recognition skills. These children were then given further 2 tests and those scoring below the 40th percentile and the 50th percentile on the subsequent test were selected		
Blythe 2006	Australia	Mean 101.5 months	Male 75% Female 25%	FSIQ-2 Mean 100.15 SD 9.38	Not reported	Not reported	Children who received group-based remedial reading instruction at school and were referred by a support teacher	After referral children completed the WISC-III FSIQ. Those who scored below the 20th percentile were excluded	Dyslexic primary school students
Ford 2009	USA	Mean 16.18 years	Male 55% Female 45%	Not reported	African-American 22%, Hispanic 67%, white 11%	Lower	Students who were enrolled in the remedial reading programme	None stated	Teenagers enrolled at an alternative high school, that is, a high

Table 4. Characteristics of participants in each study (*Continued*)

							were invited to participate. Below average reading skills were based on the ISAT		school for non-special education students or students at risk of dropping out
Hurford 1994	USA	Mean 80.35 months	Male 48% Female 52%	Mean 90.37	White 92.8, African-American 6%, Hispanic 5%, Asian-American 0.7%	Middle	Classification data from Hurford, Darrow, Edwards, Howerton, Mote, Schauf and Coffey (1993) was used with more relaxed criteria for eligibility, that is standard scores in reading of less than 91 were included rather than less than 86	None stated	Children at risk of reading disability
Hurry 2007	UK	Range 6 to 6.6 years	Male 61% Female 39%	Range 92 to 96	16% spoke English as a second language	42% of the sample were eligible for free school meals	In 63 schools the 6 poorest Year 2 readers were selected on the basis of their Diagnostic Survey (Clay, 1985) per-	The remaining children, that is, those less poor at reading then those that were selected for the experimen-	Children with reading difficulties

Table 4. Characteristics of participants in each study (*Continued*)

							formance. Of the 22 schools using Reading Recovery, the poorest scorers were offered intervention	tal condition, were assigned to a within school condition	
Levy 1997	Canada	Range 5.9 to 7.2 years	Male 48% Female 52%	Not reported	Not reported	Not reported	Children were given word reading tests, children that read fewer than 7 words on any of the screening tests were selected	None stated	All children from Grade 1 and senior kindergarten from 2 schools, whose parents consented to their participation
Levy 1999	Canada	Mean age 7.7 years	Male 56% Female 44%	Non-verbal (picture assembly) Experimental group mean 10.88 Control group mean 10.65	Mixed racial distribution	Covers all SES	Children were given a word identification test (WRAT-3), if they scored below 90 they were given another word identification test (WRMT) and if they read below half a grade below their grade level and read	None stated	17 schools participated in the screening process with permission for participation obtained from the board, schools and a parent or guardian

Table 4. Characteristics of participants in each study (*Continued*)

							no more than 15 of the training words then they were included in the sample		
Lovett 1990	Canada	Mean age 8.4 years SD 1.6 Range 7 to 13 years	Male 70.4% Female 29.6%	Verbal Mean 98.4 SD 10.6 Performance Mean 106.2 SD 12.6	Not reported	Middle	Children had to score below the 25th percentile on at least 4 of 5 reading measures used in the screening test and have at least low average intelligence	Children with English as a second language, history of extreme hyperactivity, hearing impairment, brain damage, a chronic medical condition, serious emotional disturbance, or attention deficits	Children referred to the Learning Disabilities Reading Program
Lovett 2000	Canada	Mean age 9.9 years SD 1.6 Range 7 to 13 years	Male 68.1% Female 31.9%	Verbal Mean 92 SD 13.7 Range 58 to 133 Performance Mean 98.7 SD 14.3 Range = 63 - 136	Not reported	Not reported	Children needed to demonstrate a 'substantial underachievement' on 4 of the 5 reading based screening assessments	None stated	Children with severe reading disabilities that were referred to the Clinical Research Unit for remediation

Table 4. Characteristics of participants in each study (*Continued*)

Savage 2003	UK	Mean age 5.9 years Range 5 to 6.3 years	Male 60% Female 40%	Not reported	Not reported	Not reported	Over 2 sessions a series of reading- and spelling-based assessments were used to find the poorest readers in Year 1 of the school. The lowest performers were recruited	A teacher identifying a child as being too immature to deal with working in small groups	Children with the lowest reading performance for their age within a Local Education Authority or School District
Savage 2005	UK	Not reported	Male 50% Female 50%	Not reported	Not reported	Lower	Over 2 sessions a series of reading- and spelling-based assessments were used to find the poorest readers in Year 1 of the school. The lowest performers were recruited	None stated	Children with the lowest reading performance for their age within a Local Education Authority or School District

FSIQ: Full Scale intelligence quotient; IQ: intelligence quotient; ISAT: Illinois State Achievement Test; SD: standard deviation; SES: socioeconomic status; WISC: Wechsler Intelligence Scale for Children; WRAT: Wide Range Achievement Test; WRMT: Woodcock Reading Mastery Test

Table 5. Allocation of studies to different subgroups (categories)

Subgroups		Barker 1995	Blythe 2006	Ford 2009	Hurford 1994	Hurry 2007	Levy 1997	Levy 1999	Lovett 1990	Lovett 2000	Savage 2003	Savage 2005

Table 5. Allocation of studies to different subgroups (categories) (*Continued*)

Training type	Phonics only	X	-	-	-	-	X	X	-	-	-	-
	Phonics + phoneme awareness	-	X	X	X	X	-	-	-	X	X	X
	Phonics + irregular words	-	-	-	-	-	-	-	X	-	-	-
Training intensity	< 2 hours/ week	X	X	X	X	X	X	X	-	-	X	X
	≥ 2 hours/ week	-	-	-	-	-	-	-	X	X	-	-
Training duration	< 3 months	X	X	X	-	-	X	X	X	X	X	X
	≥ 3 months	-	-	-	X	X	-	-	-	-	-	-
Training group size	1	-	X	X	X	X	X	X	-	-	-	-
	≤ 5	X	-	-	-	-	-	-	X	X	X	X
Training administrator	Human	-	-	-	-	X	X	X	X	X	X	X
	Computer	X	X	X	X	-	-	-	-	-	-	-

Table 6. Results of subgroup analyses

	Subgroups				Mean effect size			Heterogeneity			Subgroup analyses			
			N studies/ measures	N participants	SMD [95% CI]	Z	P	Chi^2	P	I^2 %	Chi^2	DF	P	I^2 %
Word reading accuracy	Training type	Phonics only	3	232	0.91 [-0.17, 1.98]	1.63	0.10	23.93	< 0.10	92				
		Phonics + phonem awareness	6	415	0. 28 [0. 00, 0. 56]	1.96	0.05	8.12	0.15	38	1.23	1	0.27	18.8
	Training intensity	< 2 hours/ week	8	559	0.48 [-0.04, 1.00]	1.80	0.07	50.65	< 0.10	86				
		≥ 2 hours/ week	2	124	0.34 [-0.03, 0.72]	1.79	0.07	1.06	0.30	6	0.17	1	0.68	0
	Training duration	< 3 months	8	498	0. 56 [0. 07, 1. 04]	2.25	0.02	39.20	< 0.10	82				
		≥ 3 months	2	185	0.12 [-0.43, 0.67]	0.42	0.67	2.8	0.09	64	1.36	1	0.24	26.3
	Training group size	1	6	419	0.62 [-0.06, 1.29]	1.78	0.07	44.35	< 0.10	89				
		≤ 5	4	264	0.25 [-0.04, 0.54]	1.67	0.10	8.78	0.29	12	0.94	1	0.33	0

Table 6. Results of subgroup analyses (*Continued*)

	Training administrator	Human	6	559	0.66 [0.08, 1.23]	2.24	0.03	46.02	< 0.10	89			
		Computer	4	124	0.15 [-0.20, 0.51]	0.85	0.40	2.63	0.45	0	2.13	1	0.14	53
Nonword reading accuracy	Training type	Phonics only	3	232	0.91 [-0.45, 2.28]	1.32	0.19	36.92	< 0.10	95			
		Phonics + phonem awareness	5	280	0.63 [0.38, 0.88]	4.86	< 0.10	1.84	0.88	0	0.16	1	0.69	0
	Training group size	1	5	284	1.06 [0.39, 1.73]	3.09	< 0.10	21.92	< 0.10	82			
		≤ 5	3	228	0.32 [-0.32, 0.96]	0.97	0.33	9.64	< 0.10	79	2.43	1	0.12	58.8
	Training administrator	Human	4	388	1.12 [0.48, 1.76]	3.42	< 0.10	22.23	< 0.10	87			
		Computer	4	124	0.31 [-0.33, 0.96]	0.96	0.34	8.65	0.03	65	3.02	1	0.08	66.8

APPENDICES

Appendix 1. Search strategies

Cochrane Central Register of Controlled Trials (CENTRAL), 2012, Issue 6, last searched 3 July 2012
#1 MeSH descriptor Reading, this term only
#2 MeSH descriptor Dyslexia, this term only
#3 (read* near/3 disorder*)
#4 (read* near/3 (abilit* or disab*))
#5 (read* near/3 impair*)
#6 (read* near/3 defic*)
#7 (read* near/3 delay*)
#8 (read* near/3 dysfunction*)
#9 (read* near/3 comprehen*)
#10 (read* near/3 accuracy)
#11(poor* near/3 read*)
#12((dysfluent or dysfluenc* or fluent or fluenc*) near/3 read*)
#13(slow* near/3 read*)
#14(remedial near/3 read*)
#15 dyslex*
#16(word NEXT blind* or wordblind*)
#17(#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16)
#18 MeSH descriptor Phonetics, this term only
#19 phonics
#20 phonem*
#21 phonolog*
#22 graphem*
#23 (lettersound* or letter NEXT sound*)
#24 letter NEXT identif*
#25 (sight NEXT word*)
#26 MeSH descriptor Remedial Teaching, this term only
#27 (remedial near/3 (teach* or method* or program*))
#28 (#18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27)
#29 (#17 AND #28)
Ovid MEDLINE(R) 1946 to June week 3 2012, last searched 3 July 2012
1 Reading/
2 (read$ adj3 disorder$).tw.
3 (read$ adj3 (abilit$ or disab$)).tw.
4 (read$ adj3 impair$).tw.
5 (read$ adj3 defic$).tw.
6 (read$ adj3 delay$).tw.
7 (read$ adj3 dysfunction$).tw.
8 (read$ adj3 comprehen$).tw.
9 (read$ adj3 accuracy).tw.
10 (poor$ adj3 read$).tw.
11 ((dysfluent or dysfluenc$ or fluent or fluenc$) adj3 read$).tw.
12 (slow$ adj3 read$).tw.
13 (remedial adj3 read$).tw.
14 dyslexia/
15 dyslex$.tw.
16 (word-blind$ or wordblind$).tw.
17 or/1-16
18 phonics.tw.

19 phonem$.tw.
20 phonolog$.tw.
21 graphem$.tw.
22 (lettersound$ or letter-sound$).tw.
23 letter identif$.tw.
24 (sight word$ or sight-word$).tw.
25 Phonetics/
26 Remedial Teaching/
27 (remedial adj3 (teach$ or method$ or program$)).tw.
28 or/18-27
29 17 and 28
30 randomized controlled trial.pt.
31 controlled clinical trial.pt.
32 randomi#ed.ab.
33 placebo$.ab.
34 drug therapy.fs.
35 randomly.ab.
36 trial.ab.
37 groups.ab.
38 or/30-37
39 exp animals/ not humans.sh.
40 38 not 39
41 29 and 40

EMBASE (Ovid) 1980 to 2012 week 26, last searched 3 July 2012

1 reading/
2 dyslexia/
3 (read$ adj3 disorder$).tw.
4 (read$ adj3 (abilit$ or disab$)).tw.
5 (read$ adj3 impair$).tw.
6 (read$ adj3 defic$).tw.
7 (read$ adj3 delay$).tw.
8 (read$ adj3 dysfunction$).tw.
9 (read$ adj3 comprehen$).tw.
10 (read$ adj3 accuracy).tw.
11 (poor$ adj3 read$).tw.
12 (read$ adj3 (fluent or fluenc$ or dysfluent or dysfluenc$)).tw.
13 (slow$ adj3 read$).tw.
14 (remedial adj3 read$).tw.
15 dyslex$.tw.
16 (word-blind$ or wordblind$).tw.
17 or/1-16
18 phonics.tw.
19 phonem$.tw.
20 phonolog$.tw.
21 graphem$.tw.
22 (lettersound$ or letter-sound$).tw.
23 letter identif$.tw.
24 (sight word$ or sight-word$).tw.
25 Phonetics/
26 (remedial adj3 (teach$ or method$ or program$)).tw.
27 or/18-26
28 17 and 27
29 exp Clinical trial/

30 Randomized controlled trial/
31 Randomization/
32 Single blind procedure/
33 Double blind procedure/
34 Crossover procedure/
35 Placebo/
36 Randomi#ed.tw.
37 RCT.tw.
38 (random$ adj3 (allocat$ or assign$)).tw.
39 randomly.ab.
40 groups.ab.
41 trial.ab.
42 ((singl$ or doubl$ or trebl$ or tripl$) adj3 (blind$ or mask$)).tw.
43 Placebo$.tw.
44 Prospective study/
45 (crossover or cross-over).tw.
46 prospective.tw.
47 or/29-46
48 28 and 47

DARE 2012(3), last searched 3 July 2012

#1 MeSH descriptor Reading, this term only
#2 MeSH descriptor Dyslexia, this term only
#3 (read* near/3 disorder*)
#4 (read* near/3 (abilit* or disab*))
#5 (read* near/3 impair*)
#6 (read* near/3 defic*)
#7 (read* near/3 delay*)
#8 (read* near/3 dysfunction*)
#9 (read* near/3 comprehen*)
#10 (read* near/3 accuracy)
#11(poor* near/3 read*)
#12((dysfluent or dysfluenc* or fluent or fluenc*) near/3 read*)
#13(slow* near/3 read*)
#14(remedial near/3 read*)
#15 dyslex*
#16(word NEXT blind* or wordblind*)
#17(#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16)
#18 MeSH descriptor Phonetics, this term only
#19 phonics
#20 phonem*
#21 phonolog*
#22 graphem*
#23 (lettersound* or letter NEXT sound*)
#24 letter NEXT identif*
#25 (sight NEXT word*)
#26 MeSH descriptor Remedial Teaching, this term only
#27 (remedial near/3 (teach* or method* or program*))
#28 (#18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27)
#29 (#17 AND #28)

ERIC (Proquest) 1966 to current, last searched 4 July 2012

Searched for:((SU.EXACT.EXPLODE("Basal Reading" OR "Beginning Reading" OR "Content Area Reading" OR "Corrective Reading" OR "Critical Reading" OR "Directed Reading Activity" OR "Early Reading" OR "Functional Reading" OR "Independent Reading" OR "Individualized Reading" OR "Music Reading" OR "Oral Reading" OR "Reading" OR "Reading Aloud to Others" OR "Read-

ing Fluency“ OR ”Recreational Reading“ OR ”Remedial Reading“ OR ”Silent Reading“ OR ”Speed Reading“ OR ”Story Reading“ OR ”Sustained Silent Reading“) OR SU.EXACT(”Dyslexia“) OR SU.EXACT(”Reading Difficulties“) OR ((slow* OR poor* OR remedial*) NEAR/5 read*) OR wordblind* OR ”word blind*“ OR word-blind*) AND (SU.EXACT(”Phonics“) OR SU.EXACT(”Phonological Awareness“) OR SU.EXACT(”Phonemic Awareness“) OR phonic* OR phonem* OR phonolog*or grapheme* OR ”letter identif*“))

ERIC (DialogDatastar) 1966 to current, last searched 31 May 2011

”(((READING#.W..DE.) OR ((READ$3 NEAR (DISORDER$ OR ABILITY OR DISABILIT$3 OR IMPAIR$4 OR DEFIC$5 OR DELAY$2 OR DYSFUNCTION$1)) .TI,AB.) OR (((DYSFLUEN$ OR FLUEN$) NEAR READ$3) .TI,AB.)OR (((SLOW$ OR POOR$ OR REMEDIAL$) NEAR READ$) .TI,AB.) OR (DYSLEXIA.W..DE.) OR(READING-DIFFICULTIES.DE.) OR ((WORDBLIND$ OR WORD-BLIND$ OR WORD ADJ BLIND$) .TI,AB.)) AND ((PHONICS.W..DE. OR PHONEMIC-AWARENESS.DE. OR PHONOLOGICAL-AWARENESS.DE.) OR ((PHONIC$ OR PHONEM$ OR PHONOLOG$) .TI,AB.) OR (GRAPHEME$.TI,AB.) OR ((LETTER ADJ IDENTIF$) .TI,AB.) OR ((SIGHT ADJ WORD$ OR SIGHT-WORD$ OR SIGHTWORD$) .TI,AB.) OR ((REMEDIAL ADJ READING) .TI,AB.) OR (REMEDIAL-READING.DE.) OR ((REMEDIAL NEAR (TEACH$ OR METHOD$ OR PROGRAM$)) .TI,AB.) OR ((LETTERSOUND$ OR LETTER-SOUND$ OR LETTER ADJ SOUND$) .TI,AB.)))

CINAHLPlus (EBSCOhost), 1938 to current**,** last searched 4 July 2012

S50 S31 and S49
S49 S32 or S33 or S34 or S35 or S36 or S37 or S38 or S39 or S40 or S41 or
S42 or S43 or S44 or S45 or S46 or S47 or S48
S48 (MH “Evaluation Research”) OR (MH “Summative Evaluation Research”) OR (MH “Program Evaluation”)
S47 (MH “Treatment Outcomes”)
S46 (MH “Comparative Studies”)
S45 (evaluat* study or evaluat* research) or (effectiv* study or effectiv* research) or (prospectiv* study or prospectiv* research) or (follow-up study or follow-up research)
S44 “cross over*”
S43 crossover*
S42 (MH “Crossover Design”) or (MH “Prospective Studies+”)
S41 (tripl* N3 mask*) or (tripl* N3 blind*)
S40 (trebl* N3 mask*) or (trebl* N3 blind*)
S39 (doubl* N3 mask*) or (doubl* N3 blind*)
S38 (singl* N3 mask*) or (singl* N3 blind*)
S37 (clinic* N3 trial*) or (control* N3 trial*)
S36 (random* N3 allocat*) or (random* N3 assign*)
S35 randomis* or randomiz*
S34 (MH “Meta Analysis”)
S33 (MH “Clinical Trials+”)
S32 MH random assignment
S31 S18 and S30
S30 S19 or S20 or S21 or S22 or S23 or S24 or S25 or S26 or S27 or S28 or S29
S29 remedial N3 teach* or remedial N3 method* or remedial N3 program*
S28 (MH “Remedial Teaching”)
S27 sight word* or sight-word* or sightword*
S26 letter identif*
S25 lettersound* or letter-sound* or letter sound*
S24 graphem*
S23 phonolog*
S22 phonem*
S21 phonics
S20 (MH “Phonetics+”)
S19 (MH “Phonology”)
S18 S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9 or S10 or S11 or S12 or S13 or S14 or S15 or S16 or S17
S17 word-blind* or wordblind* or word blind*
S16 dyslex*
S15 (remedial N3 read*)

S14 (slow* N3 read*)
S13 (slow* N3 read*)
S12 (poor* N3 read*)
S11 (read* N3 accura*)
S10 (read* N3 comprehen*)
S9 (read* N3 fluent) or (read* N3 fluenc*) or (read* N3 dysfluent) or (read* N3 dysfluenc*)
S8 (read* N3 dysfunction*)
S7 (read* N3 delay*)
S6 (read* N3 defic*)
S5 (read* N3 impair*)
S4 (read* N3 abilit*) or (read* N3 disab*)
S3 (read* N3 disorder*)
S2 (MH "Reading Disorders+")
S1 (MH "Reading+")

PsycINFO (Ovid) 1806 to June week 4 2012, last searched 4 July 2012

1 reading/ or oral reading/ or remedial reading/ or silent reading/
2 reading ability/ or reading achievement/ or reading comprehension/
3 reading development/ or reading disabilities/ or dyslexia/ or reading speed/
4 sight vocabulary/ or word recognition/
5 (read$ adj3 disorder$).mp.
6 (read$ adj3 (abilit$ or disab$)).mp.
7 (read$ adj3 impair$).mp.
8 (read$ adj3 defic$).mp.
9 (read$ adj3 delay$).mp.
10 (read$ adj3 dysfunction$).mp.
11 (read$ adj3 comprehen$).mp.
12 (read$ adj3 accuracy).mp.
13 (poor$ adj3 read$).mp.
14 ((dysfluent or dysfluenc$ or fluent or fluenc$) adj3 read$).mp.
15 (slow$ adj3 read$).mp.
16 (remedial adj3 read$).mp.
17 dyslex$.mp.
18 (word-blind$ or wordblind$).mp.
19 or/1-18
20 phonemes/ or phonetics/ or phonics/ or phonological awareness/ or phonology/
21 phonics.mp.
22 phonem$.mp.
23 phonolog$.mp.
24 graphem$.mp.
25 (lettersound$ or letter-sound$).mp.
26 letter identif$.mp.
27 (sight word$ or sight-word$).mp.
28 Remedial Reading/
29 (remedial adj3 (teach$ or method$ or program$)).mp.
30 or/20-29
31 clinical trials/
32 (randomis* or randomiz*).tw.
33 (random$ adj3 (allocat$ or assign$)).tw.
34 ((clinic$ or control$) adj trial$).tw.
35 ((singl$ or doubl$ or trebl$ or tripl$) adj3 (blind$ or mask$)).tw.
36 (crossover$ or "cross over$").tw.
37 random sampling/
38 Experiment Controls/

39 Placebo/
40 placebo$.tw.
41 exp program evaluation/
42 treatment effectiveness evaluation/
43 ((effectiveness or evaluat$) adj3 (stud$ or research$)).tw.
44 or/31-43
45 19 and 30 and 44

PsycINFO EBSCOhost, last searched 31 May 2011

S45 S30 and S44
S44 S31 or S32 or S33 or S34 or S35 or S36 or S37 or S38 or S39 or S40 or S41 or S42 or S43
S43 (evaluation N3 stud* or evaluation N3 research*)
S42 (effectiveness N3 stud* or effectiveness N3 research*)
S41 DE "Placebo" or DE "Evaluation" or DE "Program Evaluation" OR DE "Educational Program Evaluation" OR DE "Mental Health Program Evaluation" OR DE "Treatment effectiveness evaluation"
S40 (DE "Random Sampling" or DE "Clinical Trials") or (DE "Experiment Controls")
S39 placebo*
S38 crossover* or cross-over* or cross over*
S37 (tripl* N3 mask*) or (tripl* N3 blind*)
S36 (trebl* N3 mask*) or (trebl* N3 blind*)
S35 (doubl* N3 mask*) or (doubl* N3 blind*)
S34 (singl* N3 mask*) or (singl* N3 blind*)
S33 (clinic* N3 trial*) or (control* N3 trial*)
S32 (random* N3 allocat*) or (random* N3 assign*)
S31 randomis* or randomiz*
S30 S18 and S29
S29 S19 or S20 or S21 or S22 or S23 or S24 or S25 or S26 or S27 or S28
S28 (remedial N3 teach*) or (remedial* N3 method*) or (remedial* N3 program*)
S27 DE "Remedial Reading"
S26 sight word* or sight-word* or sightword*
S25 letter identif*
S24 lettersound* or letter-sound* or letter sound*
S23 graphem*
S22 phonolog*
S21 phonem*
S20 phonics
S19 ((DE "Phonics") OR (DE "Phonology")) OR (DE "Phonemes")) OR (DE "Phonetics") OR (DE"Phonological Awareness"))
S18 S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9 or S10 or S11 or S12 or S13 or S14 or S15 or S16 or S17
S17 (word-blind* or wordblind* or word blind*)
S16 DE "Dyslexia" or dyslex*
S15 (remedial N3 read*)
S14 (slow* N3 read*) or (poor* N3 read*)
S13 read* N3 comprehen*
S12 read* N3 accurac*
S11 dysfluent N3 read* or dysfluenc* N3 read* or fluent* N3 read* or fluenc* N3 read*
S10 (read* N3 dysfunction*)
S9 (read* N3 delay*)
S8 (read* N3 defic*)
S7 (read* N3 impair*)
S6 read* N3 disab* or read* N3 abilit*
S5 (read* N3 disorder*)
S4 DE "Sight Vocabulary" OR DE "Word Recognition"
S3 DE "Reading Disabilities" OR DE "Dyslexia" OR DE "Reading Speed" OR DE "Reading Development"
S2 DE "Reading Ability" OR DE "Reading Achievement" OR DE "Reading Comprehension"

S1 DE "Reading" OR DE "Oral Reading" OR DE "Remedial Reading" OR DE "Silent Reading"
Science Citation Index Expanded, Social Science Citation Index, CPCI-S, CPCI-SSH, last searched 4 July 2012
#11#10 AND #6 AND #1
#10 #9 OR #8 OR #7
9 TS=("sight word*" or sight-word*)
8 TS=(lettersound* or letter-sound* or "letter sound*" or "letter identif*")
7TS=(phonics or phonem* or phonolog* or graphem*)
6 #5 OR #4 OR #3 OR #2
5TS=(wordblind* or word-blind* or "word blind*")
4TS=(dyslexia or dyslexic*)
3TS= (READ* SAME (accuracy or comprehen* or disorder* or disab* or abilit* or impair* or defic* or delay* or dysfunction* or dysfluen* or fluen*))
2 TS= ("slow read*" or "remedial read*" or "poor read*")
1 TS=(random* or control* or trial* or group* or effectiveness or evaluation or placebo*)
ZETOC , last searched 4 July 2012
Search terms: conference: reading phonics
metaRegister of Controlled Trials, last searched 4 July 2012
(reading or dyslexia) AND (phonics or phonology or phonetics)
WHO ICTRP, last searched 4 July 2012
CONDITION reading OR dyslexia
INTERVENTION : phonics OR phonetics OR phonology
ClinicalTrials.gov, last searched 4 July 2012
phonetics OR phonology OR phonics | reading OR dyslexia
Index to Theses in UK and Ireland (Proquest), last searched 4 July 2012
(read* or dyslex* or wordblind* or "word-blind*" or "word blind*") AND (phonics or phonem* or phonolog* or graphem* OR lettersound* or "letter-sound*" or "letter sound*" or "letter identif*" or "sight word*" or "sight-word*")
ProQuest Dissertations and Theses, last searched July 2012
1. Reading/
2. (read$ adj3 disorder$).tw.
3. (read adj3 (ability or disab)) .tw.
4. (read$ adj3 impair$).tw.
5. (read$ adj3 defic$).tw.
6. (read$ adj3 delay$).tw
7. (read$ adj3 dysfunction$).tw.
8. (poor$ adj3 read$).tw.
9. (dysfluen$ adj3 read$).tw.
10. (slow$ adj3 read$).tw.
11. (remedial adj3 read$).tw.
12. dyslexia/
13. dyslex$.tw.
14. (word-blind$ or wordblind$).tw.
15. or/1-14
16. phonics.tw.
17. phonem$.tw.
18. phonolog$.tw.
19. graphem$.tw.
20. (lettersound$ or letter-sound$).tw.
21. letter identif$.tw.
22. (sight word$ or sight-word$).tw.
23. Phonetics/
24. Remedial Teaching/
25. (read$ adj3 (teach$ or method$ or program$)).tw.
26. or/16-25

27. 15 and 26
28. randomized controlled trial.pt.
29. controlled clinical trial.pt.
30. randomi#ed.ab.
31. placebo$.ab.
32. drug therapy.fs.
33. randomly.ab.
34. trial.ab.
35. groups.ab.
36. or/28-35
37. exp animals/not humans.sh.
38. 36 not 37
39. 27 and 36

DART Europe E-theses Portal (www.dart-europe.eu), Australasian Digital Theses program (adt.caul.edu.au), Education Research Theses (http://www.acer.edu.au/library/theses), Electronic Theses online (http://www.ndltd.org/serviceproviders/scirus-etd-search), Networked Digital Library of Theses and Dissertations (ETD; www.ndltd.org), Theses Canada Portal (www.collectionscanada.gc.ca/thesescanada), www.dissertation.com, www.thesisabstracts.com, last searched July 2012

1. dyslexia
2. reading disorder
3. reading disability
4. reading impairment
5. reading deficit
6. reading delay
7. reading dysfunction
8. poor reader
9. poor reading
10. dysfluent reader
11. dysfluent reading
12. slow reader
13. slow reading
14. remedial reader
15. word-blind
16. wordblind
17. phonics
18. phoneme
19. phonological
20. grapheme
21. lettersound or letter-sound
22. letter identification
23. sight word or sight-word
24. phonetics
25. remedial teaching
26. reading teaching
27. reading methods
28. reading program

HISTORY

Protocol first published: Issue 5, 2011

Review first published: Issue 12, 2012

CONTRIBUTIONS OF AUTHORS

All authors have been involved in designing the methodology; in extracting, analysing, and reporting data, and in checking and revising content.

DECLARATIONS OF INTEREST

All authors are currently involved in an quasi-RCT that is comparing the effect of phonics training versus sight-word training in children with developmental dyslexia. There are no other conflicts of interest. Funds from Australian Research Council (ARC) and National Health and Medical Research Council (NHMRC) paid a small part of the wages of all the authors except for Anne Castles. These funds were provided for research activities in general, and not specifically for doing this review.

SOURCES OF SUPPORT

Internal sources

- Macquarie University, Australia.

Funds for the salaries of McArthur, Castles, Larsen, and Marinus

External sources

- National Health and Medical Research Council (NHMRC) Project Grant (488518), Australia.

Funds the salaries of Kohnen, Jones, and Banales

- Australian Research Council (ARC) Discovery Project Grant (DP0879556), Australia.

Funds for the salaries of McArthur, Anandakumar, and Larsen

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

1. Two new review authors have been added (Pip Eve and Huachen Wang).

2. The name of the institution (ARC Centre of Excellence of Cognition and its Disorders) has been updated (previously Macquarie Centre for Cognitive Science).

3. Figure 1 of the dual route model has been removed.

4. We have revised the Background slightly to (1) improve clarity in general, (2) provide a clearer definition of phonics training, (3) provide a clearer explanation for why it is important to review simple phonics training programmes rather than complex programmes.

5. The definition of the intervention changed from "phonics training of any duration, intensity or mode of delivery, with or without concomitant training in letter identification, sight-word reading, or phoneme awareness" to "phonics program that trained reading via the letter-sound rules alone (phonics only) or with one other type of training (that is, a programme that combined phonics with phoneme awareness or irregular word reading)".

6. We modified our inclusion criteria for the intervention to include studies where the control group did the same training as the intervention group minus the phonics component (that is, phoneme awareness training or irregular word reading training).

7. We updated 'Types of participants' to include criteria relating to age, gender, and SES.

8. In the protocol for this review, we planned to calculate separate effects for accuracy for nonwords, regular words, and irregular words. We also planned to calculate separate effects for fluency for nonwords, regular words, and irregular words. However, given the paucity of data for these separate effects, this review combined measures for regular and irregular word reading accuracy into 'word reading accuracy', and combined regular and irregular word reading fluency into 'word reading fluency', to improve the power of the meta-analyses of the reading measures.

9. We had also planned to include letter identification, parsing, and blending in this review. However, no studies reported data for these measures, and so they could not be included in this review.

10. We merged 'spoken word production' and 'other phoneme awareness abilities' into 'phonological output' since these were tested with similar measures (that is, phoneme awareness tests).

11. We removed Dissertation Abstract Online from 'Searching other resources' because our institution no longer subscribes to this database.

12. We modified the list in the 'Subgroup analysis and investigation of heterogeneity' section to accurately reflect the seven subgroup analyses planned for this review. This involved (1) removing 'reading skill' as a subgroup (since this moderator was tested by comparing primary and secondary outcomes); (2) adding training duration; (3) redefining concomitant training (letter identification, irregular words, reading fluency, reading comprehension, phoneme awareness, spoken vocabulary) to 'training type (phonics alone, phonics and phoneme awareness, phonics and irregular words). We also changed the terms slightly (for example, 'training' instead of 'treatment') for the sake of clarity.

NOTES

None.

INDEX TERMS

Medical Subject Headings (MeSH)

*Reading; Dyslexia [*rehabilitation]; Language; Phonation [*physiology]; Randomized Controlled Trials as Topic

MeSH check words

Adolescent; Adult; Child; Female; Humans; Male