Research Papers in Education

ISSN: 0267-1522 (Print) 1470-1146 (Online) Journal homepage: https://www.tandfonline.com/loi/rred20

Phonics: reading policy and the evidence of effectiveness from a systematic 'tertiary' review

Carole Torgerson, Greg Brooks, Louise Gascoine & Steve Higgins

RESEARCH PAPERS IN EDUCATION
2019, VOL. 34, NO. 2, 208–238
https://doi.org/10.1080/02671522.2017.1420816

Phonics: reading policy and the evidence of effectiveness from a systematic 'tertiary' review

Carole Torgerson[a], Greg Brooks[b], Louise Gascoine[a] and Steve Higgins[a]

[a]School of Education, Durham University, Durham, UK; [b]School of Education, University of Sheffield, Sheffield, UK

ABSTRACT

Ten years after publication of two reviews of the evidence on phonics, a number of British policy initiatives have firmly embedded phonics in the curriculum for early reading development. However, uncertainty about the most effective approaches to teaching reading remains. A definitive trial comparing different approaches was recommended in 2006, but never undertaken. However, since then, a number of systematic reviews of the international evidence *have* been undertaken, but to date they have not been systematically located, synthesised and quality appraised. This paper seeks to redress that gap in the literature. It outlines in detail the reading policy development, mainly in England, but with reference to international developments, in the last 10 years. It then reports the design and results of a systematic 'tertiary' review of all the relevant systematic reviews and meta-analyses in order to provide the most up-to-date overview of the results and quality of the research on phonics.

ARTICLE HISTORY
Received 6 July 2017
Accepted 17 December 2017

KEYWORDS
Phonics; reading policy; systematic review

Introduction

Improving standards of literacy through education and schooling in particular is a shared objective for education globally. An increased policy focus on standards of literacy is also evident (e.g. Schwippert and Lenkeit 2012), as well as on methods of initial teaching. In the initial teaching of reading in languages with highly consistent orthographies (e.g. Spanish and especially Finnish), phonics is used without comment or dispute as the obvious way to give children who are not yet reading the most effective method of 'word attack', identifying unfamiliar printed words. The teaching of early reading in English, by contrast, has been highly politicised and is contentious, largely because of its notoriously complex set of grapheme–phoneme correspondences. In the United States (US), the so-called 'reading wars' have seen phonics approaches set against whole language approaches in decades of debate. While there have been what might be called 'reading skirmishes' in the United Kingdom (UK), they do not seem to have reached the same level of acrimony.

In 2007, British Government policy on how children should be taught to read changed. Until 2006, within the statutory National Curriculum (NC) for the teaching of English

CONTACT Carole Torgerson carole.torgerson@durham.ac.uk
This article was originally published with errors. This version has been corrected. Please see Erratum (https://doi.org/10.1080/02671522.2018.1429230).

in state schools in England, the National Literacy Strategy recommended the so-called 'searchlights' model for teaching reading which was a 'mixed methods' approach, including embedded phonics, but also drawing on other approaches. From 2007 onwards, exclusive, intensive, systematic, explicit synthetic phonics instruction was adopted nationally. Also, and significantly, in 2007 this sentence: 'Children will be encouraged to use a range of strategies to make sense of what they read' was removed from the NC.

In 2006, two reviews on the teaching of reading funded by the Department for Education and Skills (D*f*ES) were published using alternative designs: a systematic review (SR) undertaken by two of the authors of this paper and a colleague (Torgerson, Brooks, and Hall 2006) and an expert review undertaken by Rose (2006). The SR used explicit transparent replicable methods, with systematic identification and inclusion of studies employing strong designs which can establish causal relationships between interventions and outcomes (randomised controlled trials or RCTs), minimisation of bias at every stage in the design and methods of the review, and assessment of the quality of the evidence base *before* coming to any conclusions. In contrast, the Rose Review did not use explicit methods for identification of studies to include and did not assess the quality of the evidence base, despite acknowledging the limitations of the UK-based trials (Rose 2006, 61, paragraphs 204 and 207) included in his review.

In our systematic review, we found 12 individually randomised controlled trials; all were very small and only one was from the UK. In a meta-analysis, we found a small, statistically significant effect on reading accuracy, which we judged was derived from *moderate* weight of evidence, due to the relatively small number of trials and their variable quality. All the included studies integrated phonics with whole text level learning – in other words the phonics learning was not discrete. Our main recommendation was that systematic phonics instruction should be part of every literacy teacher's repertoire and a routine part of literacy teaching *in a judicious balance with other elements*. The difficulty of making policy recommendations for teaching reading is that such a 'judicious balance' may be disrupted by policy decisions that lack a reliable evidence base.

Background

The policy context: phonics in the NC for English in England

There have been three recognisable phases in the policy context in England since 1989. It should be noted that these apply only to England; Northern Ireland, Scotland and Wales have devolved responsibility for education.

Phase 1: making phonics statutory

A NC for English in state schools in England was introduced in 1989, and there have been three subsequent versions (1995, 1999 and 2013). All covered the compulsory education years (ages 5–16), but only the sections for the primary years (ages 5–11) are relevant here. The first edition made just one reference to phonics: 'Pupils should be able to … use picture and context cues, words recognised on sight and phonic cues in reading' (Department of Education and Science 1989, 7). This appeared to place phonics on a par with other 'cue' systems for word recognition, even though those are little better than guessing since they often lead to learners producing words other than the target (see, in particular, Stanovich

2000). Teaching children to rely on phonics to identify unfamiliar words would be more efficient.

Debate about the role and value of phonics was fuelled by the second (1989) edition of Chall's seminal *Learning to Read: The Great Debate* (1967), and by Adams' (1990) similarly comprehensive review; both concluded that phonics instruction enables children to make faster progress in (some aspects of) reading than no phonics or meaning-emphasis approaches, especially if applied to meaningful texts. Accordingly, the second edition of the NC (DfE 1995, 6–7) provided significantly more detail on phonics, while still giving a list of the 'key skills' for early reading that was essentially the same as in NC Mark 1. However, the essential terms for defining the process of phonics, namely 'phoneme' and 'grapheme', were not even mentioned, let alone the necessary underpinnings in phonetics and analysis of grapheme–phoneme correspondences.

To support NC Mark 2, the National Literacy Strategy (NLS) was rolled out from 1997. The NLS *Framework for Teaching* (D*f*EE 1998) at last introduced the term 'phoneme', but still portrayed phonics as just one of its 'searchlights' strategies for identifying words and comprehending text, the others being much the same as in NC Mark 1 and 2.

In the third edition of the NC (D*f*EE 1999, 46) the amount of detail on phonics was much the same as in the second edition, but more focused, including using 'phoneme'. Shortly afterwards, reports from the NRP (2000) and its phonics subgroup (Ehri et al. 2001) appeared in the US, and slowly began to influence research and practice in Britain.

In its report on the first four years of the NLS, the Office for Standards in Education, Children's Services and Skills (Ofsted 2002) praised some aspects of the teaching of phonics in primary schools in England but criticised others; even the fact that they could do this showed that there was more, and more focused, phonics teaching than a decade earlier. A set of support materials, *Playing with Sounds* (D*f*ES 2004), was published soon afterwards. In a period of 15 years, therefore, phonics had moved from virtual invisibility to being a central concern, with statutory backing and professional guidance.

Phase 2: which variety of phonics?

Johnston and Watson (2004) reported on two studies in Scotland comparing synthetic and analytic phonics. Experiment 1, which was not an RCT but a quasi-experiment, compared a synthetic phonics group with two analytic phonics groups and found an advantage for the synthetic phonics group, *but* this group had received training at a faster pace than the others, and 5 of the 13 whole classes involved had been allocated by the researchers to receive synthetic phonics according to their perceived greater need.

Experiment 2, which was actually conducted *before* Experiment 1, also compared synthetic phonics and analytic phonics and found a positive effect for synthetic phonics, *but* one researcher taught both groups, and the researchers did not report their method of randomisation or their sample size calculation, did not undertake intention to treat analysis (the correct analysis, keeping children in their originally allocated groups), and did not use blinded assessment of outcome.

Despite these methodological flaws, publicity for Experiment 1 (Experiment 2 received very little) led many to believe that synthetic phonics had the edge, and attracted sufficient political attention for a parliamentary committee to hold an enquiry into teaching children to read in 2004–2005; its report (House of Commons Education and Skills Committee 2005) appeared in the spring of 2005. In quick succession thereafter the British

Government: commissioned the systematic review of the research evidence on phonics (Torgerson, Brooks, and Hall 2006) which is the precursor of this 'tertiary' review; set up the Rose Review, which concentrated on good practice in the teaching of reading, including in the use of phonics, and reported in early 2006 (Rose 2006); established a pilot project on synthetic phonics to begin in 2005; and commissioned the *Letters and Sounds* framework for phonics teaching which the D*f*ES itself published (D*f*ES 2007).

In 2006, we built on the systematic review which had appeared in the US (Torgerson, Brooks, and Hall 2006). Ehri et al. (2001; see especially 393) had analysed data from both RCTs and quasi-experiments; they concluded that systematic phonics instruction enabled children to make better progress in reading than instruction featuring unsystematic or no phonics. However, they also concluded that there was no evidence to show that any particular form of phonics was superior to any other form of phonics. Using only RCTs, including the first from Britain (Experiment 2 of Johnston and Watson 2004), found firm evidence that systematic phonics instruction enables children to make better progress in *word recognition* than unsystematic or no phonics instruction, but not enough evidence to decide whether (a) systematic phonics instruction enables children to make better progress in *comprehension*, or (b) whether synthetic or analytic phonics is more effective (Johnston and Watson's Experiment 2 was one of only three relevant RCTs).

Our first conclusion was welcome to the Rose committee, but not the second or particularly the third. However, Jim Rose and colleagues who made classroom observation visits in 2005 concluded that synthetic phonics is more effective. Rose's (2006) conclusion that systematic phonics equates with synthetic phonics was seized upon by opponents as going beyond the evidence – see, for example, the debate in *Literacy*, vol.41, no.3 (Brooks et al. 2007). Though some opposition to phonics is still reported (e.g. most recently Krashen 2017), some of it based on the misapprehension that there is a forced choice between phonics and whole-language approaches, that controversy seemed to die down within a few years, and the place of phonics as part of the initial teaching of literacy now seems largely accepted in England.

The rational way to investigate the relative effectiveness of synthetic and analytic phonics would have been to conduct a large and rigorous RCT (as advocated by us in 2006: see Torgerson, Brooks, and Hall 2006, 12). Instead, the pilot project on synthetic phonics alone, known as *The Early Reading Development Pilot*, began in the school year 2005/2006 in 172 schools in 18 Local Authorities (LAs). Although no separate report on that pilot seems ever to have been published, a decision was evidently taken in central government to roll synthetic phonics out nationally, and this was carried out in successive batches of LAs between 2006/2007 and 2009/2010, under the title *The Communication, Language and Literacy Development Programme*.

The results of these programmes seem to have been analysed and published only with the appearance of a report by Machin, McNally, and Viarengo (2016), who also had access to national pupil attainment data at ages 5, 7 and 11. Using the staggered roll-out to define quasi-'treatment' and 'control' groups, the authors were able to estimate the effect of introducing synthetic phonics on children's attainment at all three ages. They concluded that there had been an across-the-board improvement at ages 5 and 7, but that at age 11 there was no average effect – however, there were lasting effects for children who could be considered as having been at risk of underachievement initially (children who entered school at risk of falling behind, those who were from disadvantaged backgrounds, and non-native speakers of English – precisely the groups one would hope would benefit) (Machin, McNally, and Viarengo 2016). This result means that there would have been a negative effect for the remaining children as there was no average overall effect.

The Rose report had contained a set of criteria for judging phonics teaching schemes, and in 2007–2010 the D*f*ES supported two different panels providing quality assurance of publishers' claims about their schemes against those criteria (see Beard, Brooks, and Ampaw-Farr forthcoming); one of the mainly initial schemes judged was *Letters and Sounds*.

The Rose review also contained, in an appendix, a version of the 'Simple View of Reading' (Gough and Tunmer 1986) by Morag Stuart, which she elaborated in Stuart (2006). This theory portrays reading comprehension as the product of language (listening) comprehension and the decoding of printed words, and holds that these dimensions can (largely) vary independently and that both decoding and comprehension require explicit teaching. In the Primary National Strategy (D*f*ES 2006), which had incorporated the NLS, this model of reading processes replaced the 'Searchlights' model.

So far, so largely similar, it would seem, to developments in other English-speaking countries. There was little remaining opposition to the use of phonics in initial literacy teaching, the Simple View of Reading had become the predominant model, and synthetic phonics had become the favoured variety, as later advocated and analysed in Stuart and Stainthorp (2016). But in England, there was to be a significant further policy turn which does not seem to have been matched elsewhere and has caused renewed controversy.

Phase 3: putting a strong official push behind synthetic phonics

There have been significant developments since the change of government in 2010. A third panel providing the DfE with quality assurance of publishers' claims about their phonics schemes operated in 2010–2012; one of the criteria was re-worded to require that schemes be synthetic. Commercial publishers had to re-submit their schemes, and some which had passed the scrutiny of the earlier panels failed this time (see again Beard, Brooks, and Ampaw-Farr, forthcoming). Almost half the roughly 100 schemes evaluated failed because they contained basic linguistic and/or phonetic errors (e.g. confusing graphemes and phonemes, or diphthongs and digraphs).

From September 2011 to October 2013, if schools ordered schemes which met the revised criteria and were therefore on an 'approved list' (in the form of a phonics catalogue on the DfE website), they could receive match funding from the DfE. In September 2014 there were just 10 full synthetic phonics schemes, and 15 sets of supplementary resources, on the DfE's approved list (DfE 2014).

The most important development after the change of government was the introduction of the 'phonics screening check' for Year 1 pupils, which was piloted in the summer term 2011 and has been implemented nationally in each summer term since 2012 (for the background, see DfE 2011). This individually administered 'check', which is a test in all but name, was promoted as 'telling parents how well their children are getting on with learning to read', and consists of 40 letter-strings to be read aloud; half are real words, the rest non-words designed to assess whether children have mastered the grapheme–phoneme correspondences (GPCs) without which they would not be able to vocalise these items. Children who score below the 'threshold' or pass mark (32 correct out of 40) receive extra instruction during Year 2, and at the end of that year are re-tested; most pass on this second attempt, but some do not, and are not re-tested again in Year 3; nor is there (apparently) any further centrally directed support for them. The test continues in force despite vocal opposition and a detailed analysis (Darnell, Solity, and Wall 2017) showing that some items require word knowledge in addition to ability to use GPCs, and that some GPCs listed in the government's specification are not in fact tested.

Meanwhile, a new version of the NC was published in 2013 for implementation in 2014. It is worth quoting its two main statements on phonics:

> [Year 1] Pupils should be taught to: apply phonic knowledge and skills as the route to decode words; respond speedily with the correct sound to graphemes (letters or groups of letters) for all 40+ phonemes, including, where applicable, alternative sounds for graphemes; read accurately by blending sounds in unfamiliar words containing GPCs [grapheme–phoneme correspondences] that have been taught … (DfE 2013, 20)

> [Other relevant information includes:] 'Skilled word reading involves both the speedy working out of the pronunciation of unfamiliar printed words (decoding) and the speedy recognition of familiar printed words. Underpinning both is the understanding that the letters on the page represent the sounds in spoken words. This is why phonics should be emphasised in the early teaching of reading to beginners (i.e. unskilled readers) when they start school.' (DfE 2013, 4)

The first of these paragraphs contains a clear and distinctive summary of synthetic phonics for reading, and both paragraphs correctly define its use as being the identification of *unfamiliar* printed words. Taken with other statements in the curriculum concerning synthetic phonics for spelling (e.g. 29) and for reading in Year 2, the notion that phonics should effectively be complete by the end of Year 2, and the comprehension and enjoyment of reading, this is a balanced view. However, the curriculum also contains an appendix (49–73) laying out in great detail the principal phoneme–grapheme and grapheme–phoneme correspondences of British English spelling relative to the RP (Received Pronunciation) accent (with a few notes on regional variation, e.g. in the pronunciation of words like *bath* and *past*), and providing a key to the International Phonetic Alphabet symbols used (73). While this knowledge appears essential for teachers to ensure accurate phonics teaching, the contrast with the exiguous earlier specifications of phonics is stark.

The overall picture of phonics in the NC for English in England is therefore of an initial tentative phase, followed by the deliberate choosing of synthetic phonics before research evidence justified this, and now firm government pressure to ensure the implementation of that variety of phonics. How accurate that implementation is remains to be investigated, as does its continued effectiveness. The Machin, McNally, and Viarengo (2016) findings are based on data from 2004 to 2011, and therefore pre-date both the Year 1 phonics test and NC Mark 4, with its highly detailed specifications. At the time of writing there is no sign that phase 3 has an end.

Rationale for the tertiary review

Ten years after the publication of our systematic review (Torgerson, Brooks, and Hall 2006), the reading skirmishes are alive and well, and the UK-based RCT we recommended has never been undertaken. However, a number of SRs and meta-analyses (and methodological re-analyses of existing meta-analyses) *have* been undertaken since 2006, and a tertiary review is particularly helpful where a number of overlapping systematic reviews have been undertaken in a given topic area (as is the case with phonics) in order to explore consistency across the results from the individual reviews. A synthesis of the findings of these studies provides a more complete picture of the evidence for the effectiveness of phonics (or alternative) reading approaches in terms of a pooled effect size or narrative synthesis of quantified outcomes of the extant SRs, and is more robust than simply looking at individual systematic reviews, small scale RCTs or a *non*-systematic synthesis of previous SRs.

Design and methods

The most scientific approach to searching for, locating, quality appraising and synthesising all the relevant systematic reviews in a tertiary review is to use systematic review design and methods: an exhaustive and unbiased search; minimisation of bias at all stages of inclusion; data extraction and quality appraisal because this increases the overall reliability in the findings. We aimed to explore the consistency (or lack) of the findings across the full range of the located reviews. In addition, we wanted to look at methodological challenges with respect to: the quality of the reviews; publication bias; and the difference in results depending on both the designs and the statistical models used in the included studies.

We used SR methods at all stages of the tertiary review, including applying strict quality assurance procedures to ensure rigour and, consequently, to increase confidence in our results.

Primary research questions

What is the effectiveness of systematic phonics instruction compared with alternative approaches, including whole language approaches or different varieties of phonics on reading accuracy, comprehension and spelling; and what is the quality of the evidence base on which this judgement is formed?

Secondary research questions

Does the evidence for effectiveness vary by design and/or statistical model for effect size calculation? Is there evidence of publication bias in the included systematic reviews, and consequently in the tertiary review itself?

Inclusion/exclusion criteria

We established inclusion criteria prior to starting the search for studies. As a minimum, included SRs had to provide evidence of the three key items of a SR for an effectiveness question, namely: a systematic search primarily using electronic databases; quality appraisal of all included studies; and a quantified synthesis or meta-analysis giving pooled effect sizes. Systematic reviews also had to include studies using a rigorous design that is able to establish causal relationships between interventions and outcomes – experimental or quasi-experimental designs (RCTs and/or QEDs). In terms of interventions, we included reviews of studies evaluating the effectiveness of phonics interventions compared with whole-language interventions or alternative approaches, including different varieties of phonics instruction (synthetic or analytic). In terms of outcomes, we included reviews of studies that included any combination of any standardised reading and spelling outcomes.

Searching

The search strings were based on relevant key words and their derivatives. For example, in ASSIA, ERIC and PsycINFO they were as follows:

> (phonic* OR phonetical* OR phonemic) AND (systematic review OR meta-analysis OR research synthesis OR research review)

See Appendix 1 for the full search strategies for all databases searched in 2014 and 2016.

We searched exhaustively (from 2001) for all the potentially relevant systematic reviews, containing meta-analyses with pooled effect sizes. The databases searched were: Applied Social Sciences Index and Abstracts (ASSIA), Education Resources Information Centre (ERIC), PsycINFO, Web of Science and World Cat. Searches were undertaken in 2014 and 2016.

Screening at first and second stages

We screened the titles and abstracts (first stage) and full papers (second stage) for inclusion using pre-established inclusion criteria. Independent double screening ensured a robust approach to this process.

Data extraction and quality appraisal

All included systematic reviews/meta-analyses were independently data-extracted and quality-appraised using specifically designed templates by two pairs of reviewers, who then conferred and agreed a final version. The template for data extraction included substantive items: details about the nature of included interventions and control conditions; number and designs of included studies; participants and settings; and outcome measures and results. The template for quality appraisal of included SRs included methodological items of the included SRs from the PRISMA checklist (Moher et al. 2009), including: methods for each stage of the review, including assessment of risk of bias within and across studies. We also extracted onto specifically designed templates data to enable us to investigate the potential for both publication bias and design bias.

Results

Results of searching

After de-duplication, there were 369 hits for the 2014 searches and 83 hits for the 2016 update. In total, we included 452 potentially relevant studies from the electronic searching. Table 1 and the PRISMA diagram in Appendix 1 show the results from searching all the databases at the two time points.

Results of screening

After screening of titles and abstracts and full papers we included a total of 12 studies. Table 2 and the completed PRISMA diagram in Appendix 1 show the results from screening at

Table 1. Results from 2014 and 2016 searches after de-duplication.

Database	2014 number of hits	2016 number of hits
Applied Social Sciences Index and Abstracts (ASSIA) (ProQuest)	11	1
Education Resources Information Centre (ERIC) (ProQuest)	132	10
PsycINFO (Ebscohost)	46	12
Web of Science (Web of Knowledge)	71	41
World Cat (First Search, OCLC)	109	19
Total	369	83

Table 2. Screening results from combined 2014 and 2016 searches.

Database searched	Number of records (Number of records after de-duplication)	Number of studies after 1st screening	Number of studies excluded	Number of studies after 2nd screening
(ASSIA)	12 (12)	7	4	3
ERIC	151 (142)	18	14	5
PsycINFO	79 (58)	7	6	1
Web of Science	167 (112)	12	9	5
World Cat	170 (128)	4	3	2
Total	579 (452)	48	36	12

both stages. We found a total of 12 studies that met our inclusion criteria for the period 2001–2016.

Results of quality assurance of screening

Initial agreement between the two authors who screened the entire database was high at both first and second stages. Any disagreements were resolved through discussion.

Results: characteristics and quality of SRs/meta-analyses

In Table 3, we summarise the main characteristics of the 12 SRs. Half (6) were undertaken in the United States, with one each in the United Kingdom and Australia, three in Germany, and one jointly in the US and Canada. Although many of the SRs focused solely on the effectiveness of phonics interventions compared with control or comparison conditions, a number looked more broadly at a range of strategies to improve reading and spelling, with phonics instruction as a sub-category (see Table 3 for specific phonics interventions).

Most of the studies provided enough detail of the interventions included to show that almost all of those labelled 'phonics' were indeed investigating approaches to the teaching of reading and spelling which focus on letter-sound relationships, i.e. the association of phonemes with graphemes. However, Adesope et al. (2011) were vague on this point, and McArthur et al. (2012) used such a narrow definition of 'pure' phonics that only three studies qualified. Galuschka et al. (2014) and Han (2010) included pedagogies which would not qualify as phonics by any reasonable professional definition – it is therefore questionable whether they should have been included in this review. Other authors may also have

Table 3. Characteristics of the included systematic reviews/meta-analyses.

Author, date, country	Aims of intervention(s) included in SR/meta-analysis	Phonics interventions	Number of studies included	Design(s) of studies	Settings and participants	Outcome measures	Results, as reported by authors	Conclusions, as reported by authors
Adesope et al. (2011), US & Canada	To improve literacy skills (via different strategies) for ESL immigrant students	'Systematic' but no further details beyond general definition of phonics	Total: 26 studies (in 20 papers)	Experimental and QE studies (do not state which are which)	ESL students in English- speaking countries.	Reading and writing (comprehension, mixed comprehension and decoding, decoding), bi-literacy, vocabulary acquisition (reading and writing). Studies where speaking was the only outcome measure were excluded. Does not state which outcome measures were specifically for phonics	Evidence to support systematic phonics instruction ($g = +0.40$), but systematic phonics instruction did not produce the largest effects (629)	Systematic phonics instruction does have the 'potential to enhance the teaching of English literacy to ESL immigrant students' (648)
		Control groups received 'traditional methods' (unspecified)	Phonics: 5 studies		Non-clinical population		'The results show that the pedagogical strategies examined in this meta-analysis produced statistically significant benefits for students in all grade levels'	Variability of moderator analysis may not be representative of the population – limiting the certainty of conclusions drawn
					Age: Kindergarten – Grade 6			
Camilli, Vargas, and Yurecko (2003), US	To improve reading and spelling skills	As Ehri et al. (2001), though eventually deconstructed & supplemented	40 (Ehri et al. 2001; 's 38 – 1 + 3)	RCTs & QEs	Schools. Children aged 5–11 (K-G6), normally achieving, at risk, reading disabled, or low achieving	Reading (decoding, word reading, text comprehension) & spelling.	Positive effect ($d = +0.24$) for systematic phonics, but also positive effect for systematic language activities ($d = +0.29$) and tutoring ($d = +0.40$). Systematic phonics instruction when combined with language activities and individual tutoring may *triple* the effect of phonics alone	Phonics, as one aspect of the reading process, should not be over-emphasised

Camilli, Wolfe, and Smith (2006), US	To improve reading and spelling skills	Reviewers assume same as Camilli, Vargas, and Yurecko (2003)	Not stated, but reviewer assumes same as Camilli, Vargas, and Yurecko (2003)	RCTs & QEs	Reviewer assumes same as Camilli, Vargas, and Yurecko (2003)	Reviewer assumes same as Camilli, Vargas, and Yurecko (2003)	Tutoring alone had significant positive effect ($d = +0.46$)	The most popular interpretations of the NRP report are not supported by the evidence collected by the panel; for the purpose of guiding instructional policy, the 'science' lacks a sound empirical grounding
							Phonics effect non-significant ($d = +0.12$)	
Ehri et al. (2001), US	To improve reading and spelling skills	Considerable discussion; all varieties included (synthetic, large unit analytic, analogy, embedded, onset-rime, phonics through spelling)	38, yielding 66 treatment/control comparisons	RCTs & QEs	Schools. Children aged 5–11 (K-G6), normally achieving, at risk, reading disabled, or low-achieving	Reading (decoding, word reading, text comprehension) & spelling	Overall effect of phonics on reading was moderate, $d = +0.41/+0.44$. Effects persisted after instruction ended. Effects were larger when phonics instruction began early ($d = +0.55$) than after first grade ($d = +0.27$). Phonics benefited decoding, word reading, text comprehension and spelling in many children. It helped low and middle SES readers, younger students at risk for reading disability (RD), and older students with RD, but not low- achieving readers who included students with cognitive limitations. Synthetic phonics and larger-unit systematic phonics programmes produced similar advantage in reading. Instruction in small groups and classes was not less effective than tutoring. Systematic phonics instruction helped children learn to read better than all forms of control group instruction, including whole language	Systematic phonics instruction proved effective and should be implemented as part of literacy programmes to teach beginning reading as well as to prevent and remediate reading difficulties
		Control groups received unsystematic or no phonics; apparently mainly whole language						

(*Continued*)

Table 3. (*Continued*).

Author, date, country	Aims of intervention(s) included in SR/meta-analysis	Phonics interventions	Number of studies included	Design(s) of studies	Settings and participants	Outcome measures	Results, as reported by authors	Conclusions, as reported by authors
Galuschka et al. (2014), Germany	To improve reading and spelling skills.	Some interventions (e.g. orally dividing words into syllables with supporting hand signals) would not fit standard definitions of phonics	22 RCTs; 49 comparisons; 29 phonics instruction	RCTs only	Studies in English-speaking countries and non-English-speaking countries (Finland, Italy, Spain, Brazil); children and adolescents whose reading performance was below 25th percentile or at least one SD, one year, or one grade below expected level, with intelligence in the 'normal range'	'Reading speed; reading comprehension; reading accuracy; pseudo-word reading accuracy; pseudo-word reading speed; non-word reading accuracy; nonword reading speed; spelling' (2)	Phonics instruction is the most frequently investigated treatment approach, and the only approach whose efficacy on reading and spelling performance in children and adolescents with reading disabilities is statistically confirmed. Effect size $g = +0.32$(CI +0.18, +0.47). The mean effect sizes of the remaining treatment approaches did not reach statistical significance	Severe reading and spelling difficulties can be ameliorated with appropriate treatment
		No details of control group instruction						In order to be better able to provide evidence-based interventions to children and adolescents with reading disabilities, research should intensify the application of blinded randomised controlled trials Cross-linguistic studies are required to explore the transferability of findings across languages

Hammill and Swanson (2006), US	To improve reading and spelling skills	As Ehri et al. (2001)	= Ehri et al. (2001)	= Ehri et al. (2001)	= Ehri et al. (2001)	= Ehri et al. (2001)	All d values = Ehri et al.'s, but when re-expressed as *r* and r^2 they become much weaker and mostly trivial because ns &/or explain too little variance	When used in tutorial settings, phonics may be slightly more beneficial than non-phonics in teaching young, low-SES, at-risk children to decode. For most other students, including both normal and problem readers, phonics is not appreciably better than non-phonics, especially when the goal is to increase comprehension, oral text reading, and spelling
Han (2010), US	To improve reading for ELL learners	Studies which taught phonemic awareness, phonics or both. No specific varieties of phonics mentioned, and of 11 teaching activities mentioned (120) only 'decoding' would meet standard definitions of phonics; all the rest are whole-word approaches, hence not phonics	29 studies, 44 independent samples (80). 44 comparisons in HLM model but only 26 citations listed in Table 1 (48)	Group experimental ($n = 25$) or quasi-experimental designs ($n = 19$)	Pre-kindergarten – 6th grade. But only phonics instructional programmes for pre-kindergarten – 2nd grade. English-speaking countries in which English is the main language of instruction in mainstream schools. Children who have not yet achieved full proficiency in the English language	Construct of reading performance: phonemic awareness ($n = 24$), phonics ($n = 26$), fluency, vocabulary, comprehension, other (34 and 52)	Phonemic awareness has the highest effect size	The fundamentality of phonemic awareness and phonics instruction at emergent (preschool to mid first grade) and beginning (kindergarten to early third grade) stages. 3 evidence-based or promising practices from 13 programmes identified. Proactive Reading and Peer-Assisted Learning Strategies both have phonics and phonemic awareness as components (Table 13)
		No details of control group instruction	Table 2, 49 –44 interventions, phonemic awareness ($n = 27$) and phonics ($n = 24$)	Not clear which are which in terms of phonics.		Quantitative measures of reading performance (standardised tests, informal reading inventories) – examples given in Appendix B	Phonemic awareness: +0.41 ($n = 26$), Phonics: .+0.33 ($n = 72$) (weighted effect sizes) 90–91: '… plausible reason of the higher effect on this measure is that ELL students show larger growth on phonemic awareness and/or the measure has greater sensitivity to students' growth'	

(*Continued*)

Table 3. (*Continued*).

Author, date, country	Aims of intervention(s) included in SR/meta-analysis	Phonics interventions	Number of studies included	Design(s) of studies	Settings and participants	Outcome measures	Results, as reported by authors	Conclusions, as reported by authors
McArthur et al. (2012), Australia	To improve literacy skills	(p. 6) 'Pure' phonics programmes that focused on learning to read via letter-sound rules alone (3 studies), vs. phonics plus phoneme awareness (PA) (7 studies), and phonics plus irregular word reading (1 study). Most of the phonics plus PA studies seem like synthetic phonics, but some had elements of onset-rime	11 studies (14 records)	All controlled trials that used randomisation or minimisation. All had phonics and control group.	English-speaking children, adolescents, and adults whose reading level was below expected (with no explanation for this)	Primary outcomes: word reading accuracy (10 studies), non-word reading accuracy (8 studies), word reading fluency (2 studies), non-word reading fluency (1 study), reading comprehension (3 studies), spelling (2 studies). Secondary outcomes: letter-sound knowledge (3 studies) and phonological output (4 studies)	Efficacy of phonics training not significantly moderated by: training type, training intensity, training duration, training group size, or training administrator	Only 3 results were statistically significant (non-word reading accuracy, word reading accuracy and letter-sound knowledge0. Significance may have been dependent on the amount of data from which they were calculated
		Control groups received no training (= business as usual) or an alternative intervention, e.g. maths (1)				Summary table on 4–5	Word reading accuracy: SMD +0.47 (95% CI +0.06 to +0.88; 10 studies). Non-word reading accuracy: SMD +0.76 (95% CI +0.25 to +1.27; 8 studies). Word reading fluency SMD −0.51 (95% CI −1.14 to +0.13; 2 studies). Reading comprehension: SMD +0.14 (95% CI −0.46 to +0.74; 3 studies). Spelling: SMD +0.36 (95% CI +0.27 to +1.00; 2 studies). Letter-sound knowledge: SMD +0.35 (95% CI +0.04 to +0.65; 3 studies) Phonological output: SMD +0.38 (95% CI −0.55 to +1.32; 1 study) See also summary table on 4–5	'Overall, 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' (26)
Sherman (2007), US	To improve reading	Synthetic, large-unit, miscellaneous (based on Ehri et al. 2001).	26, yielding 88 effect sizes, reduced to 36 comparisons	12 individual-level RCTs, 3 'random treatment' (= apparently cluster RCTs), 11 not reported	Schools (11), clinic (1), (not reported 14)	Decoding regular words & pseudo-words; word identification; spelling; reading text orally; comprehension	*d* for word identification (22 studies) = +0.53; for comprehension (7 studies) = +0.42	No main effects and no statistically significant interaction effects between or among variables of interest at the standard 95% CI

		Types of control group instruction (9): 'basal [reading schemes]; regular curriculum; whole language; whole word; miscellaneous'			US middle & high school pupils (ages 10–17), with reading level ≤25th %ile (15), 26-49th %ile (4), (not reported 7)		BUT ns ($p > 0.05$)	
							Some results significant when α level relaxed to 0.25	
Suggate (2010), Germany	To improve reading	'Explicit teaching of grapheme–phoneme correspondences' (1560). 'Letters-to-sounds: Attention to grapheme–phoneme correspondences occurring in letters' (1574)	85 studies, 116 intervention-control comparisons (32% described as phonics, 1562)	Experimental or quasi-experimental	Preschool – Grade 7.	Overall pre-reading, reading and comprehension measures. Reading outcomes expressed as standard scores	Overall effect sizes (Table 1):	Phonics interventions delivered greatest short-term benefit for reading skills (for younger children) but the utility of phonics interventions beyond Grade 1 may decline. A developmental understanding of reading remediation. PA and phonics were especially effective when pre-reading outcomes were used. At-risk status (struggling readers) was not a significant predictor
		Effectively synthetic phonics		'Random assignment of the treatment and control groups or, if the study was quasi-experimental, matching on pre-test (i.e. $p > 0.05$ and $d < 0.50$)' 1560	At risk readers – low SES or lower-performing readers, OR struggling reading at or below 15th percentile, diagnosed with reading or learning disability or at least 1 SD between intelligence quotient and achievement (1560)		Phonics ($d = +0.50$, $k = 36$, $N = 2142$). 95% CI [+0.38, +0.62]	
		'Control groups received either typical instruction or an appreciably different in-house school intervention.' (1559)					PA ($d = +0.47$, $k = 13$, $N = 731$). 95% CI not computable	

(*Continued*)

Table 3. (*Continued*).

Author, date, country	Aims of intervention(s) included in SR/meta-analysis	Phonics interventions	Number of studies included	Design(s) of studies	Settings and participants	Outcome measures	Results, as reported by authors	Conclusions, as reported by authors
							Greater effect sizes for mixed and comprehension interventions later and for phonics interventions earlier and continued into middle grades (in terms of school stage) The importance of considering interactions – grade became statistically weaker after the addition of Phonics X Grade ($\beta = 15$, $p < 0.10$) as opposed to without intervention terms ($\beta = 35$, $p < 0.01$)	
Suggate (2016), Germany	To improve reading	'Phonics interventions teach associations between phonemes and orthography.' (78) 'Phonics included letter–sound or sound–spelling relations.' (82) No further details	16, all with post-intervention follow-up data	Experimental and quasi-experimental	Preschool – Grade 6 (Risk status of samples stated but includes 'normal')	Pre-reading, reading, reading comprehension, spelling measures	Weighted effect sizes for reading by type of instruction:	p. 90 'In conclusion, this meta-analysis extends our understanding of the effectiveness of reading interventions by providing a detailed analysis of the long-term effects. Indeed, in doing so, some surprising findings emerged, namely that phonemic awareness interventions appeared better than phonics, which is inconsistent with the phonological linkage hypothesis. Comprehension interventions, on the other hand, appeared particularly effective, as did those given to older pupils'
		No details of control group instruction					(1)from Table 3 (86): • at post-test: phonemic awareness $d = +0.32$, phonics $d = +0.26$ • at follow-up: phonemic awareness $d = +0.33$, phonics $d = +0.07$ • from text (87): • at post-test: phonemic awareness $d = +0.32$, phonics $d = +0.33$ • at follow-up: phonemic awareness $d = +0.29$, phonics $d = +0.07$	

								Author's interpretation (87): 'At immediate post-test, there was little evidence that it mattered whether or not phonics or purely phonemic awareness interventions were used. However, when follow-up effect sizes were compared, there was a distinct advantage for phonemic awareness interventions, precisely the opposite of what would be predicted by the phonological linkage hypothesis' Second conclusion seems unaffected by difference between Table 3 and text, but evidence for first conclusion seems weaker in Table 3	
Torgerson, Brooks, and Hall (2006), UK	To improve reading and spelling	All systematically taught varieties, including synthetic, analytic	12 RCTs in main meta-analysis	RCTs		5 – 10.8 years (age range)	Word reading accuracy, comprehension and spelling (29)	Fixed effect: $d = +0.27$ (+0.10 to +0.45)	Systematic phonics instruction (in a broad literacy curriculum) appears to have a greater effect (ES +0.27) than unsystematic or no phonics instruction on progress in reading for children. There is uncertainty in the evidence about which phonics approach (synthetic or analytic) is most effective
		Control groups received unsystematic or no phonics; almost all whole language	20 studies (1 UK-based) in 19 papers, 14 trials			Some normally attaining, some at risk for reading disability, some 'Disabled readers', and low performers		Random effects: $d = +0.38$ (+0.02 – +0.73) (See 34, footnotes)	No evidence beyond early years for different approaches impacting on phonics in reading and writing (only 3 of the included RCTs had follow-up measures)
								Systematic phonics teaching associated with better progress in reading accuracy (across all ability levels). No significant effect for reading comprehension	Section 12 –Recommendations for teaching, teacher training and research are given

(*Continued*)

Table 3. (*Continued*).

Author, date, country	Aims of intervention(s) included in SR/meta-analysis	Phonics interventions	Number of studies included	Design(s) of studies	Settings and participants	Outcome measures	Results, as reported by authors	Conclusions, as reported by authors
							No evidence for advantage or superiority of synthetic or analytic phonics instruction (but comparison only based on 3 small RCTs) Phonics instruction did not appear to affect progress in spelling	

included non-phonics studies, but it was beyond the scope of this review to check back to every individual RCT.

A few authors (Han 2010; Suggate 2010, 2016; McArthur et al. 2012) compared phonics instruction with phonemic/phonological awareness training. Details of the instruction received by control groups were scant; where mentioned, it seemed to be 'business as usual' literacy teaching, often of a whole language variety, though McArthur et al. (2012) and Suggate (2010) hinted at alternative interventions (e.g. maths).

The number of studies included in the SRs ranged from 3 to 85, so the various SR authors were clearly using different definitions of phonics and/or inclusion/exclusion criteria. Some of the variation was due to participant selection – e.g. Adesope et al. (2011) were looking at ESL students in English-speaking countries. Only Galuschka et al. (2014) and Suggate (2010) included studies conducted in languages other than English. Participants in the studies included in the SRs range in age from pre-kindergarten children (aged 4), through children in all grades in primary (and middle) and secondary (high) schools, to adult participants in one SR. The full range of learner characteristics is represented in one or more SRs, including normally attaining and low-attaining students, those with English as a second language, or those with reading disabilities. Outcome measures in the SRs were diverse but most included studies with reading (decoding, word reading and fluency; comprehension) and spelling (writing).

Table 4 presents the results of our quality assessment of the included SRs, using the key methodological items from the PRISMA statement. The 12 SRs were of generally high, but variable quality. Most of the 12 SRs fulfilled the following criteria by providing data or text: the rationale and objectives of the SR; methods and results for searching, screening, data collection and synthesis. (The three replication SRs used the databases from the original SRs for inclusion). Having said that, a key item from the PRISMA checklist – assessment of risk of bias of included studies – was undertaken by only 7 out of the 12 SRs. In other words, 5 of the SRs did not quality appraise the studies which they included in their systematic review – and by extension, their pooled effect size – so they may have been indiscriminately including studies of high, moderate and low quality. This omission in these 5 SRs is critical and, therefore, the results from these SRs should carry lower weight of evidence in our conclusions.

Results of effect sizes for phonics

Statistically significant positive effects for phonics instruction on at least one reading outcome were found across most (10) of the SRs ranging from small to moderate effects (Ehri et al. 2001; Camilli, Vargas, and Yurecko 2003; Torgerson, Brooks, and Hall 2006; Sherman 2007; Han 2010; Suggate 2010; Adesope et al. 2011; McArthur et al. 2012; Galuschka et al. 2014; Suggate 2016). Non-significant positive effects were found in the remaining 2 SRs (Camilli, Wolfe, and Smith 2006; Hammill and Swanson 2006).

Effect size variance according to statistical model – Hedges' g or Cohen's d

The extracted effect sizes were classified according to how they were described by the authors. Most studies described or referenced the formulae for the effect size calculations and referred to this as *g* (Han 2010; Adesope et al. 2011; Galuschka et al. 2014) or *d* (Ehri

Table 4. Quality appraisal of included systematic reviews/meta-analyses, using adapted PRISMA statement (for assessment of design bias).

Study	Intro.: rationale and objectives (3 and 4)	Methods: Search (8)	Methods: Selection (9)	Methods: Data collection (10 and 11)	Methods: Risk of bias (12)	Methods: Synthesis (14)	Results: Study selection (17)	Results: Study characteristics (18)	Results: Synthesis (21)	Discussion (24, 25 and 26)
Adesope et al. (2011)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Camilli, Vargas, and Yurecko (2003)	Yes	Yes	Yes	Yes	N/S	Yes	Yes	Yes	Yes	Yes
Camilli, Wolfe, and Smith (2006)	Yes	n/a – used previous database	n/a – used previous database	Yes	No	Yes	n/a – used previous database	N/S	Yes	Partly (no discussion of limitations)
Ehri et al. (2001)	Yes	Yes	Yes	Yes	N/S	Yes	Yes	Yes	Yes	Partly (no discussion of bias)
Galuschka et al. (2014)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hammill and Swanson (2006)	Yes	n/a – used Ehri et al. (2001) database			Not discussed	n/a – used Ehri et al. (2001) database			Yes	Yes
Han (2010)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
McArthur et al. (2012)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sherman (2007)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Suggate (2010)	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Suggate (2016)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Torgerson, Brooks, and Hall (2006)	Yes	Ye	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

et al. [2001]; by cross-reference to NRP [2000] – see footnote to Table 5); McArthur et al. [2012]; Sherman [2007]; Torgerson, Brooks, and Hall [2006]). One author (Suggate 2010, 2016) followed Hunter and Schmidt's (2004) approach. Three studies used or referred to the approach adopted in the studies they were critiquing or defending (Camilli, Vargas, and Yurecko 2003; Camilli, Wolfe, and Smith 2006; Hammill and Swanson 2006).

There is some confusion in the literature about terminology, but Hedges' *g* usually refers to Hedges' bias-corrected estimator (Hedges and Olkin 1985) and *d* to Cohen's *d* (Cohen 1988). Both approaches are based on a pooled standard deviation. Cohen used the maximum likelihood estimator for the variance, which is biased with small samples, whereas Hedges used Bessel's correction ($n - 1$) to estimate the variance. In practice, for samples above 20, the difference in the effect size estimate is minimal. Estimates of effect will also vary between class and individual level analysis, and depending on whether unequal sample sizes and clustering are taken into account (Xiao, Kasim, and Higgins 2016), and on which mean scores are used (post-test or gains) and on which standard deviations are pooled (pre-test, post-test or gains). Some further details can be found in Table 5.

However, it should be noted that, of all the SRs reviewed, only Galuschka et al. (2014, 3) stated which mean scores were used in calculating ESs (post-test); they implied that the pooled standard deviations used were those of the post-test. The hidden problem when authors do not report these details is that even various results labelled as 'Cohen's *d*' or 'Hedges' *g*' may not be strictly commensurate with each other, and this may bedevil attempts to generalise from them.

Effect size variance according to design – RCT or QED

The included SRs contained both RCTs and QEDs, with two exceptions (Torgerson, Brooks, and Hall 2006; Galuschka et al. 2014) which included only RCTs. In two cases, it was not possible to determine which studies were of which designs (Adesope et al. 2011; Sherman 2007). In a number of the included SRs the authors did not report study design for the studies which investigated the effectiveness of phonics instruction. Looking at the pooled effect sizes (ES) from RCTs and QEDs, for those reviews that have included both, there are some clear differences. Some of these differences in ES are less apparent in the overall reported ES. For example, as Table 5 shows, Adesope et al. (2011) do not explicitly report ES separately for RCTs and QEDs; however, the pooled ES for random allocation is +0.31 and +0.68 for non-random allocation, a difference of +0.37. This difference is less apparent in looking at the pooled overall ESs; that for systematic phonics instruction and guided reading is +0.40 and that collapsed across all pedagogical strategies is +0.41. Suggate (2010) is similar, in that the overall ES for QEDs is larger (+0.64) than for RCTs (+0.41), with the overall mean weighted ES for phonics being +0.50. Camilli, Vargas, and Yurecko (2003) explicitly stated that there was no difference between ES for RCT and QED designs, with an overall ES of +0.24. Similarly, different ES are not stated in Camilli, Wolfe, and Smith (2006) for different designs; the overall ES reported is, however, much lower at +0.12.

Publication bias

We extracted data from each study about whether or not grey literature was searched; whether any grey literature was included; whether the issue of publication bias seemed

Table 5. Pooled effect sizes.

Study	Effect size formula stated?	Mean scores and standard deviations used stated?	Pooled overall effect size	Pooled effect size of RCTs	Pooled effect size of QEDs
Adesope et al. (2011)	Discussed on 635 Aggregate ES computed from weighted ESs. Hedges' unbiased estimate of mean ES. *Q* statistic for homogeneity of variance.	No	Systematic phonics instruction and guided reading: $g = +0.40$ ($k = 14$, $N = 1647$) (CI +0.3 to +0.5) Collapsing across all pedagogical strategies: $g = +0.41$ ($k = 26$, $N = 3{,}150$) CI +0.33 to +0.48 (636)	Not reported separately but overall, Table 5 (644) Random = +0.31	Non-random = +0.68
Camilli, Vargas, and Yurecko (2003)	Yes – detailed discussion, including Hedges' effect size adjustment, 18–19, & principles, 34	No	$d = +0.24$	'[No] evidence that randomised experiments give different results than quasi-experimental studies.' (28)	
Camilli, Wolfe, and Smith (2006)	No – presumably as Camilli, Vargas, and Yurecko (2003)	No	Phonics $d = +0.12$ (ns) Tutoring $d = +0.46$ ($p = 0.002$)	Not stated	Not stated
Ehri et al. (2001)	Cohen's *d*, stated only verbally (401). NRP report (2000, 1–10) states formula algebraically* See also critique by Camilli, Vargas, and Yurecko (2003, 18–19) Q statistic for homogeneity of variance (403)	No	$d = +0.41$ or $+0.44$	$d = +0.45$	$d = +0.43$
Galuschka et al. (2014)	Yes Hedges' *g* bias corrected (3–4)	Yes – post-test (3)	Reading: $g' = +0.322$ (95% CI [+0.177, +0.467] Spelling: $g' = +0.336$; 95% CI [+0.062, +0.610]	Reading: $g' = +0.322$ (95% CI [+0.177, +0.467] Spelling: $g' = +0.336$; 95% CI [+0.062, +0.610]	n/a n/a
Hammill and Swanson (2006)	n/a, = Ehri et al. (2001)	No	$d = +0.44$, but $r = +0.21$, $r^2 = +0.04$	$d = +0.45$, but $r = +0.28$, $r^2 = +0.08$	$d = +0.43$, but $r = +0.21$, $r^2 = +0.04$
Han (2010)	Yes, 37–45	No	Weighted ESs: phonemic awareness +0.41 ($n = 26$); phonics +0.33 ($n = 72$); fluency +0.38 ($n = 27$); vocabulary +0.34 ($n = 11$); and comprehension measures +0.32 ($n = 39$)	Not reported separately – see Table 9.	Not reported separately – see Table 9

	Formulas for transformation, adjustment and correction for small sample bias, moderator analysis, aggregation and homogeneity analysis all given. Hedges' g bias corrected			It means that studies with higher quality tended to have lower ESs (68)	It means that studies with higher quality tended to have lower ESs (68)
McArthur et al. (2012)	Yes, continuous data – 9. Mean difference (MD) used	No	SMD = +0.47 (statistically significant) (95% CI +0.06 to +0.88; Z = 2.22; p = 0.03) (Analysis 1.1)	Not reported separately	Not reported separately
	Equivalent to Cohen's d			Studies allocated participants using random allocation, minimisation or quasi-randomisation (7)	See sensitivity analysis p. 12 (unclear randomisation)
Sherman (2007)	Yes (underspecified), with discussion 23–30 Cohen's d	No	Word identification (22 studies) d = +0.53 (ns) Comprehension (7 studies) d = +0.42 (ns)	Word identification (22 studies) d = +0.53 (ns) Comprehension (7 studies) d = +0.42 (ns) (BUT difficult to tell)	n/a
Suggate (2010)	Yes p. 1562	No	Phonics – Table 1 (Mean weighted effect sizes) d = +0.50. SD = 0.06, N = 2142, k = 36, 95% CI [+0.38 to +0.62] Overall–Moderate – (d = +0.49, SD = 0.23, N = 7,522, k = 116, 95% CI [+0.04 to +0.95]) – p. 1563	Randomised-control designs (d = +0.41, SD = 0.21, k = 72, Q = 121.14, p = 0.001	Quasi-experimental studies (d = +0.64, SD = 0.19, k = 44, Q = 68.20, p = 0.01)
	Mean effect sizes (Hunter and Schmidt 2004) Seven categories (commonly occurring literacy constructs) and aggregate calculated for each				
Suggate (2016)	Yes – p. 83 (Hunter and Schmidt 2004)	No	At follow-up:	At follow-up:	At follow-up:
			Phonemic awareness overall: unweighted +0.46, weighted estimated +0.36	Unweighted +0.33, weighted estimated +0.29	Unweighted +0.40, weighted estimated +0.18

(*Continued*)

Table 5. (*Continued*).

Study	Effect size formula stated?	Mean scores and standard deviations used stated?	Pooled overall effect size	Pooled effect size of RCTs	Pooled effect size of QEDs
			Phonics overall: unweighted +0.25, weighted estimated +0.07		
Torgerson, Brooks, and Hall (2006)	Yes, effect sizes calculated based on a mean of reading accuracy, a mean of reading comprehension (where applicable) and a mean of spelling (where applicable) (25–26)	No	Fixed effect $d = +0.27$ (95% CI +0.10 – +0.45)	Fixed effect $d = +0.27$ (95% CI +0.10 – +0.45)	n/a
			Random effects $d = +0.38$ (95% CI +0.02 – +0.73)	Random effects $d = +0.38$ (95% CI +0.02 – +0.73)	

Notes: *Ehri et al. (2001) said 'The formula … consisted of the mean of the treatment group minus the mean of the control group divided by a pooled standard deviation.' The algebraic form of this is given in NRP (2000, 1–10) as $(M_t - M_c)/0.5(SD_t + SD_c)$, which is a version of Cohen's *d*. However, it fails to specify which mean scores were used (post-test or gains) and which standard deviations were used (pre- or post-test or gains). Also, simply taking the arithmetic mean of the SD's is acceptable only if they are very similar; otherwise (and it would probably be wiser to use it routinely), the formula which should be used for the pooled SD. (s) is $s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2}}$ (Hartung, Knapp, and Sinha 2008), where $n_1 + n_2$ are the sample sizes of the two groups, and $s_1 + s_2$ are their SD's. (Hedges' *g* differs only in having $n_1 + n_2 - 2$ as the denominator.).

Table 6. Information about publication bias (for assessment of potential publication bias).

Study	'Grey' literature searched?	Contains at least one item of 'grey' literature	Publication bias mentioned?	Method for assessing potential for publication bias	If publication bias was found was it addressed?
Adesope et al. (2011)	No	No	Yes	Yes	n/a
Camilli, Vargas, and Yurecko (2003)	No	No	Not clear	n/a	n/a
Camilli, Wolfe, and Smith (2006)	No	No	No	n/a	n/a
Ehri et al. (2001)	No	No	No	n/a	n/a
Galuschka et al. (2014)	Yes	Yes	Yes	Yes	Yes
Hammill and Swanson (2006)	No	No	No	n/a	n/a
Han (2010)	No	No	Yes	No	n/a
McArthur et al. (2012)	Yes	Yes	Yes	Yes	n/a
Sherman (2007)	Yes	No	Yes	No	n/a
Suggate (2010)	No	No	Yes	Yes	n/a
Suggate (2016)	No	No	Yes	Yes	Yes
Torgerson, Brooks, and Hall (2006)	Yes	Yes	Yes	Yes	n/a

to have the potential to bias the results of the study; whether a recognised method for the detection of publication bias was used (for example, funnel plot); whether any evidence for potential publication bias was found; and, if publication bias was suspected, what method was used to mitigate this bias and the results flowing from this (see Table 6).

Of the 12 systematic reviews, only 6 engaged fully with the issue of publication bias and the potential for it to bias the results of their systematic review (Torgerson, Brooks, and Hall 2006; Adesope et al. 2011; Suggate 2010, 2016; McArthur et al. 2012; Galuschka et al. 2014). The remaining 10 studies either did not mention publication bias at all (or this was unclear) or, as in the case of Han (2010), publication bias was mentioned but the author did not search for or include any grey literature, and did not use any method to assess the potential for publication bias. Sherman (2007) searched for grey literature, but had as an exclusion criterion 'not published in peer-reviewed journals' and therefore excluded those studies that they had retrieved but which were not published (total of 5). They also did not mention the issue of publication bias, in particular that the application of the exclusion criterion may have contributed to publication bias in their review.

Adesope et al. (2011) did not search for or include any grey literature. However, they did explore the issue through the use of Orwin's Fail-Safe N and Classic fail-safe N test, which suggested that the results were robust and validity was not threatened by publication bias; therefore no further analyses were undertaken.

Galuschka et al. (2014) explored publication bias for those studies which evaluated phonics instruction and used reading performance as a dependent variable (not for spelling). A funnel plot was used to explore the presence of publication bias, which displayed asymmetry with a gap on the left of the graph, indicating the possible presence of publication bias. Duval and Tweedie's trim and fill method was used to assess the extent of publication bias, and an unbiased effect size was estimated. The procedure trimmed 10 studies into the plot and led to an estimated unbiased effect size of Hedges' $g = +0.198$ (CI +0.039, +0.357), which is in contrast to a, potentially biased upwards, effect size of Hedges' $g = +0.32$ (CI +0.18, +0.47) for the main analysis.

McArthur et al. (2012) searched for and included grey literature and also undertook sensitivity analysis and a funnel plot, and concluded that their systematic review was not affected by publication bias.

Although he did not explicitly search for and include studies from the grey literature, in two meta-analyses Suggate (2010, 2016) looked at the potential for publication bias using funnel and box plots, and addressed this in the more recent meta-analysis by including only the larger studies.

In our SR (Torgerson, Brooks, and Hall 2006) we specifically searched the grey literature, and included one unpublished thesis. We used a funnel plot to investigate the potential presence of publication bias in our meta-analysis and found evidence of this, but the Egger test statistic was not significant, which reduced any certainty in the presence of publication bias.

Results of quality assurance of data extraction and quality appraisal

Initial agreement between the two pairs of authors was high; any disagreements were resolved through discussion and arbitration. The data extraction and quality appraisal of the original SR undertaken by two of the authors Torgerson, Brooks, and Hall (2006) were completed by the other two authors to minimise the potential for conflict of interest.

Discussion

The diverse range of interventions and control or comparison conditions, settings (including countries), participant characteristics, outcome measures and study designs included in the 12 SRs in our tertiary review increases the generalisability of our findings. However, there are limitations on this, in particular doubts over whether some of the interventions analysed deserve the label 'phonics', and the possible incommensurability of the overall effect sizes reported due to both under-reporting of, and differences in, methods of calculating them.

In terms of publication bias, as only 6 of the 12 meta-analyses addressed this issue and, of those, only 3 found evidence of potential publication bias, we can interpret this as an indication that publication bias is an issue in the individual meta-analyses in the tertiary review, and therefore in the tertiary review itself. The consequences of this interpretation are that we should have more caution in the findings of our review as it is likely that experimental studies have been undertaken which have found null or negative results and therefore have either not been published, or they have been published but have not been included in meta-analyses, either by design or because they were not in the public domain to be found.

The reviews were fairly consistent in demonstrating an overall positive effect of phonics teaching, with pooled estimates ranging from 0.12 to 0.5. This is probably unsurprising, given that the reviews contained many of the same studies and therefore it would be unlikely that there would be huge divergence in terms of the pooled estimate. Furthermore, there is little evidence to demonstrate the superiority of one phonics approach compared with any other instructional method – but very few individual RCTs have investigated this question, so it hardly features in the SRs. There remains uncertainty as to the overall effect given the probable presence of publication bias. Indeed, with the prevalence of so many reviews showing positive effects of phonics teaching, this means it might be less likely for null or negative results to be reported. Some of the reviews try to distinguish differential effects of phonics among educationally important subgroups. Whilst some reviews see some evidence for better or lesser effects within different types of learner, these forms of analysis should always be treated with a certain amount of caution. This is because even, within a large randomised controlled trial, there is usually very little statistical power to demonstrate meaningful subgroup differences, and within a meta-analysis the power issue is even more problematic.

Conclusions

Given the evidence from this tertiary review, what are the implications for teaching, policy and research? It would seem sensible for teaching to include systematic phonics instruction for younger readers – but the evidence is not clear enough to decide which phonics approach is best. Also, in our view there remains insufficient evidence to justify a 'phonics only' teaching policy; indeed, since many studies have *added* phonics to whole language approaches, balanced instruction is indicated. For policy, encouragement of phonics instruction within schools is justified unless and until contrary evidence emerges. Finally, in terms of research: given the uncertainties in the evidence base over publication bias, the 'phonics' status of some included studies, and how best to calculate effect sizes, there may be a case for conducting a large and even more rigorous systematic review. But what is required above all are large field trials of different phonics approaches and different phonics 'dosages'. We called for such an approach in our review of phonics teaching in 2006, and a decade later we make the same call.

In conclusion, there have been a significant number of systematic reviews of experimental and quasi-experimental research evaluating the effectiveness or otherwise of phonics teaching since 2000. Most of the reviews are supportive of phonics teaching, but this conclusion needs to be tempered by two potential sources of bias: design and publication bias. Both of these problems will tend to exaggerate the benefit of phonics teaching. Furthermore, there is little evidence of the comparative superiority of one phonics approach over any other. Ideally, each country should establish a programme of large RCTs that are adapted to local circumstances that will test different phonics approach to reading and writing acquisition. If this was adopted then we might finally end the 'reading wars'.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Carole Torgerson has been a professor of Education at Durham University since 2012. Prior to this, she was professor of Experimental Design at the University of Birmingham and Reader in Evidence-based Education at the University of York. She is an expert on randomised controlled trial and systematic review designs, having undertaken numerous experiments and reviews in various topics in education. She is also a literacy expert.

Greg Brooks worked on oracy assessment and family literacy evaluations at NFER (1981–2000). At Sheffield (2001–2007) he directed 15 adult literacy projects. In 2005–2006 he was a member of the Rose committee, and in 2008–2009 of the dyslexia subgroup of the Rose review of the primary

curriculum in England. In 2011–2012 he was a member of the EU High Level Group of Experts on Literacy.

Louise Gascoine is a research associate at Durham University. She is a former secondary school teacher and has a PhD in education (focused on metacognition). Her current research is focused on metacognition, systematic review design and the use of impact and process evaluations within randomised controlled trial design in education.

Steve Higgins is a former primary school teacher. His research interests include the effective use of digital technologies for learning in schools, understanding how children's thinking and reasoning develop, and how teachers can be supported in developing the quality and effectiveness of teaching and learning in their classrooms, using evidence from research.

References

Adams, M. J. 1990. *Beginning to Read: Thinking and Learning about Print*. Cambridge: MIT Press.

Beard, R., G. Brooks, and J. Ampaw-Farr (Forthcoming). "How Linguistically-informed Are Phonics Programmes?" *Literacy*.

Brooks, G., M. Cook, A. Littlefair, with replies from D. Wyse, and M. Styles. 2007. "Responses to Wyse and Styles' Article, "Synthetic Phonics and the Teaching of Reading: The Debate Surrounding England's 'Rose Report'" (*Literacy*, 41, 1, April 2007)." *Literacy* 41 (3): 169–176.

Chall, J. S. [1967] 1989. *Learning to Read: The Great Debate*. 2nd ed. New York: McGraw-Hill.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates.

Darnell, C. A., J. E. Solity, and H. Wall. 2017. "Decoding the Phonics Screening Check." *British Educational Research Journal* 43 (3): 505–527.

Department of Education and Science. 1989. *English in the National Curriculum*. London: Her Majesty's Stationery Office.

DfE (Department for Education). 1995. *English in the National Curriculum*. London: Her Majesty's Stationery Office.

DfE (Department for Education). 2011. *Year 1 Phonics Screening Check Pilot Evaluation*. London: Department for Education. Accessed February 5, 2017. https://www.gov.uk/government/publications/year-1-phonics-screening-check-pilot-evaluation

DfE (Department for Education). 2013. *English Programmes of Study: Key Stages 1 and 2 National Curriculum in England*. London: Department for Education.

DfE (Department for Education). 2014. *Phonics: Choosing a Programme*. London: Department for Education. Accessed February 5, 2017. https://www.gov.uk/government/collections/phonics-choosing-a-programme

D*f*EE (Department for Education and Employment). 1998. *National Literacy Strategy*. London: Department for Education and Employment.

D*f*EE (Department for Education and Employment). 1999. *The National Curriculum Handbook for Primary Teachers in England*. London: Department for Education and Employment & Qualifications and Curriculum Authority.

D*f*ES (Department for Education and Skills). 2004. *Playing with Sounds*. London: Department for Education and Skills.

D*f*ES (Department for Education and Skills). 2006. *Primary National Strategy*. London: Department for Education and Skills.

D*f*ES (Department for Education and Skills). 2007. *Letters and Sounds*. London: Department for Education and Skills.

Gough, P., and W. Tunmer. 1986. "Decoding, Reading, and Reading Disability." *Remedial and Special Education* 7: 6–10.

Hartung, J., G. Knapp, and G. M. Sinha. 2008. *Statistical Meta-analysis with Application*. Hoboken, NJ: Wiley.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-analysis*. New York: Academic Press.

House of Commons Education and Skills Committee. 2005. *Teaching Children to Read (Eighth Report of Session 2004–05)*. London: The Stationery Office Limited.
Hunter, J. E., and F. L. Schmidt. 2004. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage.
Johnston, R. S., and J. E. Watson. 2004. "Accelerating the Development of Reading, Spelling and Phonemic Awareness Skills in Initial Readers." *Reading and Writing* 17: 327–357.
Krashen, K. (2017, 2 February). Letter in *The Guardian* newspaper. https://www.theguardian.com/education/2017/feb/01/invest-in-libraries-not-phonics-tests.
Machin, S., S. McNally, and M. Viarengo. 2016. *"Teaching to Teach" Literacy*. London: London School of Economics Centre for Economic Performance Discussion Paper No 1425.
Moher, D., A. Liberati, J. Tetzlaff, and D. G. Altman. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-analyses: The PRISMA Statement." *PLoS Med* 6 (7): e1000097. doi:10.1371/journal.pmed.1000097.
NRP (National Reading Panel). 2000. *Teaching Children to Read: An Evidence-based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. Washington, DC: National Institute for Child Health and Human Development Clearinghouse.
Ofsted (Office for Standards in Education). 2002. *The National Literacy Strategy: The First Four Years 1998–2002*. London: Office for Standards in Education.
Rose, J. 2006. *Independent Review of the Teaching of Early Reading. Final Report*. London: Department for Education and Skills.
Schwippert, K., and J. Lenkeit, eds. 2012. *Progress in Reading Literacy in National and International Context. the Impact of PIRLS 2006 in 12 Countries*. Munster: Waxmann Verlag.
Stanovich, K. E. 2000. *Progress in Understanding Reading: Scientific Foundations and New Frontiers*. New York: Guilford Press.
Stuart, M. 2006. "Teaching Reading: Why Start with Systematic Phonics Teaching?" *Psychology of Education Review* 30: 6–17.
Stuart, M., and R. Stainthorp. 2016. *Reading Development & Teaching*. London: Sage.
Xiao, Z., A. Kasim, and S. Higgins. 2016. "Same Difference? Understanding Variation in the Estimation of Effect Sizes from Educational Trials." *International Journal of Educational Research* 77: 1–14.

Included systematic reviews/meta-analyses

Adesope, O. O., T. Lavin, T. Thompson, and C. Ungerleider. 2011. "Pedagogical Strategies for Teaching Literacy to ESL Immigrant Students: A Meta-analysis." *British Journal of Educational Psychology* 81 (4): 629–653.
Camilli, G., S. Vargas, and M. Yurecko. 2003. "'Teaching Children to Read': The Fragile Link between Science and Federal Education Policy." *Education Policy Analysis Archives* 11 (15). doi:10.14507/epaa.v11n15.2003.
Camilli, G., P. M. Wolfe, and M. L. Smith. 2006. "Meta-analysis and Reading Policy: Perspectives on Teaching Children to Read." *The Elementary School Journal* 107 (1): 27–36.
Ehri, L. C., S. R. Nunes, S. A. Stahl, and D. M. Willows. 2001. "Systematic Phonics Instruction Helps Students Learn to Read: Evidence from the National Reading Panel's Meta-analysis." *Review of Educational Research* 71 (3): 393–447.
Galuschka, K., E. Ise, K. Krick, and G. Schulte-Koerne. 2014. "Effectiveness of Treatment Approaches for Children and Adolescents with Reading Disabilities: A Meta-analysis of Randomized Controlled Trials." *PLoS ONE* 9 (2). doi:10.1371/journal.pone.0089900.
Hammill, D. D., and L. H. Swanson. 2006. "The National Reading Panel's Meta-analysis of Phonics Instruction: Another Point of View." *The Elementary School Journal* 107 (1): 17–26.
Han, I. 2010. *Evidence-based Reading Instruction for English Language Learners in Preschool through Sixth Grades: A Meta-analysis of Group Design Studies*. University of Minnesota, ProQuest Dissertations Publishing, 2009. 3371852.
McArthur, G., P. M. Eve, K. Jones, E. Banales, S. Kohnen, T. Anandakumar, and A. Castles. 2012. "Phonics Training for English-speaking Poor Readers." *Cochrane Database of Systematic Reviews*, CD009115 (12 December 2012).

Sherman, K. H. 2007. "A Meta-analysis of Interventions for Phonemic Awareness and Phonics Instruction for Delayed Older Readers." *University of Oregon, ProQuest Dissertations Publishing* 2007: 3285626.
Suggate, S. P. 2010. "Why What We Teach Depends on When: Grade and Reading Intervention Modality Moderate Effect Size." *Developmental Psychology* 46 (6): 1556–1579.
Suggate, S. P. 2016. "A Meta-analysis of the Long-term Effects of Phonemic Awareness, Phonics, Fluency, and Reading Comprehension Interventions." *Journal of Learning Disabilities* 49 (1): 77–96.
Torgerson, C., G. Brooks, and J. Hall. 2006. "A Systematic Review of the Research Literature on the Use of Phonics in the Teaching of Reading and Spelling." (ISBN: 1844786595 9781844786596). http://catalogue.bishopg.ac.uk/custom_bgc/files/JKEC_phonics_review.pdf.

Appendix 1. Search strategies and PRISMA diagram

Database	Search string
Applied Social Sciences Index and Abstracts (ASSIA) (ProQuest)	(phonic* OR phonetical* OR phonemic) AND (systematic review OR meta-analysis OR research synthesis OR research review)
Education Resources Information Centre (ERIC) (ProQuest)	(phonic* OR phonetical* OR phonemic) AND (systematic review OR meta-analysis OR research synthesis OR research review)
PsycINFO (Ebscohost)	(phonic* OR phonetical* OR phonemic) AND (systematic review OR meta-analysis OR research synthesis OR research review)
Web of Science (Web of Knowledge)	TOPIC: (phonic* OR phonetical* OR phonemic) *AND* TOPIC: (systematic review OR meta-analysis OR research synthesis OR research review)
World Cat (First Search, OCLC)	(kw: phonic* OR kw: phonetical* OR kw: phonemic) and ((kw: systematic and kw: review) OR kw: meta-analysis OR (kw: research and kw: synthesis) OR (kw: research and kw: review)) and la = 'eng'

PRISMA flow diagram (based on Moher, Liberati, Tetzlaff and Altman, 2009)