Teacher Professional Development and Student Literacy Growth:

A Systematic Review and Meta-analysis

Abstract

This systematic review explores the impact of teacher professional development (PD) on student reading achievement. The first part of the literature evaluates all available existing systematic reviews and meta-analyses of PD intervention studies. No quality reviews of PD and reading specifically (distinct from "attainment") were found. There was a little overlap of studies in existing reviews. The second part of the systematic review focuses on the most recent intervention studies exploring PD and student reading achievement. The results of a meta-analysis of all high-quality studies are presented in the third part of the paper. This analysis showed no strong evidence of publication bias, and an effect size for PD on student literacy of $g = 0.225$. This effect was moderated by number of hours of PD whereby studies with fewer than 30 hours of PD was significant for student reading outcomes ($g = 0.367$, $p < 0.001$) but more than 30 PD hours was not significant ($g = 0.143$, $p > .05$). Following a Weight of Evidence assessment, analysis showed that nearly all high-quality articles involved shorter PD. Weight of Evidence was a significant moderator, ($g = 0.408$, $p < 0.001$ for high quality studies, $g = 0.077$, $p > 0.5$, *n.s.*, for medium quality studies). Our review suggests that only high quality studies of short teacher PD currently provide evidence of impact on student's reading achievement.

Keywords: Systematic review, meta-analysis, teacher professional development, reading

Teacher Professional Development and Student Literacy Growth: A Systematic Review and Meta-analysis

The Rationale for Analysis

Most contemporary evaluations of teaching and education systems place teachers at the center of any attempt to produce positive change in student learning outcomes (Hattie, 2009, 2015). Teachers' learning is a complex and multi-faceted phenomenon and it is widely assumed that professional development (PD) activity influences teacher beliefs and actions and thereby improves student learning (Villegas-Reimers, 2003). This belief is explicated in conceptual models of teacher change, which suggests that opportunities to reflect on actions are essential for professional development. For example, Clarke and Hollingsworth (2002) developed a four-domain model of teaching. They argue that change is sometimes evident in certain teachers 'personal domain' following professional development activity, which may increase the value that teachers may attach to a given strategy. Any new knowledge or approach learned in PD is then explored by teachers in their 'domains of practice' and then evaluated for their efficacy there before being adopted more permanently or rejected as ineffective.

A key question then is whether PD in fact does play a cascading causal role causing change in teachers' actions that in turn causes growth in student learning outcomes. The veracity of this claim is best assessed by studies that intervene to affect outcomes through PD. Answers to this question affect models of PD and policy. Evidence from systematic empirical reviews of well-designed intervention studies potentially provides answer to such questions. Some such reviews suggest that where energetic efforts at change are undertaken, and their impact on students measured, the most productive are those that facilitate teacher change in the classroom rather than for example, focusing on technologies (Slavin, 2008), providing indirect support for a focus

on change in teacher activity as the best source of change in student learning. There are other pressing practical reasons for exploring PD. Estimates in the U.S. suggest that around 5-10% of teacher time can be spent on PD activities (Gulamhussein, 2013). In the province of Canada where this research is carried out, around 6.5% of school days are given over for PD activities. Estimates of the economic cost of PD are imprecise and complicated to calculate. Many estimates have suggested that between 1 - 3% of total state education budgets are allocated to PD in the U.S. (The Consortium of Policy Research in Education, 1995). A figure that even by the early 2000s represented over $6000 per teacher per year (Odden, Archibald, Fermanich, & Gallagher, 2002). More recent reviews by The Center for Public Education suggest that as much as 7.6% of total budgets in some U.S, districts are spent on PD (Gulamhussein, 2013). It is important to know this PD is effective and value for money in attempting to make public school systems as effective as possible. In summary, there are pressing scientific, policy, economic, and pedagogical reasons for undertaking through reviews of the effectiveness of PD on student learning outcomes.

The Need to Disaggregate Research on PD

As it currently stands, many reviews of the effects of teacher PD focus on 'attainment' (Hattie, 2009; Joslin, 1980; Timperley, Wilson, Barrar, & Fung, 2007; Wade, 1985; Yoon, Duncan, Yu lee, Scarloss, & Shapely, 2007). Such meta-analysis merge data from studies of teacher PD on a range of training foci and educational outcomes, typically in literacy, mathematics, and science. This merger of data across domains is problematic for a number of reasons. First, there is the untested assumption that PD is comparably effective across these distinct domains of attainment. This aggregation reflects multiple assumptions about the comparability of methods for teaching teachers across these domains as well as untested

assumptions about the degree to which training teachers leads to changes which can be measured in student outcomes. Merging outcomes for attainment might give a very inaccurate view of the state of the discipline-specific PD literature. For example, if any research synthesis is dominated by studies of PD in science and mathematics, and with headline effect sizes for PD thus reflecting these content areas, this might give a falsely positive effect of what we know about effective PD for literacy. For researchers with expertise in literacy this merger of outcomes also overlooks a whole range of important issues. For example, models of reading development are very well specified at the classroom level as reflecting word-level decoding and text-level comprehension skills (Savage, Burgos, Wood, & Piquette, 2015). Reading is also known to have heritable (that is, genetic) aspects to it that also quite specifically affect these same word-level decoding and text-level comprehension skills (Olson et al., 2011). Both the heritability and its specificity might or might not to be comparable to other domains of student learning. A study by Landerl, Fussenegger, Moll, and Willburger (2009) for example showed that dyscalculia is often a result of a deficit in a 'number module' and is not related to deficits linked to reading difficulties. On the other hand, direct evidence from a series of meta-analysis of well-designed reading interventions have now been shown to work at least reasonably reliably for both the word-level and text-level aspects of literacy (Savage & Cloutier, 2017). It is thus quite likely that PD targeted at those domains of literacy stands a reasonable chance of being successful if it empowers teachers to teach these elements more effectively and if it follows what we know more broadly about effective reading interventions. It is not clear whether the same evidence base exists for effective intervention in other domains of attainment. As reading researchers we are best placed to evaluate the quality of literacy PD and so undertake work in this area. For these reasons, the present paper sought to answer the following primary research question: What is the

effect of teacher professional development on student achievement in reading among elementary school students?

Recent research has shown that PD studies vary in terms of PD type (Amendum & Fitzgerald, 2013), sample size (Cohen, Manion, & Morrison, 2013), and the use of standardized testing (U.S. Department of Education, 2001). In addition, great stress in the PD literature is placed on the argument that the length of PD affects student achievement. Yoon et al., (2007) and Yoon, Duncan, Lee, Scarloss, and Shapley (2008) claim that PD averaging 49 hours increases student attainment by 21 percentile points. Guskey and Yoon (2009) later argued that only an average of 30 or more hours of PD produces positive effects on student achievement. The same argument about length of PD was adopted by Amendum and Fitzgerald (2013). For this reason, we developed sub-questions: What variations in PD moderate overall effects? Does PD of 30 hours or more moderate the effect of PD on student reading achievement? A systematic review methodology was used in conjunction with subsequent quantitative meta-analysis to answer these questions.

Method

Tertiary Systematic Review

It is often argued that systematic methods for synthesizing results in productive major fields of educational research is the most reliable way to make sense of large numbers of studies and/or studies with diverse findings (Hattie, 2009; Wright, Brand, Dunn, & Spindler, 2007). Systematic reviews along with meta-analyses are often placed at the top of hierarchies of scientific evidence due to the clarity and reliability they engender (Torgerson, 2003). Consequently, systematic reviews were used here to critically explore the state of the existing literature. This tertiary systematic review examines existing publically available meta-analysis of

studies that assess the impact of teacher professional development on student literacy achievement. The design and the quality of these meta-analyses are closely evaluated using universal and standard criteria (Torgerson, 2003). To locate any recent systematic reviews and meta-analysis, we searched in *The Campbell Collaboration Library, The What Works Clearing House, The EPPI center*, and other databases using basic key search terms "meta-analysis, "systematic reviews", "reading outcome/achievement/performance" and "Teacher PD/training." Specific criteria were then applied consistently to identify candidate reviews for our tertiary review. The inclusion criteria were based on those of *The EPPI center*, as the center is a well-known research unit in the University of London and a trustworthy source for conducting systematic reviews:

Inclusion Criteria

- Studies that focused on teacher professional development
- Studies that measured the impact of teaching/learning on students' reading measures
- Studies that focused on elementary education
- Studies that included in-service teachers
- Studies that were reported and carried out in the English Language

Criteria were also used to exclude studies, specifically:

- Studies involving pre-service teachers
- Qualitative studies
- Studies that focused on math and science
- Studies that were correlational and /or did not include control groups
- Studies that focused on students' narrative and writing outcomes

Search Procedures

We located two systematic reviews from the *EPPI Center* (Cordingely, Bell, Ishams, Evan, & Firth, 2007) and *What Works Clearing House* (Yoon et al., 2008), one meta-analysis from *What Works Clearing House* (Yoon et al., 2007), and two other meta-analysis from *Psych Info*: (Saylor & Johnson, 2014; Surrette & Johnson, 2015). We excluded the two latter meta-analyses from this review because the results were interpreted in a qualitative manner, student data was not reported, and there were no specific foci on literacy. We excluded Yoon et al. (2008) because the content entirely overlaps with Yoon et al. (2007). We also identified one book, John Hattie's *Visible Learning,* which included more than 800 meta-analyses in the field of education, including PD. A careful review of this book showed that none of these meta-analysis or systematic reviews of PD focused exclusively on student achievement in literacy (Table A1). A first finding therefore, is that we could not locate any systematic review or meta-analysis of PD that focused primarily on student literacy outcomes. Thus, the second step was to pull out individual studies from these systematic reviews and meta-analysis that measured the impact of PD on student literacy. To consider these individual studies as a possibility for a further meta-analysis, they had to meet the basic methodological criteria mentioned previously and use quantitative randomized control trials (RCTs) or quasi-experimental designs (QEDs). The result section of the tertiary systematic review will discuss the meta-analyses starting with Hattie's (2009) text. It will then discuss the systematic reviews found with a brief preface on individual studies that reported effects for literacy.

Meta-analyses. Hattie's (2009) book *Visible Learning* combined over 15 years of evidence-based research on what works in education. Hattie's (2009) synthesis provided over 800 meta-analyses on what affects student achievement. He focused on 6 areas that contribute to student learning: student, home, school, curricula, teacher, and teaching and learning approaches.

Hattie ranked all these influences on attainment using Cohen's (1988) effect sizes. The average effect size for teacher PD on student learning was derived from five meta-analyses based on a total of 537 studies in the domain of teacher professional development. The effect size was reported as .62, and was ranked 19 out of 138 ranked effects across all studies of achievement. We could only access four of these five meta-analyses. We contacted the author to retrieve the fifth meta-analysis but we received no response. Three of the meta-analyses focused on the effect of professional development on teacher change in practice and general student achievement: Wade (1985); Joslin (1980); and Harrison (1980). The fourth meta-analysis (Timperley, Wilson, Barrar, & Fung, 2007) was the only meta-analysis in Hattie's (2009) book that reported studies on student achievement in literacy separately from wider achievement, and so is discussed in more detail below before turning to the other identified meta-analyses.

The Timperley et al. (2007) meta-analysis synthesized 72 studies on the effect of teacher professional development on student achievement in the core subjects of math, science, and reading. The meta-analysis included 10 studies focusing on literacy. The studies were mostly from New Zealand, followed by the U.S., U.K. and Canada. The mean of the effect size of the included literacy studies was .34. Timperley et al. (2007) point out that low achieving students represented a large part of the sample of these 10 studies and that effects of PD were strongest in this sub-group, though details of this aspect of data analysis are sparse.

A major concern with this meta-analysis was the methodology used to include studies. Study inclusion was based on *outcome* (effect size) rather than *methodological quality*. Thus, relatively well-designed studies that had low or no measured effect of PD on attainment were considered 'supplementary studies'. These supplementary studies were used only to support conclusions drawn from the synthesis of studies with larger effect sizes. This approach to

methodology does not conform to the norms of meta-analytic review (Torgerson, 2003). For this reason, we carefully reviewed *all* of the literacy studies that were included in the Timperley et al.'s (2007) meta-analysis that fit our standard inclusion criteria above, disregarding study outcome in the selection process. On this basis, two studies were identified: Baker and Smith (1999), and Timperley and Philips (2003). Baker and Smith (1999) was excluded after careful revision because no randomization took place at the student level, and because there was no explanation as to how the professional development took place.

Yoon et al. (2007) conducted a meta-analysis on teacher professional development and its effect on student achievement. Three studies reported in this meta-analysis focused on teacher professional development and student reading achievement outcome. The overall effect size for studies in literacy and writing that was included in this meta-analysis was $ES = 0.53$. We compared the studies that we pulled from this meta-analysis with studies in Timperley et al. (2007) – none of the studies featured in Timperley et al. The three studies are: Duffy et al. (1986); McCutchen et al. (2002); and McGill-Franzen, Allington, Yokoi, and Brooks, (1999).

Systematic reviews. Cordingley et al. (2007) conducted a systematic review that looked at specialists in professional development and evaluated their impact on teachers and pupils. The researchers' criteria for inclusion were a) expertise of the specialist delivering the PD, b) the nature of the professional development, c) studies that described the intervention along with data analysis, d) studies that showed the effect of teacher change in practice on student learning and e) studies that provided evidence of reliability and validity of their data analysis. Out of 3,421 studies screened, only 22 studies were included. We reviewed all 22 studies included in this systematic review and selected the sub-set of studies that focus on literacy outcomes. There were three studies that focused on reading/language arts that fit our inclusion criteria: Fine and

Kossack (2002); Klingner, Vaughn, Arguelles, Hughes, and Leftwich, (2004); and McCutchen et al. (2002). McCutchen et al. (2002) was excluded in this part because it was discussed earlier in the meta-analysis section.

In summary, our analysis of the existing meta-analyses and systematic reviews on PD and reading achievement did not reveal the existence of a homogenous large set of quality individual studies subject to review or meta-analysis, to answer our basic questions about PD. Indeed, there exists no well-executed meta-analysis on PD and reading per se. Hattie (2009) reported five meta-analyses on PD with only one where it was possible to identify outcomes on student literacy achievement specifically (Timperley et al., 2007). Timperley et al. reported 10 out of 72 studies of PD with reading achievement outcomes published between 1991 and 2006. Their conclusions focused on positive reading academic performance in low achievers specifically. There are however, major concerns about the methodology of including studies based on outcome over methodology. Based on selection by *methodology alone* and using our own criteria, one study from Timperley was selected for analysis.

Of the two other analyses, Yoon et al. (2007) and Cordingley et al. (2007) reported three and two studies respectively that specifically focused on literacy that also met our criteria. Two meta-analyses (Saylor & Johnson, 2015; Surrette & Johnson, 2014) did not explore student achievement in detail so were excluded. In sum, from all the existing meta-analyses and systematic reviews, we could retrieve only six individual studies on teacher PD and student achievement in reading meeting our basic quality indicators up to 2007. Two studies came from Cordingley et al. (2007), one from Timperley et al. (2007) and three from Yoon et al. (2007).

These meta-analytic reviews featured studies published up to 2007. With only six selected studies, a meta-analysis of this more carefully selected data is inappropriate thus far. We

therefore extended our search beyond meta-analyses for individual studies published after 2007 in order to have a comprehensive answer to our research questions.

Research Article Search

The result from the primary tertiary review of meta-analyses identified six relevant studies published between 1986 and 2007. The purpose of this next section is to identify the most recent quality studies of PD and its impact on reading achievement published after 2007. We followed the same inclusion criteria as in the Tertiary Review section. We started with a broad literature search in an attempt to locate individual RCT and QED studies. RCTs and QEDs were targeted because these two approaches are the most reliable methods in assessing the effectiveness of intervention (Torgerson, Brooks, & Hall, 2006). Electronic searches were also conducted through main educational databases (Psych Info, ERIC, Educational Full text). A range of keywords were used (e.g. "teaching methods/strategies, skills "professional development, "teacher training", "teaching skills", "phonemic awareness", "vocabulary", "fluency") combined using the Boolean search functions of "AND", "OR" and "NOT". We read more than 1000 abstracts in our search. In some cases, we also reviewed the study's methodology and results to make sure that the study either fit our inclusion criteria or not (Figure A1).

Our comprehensive search revealed seven further studies that fit our basic criteria for inclusion. We also undertook another search by authors and list of references in each of the identified studies. Four studies of the additional 11 were added using this approach. Four of these 11 studies are RCTs: Garet et al. (2008); Gersten, Dimino, Jayanthi, Kim and Santoro (2010); Powell, Diamond, Burchinal, and Koehler, (2010); and Snow, Eadie, Connell, Dalheim, McCusker, and Munro, (2014). Four studies were Cluster RCTs: Vernon-Feagans, Kainz, Hedrick, Ginsberg, and Amendum, (2013); Vernon-Feagans, Kainz, Amendum, Ginsberg,

Wood, and Bock, (2012); Vernon-Feagans, Gallagher, Ginsberg, Amendum, Kainz, Rose, and Burchinal, (2010); Vernon-Feagans, Bratsch-Hines, Varghese, Bean, and Hedrick, (2015). Two studies were QEDs: Podhajski, Mather, Nathan, and Sammons, (2009) and Porche, Pallante, and Snow, (2012). The last study identified (Amendum, 2014), used a mixed method approach. This latter study was included because the quantitative element was based on a randomized intervention and comparison method. In total we retained 17 studies; six from the original search of meta-analyses, and 11 from the second search of individual studies from post-2007. All these studies were included for quality coding and a meta-analysis.

Results

Coding articles for quality. We first sought to further assess the quality of the 17 studies beyond the basic selection criteria. A table was adopted and modified based on CONSORT (The Consolidated Standards for reporting Trials) and on the EPPI center guidelines for assessing the quality of included studies (Table A4). The guidelines in the table include assessment of whether the studies reported method of allocation, i.e. whether the study had a comparison group based on randomization and described their method of randomization. In addition, sample justification refers to whether the study justified the sample size *n* in their study and evaluated their power estimate. Blinding refers to whether the participants in the study were unaware of the intervention. Intention to treat refers to whether the groups in the study were analyzed statistically based on how they were originally assigned disregarding any subsequent attrition (Altman, 1996; Hollis & Campbell, 1999; Moher, Schulz, & Altman, 2001). Additionally, the guidelines from the EPPI center included whether the study reported a table showing the quantitative impact of teacher on students, described in detail the process of the professional development, whether the study reported any attempt to establish reliability and validity of the

data, and lastly, if their article applied fidelity of treatment to their study (Cordingley et al., 2007).

For further analytic analysis, the Weight of Evidence (WOE) method adopted by Cordingley et al. (2007) in their systematic review was conducted. The WOE is based on three questions: WOE A: Did the reported findings in the study answer the study question and was it internally consistent? WOE B Is the research design appropriate for the review questions? And WOE C: Was the focus of the study relevant to the review question? The answer to these questions were reported by an overall WOE D rating of each study as 'High', 'Medium', or 'Low' after careful consideration of the details of the design of each study. Answering these WOE questions determined the quality of the 17 studies. To assess the studies as High, Medium, or Low, we contacted the EPPI center for guidance since it was adopted from their review. Studies that scored LOW on WOE A were deemed LOW on all WOE criteria. Studies that reported High or Medium WOE A were evaluated on all criteria and given an overall code in WOE D. For example, if a study has two High and Medium then WOE D is "High". If a study has one High, one Medium and one Low then WOE D is "Medium". The WOE coding assessment showed that six studies where coded of high quality and 11 were of medium quality (Table A5). This coding of study quality was undertaken by the primary author using coding frames reflecting the WOE criteria above. The same studies were then coded independently by a graduate research assistant who was trained to use the same coding frames as used by the primary author. Cohen's kappa was then calculated to assess the inter-reliability of the two independent codings of the 17 studies. Cohen's kappa was 0.77 for these ratings, which is broadly acceptable.

The coding assessment (TableA4) showed that only ten out of the 17 articles reported method of allocation. One study out of the 17 reported sample size justification. Six articles reported ITT (Intention to Treat), and one study reported blinding. All 17 studies reported evidence of impact on student literacy growth. All studies described in specific details their professional development. Five studies out of the 17 reported attempt to establish reliability and one study reported attempt to establish validity of data-analysis. Six studies out of the 17 reported fidelity of treatment. The student sample size in these studies ranged from very small $n = 45$ to large $n = 1530$. One study (Garet et al., 2008) included a sample of more than 1000 participants. There were good reasons to believe p values might be inflated in one study because of misalignment between the aggregated nature of intervention (PD delivered to teachers) and analysis (undertaken at the disaggregated student level within classrooms). Podhajski (2009) reported a teacher sample size of $n = 4$ and yet they calculated results based on students individual (disaggregated) gains in literacy. Notably, elsewhere, all the other studies were well executed in the specific sense of reporting their results based on aggregated classroom means. Overall, a summary of the table shows that the studies were variable in their quality with some showing strengths in all cases but many lacked at least one major feature of studies executed with the highest methodological rigour. Due to these limitations, the overall quality of these 17 studies was judged to be moderate. Such patterns are not unique in published meta-analyses (Torgerson et al., 2006).

Professional development programs in 17 studies. The next step in the analysis was to look at candidate moderators throughout the 17 studies to see if any of the moderations affect student reading outcomes. The candidates identified in the existing literature were: Kind of PD programs; sample size; the use of standardized testing; the use of technology in PD; and PD

hours. All 17 studies used different professional development programs except for the TRI (Targeted Reading Intervention) that was conducted by the same primary author in four studies: Vernon-Feagans et al. (2010, 2012, 2013, 2015), so there were 13 PD programs in our data set. Of these, only three studies compared on-site professional development versus web-cam professional development: Vernon-Feagans et al. (2013, 2015) and Powell et al. (2010). The 13 PD programs were then compared to the What Works Clearing House list of "best practices for teachers". The WWC has a list of PD programs that they had evaluated. None of the PD programs in our included studies were on WWC list, which means that we do not have any empirical evidence linking the content of PD programs to improvements in students reading achievement measures. The professional development in 15 studies can be described as 'typical' or 'traditional' PD workshops or summer institutes delivered face-to-face by a trainer on a designated PD day or days and where students are not present. Vernon-Feagans' studies uniquely used a webcam aspect to PD in addition to the 3 day summer institute. Powell et al. (2010) also used video exemplars in addition to the workshop. Two studies (Amendum, 2014; Fine & Kossack, 2002) used embedded professional development where the literacy coach provided immediate feedback while the teacher was teaching. This provided the teachers with an opportunity to be reflective in their practice. Fine and Kossack, (2002) focused on teachers using reflective journaling and peer coaching. These two studies reported the highest ES out of the 17 studies.

PD programs associated with higher effect size. The ENRICH (Early Diagnostic Reading Intervention Through Coaching) program that was adopted by Amendum (2014) reported the highest *ES* between 1.06-1.52. The PD program was based on the situated learning theory, which states that the most effective learning takes place when it is embedded in a target

activity (Amendum, 2014). In this study, the PD was delivered by an ENRICH coach who went into classrooms on a weekly basis, and who either coached and provided immediate feedback, watched the teacher teach before coaching and gave feedback, or both, depending on teachers' preferences. The instructional activities that the teachers were trained to used, were tailored around students' individual needs.

Another approach to coaching, 'Cognitive coaching', was adopted by Fine and Kossack (2002). They also reported large *ES* of between 0.90-1.02. The teachers were enrolled in masters' level courses, were coached and trained by the authors to reflect on their teaching practice in reading by using self-rubrics. The teachers were also involved in cognitive peer coaching with other colleagues and role-played and reflected on: A) themselves as teachers, B) a coach teaching other teachers what they have learnt and applied in the classroom, and C) a student learning new teaching strategies.

TRI PD in four studies. TRI (Targeted Reading Intervention) is a Tier II reading intervention to help struggling readers and was developed by the same primary author (Vernon-Feagans et al., 2010, 2012, 2013, 2015). The TRI professional development aimed to help teachers build the required essential knowledge of reading development, learn reading strategies that address children's individual needs, and apply these strategies either one-on-one or in small groups. In all four studies, the professional development was more than 30 hours in length, divided between 3 days of summer institute, bi-weekly visits, and bi-monthly visits throughout the school year. The trainers were highly trained coaches in TRI. The coaches offered more than 10 hours of training for teachers over the course of the school year. All four studies were cluster randomized control trials. Schools were first matched according to school size, eligibility for free lunch, and lastly, for percentage of minorities in each school. The schools were then randomized

into control and experiment groups. Three studies by Vernon-Feagans et al. (2010, 2012, 2013) assigned five struggling students to experimental and control groups and five non-struggling students to experiment and control groups. The last study, Vernon-Feagans et al. (2015), was a TRI PD comparison between webcam and onsite PD delivery.

PD and technology. Only three studies compared the use of technology with PD to improving students reading achievement measures. Powell et al. (2010) designed a bi-weekly semester (approximately 45 hours) including a 16-hour PD workshop. Teachers were allocated to either experiment (onsite) or control (remote) conditions. Teachers in the remote condition received the PD through video exemplars and had to submit videotapes of their teaching in classrooms for feedback and treatment integrity of the study. The results showed an *ES* of between 0.11- 0.32 for growth in reading achievement. Students in the remote condition, whose teachers had received the PD through video exemplars, showed better results than students whose teachers received onsite PD, *ES* = .32.

Vernon-Feagans et al. (2015) designed the TRI PD program to be delivered with three TRI literacy coaches on-site or with four coaches via web cam. Teachers in both conditions received the same hours of PD (approximately 35 hours). The authors concluded that students in the remote condition, whose teachers received web-cam coaching, did better than students with on-site PD with an *ES* range between -0.759 and -0.67. Vernon-Feagans et al. (2013) also delivered TRI through web-cam technology. Literacy coaches delivered a bi-weekly 50 minutes of instruction through web-cam. The coaches focused on training teachers to use TRI reading strategies, and problem solving issues about students' individual cases. The study reported *ES* between 0.37- 0.44 for students' reading outcomes.

PD and outcome literacy measurement in 17 studies. We looked into the literacy measures that were used in all 17 studies. The results showed that all 17 used standardized testing measures except for one study (Timperley & Philips, 2003) that used Clay's (1985) measures.

PD and sample size. As explained earlier, most of these studies had a small sample size. Three studies (Garet et al., 2008; McCutchen, 2002; & Snow et al., 2014), had a large sample size range between 779-1254. The other 14 studies had a sample size range between 47- 500 students.

PD hours and outcome. PD hours was an important variation among the 17 studies according to PD researchers. The average hours of PD in all 17 studies was between 10 and 70 hours. Six studies had less than 30 hours (10- 28 hours) of PD and 11 studies had PD of more than 30 hours (30-70 hours). Across all candidate moderators, we found that the largest contrast across our studies was in terms of PD hours. The other moderators did not vary often enough across studies to contrast formally as a moderator to examine its impact. Since PD hours is a sub-question in this paper, an in-depth review was taken to see if there was a relation between PD hours and the WOE quality of the article. An inspection of our literature showed that there was a very strong relationship between quality of study and length of PD. As shown in Table A5, five of the six high-quality PD studies delivered less than 30 PD hours. For this reason, we include PD hours and study quality as moderators in the meta-analysis to see if PD hours or study quality produce higher ES.

Meta-analysis

To reanalyze all 17 studies, effect sizes for all outcome measures were first calculated by the primary author to avoid possible biases in the calculations reported by the study authors.

Then the measures were converted into Hedges' *g* effect sizes based on the mean and the standard deviation given in each study for correction purposes. This analysis was undertaken using an effect size calculator in Comprehensive Meta-analysis (www.meta-analysis.com). Subsequent analyses of the studies using Hedges' *g* were conducted using a random effects model rather than a fixed effect model because the latter assumes that the effect size is comparable in all studies. Only reading measures were included in the meta-analysis. Measures of oral ability such as the PPVT (Peabody Picture Vocabulary Test) were not included in the meta-analysis. Preliminary analyses showed that the studies were heterogeneous ($Q = 131.642$, $df = 16$, $p < 0.001$). The smallest positive effect size was 0.067 and the largest effect size was 0.951. Of the 17 studies three had negative effect sizes and 14 had positive effect sizes. Two studies with negative effect sizes were statistically significant. Of the 15 studies with positive effect sizes, five were statistically non-significant. A careful analysis of each of the studies retrieved from the meta-analysis, systematic reviews, and the individual studies are presented in (Table B1).

Publication bias. To assess the validity of possible publication bias in this meta-analysis, a funnel plot was created. The funnel plot is based on the fact that the estimate of error of mean effect sizes will be more stable as the sample size in each study increases. Studies with small sample sizes will thus mostly be scattered either side of the average at the bottom of the graph while studies with large sample size will most likely cluster together creating a funnel shape when inverted. The effect size in this funnel plot was placed on the x-axis and the standard error was placed on the y-axis. The result showed that studies with small sample size are scattered around the bottom of the graph and the studies with larger sample size are closely clustered

together in symmetry. This pattern does not provide strong support for the existence of a publication bias (Figure B1).

The overall effect size in this meta-analysis was $g = 0.225$ ($CI = 0.0064$-0.385) with an associated standard error of 0.080, an effect that was significantly different from zero ($p < .05$). We undertook a further analysis to explore the extent to which length of PD affected student literacy outcomes. As explained earlier in the paper, Guskey and Yoon (2009) identified PD at or over 30 hours-duration as being of sufficient length to impact student literacy. The average PD with studies of less than 30 hours was a one-day PD. The average of studies with more than 30 hours of PD was one year long across all 17 included studies. To calculate the PD hours as a moderator, we grouped studies with 30 hours and above as high, and studies with less than 30 hours as low. We re-ran the meta-analysis across PD hours and compared effects at different levels of PD hours. Results showed that the overall effect of PD on student literacy attainment was moderated by the number of hours of PD. A significant effect was evident but the analysis also showed that fewer than 30 rather than more than 30 hours of PD was associated with increased student literacy ($g = 0.367$ versus $g = 0.091$ respectively). The studies with less than 30 hours of PD reached conventional statistical significance (significantly different from zero) but studies with higher PD hours did not reach conventional statistical significance with ($p = 0.460$). We then re-ran the meta-analysis using the WOE D criterion as a moderator to see if there was an effect of study quality on student reading outcome. The results showed that generally shorter PD studies with high quality reported $g = 0.347$, a significant effect, $p < 0.001$. Analysis of studies with medium quality and generally longer PD hours was not significant $p > 0.5$, g = 0.077. These results are consistent with the view that the effect of shorter PD length in our meta-

analysis is due to the fact that study quality in shorter PD had a higher quality than studies with longer PD.

Discussion

Overall, our analysis shows that the effect of PD on student reading to be g =. 225. The present paper took a systematic, careful and critical look at all of the research on PD and literacy specifically. This paper sought to explore the impact of regular teacher PD on student literacy outcomes in English language elementary schools. The findings in the main analysis revealed two other main findings regarding moderators of professional development programs and their effect of student literacy: Methodological quality of papers and length of PD.

Quality issues

Our analysis showed that in the selected literature only 10 articles reported method of allocation. Six studies reported ITT (Intention to Treat). Only one study reported blinding and one study reported sample size justification. These are all limitations on the quality of evidence. However, the WOE (Weight of Evidence) analysis also showed that seven out of the 17 articles are of high quality whereas the rest are of medium quality. It was noteworthy that all the 17 studies used randomization to create intervention and control trials, which is the most common design to investigate the effectiveness of a program. Yet, because of variation in other aspects of methodological quality, these studies were judged to be 'moderate' in quality overall. Such overall patterns are sometimes reported in published reviews of reading interventions (Torgerson et al., 2006).

In addition, there were no high quality meta-analytic studies on this specific issue. We did identify one study that reported literacy outcomes separate from student achievement more generally (Timperley et al., 2007). A major concern identified when reviewing Timperley et al.'s

(2007) method was their atypical strategy of including studies based on their *results*. Meta-analysis typically uses universal screening, based in significant part on methodology, to identify all studies that have adequate attention to methodology in the planning of the study, the execution of treatment, and the measures taken to evaluate the treatment (Torgerson, 2003). The results, whether positive or negative, should never be a criterion for study inclusion, per se, in meta-analyses.

The results of analyses of these selected studies showed no strong evidence of publication bias. The overall effect on student literacy (g = 0.225), was moderated by number of hours of PD. Our analyses also showed that fewer than 30 rather than more than 30 hours of PD was associated with increased student literacy (g = 0.367 versus g = 0.091 respectively). Many researchers in this field have argued that the duration of the PD is critical in order to have a positive impact on student performance. It is claimed that because teachers need time to reflect on their new understanding, to look for appropriate approaches to apply in their classrooms, and to evaluate students' performance, to be confident that any teacher professional development will have an impact, it needs to be not less than 30 hours in duration (Cordingley et al., 2007; Guskey & Yoon, 2009; Yoon et al., 2007). Our analysis showed that studies with fewer than 30 PD hours produced significantly larger effect sizes compared to studies with more than 30 PD hours. Thus, our data based on the literature, as it currently stands, clearly do not support the claim that teacher PD with more than 30 hours results in positive literacy achievement in students.

Our findings also showed that high quality studies were nearly always those with shorter PD hours. Only one study out of the 17 studies was reported to be both high in quality and PD hours. Table A5 of the WOE / PD association support the view that the length of PD is confounded with other features of study design quality. PD hours and study quality almost

perfectly overlapped and so analysis of PD length as a moderator could suggest that both PD length and PD quality are significant factors of student reading achievement. It remains unclear given the state of the field as shown in our current data, whether shorter studies are simply intrinsically easier to execute well when it comes to PD or whether well-executed longer studies might produce larger effects on attainment. If shorter PD proves to be genuinely more impactful than longer PD, it may be that it does so because shorter PD focuses on simpler, more encapsulated and specific elements, strategies, and resources in and for literacy. It may be that these elements of PD are more impactful on teachers, more readily implemented in classrooms, or address aspects of literacy that are more amenable to change through high quality regular classroom teaching (e.g. phonological awareness or early literacy versus reading comprehension or interventions for older children who have not responded to good interventions). It is possible that shorter PD produces less disruption to ongoing classroom learning processes. It may however also be that longer PD with its greater depth takes longer to impact practices and student outcomes and may be more evident only sometime after training has finished, in a delayed post-test. Certainly, well-executed future basic research and reviews on PD might usefully explore these questions, and are essential before we draw strong conclusions about the length of PD per se. In practical terms, it is however positive that relatively brief well-executed PD interventions can have small but reasonably robust measurable effects on student literacy outcomes.

In interpreting these finding it is perhaps important to first bear in mind that the overall effects of PD on ‘attainment’ more broadly construed are typically much higher than these we report here for reading specifically. An average effect size of .62 in the tertiary meta-analysis of PD on attainment more generally is reported in Hattie, (2009) for example. If one were to rank

overall effect sizes according to those in the appendix of Hattie's (2009) book Visible Learning, our *ES* = 0.225 would rank 89th where Hattie's effect of teacher PD on student achievement in general was *ES* = 0.62, and was ranked 19th. As Hattie (2009) has argued, an *ES* = .40 is also a useful baseline for interpreting intervention utility given that most interventions have *some* positive impact on attainment. We would however note that small but non-zero significant effects of PD are still important to school improvement initiatives generally. In fact, Lipsey et al. (2012) have cautioned that effect size in educational intervention should not be exclusively interpreted in terms of Cohen's *d*, but that other metrics may be relevant; for example, to compare the reported effect size to that of the annual academic outcome expected from students. In the case of reading achievement, the estimated expected outcome, according to Lipsey et al., is *ES* = 0.60 for grades 2-3 and .36 for grades 3-4. The student samples in our 17 selected studies are drawn from grades 2-4. Thus, for our reported overall effect (*ES* = .225), this could be interpreted to show that students make a 25-35% of improvement in reading achievement as a result of PD. According to Lipsey et al., this is substantially important rather than a small effect size.

What might one conclude at a finer grain of analysis from this review? We cautiously argue that the type of PD approach seems to have promise in impacting student's reading outcomes. The two studies that focused on teachers reflecting on their practice rather than traditional PD workshops produced the highest effect sizes in our review. The studies that used coaching were also reported to be high quality on the WOE criteria and used shorter PD hours. There may thus be a connection between the PD approaches used, PD hours and the study quality in order to have the expected change in student reading achievement. This specific idea could be explored in future work although we would caution that the research base on quality

studies on different kinds of PD is insufficient to draw any strong empirically-based conclusions about the most effective models of PD at this point in time.

Limitations

Most of the limitations listed here reflect the state of the literature on PD and literacy. Our review suggests that on a number of other grounds the literature on PD could be improved. As Cohen, Manion, and Morrison, (2013) explain, where randomization is used, the sample size needs to be sufficiently large to provide an unbiased estimate of error. In this paper, almost all selected studies had small samples. This observation also raises the issue of whether these studies are representative of the general population. Another issue that surfaced was the reporting of the results based on classroom level allocation followed by reporting of analyses at the student-level. Lipsey et al. (2012) argue that classroom level *ES* generally produces a larger individual effect because the denominator is bigger and ES based on student level is in fact a better representation of *ES*. On the other hand, the weighting for study *n* in meta-analysis likely offsets some of this inflation. It is important to know that our reported *ES* of 0.225 reflects studies that have aggregated mean scores and measures of variation for classroom clustering.

More generally, a central issue in interpreting our reported effects concerns whether the existing literature, when considered as a whole, is sufficiently strong to conclude that teacher PD only has a very modest effect on student literacy. Certainly, the methodologically high quality PD literature is small in size. Our extensive search of published papers since 2007 initially yielded a total number of 1505 studies. From these, we identified only 11 studies as suitable for inclusion in our review on the basis of our very basic methodological criteria. Yoon et al. (2007, 2008) expressed similar concerns in both their meta-analysis and systematic review. Their search

for high quality articles identified 3400 studies of which only nine were deemed of sufficient quality to be included. Cordingley et al. (2007) identified comparable numbers in their review. The quality of reporting of methods was variable among all studies. The content of PD programs also varied substantially among all studies. For this reason, the conclusion about the effectiveness of specific PD on student reading achievement by form and time is interpreted with caution and cannot reach any conclusion about which program is most effective. All selected studies were conducted in the U.S. with the exception of Timperley and Philips, (2003), which was conducted in Australia. Clearly, more work is needed on this issue of the effects of PD worldwide. More generally, and alongside this work, a comprehensive conceptual review of PD and teacher professional change would be valuable.

Finally, we also undertook a content analysis for all 17 studies in this review to explore whether there was a clear link between the content of PD interventions, ratings of study quality, and reported study results. We were unable to observe any obvious link between the content of the PD, the methodological quality of the study or the outcomes. Thus, we have no reason to conclude that PD content has an impact on the results of the study (Table B2). The need for alignment of quality of, PD content, methodology and exploration of study duration and impact are suggested.

Conclusion and Future Directions

Their current paper explored two research questions. The first question was "What is the effect of teacher professional development on reading measures among elementary school students?" The meta-analysis showed the reported effect size was of 0.225, which was significant at $p < 0.5$. The sub-questions referred to candidate moderators and we were able to explore "Does the length of the PD moderate this effect?" The answer through the meta-analysis

showed that shorter PD produced a larger effect size of 0.367, $p < .001$. The findings also showed that quality of the PD was more of an influence than the PD length in itself. Weight of evidence showed that high quality PD, which was generally shorter in duration, produced a larger effect size of 0.347, $p < 0.001$, while PD studies with generally longer hours in which they were of medium quality reported no significant effects ($g = 0.077$, $p > 0.5$).

For future directions, research needs to take a more rigorous approach with regards to the quality of studies that are to be conducted in terms of design quality, length, and the type and content of PD delivery undertaken. This review has shown that while most studies have used the traditional approach of workshop and summer institutes, PD studies that have produced better results took a non-traditional path, using coaching.

Finally, Lipsey et al. (2012) have estimated the cost benefit of designing and implementing PD costs $81,000 per school for every 50 students. A large amount of money is spent every year on PD. Such costs might usefully be measured against our calculated benefit from an effect size currently only evident for short well-executed PD on student outcomes. This paper thus serves as an insight for policy makers and stakeholders to understand PD research and to ensure that money is spent most effectively such that students reading levels are improving.

References

Altman, D. G. (1996). Better reporting of randomized controlled trials: The CONSORT statement. *BMJ (Clinical Research Edition), 313*, 570-571. doi:10.1136/bmj.313.7057.570

*Amendum, S. J. (2014). Embedded professional development and classroom-based early reading intervention: Early diagnostic reading intervention through coaching. *Reading & Writing Quarterly, 30*, 348-377. doi:10.1080/10573569.2013.819181

Amendum, S. J., & Fitzgerald, J. (2013). Does structure of content delivery or degree of professional development support matter for student reading growth in high-poverty settings? *Journal of Literacy Research, 45*, 465-502. doi:10.1177/1086296x13504157

Baker, S., & Smith, S. (1999). Starting off on the right foot: The influence of four principles of professional development in improving literacy instruction in two kindergarten programs. *Learning Disabilities Research & Practice, 14*, 239-253. doi:10.1207/sldrp1404_5

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*, 947-967. doi:10.1016/s0742-051x(02)00053-7

Clay, M. M. (1985). *The early detection of reading difficulties (3rd ed.)*. Portsmouth, NH: Heinemann Educational Books Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education* (7th ed.). London, NY: Routledge.

*Cordingley, P., Bell, M., Isham, C., Evans, D., & Firth, A. (2007). *What do specialists do in CPD programmes for which there is evidence of positive outcomes for pupils and teachers*. London: Eppi-Centre, Social Science Research Unit, Institute of Education.

*Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J., & Wesselman, R. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly*, *21*, 237-252. doi:10.2307/747707

*Fine, J. C. & Kossack, S. W. (2002). The effect of using rubric-embedded cognitive coaching strategies to initiate learning conversations. *Journal of Reading Education*, *27*, 31-37.

*Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . Sztejnberg, L. (2008). *The impact of two professional development interventions on early reading instruction and achievement.* NCEE 2008-4030. National Center for Education Evaluation and Regional Assistance. Retrieved from: http://files.eric.ed.gov/fulltext/ED502700.pdf

*Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher study group impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal*, *47*, 694-739. doi:10.3102/0002831209361208

Gulamhussein, A. (2013). *Teaching the teachers: Effective professional development in an era of high stakes accountability.* Retrieved from: http://www.centerforpubliceducation.org/teachingtheteachers

Guskey, T. R., & Yoon, K. S. (2009). What works in professional development. *Phi Delta Kappan, 90,* 495-500. doi:10.1177/003172170909000709

Harrison, D. (1980). *Meta-analysis of selected studies of staff development* (Unpublished Ph.D.).

University of Florida, FL.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, NY: Routledge.

Hattie, J. (2015). *What doesn't work in education: The politics of distraction.* Retrieved from: https://www.pearson.com/content/dam/corporate/global/pearson-dot-com/files/hattie/150602_DistractionWEB_V2.pdf

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal, 319*, 670-674. doi:10.1136/bmj.319.7211.670

Joslin, B. (1980). *In-service teacher education: A meta-analysis of the research* (Unpublished doctoral dissertation). University of Minnesota, Minnesota.

*Klingner, J. K., Vaughn, S., Arguelles, M. E., Hughes, M. T., & Leftwich, S. A. (2004). Collaborative strategic reading "real-world" lessons from classroom teachers. *Remedial and Special Education, 25*, 291-302. doi:10.1177/07419325040250050301

Landerl, K., Fussenegger, B., Moll, K., & Willburger, E. (2009). Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experimental Child Psychology, 103*, 309-324. doi:10.1016/j.jecp.2009.03.006

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.

*McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., & . . . Gray, A. L. (2002). Beginning literacy links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities, 35*, 69-86. doi:10.1177/002221940203500106

*McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient. *The Journal of Educational Research, 93,* 67-74. doi:10.1080/00220679909597631

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology, 1,* 2-7. doi:10.1186/1471-2288-1-2

Odden, A., Archibald, S., Fermanich, M., & Gallagher, A. H. (2002). A cost framework for professional development. *Journal of Educational Finance, 28,* 51-74.

Olson, R. K., Keenan, J. M., Byrne, B., Samuelsson, S., Coventry, W. L., Corley, R., . . . & Hulslander, J. (2011). Genetic and environmental influences on vocabulary and reading development. *Scientific Studies of Reading, 15*, 26-46. doi:10.1080/10888438.2011.536128

*Podhajski, B., Mather, N., Nathan, J., & Sammons, J., (2009). Professional development in scientifically based reading instruction teacher knowledge and reading outcomes. *Journal of Learning Disabilities, 42*, 403-417. doi:10.1177/0022219409338737

*Porche, M. V., Pallante, D. H., & Snow, C. E. (2012). Professional development for reading achievement. *The Elementary School Journal, 112*, 649-671.

*Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology, 102*, 299-312. doi:10.1037/a0017763

Savage, R. S., Burgos, G., Wood, E., & Piquette, N. (2015). The simple view of reading as a framework for national literacy initiatives: A hierarchical model of pupil-level and classroom-level factors. *British Educational Research Journal, 41,* 820-844. doi:10.1002/berj.3177

Savage, R. S., & Cloutier, E. (2017). Early reading interventions: The state of the practice, and some new directions in building causal theoretical models. In K. Cain, D. Compton, & R. Parrila (Eds.), *Theories of reading development*. Amsterdam: John Benjamins.

Saylor, L. L., & Johnson, C. C. (2014). The role of reflection in elementary mathematics and science teachers' training and development: A Meta- Synthesis. *School Science and Mathematics, 114*, 30-39. doi:10.1111/ssm.12049

Slavin, R.E. (2008). Perspectives on evidence-based research in education - What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*, 5-14. doi:10.3102/0013189x08314117

*Snow, P. C., Eadie, P. A., Connell, J., Dalheim, B., McCusker, H. J., & Munro, J. K. (2014). Oral language supports early literacy: A pilot cluster randomized trial in disadvantaged schools. *International Journal of Speech-language Pathology, 16*, 495-506. doi:10.3109/17549507.2013.845691

Surrette, T. N., & Johnson, C. C. (2015). Assessing the ability of an online environment to facilitate the critical features of teacher professional development. *School Science and Mathematics, 115*, 260-270. doi:10.1111/ssm.12132

The Consortium for Policy Research on Education (1995). *Professional development today*. Retrieved from: https://www2.ed.gov/pubs/CPRE/t61/t61c.html

*Timperley, H. S. & Phillips, G. (2003). Changing and sustaining teachers' expectations through professional development in literacy. *Teaching and Teacher Education, 19*, 627-641. doi:10.1016/s0742-051x(03)00058-1

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best Evidence Synthesis Iteration [BES]*. Retrieved from: http://www.educationcounts.govt.nz/publications/series/2515/15341

Torgerson, C. (2003). *Systematic reviews*. London, UK, New York, NY: Continuum International Publishing Group.

Torgerson, C., Brooks, G., & Hall, J. (2006). *A systematic review of the research literature on the use of phonics in the teaching of reading and spelling*. Nottingham: DfES Publications.

U.S. Department of Education (2001). *No child left behind*. (Public Law 107–110). Retrieved July 22nd, 2016 from: www2.ed.gov/nclb/landing.jhtml

*Vernon- Feagans, L., Gallagher, K., Ginsberg, M. C., Amendum, S., Kainz, K., Rose, J., & Burchinal, M. (2010). A diagnostic teaching intervention for classroom teachers: Helping struggling readers in early elementary school. *Learning Disabilities Research & Practice, 25*, 183-193. doi:10.1111/j.1540-5826.2010.00316.x

*Vernon-Feagans, L., Kainz, K., Amendum, S., Ginsberg, M., Wood, T., & Bock, A. (2012). Targeted reading intervention: A coaching model to help classroom teachers with struggling readers. *Learning Disability Quarterly, 35*, 102-114. doi:10.1177/0731948711434048

*Vernon-Feagans, L., Kainz, K., Hedrick, A., Ginsberg, M., & Amendum, S. (2013). Live webcam coaching to help early elementary classroom teachers provide effective literacy instruction for struggling readers: The Targeted Reading Intervention. *Journal of Educational Psychology, 105*, 1175-1187. doi:10.1037/a0032143

*Vernon- Feagans, L., Bratsch- Hines, M., Varghese, C., Bean, A., & Hedrick, A. (2015). The targeted reading intervention: Face- to- face vs. webcam literacy coaching of classroom teachers. *Learning Disabilities Research & Practice*, *30*, 135-147. doi:10.1111/ldrp.12062

Villegas-Reimers, E. (2003). *Teacher professional development: An international review of the literature*. Paris: International Institute for Educational Planning.

Wade, R. K. (1985). What makes a difference in in-service teacher education? A meta-analysis of research. *Educational Leadership, 42*, 48-54.

Wright, R. W., Brand, R. A., Dunn, W., & Spindler, K. P. (2007). How to write a systematic review. *Clinical Orthopaedics and Related Research, 455*, 23-29. doi:10.1097/blo.0b013e31802c9098

*Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2008). Reviewing the evidence on how teacher professional development affects student achievement: A systematic review. *Regional Educational Laboratory Southwest (NJ1)*.

APPENDIX A

Table A1

Key Search Term for Meta-analysis and Systematic Review on Reading and Literacy

Database	**Search Strategy**	**Number of hits**
Psych-info	Professional development AND teacher AND reading AND meta-analysis	0
Psych-info	Teacher Training AND reading AND meta-analysis	1
Psych-info	Professional development AND teacher AND reading AND systematic reviews	0
Eric	Professional development AND teacher AND reading AND systematic reviews	5
Eric	Professional development AND teacher AND reading AND meta-analysis	1
Education full text	Professional development AND teacher AND reading AND systematic reviews	2
Education full text	Professional development AND teacher AND reading AND meta-analysis	2
Campbell Collaboration library	Professional development OR teacher professional development	0
Campbell Collaboration library	Professional development AND teacher AND reading	0
What Works Clearing House	Professional development AND teacher AND reading AND systematic review	0
What works Clearing House	Professional development AND teacher AND reading AND meta-analysis	0

*The studies did not fit our selection criteria, thus were not included

Figure A1

PRISMA Flow Diagram

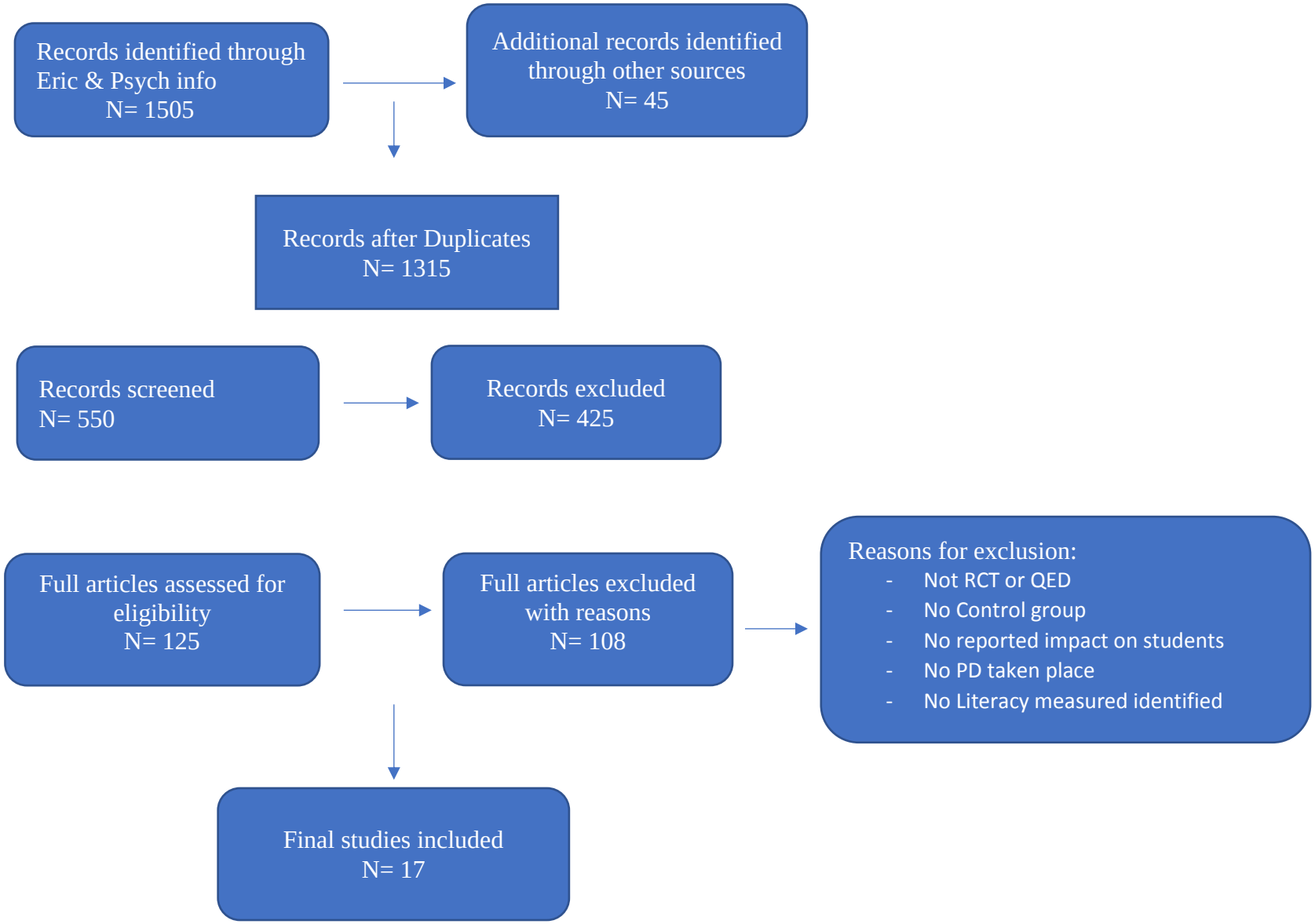

Table A2

Features of Professional Development in the 17 Studies of the Literature Review

Study	**Study design**	**Model of PD**	**Content**	**Provider of PD**	**Contact hours and duration**	**Content Group**
Amendum (2014)	Mixed method	ENRICH	Guided strategies to support students' oral reading	Author and ENRICH coach	26.5 hours	45 grade 1 students were divided into experiment and control group.
Duffy et al. (1986)	QED	N/S	Incorporating explicit instruction in teaching reading	Literacy Coaches	10 hours	Grade 5 students were stratified to teachers' treatment and group
Fine and Kossack (2002)	Clustered RCT	Cognitive Coaching	Professional talk among teachers	NS	26 hours	Grade 3 and 4 students randomly assigned
Garet et al. (2008)	RCT	LETRS PD CORE	Instructional strategies in reading	LETRS and CORES coaches	Treatment (A) 48 hours of PD Treatment (B) 48 hours of PD and 60hours of coaching	1530-second grade students from the 90 schools were randomly assigned to treatment and control group.
Gersten et al. (2010)	RCT	TSG	Implementation of new teaching strategies in classrooms	Primary researchers	20 hours	468 students from 19 reading first schools were randomly assigned
Klingher et al. (2004)	QED	CSR	How to incorporate collaborative strategic reading in the classroom	Literacy mentor	9 hours	211 5th grade students randomized into experiment and control groups

McCutchen et al. (2002)	QED	NS	Explicit instruction in phonological and orthographic awareness	Team university researchers	70 hours	779 students divided into experiment and control groups
McGill-Franzen et al. (2002)	RCT	NS	TPD using books schools	Trainers	30 hours	377 KG students
Podhajski et al. (2009)	Experiment-control method	TIME	Scientific based instructions in the 5 areas of literacy	TIME mentor	35 hours	33 1st & 2nd grade students with their teachers were assigned to experiment group. 14 1st & 2nd grade students with their teachers were assigned to control group
Porche et al. (2012)	Experiment control study	CLLIP	Assessments to diagnose and assign intervention strategies Instructional based strategies in literacy	CLLIP mentor	+50 hours	122 KG students assigned to 5-intervention and control group. 148 grade 4 students assigned to intervention and control groups
Powell et al. (2010)	RCT	Classroom Links to Early literacy	Classroom instructions to teach students phonological awareness and letter knowledge	Literacy Coach	36 hours	759 students divided to experimental and control group
Snow et al. (2014)	RCT	OLSEL	Incorporating classroom activities	OLSEL coaches	36 hours	1254 grade 1 & 2 students were

			that improves children's oral and receptive language			randomly assigned to treatment and control group
Timperley and Philips (2003)	RCT	N/S	Train teachers to understand the connection between what they teach and what children learn	Second Author	30 hours	193 students into experimental and control groups
Vernon-Feagans et al. (2010)	RCT	TRI	Teaching instruction in deciding and fluency	TRI literacy coach	+30 hours	8 experimental and 12 control classrooms
Vernon-Feagans et al. (2012)	Cluster RCT	TRI	Individualized instruction in decoding and fluency	TRI Literacy coaches	+30 hours	277 KG and Grade 1 students
Vernon-Feagans et al. (2013)	RCT	TRI) with web cam	Evidence based reading strategies to help struggling readers catch up with non-struggling readers	TRI Literacy coach	+30 hours	10 students were randomly assigned to experiment/control group from 631 students who participated in the study
Vernon-Feagans et al. (2015)	RCT	TRI	TRI was delivered face to face and through Web Cam	TRI literacy coach	+ 30 hours	271 KG and grade 1 students randomly assigned

Table A3

Effects of Professional Development Student Achievement by Study

Study	Outcome measure	Effect size
Amendum (2014)	Letter word identification	1.52
	Word attack	1.39
	Spelling of sounds	1.06
	Passage comprehension	1.52
Duffy et al. (1986)	Gates-MacGinitie Reading Test	0.07
Fine and Kossack (2002)	Degree of Reading Power (Grade 3)	0.90
	Degree of Reading Power (Grade 4)	1.02
Garet et al. (2008)	Student reading (Treatment A)	0.08
	Student reading (Treatment B)	0.03
	Follow up results	
	Student reading (Treatment A)	0.10
	Student reading (Treatment B)	0.01
Gersten et al. (2010)	**Students' outcome**	
	Passage comprehension	0.13
	Letter word identification	0.23
	ORF subtests	0.23
	California Achievement in Reading	0.20

Klinger et al. (2004)	Gates-MacGinitie Reading Test	0.49
McCutchen et al. (2002)	Gates-MacGinitie Reading Test	0.39
McGill-Franzen et al. (1999)	Concepts about Prints	1.11
	Hearing Sounds in Words	0.69
	Letter name knowledge	0.32
Podhajski et al. (2009)	**Grade 1**	
	Letter naming fluency	-0.16
	Phonemic segmentation fluency	0.97
	Nonsense word fluency	-0.16
	Oral Reading	-0.14
	Listening/ reading comprehension	2.25
	Grade 2	
	Phonemic segmentation fluency	0.56
	Sight word efficiency	-0.77
	Phonemic decoding fluency	-0.5
	Oral reading	-0.6
	Reading / listening comprehension	-0.25
*Porche et al. (2012)	**Kindergarten**	
	Letter Naming fluency	0.25
	Initial Sound Fluency	0.44
	Phoneme Segmentation Fluency	0.45
	Picture Vocabulary	0.14
	Grade 4	
	Letter Word Identification	0.09
	Passage Comprehension	0.15
	Fluency	0.26

Powell et al. (2010)	Letter-Word Identification	0.24
	Concept about Prints	0.22
	Blending	0.32
	Initial Sound Matching	0.17
Snow et al. (2014)	Reading progress test	0.60
	SPAT-R Subtest	0.35
Timperley and Philips (2003)	Letter Identification	0.38
	Concepts about Prints	0.53
	Word Knowledge	0.49
	Hearing and recording sounds in words	0.46
	Text level reading	0.61
Vernon-Feagans et al. (2010)	**Grade K**	
	Letter Word Identification	-0.81
	Word Attack	-0.12
	Grade 1	
	Letter Word Identification	-0.88
	Word Attack	-0.72
Vernon-Feagans et al. (2012)	**Kindergarten**	
	Letter word Identification	-0.15
	Word Attack	-0.07
	Grade 1	
	Letter Word Identification	-0.17
	Word Attack	-0.19

Vernon-Feagans et al. (2013)	**Kindergarten Struggling Readers Intervention group**	
	Word Attack	0.79
	Letter word Identification	0.66
	Passage Comprehension	0.57
	Sounds of Words	0.53
	Grade 1 Struggling Readers Intervention Group	
	Word Attack	0.44
	Letter word Identification	0.43
	Passage Comprehension	0.38
Vernon-Feagans et al. (2015)	Letter Word Identification	-0.76
	Word Attack	-0.86

*We thank the authors for providing us with the mean and standard deviation of their results to calculate the effect size

Table A4

Quality Code of Assessment for the 17 Studies

Author/Date	Reporting method of allocation	Sample size justification	Intention to treat analysis	Blinded assessment of outcome	Provided evidence impact on student	Described the process of professional development	Evidence made to establish reliability and validity	Evidence of treatment integrity
Amendum (2014)	Y	N/S	N/S	N/S	Y	Y	N/S	Y
Duffy et al. (1986)	N/S	N/S	N/S	N/S	Y	Y	N/S	N/S
Fine and Kossack (2002)	N/S	N/S	Y	N/S	Y	Y	Y	N/S
Garet et al. (2008)	Y	N/S	N/S	N/S	Y	Y	N/S	N/S
Gersten et al. (2010)	Y	Y	N/S	N/S	Y	Y	N/S	Y
Klingher et al. (2004)	N/S	N/S	Y	N/S	Y	Y	Y	N/S
McCutchen et al. (2002)	N/S	N/S	Y	N/S	Y	Y	Y	N/S
McGill-Franzen et al. (1999)	Y	N/S	N/S	N/S	Y	Y	N/S	N/S
Podhajski et al. (2009)	N/S	N/S	N/S	N/S	Y	Y	N/S	N/S
Porche et al. (2012)	N/S	N/S	N/S	N/S	Y	Y	N/S	Y
Powell et al. (2010)	Y	N/S	N/S	Y	Y	Y	N/S	Y
Snow et al. (2013)	Y	N/S	N/S	N/S	Y	Y	Y	N/S

Timperley & Philips (2003)	N/S	N/S	N/S	N/S	Y	Y	N/S	N/S
Vernon-Feagans et al. (2010)	Y	N/S	Y	N/S	Y	Y	N/S	Y
Vernon-Feagans et al. (2012)	Y	N/S	Y	N/S	Y	Y	N/S	Y
Vernon-Feagans et al. (2013)	Y	N/S	Y	N/S	Y	Y	N/S	Y
Vernon-Feagans et al. (2015)	Y	N/S	NS	N/S	Y	Y	N/S	Y

- Y= Yes, N/S = Not specified

Table A5

WOE (Weight of Evidence)

Author/Date	**WOE A**	**WOE B**	**WOE C**	**WOE D**	**PD hours**
Amendum (2014)	High	Low	High	High	26.5 hours
Duffy et al. (1986)	High	Medium	High	High	10 hours
Fine and Kossack (2002)	High	Low	High	High	26 hours
Garet et al. (2008)	Medium	Low	High	Medium	+48 hours
Gersten et al. (2010)	High	Medium	High	High	20 hours
Klingher et al. (2004)	High	Medium	High	High	9 hours
McCutchen et al. (2002)	High	Low	Medium	Medium	70 hours
McGill-Franzen et al. (1999)	Medium	Low	Medium	Medium	30 hours
Podhajski et al. (2009)	Medium	Low	High	Medium	35 hours
Porche et al. (2012)	High	Low	Medium	Medium	+ 50 hours
Powell et al. (2010)	High	Medium	High	High	45 hours
Snow et al. (2014)	High	Medium	Medium	Medium	36 hours
Timperly and Philips (2003)	High	Low	Medium	Medium	30hours
Vernon-Feagans et al. (2010)	Medium	Medium	Medium	Medium	+30 hours
Vernon-Feagans et al. (2012)	Medium	Medium	Medium	Medium	+30 hours
Vernon-Feagans et al. (2013)	Medium	Medium	Medium	Medium	+30 hours
Vernon-Feagans et al. (2015)	Medium	Medium	Medium	Medium	+30 hours

Appendix B

Figure B1

Funnel Plot Based on the 17 Studies in the Meta-analysis

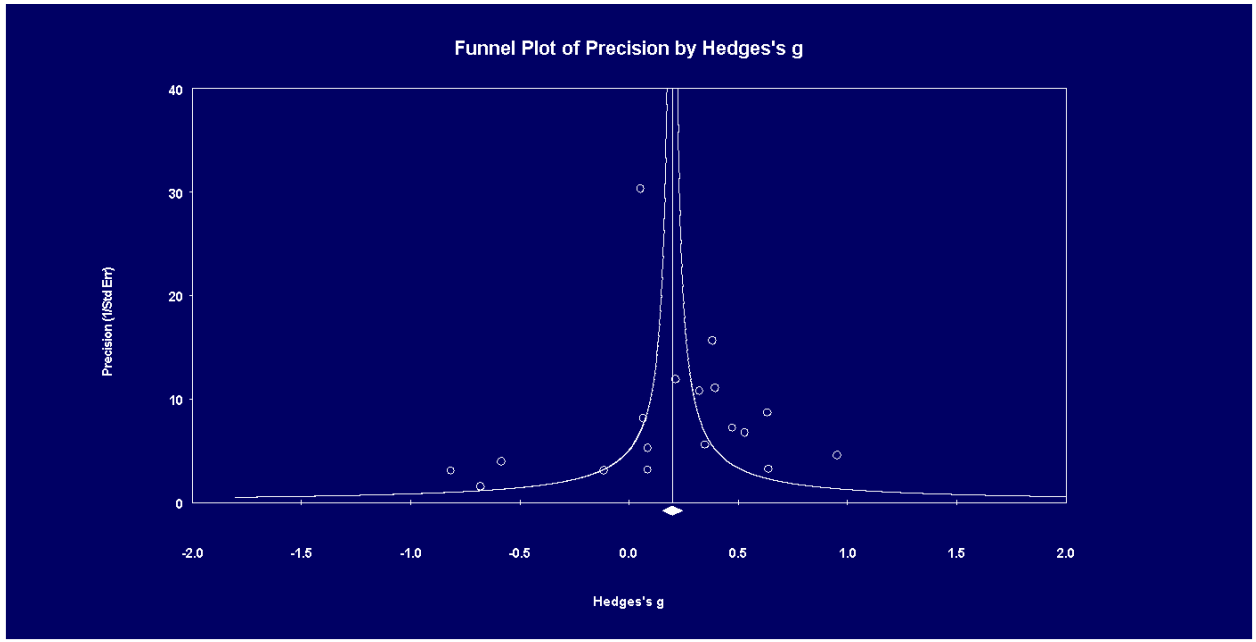

Table B1

Effect Sizes, Confidence Intervals, Z-values and p Values for All Studies

Study Name	Hedge's g	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	P-value
Amendum (2014)	0.635	0.313	0.098	0.021	1.249	2.027	0.043
Duffy et al. (1986)	0.067	0.124	0.015	-0.176	0.31	0.544	0.587
Fine and Kossack (2002)	0.951	0.221	0.049	0.516	1.385	4.292	0.000
Garet (2008)	0.055	0.033	0.001	-0.01	0.12	1.664	0.096
Gersten et al. (2010)	0.210	0.093	0.009	0.028	0.392	2.259	0.024
Klinger et al. (2004)	0.475	0.139	0.019	0.201	0.748	3.403	0.001
McCutchen et al. (2008)	0.395	0.091	0.008	0.217	0.574	4.35	0.000
McGill-Franzen (1999)	0.695	0.116	0.013	0.468	0.922	5.991	0.000
Podhajski et al. (2009)	0.084	0.321	0.103	-0.544	0.713	0.263	0.793
Porche et al. (2012)	0.211	0.192	0.037	-0.165	0.587	1.099	0.272
Powell et al. (2010)	0.214	0.084	0.007	0.048	0.379	2.531	0.011
Snow et al. (2014)	0.412	0.064	0.004	0.286	0.537	6.41	0.000
Timperly and Philips (2003)	0.496	0.149	0.022	0.205	0.787	3.336	0.001
Vernon-Feagans et al. (2010)	-0.570	0.323	0.104	-1.203	0.064	-1.763	0.078
Vernon-Feagans et al. (2012)	-0.146	0.242	0.059	-0.62	0.328	-6.604	0.546
Vernon-Feagans et al. (2013)	0.358	0.174	0.03	0.017	0.70	2.055	0.040
Vernon-Feagans et al. (2015)	-0.978	0.156	0.024	-1.285	-0.672	-6.262	0.000
Random	0.225	0.082	0.007	0.064	0.385	2.741	0.006

Table B2

Content Analysis of All 17 Studies

Author Name / Date	**PD hours**	**Effect size**	**Quality of study**	**Content of PD**
Amendum (2014)	26.5	0.635	High	Familiar rereading Word Study Teacher guided reading
Duffy et al. (1986)	10	0.067	High	Not enough information on PD content reported to document
Fine and Kossack (2002)	26	0.951	High	Rubrics, reflective journaling, and cognitive peer coaching
Garet et al. (2008)	>48	0.055	Medium	5 areas of literacy (Fluency, Phonics phonemic awareness, vocabulary, and reading comprehension
Gersten et al. (2010)	20	0.21	High	comprehension strategies vocabulary instruction phonemic awareness decoding phonics fluency
Klinger et al. (2004)	9	0.475	High	No enough information on PD content reported to document
McCutchen et al. (2002)	70	0.295	Medium	Phonology Phonological awareness Orthographic Awareness Spelling

McGill Franzen et al. (2002)	30	0.695	Medium	Physical design of the classroom effective book display Importance of reading aloud to children Environment Print Author genre, content, and theme
Podhajski et al. (2009)	35	0.084	Medium	Phonemic Awareness Phonics Fluency
Porche et al. (2012)	>50	0.212	Medium	Phonemic Awareness Phonological Awareness Alphabetic Principle Phonics Instruction Fluency Vocabulary Writing Reading Comprehension
Powell et al. (2010)	45	0.214	Medium	Phonological Awareness Letter Knowledge
Snow et al. (2014)	36	0.41	Medium	Phonological Awareness Phonemic Awareness Vocabulary Knowledge Story Grammar Comprehension
Timperly and Philips (2003)	30	0.496	Medium	Creating a link between saying/seeing in context of the semantic intent of the author
Vernon-Feagans et al. (2010)	> 30	-0.57	Medium	Rereading for Fluency Word Work Guided Oral Reading
Vernon-Feagans et al. (2012)	>30	-0.146	Medium	Rereading for Fluency Word Work Guided Oral Reading

Vernon-Feagans et al. (2013)	>30	0.358	Medium	Rereading for Fluency Word Work Guided Oral Reading
Vernon-Feagans et al. (2015)	>30	-0.978	Medium	Rereading for Fluency Word Work Guided Oral Reading