International Journal of Educational Research 76 (2016) 190–210

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

Cross-age tutoring in kindergarten and elementary school settings: A systematic review and meta-analysis

Yulia Shenderovich [a,*], Allen Thurston [b], Sarah Miller [b]

[a] Institute of Criminology, Sidgwick Avenue, Cambridge CB3 9DA, UK
[b] Centre for Effective Education, Queen's University Belfast, Belfast , Northern Ireland, UK

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form 13 March 2015
Accepted 19 March 2015
Available online 27 May 2015

Keywords:
Tutoring
Systematic review
Literacy
Peer learning
Volunteer effectiveness
Cooperative learning

ABSTRACT

This systematic review summarizes effects of peer tutoring delivered to children between 5 and 11 years old by non-professional tutors, such as classmates, older children and adult community peer volunteers. Inclusion criteria for the review included tutoring studies with a randomized controlled trial design, reliable measures of academic outcomes, and duration of at least 12 weeks. Searches of electronic databases, previous reviews, and contacts with researchers yielded 11,564 titles. After screening, 15 studies were included in the analysis. Cross-age tutoring showed small significant effects for tutees on the composite measure of reading ($g = 0.18$, 95% CI: 0.08, 0.27, $N = 8251$), decoding skills ($g = 0.29$, 95% CI: 0.13, 0.44, $N = 7081$), and reading comprehension ($g = 0.11$, 95% CI: 0.01, 0.21, $N = 6945$). No significant effects were detected for other reading sub-skills or for mathematics. The benefits to tutees of non-professional cross-age peer tutoring can be given a positive, but weak recommendation. *Effect Sizes* were modest and in the range −0.02 to 0.29. Questions regarding study limitations, lack of cost information, heterogeneity of effects, and the relatively small number of studies that have used a randomized controlled trial design means that the evidence base is not as strong as it could be. Subgroup analyses of included studies indicated that highly-structured reading programmes were of more benefit than those that were loosely-structured. Large-scale replication trials using factorial designs, reliable outcome measures, process evaluations and logic models are needed to better understand under what conditions, and for whom, cross-age non-professional peer tutoring may be most effective.

1. Introduction

Individualized tutoring is considered to be one of the most effective ways to promote improved educational outcomes (Bloom, 1984; Elbaum, Vaughn, Hughes, & Moody, 2000). Non-professional peer tutors can deliver tutoring programmes at schools with reduced costs compared to professional teachers or tutors (Goodlad & Hirst, 1990; Leung, Marsh, & Craven, 2005). Our review considers tutoring schemes, in which children, university students and community volunteers tutor kindergarten and elementary school pupils. These non-professional tutors are considered peer tutors here because they do not have the status of professional educators and are either close in age to the tutees (in the case of school or university student volunteers), or close in terms of background and spatial proximity (in the case of community peer volunteer

* Corresponding author. Tel.: +44 07774955977.
E-mail addresses: y.shenderovich@gmail.com, ys416@cam.ac.uk (Y. Shenderovich).

http://dx.doi.org/10.1016/j.ijer.2015.03.007
0883-0355/

tutors), and share the local environment with tutees. Therefore, we take a wide, inter-generational view of what constitutes a peer within a community.

Tutoring by school pupils, university students and community volunteers has been reported as an effective intervention for improving academic and attitudinal outcomes among school-aged children (Medway, 1995; Ritter, Denny, Albin, Barnett, & Blankenship, 2006; Higgins et al., 2013). Conversely, several studies have found null or negative effects for non-professional tutoring on academic results of tutees (Jensen, 1991; McKinney, 1995; Ritter, 2000). Therefore there is need for a systematic review to assess what high quality studies report in terms of the efficacy of peer tutoring.

1.1. Theoretical background

There is no single dominant theory of change for peer tutoring. Students are expected to improve academic outcomes through elaborating thoughts in the tutoring process, thus cooperatively constructing knowledge within the so-called *zone of proximal development* (ZPD). The ZPD is loosely defined as the distance between child's independent level of problem solving and the level of problem solving under the guidance of a more advanced peer or an adult (Vygotsky, 1978; Chi et al., 2001; Webb, 1989). In this manner peer tutoring is often reported as being a form of cooperative learning (Pesci, 2015). Peer tutoring can provide students with timely feedback (Bloom, 1984; Merrill, Reiser, Merrill, & Landes, 1995), increased time on task (Delquadri, Greenwood, Whorton, Carta, & Hall, 1986) and more appropriate pacing (Shanahan, 1998).

Tutoring programmes are also expected to improve socio-emotional outcomes, such as self-efficacy (Elliott, Arthurs, & Williams, 2000), self-confidence (Margolis, 2005), and child's confidence in the academic subject tutored (Koh, Sanders, & Meyer, 2012). Peer tutoring is reported to result in improved social ties between tutees and tutors (Goodlad & Hirst, 1989), strengthened attachment to the school, and improved attendance at school (Pridmore, Stephens, & Stephens, 2000). Many authors have also suggested that peer tutors can serve as role models for the tutees (Potter, 1994; Topping & Hill, 1995). In this way, peer tutoring by non-professional educators is expected to be qualitatively different from tutoring delivered by professionals and employed teaching staff.

1.2. Ongoing programmes

In the USA since the late 1990s America Reads Challenge has mobilized tens of thousands of college students as volunteer reading tutors for children in Kindergarten through Third Grade (Fitzgerald, 2001). In this context, several manualized programmes were developed, such as Book Buddies which involved 45-minute biweekly sessions consisting of rereading a familiar book, word studies, writing, and reading a new book (Meier & Invernizzi, 2001). In India, a programme called India Reads was managed by the largest educational non-governmental organization, Pratham. The programme is reported to have enabled communities to mobilize and train volunteers to work in schools both during and after school hours. The initiative involved nearly 450,000 community volunteers acting as tutors using techniques described in programme manuals (Poverty Action Lab, 2009). Other programmes have less formal structures for tutoring interactions. The UK literacy charity Beanstalk connected adult community volunteer tutors with 6400 primary school children in England during the 2011–2012 academic year. It provided community volunteers general guidance, such as "Use open-ended sentences to encourage conversation" and "Be generous with your praise" (Beanstalk, 2013).

Most reports available in English have described tutoring programmes in high-income English-speaking countries, such as USA, UK and Australia, but there are also reports of similar projects in other countries, such as China, India, Jamaica, Lithuania, South Africa, Tanzania and Thailand (Goodlad, 1995, 1998). Banerjee and Duflo (2011) reported that tutoring programmes involving community volunteers are currently being tested in Ghana, with plans for similar programmes drafted in Senegal and Mali.

1.3. Existing studies and reviews

Following a number of narrative reviews (Rosenshine & Furst, 1969; Devin-Sheehan, Feldman, & Allen, 1976), Hartley (1977) carried out the first meta-analysis on the topic, identified by this review. Hartley summarized peer tutoring studies in mathematics with child tutors and found a mean Cohen's *d* of 0.6. The widely cited Cohen, Kulik, and Kulik (1982) review examined 65 randomized and matched studies based in elementary and secondary schools with schoolchildren as tutors. It reported significant overall Cohen's *d Effect Sizes* of 0.29 for reading (95% CI 0.17, 0.41) and significant *Effect Sizes* of 0.6 (95% CI 0.29, 0.91) for mathematics. However, Rohrbeck, Ginsburg-Block, Fantuzzo, and Miller (2003) reported that older meta-analyses may have serious methodological limitations, such as 'lax and 'non-transparent' study inclusion criteria. More recent reviews (Wasik & Slavin, 1993; Shanahan, 1998; Wasik, 1998; Elbaum et al., 2000) looked at one-to-one tutoring undertaken by adults, including professional tutors. It was reported that, "college students and trained, reliable adult community volunteers were able to provide significant help to struggling readers" (, p. 616).

More recently, Slavin and Lake (2008), Slavin, Lake, Chambers, Cheung, and Davis (2009a), Slavin, Lake, Cheung, and Davis (2009b), Slavin, Lake, Chambers, Cheung, and Davis (2009c), Slavin, Lake, Davis, and Madden (2010) Slavin, Lake, Davis, and Madden (2011), Slavin and Madden (2011) carried out large Best Evidence Encyclopedia syntheses of various reading programmes in Kindergarten to Fifth Grade. The reviews reported significant standardized mean difference *Effect Sizes* of 0.26 for cross-age tutoring. Leung et al. (2005) conducted a meta-analysis of 68 published studies, in which children

and university students acted as tutors. It was reported that there were significant *Effect Sizes* of 0.65 for overall academic achievement (95% CI: 0.59, 0.71) and 0.88 for self-concept (95% CI: 0.69, 1.07). In contrast, Torgerson and King (2002) and Ritter et al. (2006) focused on randomized controlled trials (RCTs) including only adult non-professional tutors. Torgerson and King (2002) summarized four trials, finding a mean *Effect Size* of 0.19 that was not statistically significant (95% CI: −0.31, 0.68). Ritter and colleagues included 21 USA based studies, finding a significant mean *Effect Size* of 0.3 (95% CI: 0.18, 0.42) for the composite measure of reading and a non-significant mean *Effect Size* of 0.27 (95% CI: −0.18, 0.72) for mathematics. A recent review of 76 randomized experiments in education conducted in low and middle income countries found an average effect of 0.10 for community volunteer teaching (McEwan, 2013). These *Effect Size* estimates are lower than those reported by Leung et al. (2005). Thus results of previous meta-analyses ranged from null to small and medium positive significant effects.

Given the wide diversity of effects identified in previous research, the current review was deemed necessary to systematically identify randomized studies in this area, including the recent research evidence, critically appraise the findings and provide a more precise estimate of the effect of tutoring on academic outcomes. Given the wide use of tutoring programmes, this review is needed to make suggestions for teaching as well as inform possible directions for future research.

2. Method

2.1. Inclusion criteria

To develop inclusion criteria for the review and ensure that only studies with high methodological rigour were included, current criteria published by What Works Clearing House (2010), Cochrane Collaboration (Higgins & Green, 2011) and Best Evidence Encyclopedia (2013) were examined. After close examination and discussion within the review team, a full list of inclusion criteria for this review was developed as follows.

2.1.1. Sample size included at least two classrooms per treatment group

Contextual factors in education research are important (McCartney & Ellis, 2008). In small-scale studies, intervention effects are likely to have confounds with particular schools, classes, or teachers, dramatically limiting generalizability of the results. There will be some common attributes of the 'cluster', and there is a danger in single classroom/context studies. For example teacher quality, school quality or socio-economic status of participants may be more powerful than the effects of the intervention (Slavin & Smith, 2009). Therefore, in agreement with What Works Clearinghouse guidelines (What Works Clearing House, 2010), studies with only one classroom per treatment were not included due to the risk of single context effects biasing reported outcomes.

2.1.2. Randomization was used to assign to treatment or control condition

Randomized controlled trials (RCTs) are studies, in which participants, or groups of participants, are randomly assigned to experimental and control groups. The experimental participants receive treatment, while control participants receive treatment as usual, an alternative treatment or no treatment at all (Bowling, 2009). Randomized controlled trials are widely recognized as the most reliable research design to assess the effectiveness of an intervention as they create two equivalent groups to identify intervention effects (Guyatt et al., 2000; Glazerman, Levy, & Myers, 2003; Petticrew & Roberts, 2003; Agodini & Dynarski, 2004; Wilde & Hollister, 2007). Although randomized controlled trials and high-quality matched studies may identify similar *Effect Sizes* (Torgerson, 2006), randomized controlled trials and matched studies do not always lead to same conclusions (Shadish & Ragsdale, 1996; Glazerman et al., 2003). This review relies exclusively on studies with an RCT research design so that outcomes were not unduly affected by research design.

2.1.3. Outcome measures did not bias treatment over control condition

The review included studies with measures that were reliable and valid. A measure is inherent to the experimental treatment if it assesses particular skills or concepts that have been taught only to the experimental group. Miller, Maguire, and Macdonald (2012) reported that measures described as directly related to the programme's goals may be inherent to the treatment and thus bias any comparison in favour of the intervention group. It follows that findings of a study are determined not only by the intervention investigated and the nature of the comparison group, but also by the quality and independence of measures used. Gersten, Baker, and Lloyd (2000) highlighted that when experimental design was undertaken in education, it was important to distinguish experimenter-developed and external measures. This review included studies that used attainment scales in which the reliability and validity of measures could be ascertained, e.g. where a standardized instrument was used or at least a full description of the psychometric properties of the scale and its scoring were available. Pre-test differences between control and treatments groups had to be reported as non-significant, or any pre-test differences controlled for during analysis.

2.1.4. Outcome measures of academic or socio-emotional ability

Secondary outcomes are outcomes that are not priority of the review, but are important for explaining intervention effects (O'Connor, Green, & Higgins, 2008). Tutoring is theorized to rely not only on cognitive, but also socio-emotional

outcomes (Robinson, Schofield, & Steers-Wentzell, 2005), such as confidence in the academic subject (Koh et al., 2012), self-efficacy (Elliott et al., 2000) and self-confidence (Margolis, 2005). Therefore, although academic outcomes were the primary aim of the review, socio-emotional results, if available, were included as secondary outcomes.

2.1.5. Intervention length was 12 weeks or longer

The review focused on "practical programmes that can be used over extended time periods, not theoretically interesting but impractical procedures that could never be replicated for extended periods" (, p. 11). Consequently, to achieve higher external validity and relevance to school practice, the minimum length for a study to be included in this review was 12 weeks between pre-test and post-test, following Best Evidence Encyclopedia standards (Center for Data-Driven Reform in Education, 2013) on this issue. In contrast, very short programmes may not lead to forming sustainable habits (Lally, van Jaarsveld, Potts, & Wardle, 2010).

2.1.6. Nature of tutoring

(1) School-based programmes using individualized instruction in dyads or small groups, involving a more academically advanced tutor and one or more less advanced tutees (Medway, 1995; Topping, 1998).
(2) Tutor and tutee had fixed roles, i.e. tutoring was non-reciprocal, and tutors and tutees remained in those roles for the duration of the programme.
(3) Tutoring was delivered by classmates or older students, parents, university students, or other adults (for example community volunteers) acting in a non-professional peer tutoring role. Paraprofessional and professional teachers, and professional tutors were excluded[1].
(4) Tutoring took place in a face-to-face setting (this was used as an inclusion criteria as the differences between face-to-face and on-line tutoring have not yet been fully explored in the research literature).
(5) Tutoring was carried out within the school context of the tutee.
(6) The recipients of the tutoring were tutees in a kindergarten, primary, or elementary school setting, which corresponds to the age bracket of five to eleven years old.
(7) Tutoring had an academic focus in any subject area.
(8) Outcome measures for the tutees included attainment tests, and information was provided that allowed *Effect Sizes* to be calculated from the reported data.
(9) Intervention tested tutoring on its own without significant additional components, such as scholarships.
(10) The duration of the tutoring intervention was not less than 12 weeks long.

2.3. Search strategy for identification of studies

Given the spread of published educational intervention research over many resources (Newman, 2003), a wide range of databases were identified to reduce the possibility of missing studies. Modifications of the search string *tutor* AND (peer* OR cross-age OR volunteer*) AND (evaluation* OR program* OR experiment* OR random*) NOT technolog** was used on ASSIA, Australian Education Index, British Education Index, ERIC, International Bibliography of Social Sciences, JSTOR, PsycINFO, PRISMA, ProQuest Dissertations & Theses, Web of Knowledge, Social Services Abstracts, and Sociological Abstracts[2]. In addition to databases, organizations' websites, and bibliographies of key studies, literature reviews and meta-analyses were searched for review titles. Furthermore, 104 researchers who have published studies on tutoring were contacted by email to identify unpublished studies.

Data presented in Fig. 1 shows the flow diagram of identification and screening of studies. A total of 11,564 titles were retrieved through the review searches. Citations were imported into Microsoft Excel, which was used to remove duplicated records, leaving 10,910 unique titles and abstracts. Initial screening of titles and abstracts by the first author left 183 studies for further review. Shenderovich and Thurston also examined the full list of titles to discuss any studies that caused uncertainty as to whether further screening would be required and made decisions in each case. Full texts of the 183 studies were obtained and assessed for eligibility by Shenderovich. Both Shenderovich and Thurston further screened a randomly selected 20% of studies with no disagreements. After the screening, fifteen studies (reporting data from 16 cohorts of participants) fulfilled all inclusion criteria as determined by two authors[3]. All studies were fully coded by Shenderovich, and half were blind double-coded by both Shenderovich and Thurston. The other half of included studies were checked by Thurston for coding accuracy and to ensure inclusion criteria were met.

[1] To distinguish peer volunteers and paraprofessionals, this review considered tutors to be volunteers if they received no payment at all or if they were only reimbursed for travel to the school (Lee, Morrow-Howell, Jonson-Reid, & McCrary, 2012) and other participation costs incurred (Cabezas et al., 2011).
[2] Using * (wildcard) at the end or in the middle of a word will return searches of all letter strings/spellings that are contained in the string. For example randomi*ed would return all search items with spelling of both randomized (USA spelling) and randomised (UK spelling).
[3] One study (Allor & McCathren, 2004) included two separate cohorts of students in two consecutive years.

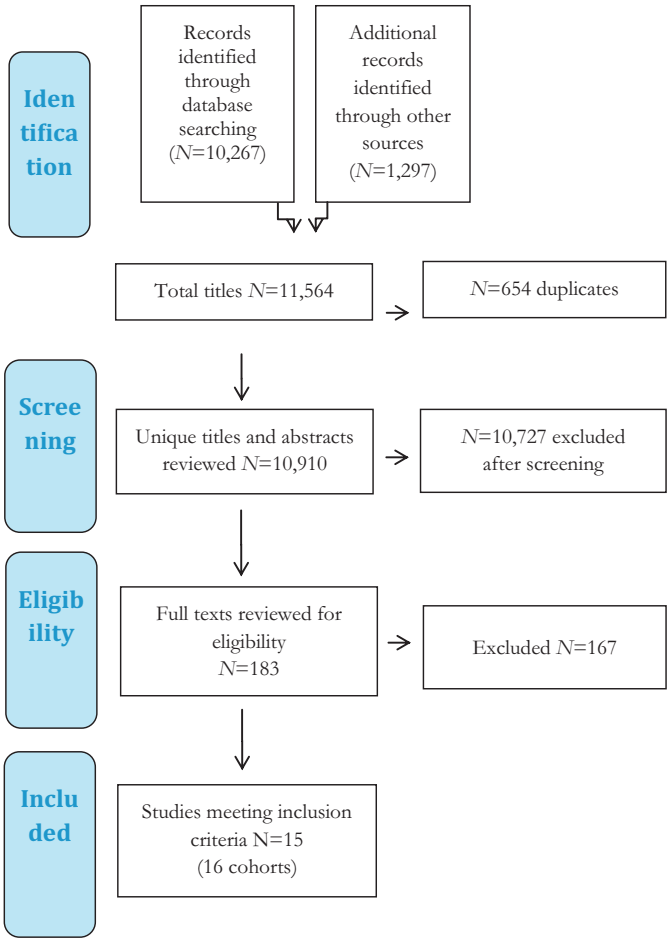

Fig. 1. Flowchart of study selection adapted from PRISMA Statement (Moher, Liberati, Tetzlaff, & Altman, 2009).

2.4. Effect size calculations

To determine if tutoring had greater effect in any area of reading sub-skills, reading outcomes were categorized under the following categories for separate meta-analyses: comprehension, fluency, decoding, writing and overall reading ability, using the approach adopted by Ritter et al. (2006). As mathematics outcomes are categorically different from reading outcomes, reading and mathematics outcomes were maintained as separate variables. In cases where several measures within a study assessed the same construct, *Effect Sizes* and their confidence intervals were averaged to make sure that no study was unduly weighted (Becker, Hedges, & Pigott, 2004), assuming a correlation of 0.5 between related scores (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Analyses were carried out using Comprehensive Meta-Analysis software, version 2 (Biostat Englewood, NJ). Standardized mean difference (Cohen's *d*) is the appropriate *Effect Size* metric to contrast two groups on continuous variables, such as test performance (Lipsey & Wilson, 2001). Standardized mean difference is calculated as difference in mean outcomes between groups divided by pooled standard deviation of outcome among participants. *Effect Sizes* and confidence intervals were divided by Hedges's approximation (Hedges & Olkin, 1985; Lipsey & Wilson, 2001). Given the diversity of tutoring programmes, random-effects model was pre-selected in the review protocol to make studies more equally weighted (Sterne, Egger, & Smith, 2001) and results more generalizable (Field, 2001). Manuscript authors were contacted directly if any missing information was needed to calculate *Effect Sizes*.

In educational research it is common to assign groups of children, such as classes or schools, to treatment and control groups (Boruch et al., 2004; Campbell, Elbourne, & Altman, 2004). The effective sample size in a cluster-randomized trial is the original sample size divided by the "design effect", which equals $1 + (M - 1) \times ICC$, where M is the average cluster size and ICC is the intra-cluster correlation coefficient (Higgins et al., 2008). ICC adjustment was applied for the Elliott et al. (2000) study, the only included cluster-randomized trial. We used ICC of 0.15, the value suggested by a recent compilation of research on intra-class correlation values of academic achievement in the USA (Hedges & Hedberg, 2007).

3. Results

3.1. Description of the included studies

As described in Table 1, eleven of the investigations were carried out in USA, four in the UK and one in Chile. The majority of tutoring programmes focused on low-achieving children, indicated either by their classroom teacher or test assessment. In respect to external validity, it is important to point out that the majority of studies recruited what appeared to be a convenience sampling of classrooms and schools, and are therefore not necessarily generalizable to other settings. However, some studies used representative samples, either of local schools (Miller & Connolly, 2012) or of the tutoring programme's participants (Lee et al., 2012). All studies focused on schools with disadvantaged socio-economic profiles. Several programmes targeted one age group (Pullen, Lane, & Monaghan, 2004; Allor & McCathren, 2004—Gr 1; Cabezas, Cuesta, & Gallego, 2011—Gr 4), while others included a variety of primary school grades (Ritter, 2000—Gr 2–5; Lee et al., 2012—Gr 2–3).

Study sizes ranged from small-scale trials with 42 (Rimm-Kaufman, Kagan, & Byers, 1998) and 47 children (Pullen et al., 2004), to large-scale studies with 734 (Miller & Connolly, 2012), 883 (Lee et al., 2012) and 6136 children (Cabezas et al., 2011) enrolled respectively. In total studies involved 9484 participants. Following the approach of Best Evidence Encyclopedia, this review defines large studies as those with greater than 250 participants (Slavin, 2008). Five included studies with samples over 250 looked at on-going programmes (Experience Corps, West Philadelphia Tutoring Project, Time to Read, Servicio País en Educación) in multiple locations and, thus, were effectiveness – as opposed to efficacy – studies (Flay, 1986; Flay et al., 2005).

Most included studies focused on reading, and two studies involved tutoring in mathematics. Ham (1977) assessed the "halo effect" of tutoring in reading on achievement in mathematics. The observed emphasis on reading focused studies could be reflective of the importance of reading in primary school, as well as of the more complex nature of designing tutoring procedures in mathematics (Topping, 2004). Studies identified by this review did not target any other academic subjects.

Two cohorts included in the review utilized older schoolchildren as tutors (Jensen, 1991; Policy Studies Associates, 2007), and fourteen investigated tutoring by adults (eight of them involved adult community volunteers, and six with university student volunteers). All studies except one involved English-language instruction (Cabezas et al., 2011 studied reading in Spanish language in Chile). In addition to tackling outcomes of primary school tutees, some of programmes aimed to improve achievement of tutors who were school or university students (Policy Studies Associates, 2007) or to contribute to social wellbeing of older tutors (Lee et al., 2012).

Seven studies examined programmes that prescribed specific tutoring lessons and materials or specified time allocated for various activities. This review characterizes such programmes as "highly structured"—incorporating standardization by precise activities or by functions and processes (Backer, 2001). More structured programmes also had more extensive tutor training. For instance, Pullen et al. (2004) provided university student volunteers with step-by-step lesson guides, and the tutoring sessions were observed by supervisors. On the other hand, nine studies provided only general advice to tutors and are therefore classified as "loosely structured". For example, in Northern Ireland the Time to Read programme, evaluated by Miller et al. (2012a), adult community volunteers did not receive a pre-set tutoring session structure. In Baker, Gersten, and Keating (2000), adult community volunteers were "provided with a broad framework to use during sessions, rather than specific techniques" (p. 497). Similarly, in the Ritter (2000) evaluation of West Philadelphia Tutoring Project, tutors (University of Pennsylvania volunteer students) had general guidance on working with their tutees, and curriculum guides were only provided in some of the participating schools. There was no structured process evaluation, but anecdotal reports suggested that during sessions tutors helped pupils with homework tasks or made up their own exercises in reading and mathematics.

3.2. Description of excluded studies

Most studies were excluded due to lack of randomization. In addition, to examine sustainability, a minimum of 12 weeks length was set for inclusion, as discussed above, which left out several otherwise eligible studies. For instance Spörer, Brunstein, and Kieschke (2009), randomized 210 elementary school children from 4 classes in a medium-sized German town to four groups: instructor-guided small groups; direct instruction followed by reciprocal tutoring; a mix of direct instruction and reciprocal tutoring; and a no-intervention control group. However, the study only lasted seven weeks. In addition, several studies were excluded because of a lack of eligible comparison groups.

In another excluded paper, an unpublished study based in migrant schools in Beijing, China (Li, Han, Rozelle, & Zhang, 2010), all study groups were paid for grades, and, in addition, a third of the 850 students received tutoring from classmates and a third tutoring from classmates, plus a parental communication intervention. Thus, there was not a tutoring only group where no payment was made available. It was reported that tutoring and pay showed an *Effect Size* of 0.14 on reading and the group with tutoring and pay plus parental communication had an *Effect Size* of 0.2. Another study (Banerjee, Banerji, Duflo, Glennerster, & Khemani, 2010) describes a set of interventions evaluated in 65 randomly assigned villages in India in 2005. Similarly, none of the interventions tested tutoring on its own, so the study was not included. All three interventions involved sharing information on educational resources with communities through small-group discussions. A second intervention also included offering communities testing tools to assess children's reading and mathematics results, and the third facilitated community volunteer tutors providing afterschool reading.

Table 1
Overview of key features of the included studies.

Authors of study	*N*	Description of tutees	Description of tutors	Total (In hours)	Per week (In hours)	Length (weeks)	Fidelity	Location	Intervention description
Allor and McCathren (2004)	86 year 1	Gr.1 *M* = 6.7 y.o.	University education major student volunteers	12	1	26	Used a checklist *M* = 86.98% (*SD* = 5.67)	8 Underachieving schools, urban south USA	Outside class during school day Remedial tutoring for low-achieving children Tutor training: America Reads tutor training, 3 1-hour trainings, monthly training, and on-site assistance Scripted lessons with progressively challenging lessons, containing games on phonemic awareness, letter-sound correspondence, word-study activities and reading of levelled books 3 Research assistants observed and supported tutors
Allor and McCathren (2004)	157 *year 2*	Gr.1 *M* = 6.6 y.o.		13	1	26	*M* = 86.53% (*SD* = 4.80)	10 underachieving schools	
Baker et al. (2000)	84	Gr.1	Adult community volunteers (33% 30–45 y.o., 29% 45–65, 20% > 65)	37	1	72	Not reported	6 Title-1 schools, Oregon, USA	Outside class during school day Remedial tutoring for low-achieving children Tutees selected based on reading difficulties and need for relationship with a caring adult Tutor training: 1-2 hour training and community volunteer handbook Tutoring focused on increasing children's interest in reading, program providing books for children to take home.
Cabezas et al. (2011)	4903	Gr.4 9–10 y.o.	University student volunteers	18	1.5	12	High volunteer turnover	85 Vulnerable schools in 10 counties in Biobio and Great Santiago regions, Chile	After class School-wide one to small group tutoring (5–6 students assigned to a tutor) Tutoring focused on "shared-reading … of traditional stories and informative texts, which are age-and interest appropriate for students" Volunteers supported by an employee of "Fundación para la Superación de la Pobreza" at each school Volunteers received stipends for travel
Elliott et al. (2000)	30	Reception class 4–5 y.o.	Adult community volunteers	19	1	19	Didn't measure	3 Low-SES schools, Northeast England, UK	"Time for Reading" During school day, both in and outside classroom Class-wide tutoring one to small group tutoring Tutor training: 6 hours over 3 weeks Tutors worked alongside classroom teacher, providing "individual assistance … The focus of the work was reading for meaning and most of the training sessions involved the child reading to the helper from a fiction text and discussing elements of the story"

Ham (1977)	147	Gr. 1, 2, 3	Adult community volunteers	36	2	22	Record keeping failed, high tutor turnover	4 Schools with low SES & minority students, Sumter County, rural USA	During language arts classes, outside class One-to-one and small groups tutoring Remedial tutoring for low-achieving children Tutors worked following teachers' recommendations, "because of the turnover in volunteers and because volunteers as persons are difficult to program or control, plans for standardization of instructional approach had to be abandoned" p. 63
Jensen (1991)	93	Gr.2	Gr. 5	46	2	23	Not reported	7 Elementary schools, Cache Valley, Utah, USA	One-to-one tutoring Remedial tutoring for low-achieving children Tutor training: weekly sessions on "effective tutoring techniques, error correction procedures, and proper prompting techniques"; effects on tutors also assessed Tutoring focused on timed reading aloud, reading passages assigned by paraprofessionals; tutors corrected mistakes and feedback for correct reading, asked comprehension questions
Lee (1980)	40	Gr.3–6	University volunteers, juniors and seniors	76	4	19	Not reported	4 Schools, low SES & minority, urban USA	After class One to small group tutoring Remedial tutoring for low-achieving children, or based on minority status or residence Tutoring focused on homework assignments, improving reading and maths skills, addressing personal concerns Tutor training: 7 training modules; tutors supervised by two graduate counselling students
Lee et al. (2012)	881	Gr.1, 2, 3 $M = 7.09$ y.o.	Adult community volunteers, 50 to 93 y.o., mean 65	21	1.75	36	Not reported	81 Schools in Boston, 52 in New York, and 41 in Port Arthur, USA	"Experience Corps" One-to-one tutoring Remedial tutoring for low-achieving children Tutor training: 15 to 32 hours NY: Book Buddies (phonics, rereading familiar books, word study, writing, and reading a new book) Boston: Reading Coaches (building student's oral vocabulary and increasing reading comprehension by asking prediction questions, discussing, and writing about the story) Port Arthur: Brigance Inventory of Basic Skills materials (word recognition, comprehension, and word analysis) Nationally, 43% of community volunteers have high school diplomas, and 75% –some college education, some are former teachers

Table 1 (*Continued*)

Authors of study	*N*	Description of tutees	Description of tutors	Total (In hours)	Per week (In hours)	Length (weeks)	Fidelity	Location	Intervention description
Loenen (1989)	81	7–11 y.o., *M* = 8.8 y.o.	Adult community volunteers	24	1	26	Observed 15 tutors, low fidelity to the training	13 Schools in inner London, UK	"Volunteer Reading Help" Outside class during school day One-to-one tutoring Remedial tutoring for low-achieving children Tutor training: short compulsory training course (3 1.5- sessions on reading & practical tips) Volunteers encouraged to talk to teachers, but no formal structure
Miller, Connolly, Odena, and Styles (2009)	734	8–9 y.o.	Adult community volunteers	13	0.5	58	High tutor turnover, "variation in delivery"	Northern Ireland, UK 50 schools	"Time to Read" Outside class during school day One-to-one tutoring Remedial tutoring for below-average performing children Tutor training: half-day tutor training in paired reading strategies to improve reading fluency, word recognition, meaning, and comprehension for tutors, emphasizing repetition, alternate reading, word recognition, word meaning and comprehension, no structure provided for the sessions but a set of books. Some children received a workplace visit
Miller et al. (2012a)	483	8–9 y.o.	Adult community volunteers	29	1	29	Not recorded	50 Schools in Northern Ireland	"Time to Read" See above (note increased intensity/dose)
Policy Studies Associates (2007)	124	Gr.2	Gr. 4–5	72	2	36	Not recorded	Irving, TX, and Mongomery County, Maryland, US	"Reading Together" Outside class during school day One-to-one tutoring Remedial tutoring for students at risk of reading failure, Tutor training: 9 hours Tutoring focused on a curriculum on "reading comprehension, reading fluency, vocabulary, and writing... to move students from decoding to comprehending"
Pullen et al. (2004)	47	Gr.1	University student volunteers, majors related to education	10	0.75	12	Used a checklist *M* = 92%	North-central Florida, US 10 schools	Outside class during school day One-to-one tutoring Remedial tutoring for students below 30th percentile Tutor training: 4 h Three-step tutoring model: repeated reading of familiar text, explicit coaching in decoding and word-solving strategies, and reading new books during each session

Rimm-Kaufman et al. (1998)	42	Gr.1	Community volunteers	72	2.25	35	Not reported	Cambridge, MA, US 6 schools	Outside class during school day One-to-one tutoring - Remedial tutoring for students below 30th percentile Tutor training: 5 sessions and bimonthly meetings Prescribed tutoring session schedule: reading for meaning associations between print and pictures, phonetics taught within the context of stories). "The tutors used games, drawing, writing, and related activities to engage the children in learning"
Ritter (2000)	319	At-risk Gr.2,3,4, 5	University volunteers	21	1	21	Not reported	Philadelphia, PA, US 11 schools	"West Philadelphia Tutoring Project" Outside class during school day One-to-one tutoring Remedial tutoring for students below 30th percentile Tutor training: minimal training and supervision Limited tutoring structure - "variety of tasks ... spelling, reading, math problems, games, puzzles, crafts, and storytelling"

*SES-socioeconomic status, y.o.- years old.

3.3. Overall effects

The review suggested small (as defined in Cohen, 1988) statistically significant positive effects, with high heterogeneity, of cross-age tutoring programmes on reading overall, as well on decoding and comprehension skills, while outcomes on other reading measures and mathematics were non-significant. The high heterogeneity of findings for many of the outcomes indicates that the studies, populations and interventions included are diverse.

Outcome measures were grouped into seven categories, following the example of the Ritter et al. (2006) systematic review:

- Composite measure of reading: measure combining all reading scales available in each study (see Forest plot in Fig. 2).
- Overall reading: overall batteries in reading achievement tests.
- Decoding: subtests on decoding of words and knowledge of words, consonant sounds, short vowels, digraphs and combinations, sight words, and non-word decoding.
- Comprehension: reading comprehension subtests.
- Fluency: fluency subtests.
- Writing: writing subtests.
- Mathematics: mathematics outcomes.

These seven categories covered the reported attainment measures of all included studies and therefore form an all-inclusive set of outcome descriptors. Fig. 2 shows the composite measure of reading, with upper and lower *Effect Sizes* for the battery of reading tests reported by each manuscript.

3.4. Homogeneity analysis

Table 2 lists several measures of homogeneity. Q represents a standardized measure of total variation, and df, the expected variation. Thus Q minus df is the excess variation. The Q statistic and its p-value are a test of significance of the viability of the null hypothesis of zero true dispersion. I^2 is the percentage of the dispersion that is real and not due to sampling error. Higgins, Thompson, Deeks, and Altman (2003) tentatively suggest that I^2 values of 25%, 50%, and 75% are respectively low, moderate, and high, with about a quarter of meta-analyses having I^2 over 50%. Finally, T^2 is the variance and T the standard deviation of true effects, measured on the same scale as effects. The level of heterogeneity for decoding, fluency and composite measure of reading was high. Nevertheless, Ioannidis, Patsopoulos, and Evangelou (2007) suggest that overall meta-analysis is usually desirable, even with high statistical heterogeneity. Although statistical homogeneity tests are weak and not very precise (Ioannidis et al., 2007; Thorlund et al., 2012), statistical heterogeneity can be a useful tool (Berlin, 1995) as it points to the presence of clinical or methodological diversity, or both (Deeks, Higgins, Altman, 2011).

3.5. Sensitivity analysis

Sensitivity analysis is necessary to assess potential bias that may be associated with individual *Effect Sizes* and distort the aggregated effects (Hedges & Olkin, 1985). "One Study Removed" analysis allows to assess if any single study has disproportionate influence. In this set of studies, several very large samples are present. In particular, the large sample ($N = 4903$) in Cabezas et al. (2011) made up 59% of all reading studies' participants. Using a random effects model, all

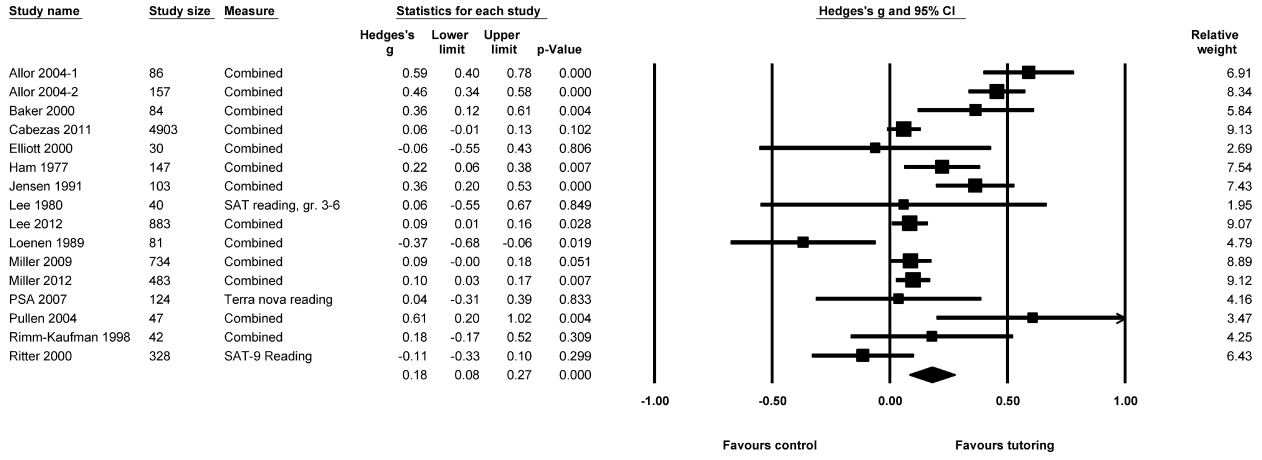

Fig. 2. Forest plot of comparison between control group and tutoring on the composite measure of reading.

Table 2
Effect Sizes and random effects of included studies.

Outcome area	*N* cohorts	*N* students	Hedges' *g* (random effects)	95% CI	*p*-Value	Heterogeneity
Composite measure of reading	16	8251	**0.18***	0.08, 0.27	< 0.001	$Q = 97.8$; $df = 15$; $p = 0.000$; $I^2 = 84.663$; $T = 0.155$; $T^2 = 0.024$
Overall reading ability measure	6	1457	0.07	−0.06, 0.20	0.299	$Q = 7.903$; $df = 5$; $p = 0.162$; $I^2 = 36.737$; $T = 0.095$; $T^2 = 0.009$
Decoding measure	9	7081	**0.29***	0.13, 0.44	**0.000**	$Q = 60.095$; $df = 8$; $p = 0.000$; $I^2 = 86.688$; $T = 0.208$; $T^2 = 0.043$
Comprehension measure	10	6945	**0.11***	0.01, 0.21	0.025	$Q = 15.223$; $df = 9$; $p = 0.085$; $I^2 = 40.877$; $T = 0.091$; $T^2 = 0.008$
Fluency measure	4	687	0.11	−0.21, 0.44	0.494	$Q = 13.104$; $df = 3$; $p = 0.004$; $I^2 = 77.106$; $T = 0.275$; $T^2 = 0.075$
Writing measure	3	4975	0.01	−0.07, 0.09	0.774	$Q = 0.281$; $df = 2$; $p = 0.869$; $I^2 = 0.000$; $T = 0.000$; $T^2 = 0.000$
Mathematics measure	3	506	−0.02	−0.18, 0.13	0.778	$Q = 1.774$; $df = 2$; $p = 0.412$; $I^2 = 0.000$; $T = 0.000$; $T^2 = 0.000$

* Significantly different from zero, $p < 0.05$, favouring tutoring over the control.

estimates with one study removed fell inside the 95% confidence interval of the overall estimate with all available studies. Therefore no study was found to have an excessive influence on results.

3.6. Publication bias

Five of the included studies have not been published in academic journals. Three were dissertations and two were reports. Non-significant or negative results, especially in small-sample studies, are often not submitted or not accepted for publication, although they may be of equal quality as published work (Iyenger & Greenhouse, 1988; Hopewell, Loudon, Clarke, Oxman, & Dickersin, 2009). To assess the possibility of publication bias, the "trim and fill" procedure (Duval & Tweedie, 2000) was conducted for each outcome to identify and correct funnel plot asymmetry (see Fig. 3 for composite measure of reading funnel plot). The "trim and fill" procedure for the composite measure of reading did not indicate any missing studies. However, there was an indication of studies missing to the left of mean effect sizes for the overall reading ability, comprehension, decoding, and mathematics measures, suggesting possible publication bias. The impact of publication bias still may be trivial as at least 8–10 studies are required for trim-and-fill test to have sufficient power (Sutton, Duval, Tweedie, Abrams, & Jones, 2000a; Sutton, Duval, Tweedie, Abrams, & Jones, 2000b). In addition, Egger's regression testing asymmetry of the funnel plot was not significant ($p > 0.05$) for any measure, indicating low risk of publication bias, although the small number of studies does not allow for definitive conclusions.

3.7. Moderator analyses and meta-regressions

Several programme features were examined through subgroup analyses and meta-regressions. Grouping of studies was used to assess the possibility of varying reading outcomes of different types of programmes to analyse possible sources of heterogeneity (see Table 3 for a summary). Mixed effects analysis was used, meaning that random-effects model is used within groups and fixed effects across subgroups with pooled estimates of T^2. Studies were grouped by the variable of

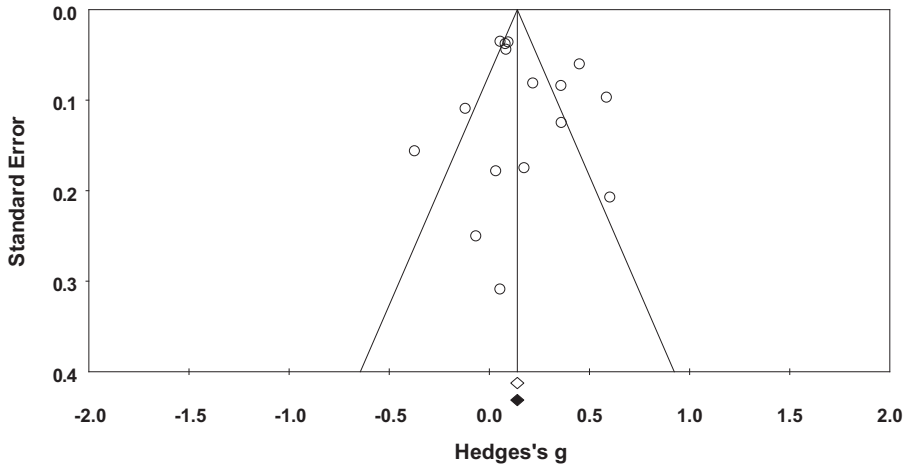

Fig. 3. Funnel plot of standard errors by Hedges's *g* for composite measure of reading, random-effects.

Table 3
Reading Effect Sizes by moderator.

Study feature	*N* cohorts	Hedges' *g* (random effects)	95% CI	Homogeneity between groups (random effects)
Study size				
Large	5	0.08	0.04, 0.11	$Q = 9.771^*$, $df = 2$, $p = 0.008$
Small	11	0.23	0.07, 0.39	
Publication status				
Journal article	11	0.21	0.08, 0.34	$Q = 0.619$, $df = 0.6$, $p = 0.431$
Report or dissertation	5	0.13	−0.03, 0.28	
Type of tutor				
Older child peer tutor	2	0.24	−0.07, 0.55	
University student	6	0.28	0.03, 0.53	$Q = 2.230$, $df = 2$, $p = 0.328$
Adult community volunteer	8	0.11	0.03, 0.18	
Tutoring structure				
Loosely structured	9	0.33	0.14, 0.52	$Q = 5.903^*$, $df = 1$, $p = 0.015$
Highly structured	7	0.08	−0.01, 0.16	

interest, and subgroup effects were compared using significance of *Q* to see if *Effect Sizes* between groups were statistically different.

3.7.1. Study size

Eleven studies had samples of 30 to 157 children, and were coded as "small", while five studies with samples of 328 to 4903 were coded as "large". Difference between two groups was statistically significant for Composite measure of reading ($p < 0.01$) and Decoding ($p < 0.001$), with larger studies showing significantly smaller effects than smaller studies. This is a common feature when reporting data in systematic review and comparing studies. Similarly to previous studies, there were much higher levels of heterogeneity among smaller studies ($Q = 41.176$, $df = 10$, $p < 0.001$, $I^2 = 75.714$) than among larger studies ($Q = 3.714$, $df = 4$, $p = 0.446$, $I^2 = 0.000$). Smaller studies are subject to higher sampling variation (Higgins & Altman, 2008) and have lower statistical power, increasing likelihood of a false positive result (Christley, 2010). Larger studies produce more precise estimates as they are generally better powered to detect effects (Ginsburg-Block, Rohrbeck, & Fantuzzo, 2006). Method of moments meta-regression suggests no significant correlation between study size and composite measure of reading ($p_{slope} = 0.315$).

3.7.2. Tutoring structure

Highly structured programmes (9 studies, $g = 0.33$, 95% CI: 0.14, 0.52, $N = 1388$) had a significant advantage over low-structure programmes (7 studies, $g = 0.08$, 95% CI: −0.01, 0.16, $N = 6863$) on the Composite measure of reading outcome. Comparing groups with the *Q*-test, $Q = 5.903$, $p = 0.02$, thus *Q* is statistically significant, and *Effect Size* is related to the level of structure.

3.7.3. Type of tutor

Subgroup differences by type of tutor comparing tutors who were university students, adult community volunteers or school children did not indicate significant differences in random effect analysis.

3.7.4. Publication status

Subgroup differences depending on publication status (published or unpublished report or thesis) did not indicate significant differences in random effect analysis.

3.7.5. Amount of tutoring

Method of moments meta-regression examines differences in the effect of tutoring on composite measures of reading, depending on 'dose' of tutoring, as measured by the number of tutoring hours. Amount of tutoring did not give a good explanation of effectiveness of tutoring in included studies ($p_{slope} = 0.584$).

3.8. Social, self-concept and behavioural outcomes

Few studies included in this review tested non-academic outcomes alongside academic skills. Due to their diversity and small number, non-academic results were not meta-analysed but are summarized in Table 4, and all were non-significant except one.

3.9. The quality of evidence

Littell, Corcoran, and Pillai (2008), p. 72 propose that "Even when a review is limited to randomized controlled trials, a deeper assessment is needed to judge variations in quality of those studies that may be associated with bias." This is particularly important because randomized controlled trials in school and educational settings are reported to have lower

Table 4
Non-academic outcomes in the included studies.

Study	Outcome	Scale	Hedges *g* (95% CI)
Lee (1980)	Self-concept of reading	Piers-Harris children's self-concept scale (Piers & Harris, 1969)	0.31 (−0.32, 0.93)
	Classroom behaviour	Devereaux elementary school behavior rating scale (Spivack & Swift, 1967)	−2.12 (−2.90, −1.35)
Loenen (1989)	General self-concept	McDaniel-Piers young children's self-concept scale (McDaniel & Leddick, 1978)	0.06 (−0.39, 0.51)
	Composite classroom behaviour	Rutter B-scale for teachers (Rutter, 1967)	−0.10 (−0.58, 0.39)
Miller et al. (2009)	Future aspirations	Future aspirations measure (Loeber et al., 1991)	0.17* (0.02, 0.33)
	Enjoyment of learning	Enjoyment of learning (Pell and Jarvis's 2001)	−0.09 (−0.22, 0.03)
	Self-esteem	Global self-worth scale of the self-perception profile for Children (Harter, 1985)	−0.04 (−1.87, 0.10)
	Locus of control	Rotter's locus of control scale (Rotter, 1966)	−0.05 (−0.31, 0.21)
Miller et al., (2012a)	Enjoyment of reading	The Garfield elementary reading attitudes scale (McKenna & Kear, 1990)	0.03 (−0.11, 0.17)
	Reading confidence	The reader self-perception scale (Henk and Melnick, 1995)	0.03 (−0.13, 0.22)
	Aspirations for the future	Aspirations for the future scale (Loeber et al., 1991)	0.03 (−0.11, 0.17)

* $p < .05$.

quality than in healthcare (Torgerson, Torgerson, Birks, & Porthouse, 2005). Assessments of domains of bias specified in Cochrane Collaboration Risk of Bias Tool (Higgins & Altman, 2008) are outlined below. As reported in Table 5, the included studies did not address many areas of potential bias.

3.9.1. Selection bias

Only four studies specified their approach to generation of randomization sequence, and all four used computer-generated sequences. Two studies, Loenen (1989) and Ritter (2000) discussed practical challenges surrounding gaining cooperation from schools for randomization. Therefore, it is not possible to rule out selection bias as a contributing factor to effects in some studies due to sequence generation and allocation concealment.

3.9.2. Performance and detection bias

Although blinding of study participants and intervention personnel (such as class teachers and tutors) is not possible in a tutoring intervention, it may be possible to blind the assessors. Six of the studies did this. Rimm-Kaufman et al. (1998) reported that classroom teachers were blinded to which children were assigned to the control group.

3.9.3. Attrition bias

The studies described a wide range of attrition levels, some as high as 35%. There was no standard approach to intention to treat analysis and so it was not possible to assess attrition risk in a quantifiable manner.

3.9.4. Reporting bias

The presence of differences between reported and unreported findings could not be assessed due to lack of study protocols.

Table 5
Cochrane collaboration risk of bias tool application in the included studies.

Study	Selection bias: sequence generation	Detection bias: blinding of outcome assessors	Attrition bias: incomplete outcome data
Allor and McCathren (2004)	?*	?	−*
Allor and McCathren (2004)	?	?	−
Baker et al. (2000)	?	+*	+
Cabezas et al. (2011)	?	?	+
Elliott et al. (2000)	?	?	+
Ham (1977)	+	?	+
Jensen (1991)	?	+	+
Lee, 1980	?	?	−
Lee et al. (2012)	?	?	−
Loenen, 1989	?	?	−
Miller et al. (2009)	+	+	−
Miller et al. (2012a,b)	+	+	−
Policy Studies Associates (2007)	?	?	−
Pullen et al. (2004)	?	+	−
Rimm-Kaufman et al. (1998)	?	+	?
Ritter (2000)	+	?	−

* Note, + low risk of bias, − high risk of bias,? unclear risk of bias.

3.9.5. Other biases

(1) There were significant pre-treatment (baseline) differences between treatment and control groups (either due to chance or problems with randomization) in two studies (Jensen, 1991; Pullen et al., 2004), but it was reported that differences were accounted for in ANCOVA analyses.
(2) There was a lack of long-term follow up measurements in the included studies. A possible explanation for this may be sue to ethical and practical difficulties of having a no-intervention control group in schools. Only the Policy Studies Associates (2007) and Elliott et al. (2000) studies had follow-up assessments. Thus the review is primarily based on post-test (tests at the end of interventions) rather than on follow-up measures. Longevity of change was therefore difficult to assess.
(3) Five large studies used multilevel modelling to account for classroom and school effects. However, smaller studies did not adjust for clustering effects within classrooms and schools, and as , p. 12 note, "clustered nature of data" is present when children come from the same classrooms and schools, violating statistical assumptions of independence.

4. Discussion

Whilst publication bias was not apparent, evidence presented by the review must be viewed with caution due to high heterogeneity, quality limitations and small number of included studies. The review suggested that tutoring programmes had small positive effects on combined measures of reading as well as specifically on decoding and comprehension. However, Chall's synthesis of theories of reading concludes that both decoding and fluency skills are necessary for comprehension skills to develop (Chall, 1989). One explanation is that decoding and comprehension measures had more eligible large and well-powered studies included in the synthesis, and thus the meta-analyses for these measures had more power to detect effects (Borenstein et al., 2009).

In-line with previous reviews on tutoring (Fitz-Gibbon, 1977; Palincsar & Brown, 1989; Wasik & Slavin, 1993; Ginsburg-Block et al., 2006; Ritter et al., 2009; McEwan, 2013), studies with a pre-set structure of tutoring report greater *Effect Sizes*. This could support the idea that "open-ended discussions and explanations are problematic, confusing and ineffective" (, p. 16). Non-trained tutor behaviours have been reported to use 'knowledge-telling' rather than 'knowledge-building' explanations (Roscoe & Chi, 2007). However, findings of subgroup analyses are observational and should be treated with caution as we cannot account for potential confounders. For example, it is also possible that more structured programmes were better organized in other respects, such as better tutor training. Moderator analyses suggested that using different types of reading tutors, depending on who is available in the given community, could produce similar results, if a structured tutoring programme was established. However, the number of studies is small, and only two eligible studies with child tutors were identified.

Based on meta-regression results, there was no difference in reading outcomes by dose of tutoring, as measured by number of hours. It should be noted that meta-regressions have very weak statistical power a low number of studies. Regarding this apparent lack of dose-response relationship in tutoring, the findings of this review are in line with results of recent large-scale randomized trial of peer tutoring study in Scotland, The Fife Peer Learning Trial (Tymms et al., 2011). A no-intervention control group was absent, and the different groups served as controls to each other (e.g. reading tutoring children served as controls for mathematics and vice-versa), so the study was not included in this review. The study was a large-scale district-wide effectiveness trial involving two15-week tutoring periods spread out over two years (129 elementary schools, nearly 9000 pupils). The factorial design examined effects of intensity (once per week against three times per week), cross-age (10 year olds tutoring 8-year olds) against same-age tutoring (8-year olds) and tutoring in maths only, reading only and both reading and maths. HLM analysis indicated that intensity did not have a significant effect on outcomes in Performance Indicators in Primary Schools standardized tests, but that *Effect Sizes* for cross-age tutoring were significantly greater than for same-age tutoring (0.25 as compared to 0.02).

On the other hand, Vadasy, Jenkins, Antil, Phillips, and Pool (1997) compared a group of paraprofessional tutors who came to each session and tutored the full amount of time to a group who did not follow time commitments as closely. The study found much higher *Effect Sizes* for tutees whose tutors attended regularly, suggesting that quantity of tutoring may have an impact on student outcomes. However, it should be noted that the study had a very small sample of 20 students. Similarly, in Lee et al. (2012) reported gains were slightly stronger (*Effect Size* 0.01–0.04) on three out of four decoding measures for students who received at least 35 tutoring sessions. However, it is possible that the Fife Peer Learning Project gives better comparability as students received fewer sessions by design and findings were unlikely to be biased by clustering effects of the quality of implementation.

There was not a significant correlation between study size and *Effect Size*, but the five large tutoring studies had significantly lower effects than the smaller studies. Thus, the large studies seemed to disagree with the smaller ones. Four out of five of the largest cross-age tutoring studies also had low-structure sessions, so differences could have been an artefact of low structure of sessions in the large studies. Still, this difference could point to super-realization bias as smaller studies offer the potential to be closely overseen by researchers (Cronbach et al., 1980). LeLorier, Gregoire, Benhaddad, Lapierre, and Derderian (1997) reviewed 12 clinical medical interventions and reported that outcomes of larger studies (1000 patients or

more) were not predicted accurately 35% of the time by earlier meta-analyses on the same topics. Based on included studies in this review, it appears likely that "the larger studies tend to be those conducted with more methodological rigour, or conducted in circumstances more typical of the use of the intervention in practice" (, p. 321), so evidence from large trials needs to be given priority when using systematic reviews to report results that may be generalizable.

4.1. Implications for research

4.1.1. Protocol registration and rigorous study design and reporting

One of the important observations from this review is the need for standardized publication of research protocols. Ideally this should take place prior to research being conducted. Protocols should make particular note of procedures for randomization. In addition it is vital that data is given on demographics of research participants. Some of the key demographic information about participating children, such as their gender and socioeconomic background, was not reported in detail in the majority of studies. Participant demographic information allows for moderator analyses (Gardner, Burton, & Klimes, 2006; Gardner, Hutchings, Bywater, & Whitaker, 2010) to help better understand what works for whom and under which conditions (Hargreaves, 1996). For instance, Cabezas et al. (2011) reported that overall programme effects were not significant, but subgroup analyses indicated a significant positive impact on reading in low socio-economic status public schools in Bio Bio Region. In addition, the ultimate purpose of interventions are "important gains [. . .] generalized and maintained over time" (, p. 85). Studies with long-term follow-up are needed (Flay et al., 2005), particularly in mathematics as only two mathematics tutoring programmes were identified by the review.

The implementing organizations also merit more description in future research, given recent evidence suggesting that it can also be very important to student outcomes in educational programmes. It is reported that short-term teacher contracts increased student attainment in Kenya when implemented by non-governmental organization World Vision Kenya, but showed no effects in provinces randomly allocated to condition where there was implementation by government officials (Bold, Kimenyi, Mwabu, Ng'ang'a, & Sandefur, 2013). Findings in this study were reported to be due to differences in fidelity of implementation, although fidelity was not formally assessed. It was concluded that the influence of an implementing organization is so significant, that even findings from effectiveness studies may not be directly relevant to programme implementation in real-world settings, if the implementation agent is different from the one researched. For instance, organizations undertaking RCTs might have "stronger drive for performance or generally stronger capability" (Pritchett & Sandefur, p. 31).

4.1.2. Emphasis on theory of change

Previous reviews discussed that tutoring programmes need stronger theoretical grounding (Devin-Sheehan et al., 1976; Rohrbeck et al., 2003). As , p. 140 reported "it remains relatively unknown how or why volunteer mentoring programmes are effective". For instance, only Lee (1980) and Ritter (2000) studies discussed matching tutors and tutees, although matching has been described as an important programme element by many authors (Reisner, Petry, & Armitage, 1989; Topping & Whiteley, 1993; University of Barcelona, 2007; Naidoo, 2009).

Every intervention is based on a theoretical model (Weiss, 1997; Bickman, 2000). To be tested effectively, theories need to be expressed, for example in a logic model (Zief, Lauver, & Maynard, 2006; Cooksy, Gill, & Kelly, 2001) or Causal Chain Analysis (Loyalka et al., 2013). This allows process implementation to focus on the underpinning logic of the theoretical base. In particular, tutoring is theorized to also rely on socio-emotional processes, but "tutoring programmes have placed greatest emphasis on cognitive processing" (p. 231). Similarly to previous reviews (e.g., Cohen et al., 1982; Ritter, Barnett, Denny, & Albin, 2009), this systematic review identified few studies measuring socio-emotional outcomes. Developing and testing logic models for peer tutoring programmes could also help to distinguish between elements that are essential and variable in the intervention (Craig et al., 2008). Perhaps the best way to compare components of an intervention is within a randomized controlled factorial design (Deeks et al., 2011). If sufficient sample sizes are recruited, it would allow comparison of several types of tutoring and explore variables individual contribution to outcomes e.g. the effectiveness of different types of tutors. Otherwise, there is danger of being unable to detect how variables such as tutor competence or training may predict outcomes.

4.1.3. Process evaluation

Even potentially effective programmes may fail to improve outcomes due to how treatment was delivered (Dobson and Cook, 1980; Hawe, Shiel, & Riley, 2004; Mihalic, 2004). Process evaluations add crucial insights to study results (Linnan & Steckler, 2002; Moore et al., 2015). Loenen (1989, p. 310) reported observations of 30 tutoring sessions and characterized them as "different from VRH [Volunteer Reading Help charity, currently Beanstalk] presented in the initial VRH training course". What happened in practice was not what the designers had planned. Topping, Miller, Murray, and Conlin (2011) undertook process observations in the Fife Peer Learning Trial and data suggested that "tutoring technique was only partly implemented". Lack of assessment of implementation fidelity may produce descriptive ambiguity (Rychetnik, Frommer, Hawe, & Shiell, 2002), and result in researchers "evaluating a programme that has not been adequately implemented" (Basch, Sliepcevich, Gold, Duncan, & Kolbe, 1985, p. 316). Process observations can further illuminate the theory of change through testing correlation between implementation variables and attainment (Topping, Thurston, McGavock, & Conlin, 2012).

As part of the process evaluation, intervention cost should be recorded and reported as it informs subsequent recommendations about using an intervention, along with the quality of evidence (Guyatt et al., 2008a; Krishnaratne, White, & Carpenter, 2013). Resource scarcity is a notorious issue in education, and it is important to record all resources, including personnel and materials required (McEwan, 2012). Although many programmes mention that they are less costly then employing professional tutors, only Ham (1977) has given the actual programme costs, although Cabezas et al. (2011) provided a cost-benefit analysis.

4.2. Implications for implementation of tutoring programmes

Based on the limited sample of included studies, it appears that using highly structured interactions between tutor and tutee is important. In the West Philadelphia Tutoring Programme, Ritter and Maynard (2008) highlighted the lack of tutor training and tutoring session structure to explain the absence of positive effects. Ritter and Maynard also concluded that highly structured tutoring programmes are more likely to lead to improved reading. Similar phenomenon was observed in the Fife Peer Learning study, which reported *Effect Sizes* of 0.2–0.25 for highly structured peer tutoring in mathematics (Tymms et al., 2011). The impact of structure shows the important role that an educator has in designing tutoring programmes to ensure that interactions maximize the behaviours seen as providing effective learning.

This review included only 16 study cohorts, so any findings must be treated with some degree of caution. Nevertheless, as the lack of statistically significant student improvements on some measures indicated, cross-age tutoring may not always increase academic outcomes as intended. While this review focused on benefits to tutees, some evidence suggested that children benefit to a greater extent when acting in the role of peer tutor rather than tutee (Robinson et al., 2005). Therefore, this review does not assess the overall benefit of tutoring programmes. This is one of the limitations of the review. Although a transparent and rigorous search strategy was employed, study selection and quality appraisal was intentionally set to a level whereby findings may have been generalised to different educational contexts. However, the small number yet wide diversity of eligible studies limits the strengths of conclusions. The authors are currently undertaking a large-scale (128 class) cluster randomized trial of cross-age peer tutoring where the differential benefits to tutors and tutees of tutoring programmes will be assessed.

In conclusion there are lessons and messages for both practitioners and researchers from the review. Practitioners need to be aware that studies are not consistent in the definitions of "tutoring", "mentoring" and "volunteering", so it is important to obtain the specific programme descriptions to be clear about the structure and form/function of interactions. In addition practitioners still need to undertake some form of assessment within their specific educational context to ensure that the tutoring that is implemented transfers to their setting. Research on peer tutoring suggested that it has potential to produce consistent positive effects if used in reading with a structured approach, but that studies are not robust enough to ensure that findings transfer and generalise to all contexts. There are also lessons for researchers. Researchers may not have a shared definition of what constitutes a peer tutor, a student tutor, a non-professional tutor, or a community volunteer. However, if manuscripts define how the authors have interpreted these terms then it is possible to synthesise common research in cognate groups, even if original manuscripts have used differing terms and descriptors initially. There are also methodological issues in design and reporting. Medical RCTs generally follow CONSORT guidelines to ensure consistency of approach and that all appropriate variables are reported (Campbell et al., 2004). There may be a need to develop trial and reporting criteria specifically for education RCTs or utilize guidance on reporting social and psychological interventions (Montgomery et al., 2013), otherwise future reviews will be similarly limited in their ability to provide a definitive evidence base to educational professionals.

Acknowledgments

We would like to thank the following researchers who supplied the references and advice crucial to completing this review: Sinclair Goodlad, Gary Ritter, Keith Topping, Lucie Cluver, Paul Montgomery, Sean Grant, G.J. Melendez-Torres.

References

Agodini, R., & Dynarski, M. (2004). Are experiments the only option?. A look at dropout prevention programmes. *The Review of Economics and Statistics, 86*(1), 180–194.

Allor, J., & McCathren, R. (2004). The efficacy of an early literacy tutoring programme implemented by college students. *Learning Disabilities Research and Practice, 19*(2), 116–129.

Backer, T. (2001). *Finding the balance—Programme fidelity and adaptation in substance abuse prevention: A state-of-the-art review*. Rockville, MD: Center for Substance Abuse Prevention.

Baker, S., Gersten, R., & Keating, T. (2000). When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring programme requiring minimal training. *Reading Research Quarterly, 35*(4), 494–519.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programmes: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy American Economic Association, 2*(1), 1–30.

Banerjee, A., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. New York, NY: Public Affairs/Perseus Books Group.

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, D. F., & Kolbe, L. J. (1985). Avoiding type III errors in health education programme evaluations: A case study. *Health Education Quarterly, 12*(4), 315–331.

Beanstalk (2013). *Tips for reading helpers.*. Retrieved from ⟨www.beanstalkcharity.org.uk/reading-helpers/tips⟩.

Becker, B. J., Hedges, L. V., & Pigott, T. D. (2004). *Campbell Col*la*bor*ation statistical analysis policy brief*. Oslo, Norway: Campbell Col*lab*ora*tion Retrieved from ⟨www.campbellcol*la*bor*ation.org/ECG/policy_statasp⟩.
Berlin, J. A. (1995). Invited commentary: Benefits of heterogeneity in meta analysis of data from epidemiologic studies. *American Journal of Epidemiology, 142*, 383–387.
Best Evidence Encyclopedia (2013). *Review methods: Criteria for inclusion in the Best Evidence Encyclopedia*. Retrieved from ⟨www.bestevidence.org/methods/criteria.htm⟩.
Bickman, L. (2000). Summing up programme theory. *New Directions for Evaluation, 87*, 103–112.
Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4–16.
Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). Scaling-up what works: Experimental evidence on external validity in Kenyan education. In *Center for Global Development. No. WPS/2013-04.* Retrieved from ⟨www.cgdev.org/publication/scaling-what-works⟩.
Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.
Boruch, R., May, H., Turner, H., Lavenberg, J., Petrosino, A., De Moya, D., et al. (2004). Estimating the effects of interventions that are deployed in many places place-randomized trials. *American Behavioral Scientist, 47*(5), 608–633.
Bowling, A. (2009). *Research methods in health: Investigating health and health services0*. Milton Keynes, UK: Open University Press.
Cabezas, V., Cuesta, J. I., & Gallego, F. A. (2011). *Effects of short-term tutoring on cognitive and non-cognitive skills: Evidence from a randomized evaluation in Chile*. Poverty Action Lab. Retrieved from ⟨www.povertyactionlab.org/evaluation/impact-short-term-tutoring-cognitive-and-non-cognitive-skills-chile⟩.
Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: Extension to cluster randomised trials. *British Medical Journal, 328*(7441), 702–708. Retrieved from ⟨www.bmj.com/content/328/7441/702⟩.
Center for Data-Driven Reform in Education (2013). *About the Best Evidence Encyclopedia*. Retrieved from ⟨www.bestevidence.org/aboutbee.htm⟩.
Chall, J. S. (1989). Learning to read: The great debate 20 years later. *Phi Delta Kappan, 70*, 521–538.
Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471–533.
Christley, R. M. (2010). Power and error: increased risk of false positive results in underpowered studies. *Open Epidemiology Journal, 3*, 16–19.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Laurence Erlbaum Publishers.
Cohen, P. A., Kulik, J. A., & Kulik, C-L.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*, 237–248.
Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The programme logic model as an integrative framework for a multimethod evaluation. *Evaluation and Programme Planning, 24*(2), 119–128.
Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *British Medical Journal*, *337*(1655).. Retrieved from ⟨www.bmj.com/content/337/bmj.a1655.full?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=1&author1=macintyre&andorexacttitle=and&andorexacttitleabs=and&andorexactfulltext=and&searchid=1&FIRSTINDE=⟩.
Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. O., Hornik, R. C., & Phillips, D. C. (1980). *Toward reform of programme evaluation: Aims, methods, and institutional arrangements*. San Francisco, CA: Jossey-Bass.
Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2011). Chapter 9: Analysing data and undertaking meta-analyses. In J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011)*. The Cochrane Collaboration Available from www.cochrane-handbook.org.
Delquadri, J., Greenwood, C. R., Whorton, D., Carta, J. J., & Hall, R. V. (1986). Classwide peer tutoring. *Exceptional Children, 52*, 535–542.
Devin-Sheehan, L., Feldman, R. S., & Allen, V. L. (1976). Research on children tutoring children: A critical review. *Review of Educational Research, 46*, 355–385.
Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning, 3*(4), 269–276.
Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.
Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programmes in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*(4), 605–619.
Elliott, J., Arthurs, J., & Williams, R. (2000). Volunteer support in the primary classroom: The long-term impact of one initiative upon children's reading performance. *British Educational Research Journal, 26*(2), 227–244.
Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random effects methods. *Psychological Methods, 6*, 161–180.
Fitzgerald, J. (2001). Can minimally trained college student volunteers help young at-risk children to read better? *Reading Research Quarterly, 36*, 28–46.
Fitz-Gibbon, C. (1977). An analysis of the literature of cross-age tutoring. In *ERIC document reproduction service no. ED 148 807*. Washington, DC: National Institute of Education.
Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programmes. *Preventive Medicine, 15*(5), 451–474.
Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151–175.
Fuchs, D., Fuchs, L. S., Thompson, A., Svenson, E., Yen, L., Al Otaiba, S., et al. (2001). Peer-assisted learning strategies in reading extensions for kindergarten, first grade, and high school. *Remedial and Special Education, 22*(1), 15–21.
Gardner, F., Burton, J., & Klimes, I. (2006). Randomised controlled trial of a parenting intervention in the voluntary sector for reducing child conduct problems: Outcomes and mechanisms of change. *Journal of Child Psychology & Psychiatry, 47*, 1123–1132.
Gardner, F., Hutchings, J., Bywater, T., & Whitaker, C. (2010). Who benefits and how does it work? Moderators and mediators of outcomes in a randomised trial of parenting interventions in multiple 'Sure Start' services. *Journal of Clinical Child & Adolescent Psychology, 39*, 568–580.
Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high-quality research in special education group experimental design. *The Journal of Special Education, 34*(1), 2–18.
Ginsburg-Block, M. D., Rohrbeck, C. A., & Fantuzzo, J. W. (2006). A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. *Journal of Educational Psychology, 98*(4), 732–749.
Glazerman, S., Levy, D. M., & Myers, D. (2003). Noexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science, 589*(1), 63–93.
Goodlad, S., & Hirst, B. (1990). *Explorations in peer tutoring*. Oxford: Blackwell Education.
Goodlad, S., & Hirst, B. (1989). *Peer tutoring: A guide to learning by teaching.*. London, UK: Kogan Page.
Goodlad, S. (1995). *Students as tutors and mentors*. London, UK: Kogan Page.
Goodlad, S. (1998). *Mentoring and tutoring by students*. London, UK: Kogan Page.
Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., et al. (2000). Users guide to the medical literature XXV. Evidence-based medicine: Principles for applying the users guides to patient care. *Journal of the American Medical Association, 284*, 1290–1296.
Guyatt, G. H., Oxman, A. D., Kunz, R., Falck-Ytter, Y., Vist, G. E., Liberati, A., et al. (2008). Rating quality of evidence and strength of recommendations: Going from evidence to recommendations. *British Medical Journal, 336*(7652), 1049–1051.
Ham, W. (1977). *Effects of a volunteer tutor programme on self-esteem and basic skills achievement: In the primary grades of a southern rural school system*. University of Florida (Unpublished dissertation).
Hargreaves, D. (1996). Teaching as a research-based profession: Possibilities and prospects. In *The teacher training agency annual lecture 1996* Retrieved from ⟨eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/TTA%20Hargreaves%20lecture.pdf⟩.
Harter, S. (1985). *The Self-Perception Profile for Children: Revision of the Perceived Competence Scale for Children*. Denver, CO: University of Denver.
Hartley, S. S. (1977). *Meta-analysis of the effects of individually paced instruction in mathematics* (38(7-A)). University of Colorado. Dissertation Abstracts International, University Microfilms 4003 (77-29, Doctoral dissertation).
Hawe, P., Shiel, A., & Riley, T. (2004). Complex interventions: How "out of control" can a randomized controlled trial be? *British Medical Journal, 328*, 1561–1563.
Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60–87.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
Henk, W. A., & Melnick, S. A. (1995). The reader self-perception scale (RSPS): a new tool for measuring how children feel about themselves as readers. *The Reading Teacher, 48*(6), 470–482.
Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557–560.
Higgins, J. P., & Altman, D. G. (2008). Assessing risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: John Wiley & Sons (Chapter 8).
Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions (Version 5.1.0)*. Retrieved from ⟨handbook.cochrane.org/⟩.
Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L. E., & Coe, R. (2013). *The Sutton Trust-Education Endowment Foundation teaching and learning toolkit*. London, UK: Education Endowment Foundation Retrieved from ⟨educationendowmentfoundation.org.uk/toolkit⟩.
Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews, 1*(MR000006) http://dx.doi.org/10.1002/14651858.MR000006.pub3
Ioannidis, J. P., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal, 335*(7626), 914–916.
Iyenger, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem (with discussion). *Statistical Science, 3*, 109–135.
Jensen, R. J. (1991). *The effects of cross-age tutoring on the reading achievement of underachieving second and fifth-grade students*. Brigham Young University Retrieved from ⟨search.proquest.com/docview/303975464?.accountid=⟩ (Doctoral dissertation, ProQuest Dissertations and Theses, 208–208).
Koh, S., Sanders, K., & Meyer, J. (2012). Roles of active learning and tutor input in students' perception of learning. *Teaching and Learning Forum*, 1–9 Retrieved from ⟨www.roger-atkinson.id.au/tlf2012/refereed/koh.pdf⟩.
Krishnaratne, S., White, H., & Carpenter, E. (2013). Quality education for all children?. In *What works in education in developing countries. 3ie* Retrieved from ⟨www.3ieimpact.org/en/evaluation/working-papers/working-paper-20/⟩.
Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology, 40*, 998–1009.
Lee, C. C. (1980). The homework helper programme: Volunteer service for academic and social enrichment in the elementary school. *The School Counselor, 28*, 11–21.
Lee, Y. S., Morrow-Howell, N., Jonson-Reid, M., & McCrary, S. (2012). The effect of the experience corps[R] programme on student reading outcomes. *Education and Urban Society, 44*(1), 97–118.
LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine, 337*(8), 536–542.
Leung, K. C., Marsh, H. W., & Craven, R. G. (2005). Are peer tutoring programmes effective in promoting academic achievement and self-concept in educational settings: A meta-analytical review. *International Conference of the Australian Association for Research in Education*.
Li, T., Han, L., Rozelle, S., & Zhang, L. (2010). *Cash incentives, peer tutoring, and parental involvement: A study of three educational inputs in a randomized field experiment in China*. Beijing, China: Peking University Retrieved from ⟨mitsloan.mit.edu/neudc/papers/paper_223.pdf⟩.
Linnan, L., & Steckler, A. (2002). Process evaluation for public health interventions and research: An overview. In A. Steckler & L. Linnan (Eds.), *Process evaluation for public health interventions*. San Francisco, CA: Jossey-Bass.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage: Thousand Oaks, CA.
Littell, J. H., Corcoran, J. C., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford, UK: Oxford University Press.
Loeber, R., Stouthamer-Loeber, M., Van Kammen, W. B., & Farrington, D. P. (1991). Initiation escalation and desistance in juvenile offending and their correlates. *Journal of Criminal Law and Criminology, 82*, 36–82.
Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J., et al. (2013). Can information and counseling help students from poor rural areas go to high school? *Evidence from China, Journal of Comparative Economics, 36*, 26–40.
Loenen, A. (1989). The effectiveness of volunteer reading help and the nature of the reading help provided in practice. *British Educational Research Journal, 15*, 297–316.
Margolis, H. (2005). Increasing struggling learners' self-efficacy: What tutors can do and say. *Mentoring and Tutoring: Partnership in Learning, 13*(2), 221–238.
McCartney, E., & Ellis, S. (2008). Open dialogue peer review: a response to Tymms Merrell & Coe. *The Psychology of Education Review, 32*(2), 11–12.
McDaniel, E. D., & Leddick, G. R. (1978). *Elementary Children's Self-Concepts Factor Structures and Teacher Ratings*. Toronto: Paper presented at the Annual Meeting of the American Psychological Association.
McEwan, P. J. (2012). Cost-effectiveness analysis of education and health interventions in developing countries. *Journal of Development Effectiveness, 4*(2), 189–213.
McEwan, P. J. (2013). *Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments*. Retrieved from ⟨academics.wellesley.edu/Economics/mcewan/PDF/meta.pdf⟩.
McKinney, A. D. (1995). *The effects of an after school tutorial and enrichment programme on the academic achievement and self-concept of below grade level first- and second-grade students*. University of Mississippi (Unpublished doctoral thesis).
McKenna, M. C., & Kear, D. J. (1990). Measuring attitude toward reading: A new tool for teachers. *The Reading Teacher, 43*, 626–639.
Medway, F. J. (1995). Tutoring. In L. W. Anderson (Ed.), *International encyclopedia of teaching and teacher education*. Cambridge, UK: Pergamon.
Meier, J. D., & Invernizzi, M. (2001). Book buddies in the Bronx: Testing a model for America Reads. *Journal of Education for Students Placed at Risk, 6*(4), 319–333.
Merrill, D. C., Reiser, B. J., Merrill, S. K., & Landes, S. (1995). Tutoring: Guided learning by doing. *Cognition and Instruction, 13*(3), 315–372.
Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth, 4*(83–86), 99–105.
Miller, S., Connolly, P., Odena, O., & Styles, B. (2009). *A randomised controlled trial evaluation of business in the community's time to read pupil mentoring programme*. Centre for Effective Education, Queen's University Belfast Retrieved from ⟨www.qub.ac.uk/cee⟩.
Miller, S., & Connolly, P. (2012). A randomised controlled trial evaluation of time to read, a volunteer tutoring programme for 8- to 9-year-olds. *Educational Evaluation and Policy Analysis, 35*(1), 23–37.
Miller, S., Connolly, P., & Maguire, L. K. (2012). The effects of a volunteer mentoring programme on reading outcomes among eight- to nine-year-old children: A follow up randomized controlled trial. *Journal of Early Childhood Research, 10*, 134–144.
Miller, S., Maguire, L. K., & Macdonald, G. (2012). Home-based child development interventions for preschool children from socially disadvantaged families. *Campbell Systematic Reviews, 1* http://dx.doi.org/10.4073/csr.2012.1
Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. http://dx.doi.org/10.1371/journal.pmed.1000097. Retrieved from ⟨www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000097⟩
Montgomery, P., Grant, S., Hopewell, S., Macdonald, G., Moher, D., Michie, S., et al. (2013). Protocol for CONSORT-SPI: An extension for social and psychological interventions. *Implementation Science, 8*(1), 99.
Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., & Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *British Medical Journal 350* Retrieved from www.bmj.com/content/350/bmj.h1258.
Mullen, E. J. (2006). Choosing outcome measures in systematic reviews: Critical challenges. *Research on Social Work Practice, 16*(1), 84–90.
Newman, M. (2003). *A pilot systematic review and meta-analysis on the effectiveness of problem based learning. Campbell Collaboration Systematic Review Group on the effectiveness of problem based learning. Learning and Teaching Support Network*. Newcastle, UK: University of Newcastle.
O'Connor, D., Green, S., & Higgins, J. P. (2008). Defining the review question and developing criteria for including studies. In F. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: John Wiley & Sons (Chapter 5).
Palincsar, A. S., & Brown, A. L. (1989). Classroom dialogues to promote self-regulated comprehension. In J. Brophy (Ed.), *Teaching for meaningful understanding and self-regulated learning*. Greenwich, CT: JAI Press.
Pell, T., & Jarvis, T. (2001). Developing attitudes to science scales for use with children of ages from five to eleven years. *International Journal of Science Education, 23*(8), 847–862.

Pesci, A. (2015). *Cooperative learning and peer tutoring to promote students' mathematics education*. Retrieved from ⟨math.unipa.it/∼grim/21_project/Pesci486-490.pdf⟩.
Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiology and Community Health, 57*(7), 527–529.
Piers, E. V. (1969). *Manual for the Piers-Harris Children's Self Concept Scale*. Nashville, Tenn: Counselor Recordings and Tests.
Policy Studies Associates (2007). *Evidence of long-term learning outcomes among reading together tutees*. Washington, DC: Policy Studies Associates.
Potter, J. (1994). 'No Limit' a blueprint for involving volunteer tutors in primary schools. *Mentoring & Tutoring: Partnership in Learning, 2*(2), 61–62.
Poverty Action Lab (2009). *Read India: Helping primary school students in India acquire basic reading and math skills*. Poverty Action Lab Retrieved from ⟨www.povertyactionlab.org/evaluation/read-india-helping-primary-school-students-india-acquire-basic-reading-and-math-skills⟩.
Pridmore, P., Stephens, D., & Stephens, J. (2000). *Children as partners for health: A critical review of the child-to-child approach*. London, UK: Zed Books.
Pullen, P. C., Lane, H. B., & Monaghan, M. C. (2004). Effects of a volunteer tutoring model on the early literacy development of struggling first-grade students. *Reading Research and Instruction, 43*(4), 21–40.
Reisner, E., Petry, C., & Armitage, M. (1989). *A review of programmes involving college students as tutors or mentors in grades k-12*. Washington, DC: Policy Studies Associates Inc Department of Education.
Rimm-Kaufman, S. E., Kagan, J., & Byers, H. (1998). The effectiveness of adult volunteer tutoring on reading among "at risk" first-grade children. *Reading Research and Instruction, 38*(2), 143–152.
Ritter, G. W. (2000). *The academic impact of volunteer tutoring in urban public elementary schools: Results of an experimental design evaluation* (61). University of Pennsylvania 3A (Doctoral dissertation, Retrieved from: Dissertation Abstracts International).
Ritter, G. W., & Maynard, R. A. (2008). Using the right design to get the "wrong" answer?. Results of a random assignment evaluation of a volunteer tutoring programme,. *Journal of Children's Services, 3*(2), 4–16.
Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programmes for elementary and middle school students: A meta-analysis. *Review of Educational Research, 79*(1), 3–38.
Ritter, G. W., Denny, G., Albin, G., Barnett, J., & Blankenship, B. (2006). *The effectiveness of volunteer tutoring programmes: A systematic review* (7). Campbell Collaboration http://dx.doi.org/10.4073/csr.2006.7
Robinson, D. R., Schofield, J. W., & Steers-Wentzell, K. L. (2005). Peer and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review, 17*(4), 327–362.
Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 95*, 240–257.
Roscoe, R. D., & Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534–574.
Rosenshine, B., & Furst, N. (1969). The effects of tutoring upon pupil achievement: A research review. In *ERIC document reproduction service, ED 064462*. Washington, DC: Office of Education.
Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcements. Psychological Monographs, 80, Whole No. 609.
Rutter, M. (1967). A children's behaviour questionnaire for completion by teachers: preliminary findings. *Journal of child Psychology and Psychiatry, 8*(1), 1–11.
Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. (2002). Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology and Community Health, 56*(2), 119–127.
Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of consulting and clinical psychology, 64*(6), 1290.
Shanahan, T. (1998). On the effectiveness and limitations of tutoring in reading. *Review of Research in Education, 23*, 217–234.
Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*(4), 500–506.
Slavin, R. E., & Lake, C. (2008). Effective programmes in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427–515.
Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational programme evaluations. *Educational Researcher, 37*(1), 5–14.
Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). *Effective beginning reading programmes. Center for Data-Driven Reform in Education*. Baltimore, MD: Johns Hopkins University.
Slavin, R. E., Lake, C., Cheung, A., & Davis, S. (2009). *Beyond the basics: Effective reading programmes for the upper elementary grades. Center for Data-Driven Reform in Education*. Baltimore, MD: Johns Hopkins University.
Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programmes for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391–1466.
Slavin, R. E., Lake, C., Davis, S., & Madden, N. (2010). Effective programmes for struggling readers: A best evidence synthesis. *Educational Research Review* Retrieved from ⟨dx.doi.org/10.1016/j.edurev.2010.07.002⟩.
Slavin, R. E., & Madden, N. A. (2011). Measures inherent to treatments in programme effectiveness reviews. *Journal of Research on Educational Effectiveness, 4*(4), 370–380.
Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programmes for struggling readers: A best-evidence synthesis. *Educational Research Review, 6*(1), 1–26.
Spörer, N., Brunstein, J. C., & Kieschke, U. (2009). Improving students' reading comprehension skills: Effects of comprehension instruction and reciprocal teaching. *Learning and Instruction, 19*, 272–286.
Spivack, G., & Swift, M. S. (1967). *Devereux elementary school behavior rating scale*. Devereux Foundation.
Sterne, J. A., Egger, M., & Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal, 323*(7304), 101–105.
Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000a). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal, 320*, 1574–1577.
Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000b). High false positive rate for trim and method. *British Medical Journal, 320*, 1574–1577 Retrieved from ⟨www.bmj.com/rapid-response/2011/10/28/high-false-positive-rate-trim-and-fill-method⟩.
Thorlund, K., Imberger, G., Johnston, B. C., Walsh, M., Awad, T., Thabane, L., et al. (2012). Evolution of heterogeneity (I^2) estimates and their 95% confidence intervals in large meta-analyses. *PLoS ONE, 7*(7), e39471.
Topping, K. J. (1998). Commentary: Effective tutoring in America Reads: A reply to Wasik. *The Reading Teacher, 52*(1), 42–50.
Topping, K. J. (2004). Tutoring in mathematics: A generic method. *Mentoring and Tutoring: Partnership in Learning, 12*(3), 353–370.
Topping, K., & Whiteley, M. (1993). Sex differences in the effectiveness of peer tutoring. *School Psychology International, 14*(1), 57–67.
Topping, K. J., & Hill, S. (1995). University and college students as tutors for schoolchildren: A typology and review of evaluation research. In S. Goodlad (Ed.), *Students as tutors and mentors* (pp. 13–31). London, UK: Kogan Page.
Topping, K. J., Miller, D., Murray, P., & Conlin, N. (2011). Implementation integrity in peer tutoring of mathematics. *Educational Psychology: An International Journal of Experimental Educational Psychology, 31*(5), 575–593.
Topping, K. J., Thurston, A., McGavock, K., & Conlin, N. (2012). Outcomes and process in reading tutoring. *Educational Research, 54*(3), 239–258.
Torgerson, C. J., Torgerson, D. J., Birks, Y. F., & Porthouse, J. (2005). A comparison of randomised controlled trials in health and education. *British Educational Research Journal, 31*(6), 761–785.
Torgerson, C. J. (2006). The quality of systematic reviews of effectiveness in literacy learning in English: A 'tertiary' review. *Journal of Research in Reading, 30*(3), 287–315.
Torgerson, C. J., & King, S. (2002). Do volunteers in schools help children learn to read?. A systematic review of randomised controlled trials. *Educational Studies, 28*(4), 433–444. Retrieved from ⟨www.tandfonline.com/doi/abs/10.1080/0305569022000042435⟩.

Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping, K., & Miller, D. (2011). Improving attainment across a whole district: School reform through peer tutoring in a randomized controlled trial. *School Effectiveness and School Improvement, 22*(3), 265–289.
Vadasy, P. F., Jenkins, J. R., Antil, L. R., Phillips, N. B., & Pool, K. (1997). The Research-to-practice ball game, Classwide peer tutoring and teacher interest, implementation, and modifications. *Remedial and Special Education, 18*(3), 143–156.
Vygotsky, L. S. (1978). *Mind in Society*. Cambridge, MA: Harvard University Press.
Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programmes. *Reading Research Quarterly, 28*, 179–200.
Wasik, B. A. (1998). Volunteer tutoring programmes in reading: A review. *Reading Research Quarterly, 33*(3), 266–291.
Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Education Research, 13*, 21–39.
Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation, 76*, 41–55.
What Works Clearing House (2010). *Procedures and standards handbook (Version 2.1)*. Washington, DC: What Works Clearing House.
Wilde, E. T., & Hollister, R. (2007). How close is close enough?. Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis & Management, 26*(3), 455–477.
Zief, S. G., Lauver, S., & Maynard, R. A. (2006). *Impacts of after-school programmes on student outcomes. Campbell library of systematic reviews*. Oslo, Norway: Campbell Collaboration.