

The Effects of Professional Development Schools in Initial Teacher Education: A Review of the Evidence



by Sean Thomas Kaczmarek

September 2017

The Effects of Professional Development Schools in Initial Teacher Education: A Review of the Evidence

Sean Thomas Kaczmarek

September 2017

This dissertation may be made available to the general public for borrowing, photocopying or consultation without the prior consent of the author. This paper has been slightly adapted from the Masters dissertation submitted September 2017 for the MSc Social Policy and Social Research at the UCL Institute of Education.

© Copyright

Authors of the systematic reviews on the EPPI-Centre website (<http://eppi.ioe.ac.uk/>) hold the copyright for the text of their reviews. The EPPI-Centre owns the copyright for all material on the website it has developed, including the contents of the databases, manuals, and keywording and data-extraction systems. The centre and authors give permission for users of the site to display and print the contents of the site for their own non-commercial use, providing that the materials are not modified, copyright and other proprietary notices contained in the materials are retained, and the source of the material is cited clearly following the citation details provided. Otherwise users are not permitted to duplicate, reproduce, re-publish, distribute, or store material from this website without express written permission.

Table of Contents

Figures.....	2
Definitions.....	4
1. Introduction.....	5
1.1 Rationale for Research.....	6
1.2 Background and Literature Review.....	8
1.3 Research Question.....	15
2. Methodology.....	16
2.1 Conceptual Framework.....	16
2.2 Identification of Relevant Studies.....	18
2.3 Data Extraction, Assessment, and Management.....	21
2.4 Data Synthesis.....	23
3. Results.....	27
3.1 Study Descriptions and Quality Assessment.....	27
3.2 Aggregated Results for Indicators of Teacher Preparedness.....	38
3.3 Funnel Plots.....	54
4. Discussion.....	56
4.1 Limitations of this Study.....	57
4.2 Implications for Research.....	58
4.3 Implications for Policy.....	60
4.4 Conclusion.....	64
References.....	65
Appendix 1: Data Extraction Tool.....	70
Appendix 2: Weight of Evidence Rubric.....	86

Figures

- Figure 3.1. Fixed Effects Meta-Analysis for Urban Environments and Diversity
- Figure 3.2. Fixed Effects Meta-Analysis for UED without McKinney et al. (2004)
- Figure 3.3. Fixed Effects Meta-Analysis for Self-Efficacy (High SD)
- Figure 3.4. Fixed Effects Meta-Analysis for Self-Efficacy (Low SD)
- Figure 3.5. Fixed Effects Meta-Analysis for Reflective Educator
- Figure 3.6. Fixed Effects Meta-Analysis for Classroom Management & Org'n
- Figure 3.7. Fixed Effects Meta-Analysis for Lesson Planning (High SD)
- Figure 3.8. Fixed Effects Meta-Analysis for Lesson Planning (Low SD)
- Figure 3.9. Funnel Plot A
- Figure 3.10. Funnel Plot B

Abstract

In the United States, initial teacher education (ITE) has seen many debates and attempted reforms over recent years. While some advocate for shorter, more intensive programmes, the Professional Development School (PDS) movement stands at the opposite end of reform, promoting more thorough, and potentially longer, teacher education. While many studies have discussed the potential value of Professional Development Schools, few have presented systematic reviews of the evidence from which to derive judgement. A systematic review is important because the effectiveness of education reforms can vary based on context, and a review would provide better insights into contextual and aggregate success than a single study could provide. Using an abridged systematic review, a systematic rapid evidence assessment, this study examines the impact of Professional Development Schools compared to more traditional teacher education across five sub-categories: Interaction with Diverse Students in Urban Environments (hereafter referred to as Urban Environments and Diversity), Self-Efficacy, Reflective Educator, Classroom Management and Organisation, and Lesson Planning. Included in this assessment are articles found on the ERIC (ProQuest) database showing comparative data on both Professional Development Schools and more traditional ITE forms. A Quality Assessment for each study was conducted. Quantitative data were entered into the MetaLight meta-analysis programme to develop pooled effect sizes for each sub-category. Qualitative data, if found for a sub-category, were used to augment the quantitative analysis. A total of 10 studies provided data which was included in this review. All pooled effect sizes were positive, although some were found to be insignificant. Qualitative data predominantly showed support for Professional Development Schools, although some did indicate otherwise. Overall, the findings suggest Professional Development Schools may provide positive effects, although the time constrained nature of this study suggests more research ought to be conducted. The shortcomings identified in this study point to the need for educational research to strengthen data collection methods and the evidence on which policy is based.

Definitions

Initial Teacher Education (ITE) – Also called teacher training or teacher education. It is through these types of programmes which teacher certification/licensure is obtained.

Pre-Service Teachers – Also called teachers-in-training, student-teachers, or a variation of ITE students. This group is enrolled in an ITE programme.

In-service Teachers – Teachers who are currently classroom teachers.

Co-operating Teachers/Mentor Teachers – Teachers at school sites where pre-service teachers undertake their clinical field experiences. These teachers act as supervisors and/or mentors to the pre-service teachers. Their exact duties and responsibilities vary depending on the type of teacher preparation programme of the student-teacher.

Professional Development School (PDS) – A specific type of school-university partnership which involves collaboration and restructuring of roles between university faculty and school faculty. The broad aims of PDSs are to improve ITE through longer clinical field experiences with more reflection, more support from co-operating teachers, and more consistent performance feedback.

Traditional Teacher Preparation/Education – While teacher education programmes considered more traditional may vary in many ways, there are common differences with Professional Development Schools. A traditional teacher education programme does not include a formal partnership between a university teacher education program and a local school site. This results in less integration between the programme and the school sites of its student-teachers and their co-operating teachers. Because of this, supervision is not as structured; university educators and in-service/co-operating teachers keep to their own sites, except for the occasional student-teaching observation required of the university educator. Coursework is typically undertaken at a university campus, with attendance at a school site reserved for required observations and/or student-teaching field experiences.

1. Introduction

In his 2011 State of the Union Address, United States President Barack Obama called on the U.S. to value its teaching force as highly as South Korea. “In South Korea,” the president said, “teachers are known as nation-builders.” This statement, and many others like it, praise teachers by recognizing the responsibility they hold. Many will readily agree that, as nation-builders, teachers are not merely dispensers of knowledge, but should look to instil in their students a desire to learn and help their students’ personal development. Teachers ought to, and do, push students to achieve at their highest potential. While some outwardly praise teachers for choosing this profession, others are openly critical toward teachers when students do not meet arbitrary performance expectations.

If teachers are to have the impact that is anecdotally claimed, the training which teacher candidates receive must properly prepare them. However, evidence of problems with initial teacher education (ITE) is anything but anecdotal. In one example, Haberman (1987, 1995) discusses problems faced by urban districts, specifically pointing toward accepted ITE applicants and the structure of ITE programmes. Haberman argues that admission to ITE programmes should be a more rigorous process, especially those specifically targeted toward urban environments. Requirements for admission would include a minimum number of hours spent in classrooms, an interview measuring ability to appropriately interact with at-risk youth (which is possible to fail), and continuous evaluations of students prior to and during their clinical field experiences. In terms of programme structure, Haberman claims ITE teacher educators often have little experience teaching, especially with youth from urban and/or low-income backgrounds. It is reasonably questioned how well these teacher educators could teach others to work in this environment if they never have themselves.

While the issues with traditional teacher preparation must be discussed and debated along with potential solutions, it is also important to acknowledge and understand the nature of teacher supply and demand. A study by Ingersoll (2002) found that, although national teacher turnover is higher than the national turnover average for other professions, the aggregate number of teachers is not

an issue. In fact, some districts have waiting lists of qualified teachers who want to join their districts. However, within public school systems, districts in high-poverty areas have the highest national turnover rates. Half of teacher departures are cases of migration to other districts rather than permanent retirement. In a follow-up study in 2003, Ingersoll sums up the problem by stating it is not just a basic shortage, but a shortage of teachers who desire to continue teaching in specific school environments (that of urban and high-poverty districts) at a specific salary. This context-specific shortage may result from numerous factors, but the dearth of university teacher education programmes which collaborate and are integrated with less affluent districts does not ameliorate this problem.

Overall, Haberman (1986) claims it has become too easy to become a teacher. He alleges the easy ITE pathway is exploited by many who are not actually interested in teaching as a profession: graduate students with no funding sources, those wanting to improve their resumes, or people who do not know what else they should do and use it as time to personally grow and learn about themselves. If this path were to continue, especially with teacher shortages, he asserts teachers could be relegated to job-holders, not professionals. As a result, it is argued that teacher education programmes need to be more selective in accepting applicants who are committed to the teaching profession.

1.1 Rationale for Research

Teacher Effects

While arguments have been made for ITE improvement, questions of a teachers' ability to significantly affect a students' academic outcomes have existed for decades. If a teacher has minimal effect on student achievement, is ITE worth the investment and reform efforts? A basis for these questions is provided by a seminal education report conducted by sociologist James Coleman, *Equality of Educational Opportunity* (the Coleman Report) (1966). The report found that non-school factors such as family background and socioeconomic status are the most significant determinants of student achievement outcomes. The effects of these non-school factors on outcome

equity can be seen when viewing the educational achievement gap by race and income in the United States.¹

Although education-specific policy interventions cannot change already-established student backgrounds, they are able to affect components of the school system including teachers, curriculum, class size, funding, and school-type availability (public or private schools – the United States' counterpart to U.K. state or public schools, respectively). Within these school system components that education policy can directly affect, much research conducted after the Coleman Report has focused on teacher preparation and/or teacher quality as they relate to student achievement. The importance of teacher quality, alone, has been the subject of hundreds of reports (Hanushek, 2011).

Goldhaber (2002) acknowledges the Coleman Report's finding that factors outside the purview of the formal education system are the most influential factors on student achievement. However, he also claims teachers may be able to account for 8.5% of the variation in student achievement, making teacher quality the largest in-school factor affecting student achievement outcomes. This finding appears to corroborate Darling-Hammond's prior research (2000b) which claims teacher quality is more highly correlated with student outcomes than any other factor. Darling-Hammond's study further demonstrated a correlation between state-wide student achievement levels and state spending on teacher quality. While Rockoff (2004) and Hanushek (1986) argue that many qualities of good teachers are unobservable and thus cannot be empirically studied, they both claim that measures to raise teacher quality may be critical for improving student outcomes.

In Tennessee, where a value-added system² specifically tracks teacher performance, it has been found that students with teachers in the top quintile and bottom quintile of performance predominantly made desirable and unsatisfactory academic gains, respectively. These phenomena was also found

¹ See (Hemphill & Vanneman, 2011), (Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009), and (Reardon, 2012) for Hispanic-White, Black-White, and income-level achievement gaps, respectively.

² Value-Added systems incorporate value-added measurements into teacher evaluation methods. Value-Added measurements are econometric tools used to measure the effect of teachers on student outcomes. Originally developed by Eric A. Hanushek (1971).

to be cumulative, with observable, residual effects two years afterward for both effective and ineffective teachers (Sanders & Rivers, 1996). Wright, Horn, and Sanders (1997), also focused on the Tennessee value-added system, found the dominant in-school factor affecting student gains is teachers, aligning with previously-mentioned research. This effect was more strongly correlated with student outcomes than either class size or class heterogeneity. Both papers found that lower achieving students made the most gains, and were the first to benefit from high-quality teachers.

It follows that teacher education ought to be a natural focus of research and policy. Researching, understanding, and developing the best methods for teacher education should be able to enhance and widen the positive effects that teachers can have on their students. However, differing opinions on how to develop quality teachers and quality teacher education programmes have sparked fierce debates.

1.2 Background and Literature Review

Traditional Teacher Training

Traditional ITE in the United States consists of various components. Among these are educational coursework taken at university-based schools of education, subject matter coursework set in specific liberal arts departments, and practice teaching in schools (Sykes, Bird, & Kennedy, 2010). Because these three components take place in different locations, Sykes, et al. (2010) claims they are difficult to unite into a coherent programme. Darling-Hammond (2000a) discusses further issues with traditional teacher education. Four years is often not enough time for an undergraduate student to sufficiently study both pedagogy and subject matter. In addition, many content courses may be structured around anecdotes of teaching experience, rather than methods to properly develop knowledge about curriculum, pedagogy, and educational psychology. A significant disconnect between the coursework undertaken at the university and the field-experience undertaken at a school has become a further common criticism. Complicating this disconnection is the inadequacy of clinical training due to poor relationships and a lack of integration between the

university and the school site. A lack of resources continues to ingrain many of these problems.

Even for those who are dedicated to the teaching profession, teacher education's short duration, poor resources, its structure of content coursework, and insufficient supervision and management of clinical field experiences are still problems. These issues are reiterated by Sykes, Bird, and Kennedy (2010) before they note what is, possibly, one of the largest problems with teacher training programmes: the success of a teacher (and, ergo, their students) can often be due to the setting in which they work. If teacher training programmes do not effectively prepare students for multiple settings and cultural contexts, it is unlikely they will find success outside of the contexts of their own experience. Specifically, it has been argued that urban and high-needs districts bear a large portion of the consequences of this lack of preparation.

Importance of Preparation and Certification

Research has demonstrated that preparation leading to full teacher certification is beneficial for student achievement (for example, Hawk, Coble, and Swanson (1985)). Boe, Shin, and Cook (2007) come to similar conclusions in their study, which finds that pre-service teachers with extensive preparation felt better prepared in their pedagogical skills and more able to teach effectively (this included both special education teachers and general education teachers). Results also showed that pre-service teachers who underwent more extensive preparation were more likely to plan lessons effectively, use a variety of instructional methods, and be better classroom managers.

Despite the aforementioned issues that plague more traditional teacher preparation programmes, Darling-Hammond (1996, 2000a) discusses promising new developments which attempt to address these problems. Included within these developments are more preparation in learning methods and cognition, the development of a reflective and problem-solving skillset, the promotion of inquiry, and consistent supervision within clinical field experiences. These components have become an integral part of formal school-university partnerships called Professional Development Schools (PDSs), which seek to fundamentally change the way teachers are prepared. However, arguments for

the expansion of shortened, alternative programmes are still made, and among their advocates are prominent members of the education system.

Alternative Pathways

Alternative and shortened teacher education programmes have been promoted in recent decades as a way to solve issues with traditional ITE programmes and the lack of equity throughout the education system. Organisations such as Teach for All (developed from Teach for America (TFA) and later including programmes such as Teach First) and The New Teacher Project (located in the United States) are programmes that have promoted shorter, intensive teacher education curriculum. Teacher education is completed during one summer, and cohort members become a certified teacher with their own classroom that autumn. These programmes were conceived to address the problems within traditional teacher education, location-specific teacher shortages, and educational inequity and inequality.³

Notable education leaders, such as U.S. Secretary of Education Rod Paige, have called for the expansion of such programmes to provide both quicker routes into teaching and more highly-qualified teachers (HQTs) as defined by the No Child Left Behind Act of 2001 (2003). The Secretary argued there was no known relationship between full teaching certification and HQTs, and that traditional ITE had become too cumbersome. The research used to support these positions has been called into question by subsequent research. Two prominent examples are papers by Goldhaber and Brewer (2000) and Monk (1994). Monk's data found that the amount of pedagogy training was positively related to student learning, and was, in some cases, more powerful than additional content area preparation. Goldhaber and Brewer's paper also showed positive effects of full certification on student achievement levels. Despite the evidence in their respective papers, both make ill-founded conclusions that pedagogy training and certification are not linked to student achievement.⁴

³ Historical information can be found on Teach for America, Teach for All, and The New Teaching Project websites.

⁴ For more discussion and detail on the conclusions of Goldhaber's and Brewer's article, please see Darling-Hammond, Berry, and Thoreson (2001)

Further comparisons between TFA and non-TFA teachers who had full certification, alternative certification, or no certification found those with full certification were generally more effective, regardless of their entrance programme (Darling-Hammond, Holtzman, Gatlin, & Heilig, 2005). In addition, they found no circumstance where students of an uncertified TFA teacher outperformed a certified teacher of similar experience levels. In addition, TFA's two-year attrition rate is nearly double that of the national average for teachers after two years. While there may be positive effects of such programmes in high-turnover districts that otherwise would not have permanent teachers (Darling-Hammond et al., 2005), this certainly casts doubts on the ability of TFA and other shortened, alternative programmes to promote long-term stability in the teaching supply (Donaldson & Moore Johnson, 2011). Ideally, fully-certified teachers who were thoroughly and fully prepared, as is proposed by Professional Development Schools, would make up most of the teaching force.

Philosophy of Professional Development Schools

Two decades ago, an association of university education faculty established The Holmes Group, which set out to solve the problems they saw in current teacher education programmes. Throughout three reports – *Tomorrow's Teachers* (1986), *Tomorrow's Schools* (1990), and *Tomorrow's Schools of Education* (1995), they put forth a view of teacher education programmes which integrate subject matter coursework, educational coursework, and practice teaching in school settings (2007).⁵ Professional Development Schools, as named by the Holmes Group, are meant to be a partnership between university-based teacher education programmes and local primary, middle, or secondary schools.

Darling-Hammond (1994) notes PDSs are designed to address problems that have been identified within traditional ITE. Among these is a long-standing learn-by-observation mentality without significant direct practice, lack of a supportive environment for career-long learning by school staff, and the previously mentioned issue of insufficient connection between university

⁵ The cited work is a compilation of the three reports produced by the Holmes Group, published as a trilogy in 2007, rather than each individual report.

schools of education and local schools. Through continual feedback, shared decision-making, research inquiry, and exposure to multiple school environments, PDSs have the ability to significantly alter teacher education, and the education system as a whole. Not only would they support the education of future teachers, but the increase of engagement between universities and schools would expand the knowledge-base of education and encourage consistent professional development. As summed up by the Holmes Group (2007), PDSs are intended to serve as the premier site of teacher education, and, in order to address concerns with teacher education, should be based on six principles.

The first two principles are based on continual learning for understanding and the creation of ITE programmes which integrate universities, schools, and local communities. These principles are further rooted in the belief that teachers-in-training ought to understand theories and concepts, and how they will be applied to classroom situations. Continual learning results from active curriculum that develops skills of problem-solving, questioning, arguing, and reasoning. It involves not merely memorisation of facts, but plentiful classroom exposure, allowing pre-service teachers to learn appropriate reactions in various situations that may arise. When integrated with in-service teachers and university researchers, and ITE programme allows its pre-service teachers to discuss and debate with people of multiple viewpoints and different experiences, creating these integrated learning opportunities. Through observation and responses followed by discussion, students will be able to hone their own knowledge and more thoroughly comprehend not only what they are seeing in a classroom, but how to respond to it in the most effective way.

The third principle is the need to recognise the importance of a quality education for all students, regardless of community and socio-economic contexts. This recognition involves overcoming barriers to academic success such as structural racism and ethnocentric attitudes (2007). By providing experiences to pre-service teachers in various (multi-)cultural settings, a better understanding of these issues can be developed. Including discussions of racism, xenophobia, and discrimination in ITE programmes will only better prepare future teachers to enter a classroom with the necessary cultural capital.

The values enshrined in this principle may also be related to teacher shortages, which, as discussed previously, predominantly occur in urban and low-income environments. With better preparation for multi-cultural and high-poverty environments, teacher turnover may become less of an issue.

While the fourth principle focuses on the continued learning of teachers, teacher educators, and administrators, such a focus would enhance teacher preparation by ensuring those whom the student-teachers learn from are up-to-date and well-informed of contemporary education's best practices. These groups are also part of the fifth principle, the development of long-term inquiry and reflection. With reflection and inquiry being at the centre of teacher education, teachers-in-training would be able to study the problems they see or are having in the classroom more in depth, reflecting with their co-operating teacher and university faculty when discussing potential solutions.

The sixth and final principle is the establishment of new institutions that put these principles at the centre of their education programmes. Within this principle are five guideposts for the structure of these organisations: coordination of instructional and community services to connect students' academic and personal lives; flexible staffing assignments which allow collaboration, research, reflection, and inquiry by each staff member; the creation of assessments based on the specific context of the school and development of professional standards for teachers (including technical, social, and reflective practices); collaboration for staff development and curriculum development; and reciprocity agreements between universities and schools that engage both professors and teachers.

Effectiveness Claims of Professional Development Schools

Darling-Hammond (1994) argues PDSs have potential to implement beneficial ITE reforms. Providing one example, she discusses how some schools have integrated regular staff collaboration into their schedules, helping to shape understanding and promote knowledge-sharing among colleagues. Such collaboration is recognised as important by staff-members and may help pre-service teachers at the beginning of their field experiences.

Aside from increased collaboration, Darling-Hammond notes other potential benefits of PDSs, including diversity of learning environment, connecting pre-conceived notions of good teaching with actual practice, translating educational theory to the classroom, and supporting the most up-to-date education methods (1994, 1996). Citing the work of Frakes, Valli, and Cooper (1998), Byrd and McIntyre (1999) point out that PDSs appear to promote further improvements in the education system, such as research-based efforts among university and school site staff, teacher leadership development, enhanced mentoring for pre-service teachers, and an increased frequency of university faculty teaching in schools and school site teachers directing university courses.

A study by Reinhartz and Stetson (1999) displayed results that indicated PDS-trained teachers predominantly felt better, more prepared, and/or knowledgeable than non-PDS trained peers on measures such as quality of teacher education programme, knowledge of school functions, instructional ability, classroom management, and leadership skills. Comments from school site principals confirmed these personal assessments. The principals indicated their belief that the PDS-based pre-service teachers appeared better prepared, less anxious, were better at working with a diverse set of students, and had more self-confidence. Kochan's study (1999) found the largest benefit of PDSs from the view of stakeholders (including university and school site faculty and administration) was improved teacher education. Among the reasons cited for this view was more realistic experiences for pre-service teachers, more exposure to diversity, and more bona fide preparation. Collaboration, including trust, the translation of theory into practice, professional growth, and school-university collaboration followed up improved teacher education as the second-largest benefit from the stakeholders' point of view.

However, along with the potential benefits of PDSs, there are challenges that must be recognised. The same Kochan study based on stakeholders' views also indicated those stakeholders felt time-intensity, labour-intensity, long-term commitment, and the slow pace of progress were challenges to which a PDS had to overcome and adapt. It is these concerns that has led some to question whether PDSs have become responsible for more than they can handle (Olson,

1989). The future development and effectiveness of PDSs lay in their ability to overcome these challenges and fulfil their promise.

1.3 Research Question

PDS advocates argue that PDSs have the potential to reform teacher education in significant ways. However, perhaps due to lack of resources, PDS ITE is not the primary form of teacher education in the United States. Given the noted problems with current ITE programmes and teacher retention, especially in urban environments, it would be useful to investigate the potential benefits of PDS sites in addressing these issues. A systematic review of current research on PDS impacts in ITE could provide valuable information. Thus, the question being asked is:

What is the impact of Professional Development Schools on outcomes of teacher preparedness in comparison to more traditional teacher preparation programmes?

2. Methodology

2.1 Conceptual Framework

PDS advocates argue that PDS forms of ITE develop teachers who are better prepared. This claim would inherently rule out the possibility that other, more traditional forms of ITE prepare teachers better than PDS sites. If this is true, one would expect PDS sites to consistently outperform more traditional ITE programmes on indicators of teacher preparedness. A comparative study between PDS ITE and more traditional ITE can explore PDS advocates' claim while measuring the effects of more traditional ITE forms on teacher preparedness. Without data on both PDSs and more traditional ITE approaches, we would have difficulty determining the validity of the causal claim argued by PDS advocates.

In order to measure the effects of PDS interventions, this study will compare the reported preparedness results of pre-service teachers who were trained at a PDS site and pre-service teachers who were prepared in a more traditional ITE programme. A systematic rapid evidence assessment will be used. As an abridged systematic review, a systematic rapid evidence assessment will allow data on PDS ITE and non-PDS ITE to be gathered and assessed from multiple sources. Meta-analyses will be conducted for quantitative purposes. Any qualitative findings or other data that is relevant, but not able to be included in the meta-analysis, will be used to augment the study.

In order to provide a basis for a causal claim, any meta-analysis should be significant and demonstrate a pooled effect size which favours the PDS intervention. When analysing meta-analysis results, it is likely context will need to be taken into account. Without studying the specific context of the results, understanding why an intervention succeeded or failed is not possible. This is particularly important in an urban environment, as research has shown that effective teaching in urban schools requires different skills than effective teaching elsewhere (Haberman, 1987; McKinney & Finke, 2005; McKinney, Haberman, Stafford-Johnson, & Robinson, 2008; McKinney, Robinson, & Spooner, 2004). When context is accounted for, it is possible that a small effect

size represents a large impact. However, we do not have data on how effect sizes of various indicators on teacher preparedness impact student academic gains. Rather than assigning an arbitrary effect size which would have a large impact, it is best to view the results in their contexts and on a case-by-case basis. Both the findings of the meta-analysis and those studies not included in the meta-analysis will be taken into account.

Population

Students enrolled (or previously enrolled) in ITE programmes are the population of interest. Some studies also include in-service teachers.

Intervention

The PDS model of ITE is the intervention. Therefore, students enrolled in PDS ITE programmes may be considered a treatment group, while those enrolled in a traditional form of ITE act as a control group.

Comparison

Comparisons between PDS ITE students' and more traditional ITE students' preparedness outcomes for teaching will be the comparative focus of this assessment.

Outcome

The overall outcome sought is teacher preparedness. Initially, no specific indicators/sub-categories of teacher preparedness (e.g., classroom management, lesson planning, etc.) were included when conducting a literature search. After the literature search and data extraction, five specific sub-categories of outcomes were identified (the development of these categories is explained in more depth during the "Data Synthesis" section of this paper). These categories are Urban Environments and Diversity, Self-Efficacy, Reflection, Classroom Management and Organisation, and Lesson Planning. Both quantitative and qualitative outcomes were gathered, which included some graphs displaying trends over time.

2.2 Identification of Relevant Studies

The search process needed to produce sufficient information, but also a manageable amount of information to be systematically assessed within the given time frame. To focus this evidence assessment on the effects of PDS ITE interventions on teaching preparedness outcomes, specific inclusion and exclusion criteria were developed for articles to be assessed. The search process outlined below was the initial search criteria used for a literature search on ERIC (ProQuest), accessed via University College London. Any change in criteria for inclusion or exclusion are noted in the appropriate section of this paper.

Exclusion and Inclusion Criteria

Topic

Each study must have reported on the impact of PDS versus the impacts of more traditional teacher education programme. While the specificities of PDSs may differ at each site, papers were only included in this study if the PDS site in the research included an elongated field experience (or multiple field experiences) with more integrated supervision of that experience for pre-service teachers. More integrated supervision includes increased feedback and supervision by both the assigned co-operating teacher and university staff. Most papers self-identify certain programmes within their research as a PDS site, or a non-PDS site.

Reported Data

To have been included, a paper must have contained data which displayed the impact of a PDS site compared to the impact of a more traditional ITE programme on teacher preparedness.

Language

Only studies published in English were included in this paper.

Date of Publication

The final returns from search databases did not include any restriction on date of publication. More information on this aspect of the review process is described in the Search Strategy section.

Search Strategy

Sources

Searches were conducted on the electronic databases ERIC (ProQuest).

Use of Search Strings and Filters

Search strings were developed to identify the population of interest, intervention of interest, and comparisons of interest. Outcomes were not specified in search strings. Searching for outcomes with pre-defined terms would leave open the possibility that potential outcomes will be missed, and excluding such terms will allow a more thorough search of all potential outcomes. Filters for “scholarly journals,” “peer-review,” and “English Language” were used throughout the search process. It should be noted that “scholarly journals” and “peer-review” filters may limit the number of the returned pieces of research, and subsequently not include pieces published elsewhere or not peer-reviewed. Recognising this possibility, a funnel plot of the final included studies was produced to check for publication bias.

While using ERIC (ProQuest), developed search strings consistently returned an unmanageable number of articles. When this occurred with the first several search attempts, a filter for years 1986-2017 was included to restrict the amount of returns to a manageable number. These years were used as a filter because *Tomorrow's Teachers*, The Holmes Group report which formed the basis for future PDS reports and proposals, was published in 1986. However, this year range did not result in a sufficient reduction of the amount of articles returned.

As different search strings continued to return an unmanageable number of articles, a filter of years 2002-2017 was used for restriction. The year 2002

coincided with the implementation of *The No Child Left Behind Act*, considered a major shift in American education policy towards students and teachers.

Because search strings continued to return a large and unmanageable number of articles, a simplified search string was used due to the concern that search strings with too many terms were returning too much irrelevant literature. The simplified search string of "*professional development school*" AND student was used, and returned 102 articles. This search string may be simplified, but, as discussed earlier, more elaborate and specific search strings resulted in an unmanageable number of returns, sometimes returning hundreds of thousands of articles. Once the publication date limitation was removed, the search returned 143 articles, which formed the basis of the systematic assessment process. Of course, the possibility that the simplicity of this search string resulted in the exclusion of relevant articles must be recognised. However, due to the time constraint on this paper, the benefit of this simplified search string (a manageable number of articles) outweighed the cost (more time spent developing search strings and less time spent analysing the results).

Screening

The inclusion and exclusion criteria was applied to each title and abstract returned by the final search string to determine which articles were relevant for this assessment. If, based on the title and abstract, it was clear an article was not relevant, it was not included in any further screening or data collection processes.

Studies which appeared to satisfy the initial inclusion criteria, and studies with insufficient information for that determination, were then moved onto the full-text screening process to determine their relevance for this assessment. Out of the 143 articles returned, 79 articles were moved on to full-text screening.

Full-text screening provided a determination on whether pre-service teacher outcomes were reported. If it was determined that pre-service teacher outcomes were not reported on, the article was not included in any further screening or data collection. Out of the 79 articles that received full-text screening, 19 were determined to have reported on pre-service teacher

outcomes at PDS-based teacher education programmes. Of these 19 articles, it was noted that 10 studies were comparisons involving a more traditional, non-PDS programme. Out of these 10 studies, eight provided quantitative measures used in meta-analyses (one out of these eight studies measured both PDS vs. non-PDS pre-service teachers and PDS vs. non-PDS in-service teachers. Both pre-service measurements and in-service measurements were used for the meta-analyses. As a result, a total of nine different measurements from eight different papers are included throughout the meta-analyses.), and two were restricted to the non-quantitative analysis.

2.3 Data Extraction, Assessment, and Management

Data Extraction

Data extraction was conducted on the 10 articles which were determined relevant for the aims of this study. The data extraction tool used was adapted from a tool developed by the EPPI-Centre for systematic reviews (2003). Please see Appendix 1 to view the tool used.

Quality and Relevance Assessment

After data was extracted, each study was judged for quality and relevance.⁶ For the quality and relevance assessment (QRA) a Weight of Evidence (WoE) framework was used. As discussed by Gough (2004) the framework allows for an overall judgment of an individual study (high, medium, or low) based on its overall coherence and trustworthiness, and the suitability and relevance for the present research. Consisting of three sections and the overall judgment (sections A, B, C, and D, respectively), separate criteria were developed for primary research studies and reviews under each WoE section. To determine the criteria under each section the TAPUPAS framework (Pawson, Boaz, Grayson, Long, & Barnes, 2003) consisting of Transparency, Accuracy, Purposivity, Utility, Propriety, Accessibility and Specificity, was incorporated, as

⁶ While all studies should be relevant because they have met inclusion criteria, there may still be variation in relevance within included studies (e.g., two studies with a sole focus on teacher preparedness outcomes and partial focus on teacher preparedness outcomes, respectively, may both be included, but assigned a different level of relevance).

shown in Gough (2007). The content of each section's evaluation is described below. For a rubric chart of weight of evidence ratings, please see Appendix 2.

Weight of Evidence A (WoE-A)

This section specifically focuses on the individual study. Judgments for WoE-A were made about the clarity of purpose, clarity of methodological and data collection details, the validity of the conclusion drawn from that data, and the coherence of the overall report. Any Ethics considerations will fall under this section. In terms of the TAPUPAS framework, the categories under this section are transparency, accuracy, propriety, accessibility, specificity.

Weight of Evidence B (WoE-B)

This section relates to the methodology used for each collected report, how well the methodology fits the purpose of the current assessment, and the usefulness of the report. Since this study looks at measures of teacher preparedness at the end of initial teacher education, data is primarily gathered through questionnaires, interviews, and/or evaluations about preparedness given to student-teachers or superiors toward the end of training.⁷ In terms of the TAPUPAS framework, the category under this section is purposively.

Weight of Evidence C (WoE-C)

This section concentrates on the relevance of the content for each collected piece of research. Not only is the topical focus included in the evaluation for WoE-C, but the nature of the evidence gathered (e.g., if gathered through sampling, how large were the samples, from where were the samples drawn, etc.) is also included. In terms of the TAPUPAS framework, the category under this section is utility (i.e. did the study include comparisons between PDS and non-PDS groups?).

⁷ Superiors refers to university faculty and/or administration, school faculty and/or administration, or any other person who supervised study subjects.

Weight of Evidence D (WoE-D)

This is the overall judgment of the piece of research taking WoE-A, WoE-B, and WoE-C into account.

Data Management

All 143 results were imported into EndNote, screened, appropriately labelled, and included any additionally relevant notes.⁸ At each point, the rationale for including or not including an article in the next phase of screening was recorded. The titles and abstracts were also printed out and manually and marked accordingly. For data extraction, an Excel file was used to record all data pulled by the extraction tool.

2.4 Data Synthesis

The topic of all measurements for each of the ten articles identified were recorded in an excel file. Most papers reported multiple measurements across multiple indicators (i.e., some included measurements of self-efficacy *and* urban environments, not just one or the other), each of which was recorded. Each measurement was also labelled as quantitative or qualitative (one study prominently included graphs with little explicit numerical data, so these outcomes were listed as “graph”). After all measurements were recorded, the excel sheet was analysed to determine if there were prominent themes/sub-categories across all ten papers. Five sub-categories of teaching preparedness were common among all measurements. These sub-categories were Urban Environments and Diversity), Self-Efficacy, Reflection, Classroom Management and Organisation, and Lesson Planning. These sub-categories form the basis of this study.

The following is an example of the process for categorising the measurements found in the accumulated studies. If three papers, Paper A, Paper B, and Paper C, reported on measurements which they called self-efficacy, teaching effectiveness, and impact on students, they would be grouped into the meta-

⁸ Initially, only 142 of the 143 identified articles were imported. Manual screening determined which article had not been imported, and the article was subsequently imported and evaluated just as the other 142 articles were evaluated.

analysis for self-efficacy because each measured a type of self-efficacy (i.e., these measurements represent different aspects of self-efficacy). Similarly, if Paper A and Paper B included measurements on what they called planning & organisation and classroom management, both measurements would be included in the meta-analysis for classroom management because they both measured a type of classroom management (i.e., these measurements represent different aspects of classroom management). Because there were five sub-categories of teaching preparedness, there was a meta-analysis performed for each of the five sub-categories.⁹ Each reported measurement was of pre-service teachers' competency in that specific classification, except where the measurement explicitly states in-service teachers were the measurement.¹⁰

Once the reported measurements in the ten studies were grouped into categories, quantitative data was analysed with MetaLight.¹¹ This involved aligning each quantitative measure in one of the five specific categories with its corresponding number of participants, mean, and standard deviation. MetaLight then creates aggregate data plots (each category's plot is attached in the "Results" section of this paper) which illustrate the pooled effect of the intervention. This pooled effect can help determine if the intervention had a positive, negative, or no impact on the outcome being studied.

Qualitative data, when provided, were used to augment the quantitative analysis for the data's respective category. Most qualitative data were procured through interviews conducted in the ten studies. No new interviews were conducted for this study. Some studies included measurements which fit into a

⁹ As discussed in more detail in the respective sub-category's meta-analysis section, three sub-categories (Urban Environments and Diversity, Self-Efficacy, and Lesson Planning), included two meta-analyses. This was due to either 1) a possible outlier which was included in one meta-analysis and excluded in the other or 2) assumed levels of standard deviation, which were high in one meta-analysis and low in the other meta-analysis. This is discussed and explained in Section 3.2, "Aggregated Results for Indicators of Teacher Preparedness."

¹⁰ It was thought to be important to include data on in-service teachers, where available, as this may help see the effects of PDS interventions after graduation and continuing into a teaching career

¹¹ This is a tool developed by the EPPI-Centre at University College London's Institute of Education for the purpose of quantitative meta-analysis. For more information, please see the following: <https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3086>

sub-category, but were presented as frequency distributions or graphs. Each measurement of this type was placed in its corresponding sub-category.

The Results section discusses each quantitative meta-analysis's outcomes. Studies that reported number of participants, mean, and standard deviation for a particular measurement were included in the appropriate category of meta-analysis as indicators of teaching preparedness. Through MetaLight, the measurement of Hedge's G was used for the meta-analysis, and a fixed effects model is displayed for each sub-category. Results of the random effects model are also discussed. Any statement that refers to a confidence interval or effect size refers to the pooled estimate of each measurement, except where otherwise explicitly stated. When effect sizes are called small, moderate, or large, the term is based on Cohen's effect size recommendations, which state that 0.2 is small, 0.5 is moderate, and 0.8 is large (Cohen, 1988). Further, I^2 and Q measurements, as determined through MetaLight, were used for tests of heterogeneity.

Some studies included in the meta-analysis did not report standard deviation, but did include number of participants and mean. In such cases, two meta-analyses were conducted under distinct assumptions. The first assumption was that the studies which did not report standard deviation on their measurement had pooled standard deviations equivalent to the highest-reported pooled standard deviation of the studies included in the same meta-analysis. For example, if there were four studies included in a meta-analysis, and the first three reported standard deviations of 0.25, 0.50 and 1.0 while the fourth study did not report a standard deviation, the first meta-analysis would be conducted under the assumption that the fourth study had a standard deviation of 1. Another meta-analysis would be conducted under a second assumption, that the non-reported standard deviation would be equivalent to the lowest pooled standard deviation of the studies in the same meta-analysis. Under this assumption, the fourth study would take a standard deviation value of 0.25. These two meta-analyses (figures labelled High SD or Low SD) were then compared with differences discussed.

Additionally, some studies involved Likert scales in which a lower numbered response was a more positive indicator of teacher preparedness. This was the

opposite of the majority of studies collected. In such cases, the scale was reversed. For example, if a study used a Likert scale of ratings one through five, with one being the most positive response and five being the most negative, the scale was inverted. This allowed for the distance between the most positive response and the reported mean remaining the same. If, in this example, the reported mean was 2.5, the distance between the best response, 1, and the mean, 2.5, is 1.5. When the scale was reversed, the difference between the best response and mean, 1.5, was then subtracted from the adjusted best response, 5. This would result in an adjusted mean of 3.5, which would fit with other Likert scale data in which the higher number is a more positive response.

3. Results

3.1 Study Descriptions and Quality Assessment

The 10 studies which were included in the final analysis are discussed below with their quality assessment. Because of the natural weeding out, per se, of irrelevant studies through the review process, a mark of “high” on Weight of Evidence measurements was common. As mentioned before, detail on each Weight of Evidence measurement and what qualified as “high,” “medium,” or “low,” are included as Appendix 2.

Holbein, Woong, Annis, & Doll (2016)

WoE-A	WoE-B	WoE-C	WoE-D
High	High	High	High

The purpose of this study was to evaluate student competency in urban environments and self-efficacy. Students in the PDS option attended courses at the PDS site with co-instruction delivered by school teachers and university faculty. The PDS programme also included a year-long internship at the PDS site. This involved PDS students co-teaching with school teachers and coaching visits (coaching the pre-service teachers through teaching skills) which occurred in addition to supervision by the mentor teacher. The non-PDS students attended their courses at the university, and these courses were delivered solely by university faculty; they were not co-taught by current school teachers. In addition, no coaching visits occurred.

Three online, Likert scale surveys were given to pre-service teachers and current teachers: The Teacher Attitudes about Students in Urban Schools Survey (AU), the Self-Efficacy for Urban Schools Survey (SEUS), and the Professional Development School Survey (PDS). Independent t-tests were employed to examine the differences between all three groups. All surveys were given during the school year. Since the PDS is, itself, an intervention, this survey can be categorised as being given mid-intervention for pre-service students. Because in-service teacher data was gathered only from those educated at a PDS site (with no comparison group), no in-service teacher data was used in this analysis. Measurements from this paper were included in the

meta-analyses for indicators of Urban Environments and Diversity and Self-Efficacy. With each individual Weight of Evidence measurement being rated as “high,” this study contains clarity of purpose and details, appropriate methodology, and relevant findings.

Castle, Fox, & Fuhrman (2009)

	WoE-A	WoE-B	WoE-C	WoE-D
Qualitative Data	High	High	High	High
Quantitative Data	Low	Low	High	Low

The purpose of this study was to evaluate differences between PDS-based teacher candidates and non-PDS-based teacher candidates across three different cohorts. In addition to focusing on teacher preparation and development, the PDS programme also incorporated research. Students in the PDS programme participated in a year-long internship, during which students undertook coursework simultaneously. University faculty attended the PDS school site once per week to observe, provide professional development (in which PDS students participated), supply research training to develop and study student learning, and conduct seminars. Pre-service teachers taught classes throughout the school (sometimes serving as substitutes) and were not limited to one classroom. Contrarily, the non-PDS programme was solely focused on teacher preparation and included only one 15-week internship. Faculty of the non-PDS programme only attended the school site 4-6 times throughout the 15-week period. In addition to attending all courses at the university, non-PDS students were not substitute teachers and were only limited to one classroom during their internship.

The Student Teaching Evaluation Form was given to all participants in the project, and one-way ANOVA was used to determine any significant differences between groups. Qualitative analysis in the form of end-of-programme audiotapes for portfolios (first and second cohorts) and two written end-of-programme reflections (third cohort) consisting of an autobiographical and critical incident reflection sought to capture reflections about teaching and

looked for emergent themes. Occurring at the end of their programme, these evaluations took place post-intervention. Qualitative data was used for the topics of Reflective Educator and Integrated Curriculum and Practice.

The Weight of Evidence assessments for this study were different than most studies included in the results. Due to the nature of the reported data, I felt it necessary to separate assessments for quantitative and qualitative data. The qualitative data received “high” marks in each category, indicating clarity of purpose, appropriate methodology, and useful findings. However, the quantitative measures received “low” marks due to unclear reporting and methodology. The Student Teacher Evaluation Form was not discussed in any detail, and the reporting was not useful, as few individual indicators were discussed. The reported data stated there were ten specific measurements with significant differences between PDS students and non-PDS students in the first cohort, and then one significant difference each in the second and third cohorts; each difference favoured PDS students. The topic of the significant differences were only mentioned if they were common across multiple cohorts (i.e., the only two differences mentioned by name were the one significant difference found in cohort 2 because it was shared with cohort 1 – but not with cohort 3 – and the one significant difference found in cohort 3 because it was shared with cohort 1 – but not with cohort 2). To reiterate, only these indicators were mentioned (the other 8 significant differences in cohort 1 were never stated by name) and no means or standard deviations were reported for any indicator of the Student Teacher Evaluation Form. Due to the low quality of the quantitative findings, no quantitative measurements were able to be used for the meta-analysis.

McKinney, Haberman, Stafford-Johnson, & Robinson (2008)

WoE-A	WoE-B	WoE-C	WoE-D
High	High	High	High

This study included two groups of pre-service teachers, one PDS and one non-PDS. Each group went through their internship experiences in high-poverty, urban school districts. Students in both the non-PDS programme and PDS programme participated in an initial field experience. This experience involved supervisors who conducted bimonthly observations and weekly seminars (topic choice was up to the seminar leader). The building administrator also

conducted one formal observation. The pre-service teachers slowly moved into teaching the entire class, adding more subjects/responsibilities each week, culminating in a two-week experience where they taught all material and handled all classroom management. The partnership between the university and the school did not go beyond the input/feedback of the university supervisor given to pre-service teachers.

Following this more-traditional field experience, PDS students then completed another field experience at a PDS site. In this portion of the PDS programme, pre-service teachers received training in consistency management and co-operative discipline to promote better student co-operation and self-discipline. Additionally, there were regular seminars and workshops which focused on the demands of teaching in urban schools, co-operating teachers and university faculty worked in conjunction to provide regular feedback to pre-service teachers, and pre-service teachers participated in community experiences (not being limited to the single school site).

The Urban Teacher Selection Interview (developed by Haberman) was used to measure the development of 10 different, specific characteristics of effective urban teachers ((1) Persistence, (2) Values Children’s Learning, (3) Theory to Practice, (4) Work With At-Risk Students, (5) Approach to Children, (6) The Bureaucracy, (7) Admit Mistakes, (8) Teacher Success, (9) Student Success, and (10) Planning/Organisation) prior to and after the internship experience. Aside from descriptive statistics, paired sample t-tests were used and $p > .05$ was set for pre-test and post-test scores to determine any significant difference. Because this study focused specifically on urban education, the only category for which this study provided a measurement was Urban Environments and Diversity. Each individual Weight of Evidence measurement was rated as “high,” indicating this study contains clarity of purpose and details, appropriate methodology, and relevant findings.

Ridley, Hurwitz, Hackett, & Miller (2005)

WoE-A	WoE-B	WoE-C	WoE-D
High	High	High	High

This study, unique among those included in the results, includes two phases: pre-service and in-service; PDS and non-PDS groups existed for each phase of the study. Seven PDS pre-service teachers who participated in the pre-service phase of the study also participated during the in-service phase of this study; this is 50% of the total PDS in-service participant group. No pre-service teacher participants from the non-PDS group participated during the in-service portion of the study.

The PDS programme featured a year-long (three semesters) field experience. In the first semester, students were at the PDS site Monday through Friday, 7:30am until 5:30pm. Pre-service teachers spent full days in the classroom with a co-operating teacher Monday, Wednesday, and Friday, and completed coursework Tuesday and Thursday. During the second semester (summer, in which U.S. compulsory school students are typically on summer break unless attending summer school due to failed classes or the desire to take extra credits) pre-service teachers were put into pairs and taught during a seven-week, half-day summer programme. During this semester they received constant supervision, feedback, and support from university faculty and school teachers. During the afternoons (the half-day portion when they were not teaching) they completed methods coursework. In the third semester, each pre-service teacher co-taught with their assigned co-operating teacher for the first six weeks, after which they took over full classroom and teaching responsibilities until the end of the semester. During this time, the level of feedback and supervision remained the same. The non-PDS programme provided a very different experience. This programme was two academic years (four semesters) in length. During the first three semesters, pre-service teachers spent approximately 5 hours per week at school sites. The fourth semester was a full-time student teaching experience at the school site with additional courses taught at the university campus. The only course not taught at the university campus was a methods course (taught at the school site) which was taken during the third semester of the programme.

In both phases of the study, participants completed a survey for professional teaching knowledge, a written lesson plan to determine reasoning-based application of instructional theory, a video-recorded lesson to determine skill-

based application of instructional theory, and a post-lesson self-reflection, each of which were scored with a point-based rubric. Student-teachers completed these as part of their programme's coursework, and in-service teachers were contacted during the second year of the study and offered a stipend for participation. A general linear model and MANOVA procedure was used for quantitative analysis. Measurements from this study were included in meta-analyses for Self-Efficacy, Reflective Educator, and Lesson Planning, all of which include measurements for both pre-service and in-service teachers. Because each individual Weight of Evidence measurement was rated as "high," this study was determined to contain clarity of purpose and details, appropriate methodology, and useful findings.

McKinney & Finke (2005)

WoE-A	WoE-B	WoE-C	WoE-D
High	High	High	High

This study was conducted to determine the results of student teaching experience on a variety of indicators including commitment to teaching, teaching for real/context based learning, reflectivity, approach to urban/at-risk learners, and self-efficacy. Focused specifically on field experiences in urban environments, the paper discusses the differences between PDS and more-traditional field experiences. For PDS sites, collaboration between the university faculty, school teachers, and pre-service teachers is vital, as university faculty and school teachers provide constant feedback and support to pre-service teachers. Included in practices of feedback is the development of new ideas and approaches to teaching, and addressing the current pre-dispositions and beliefs about urban schools that may be held by the pre-service teacher. On the other hand, non-PDS field experiences do not involve such collaborative and regular feedback as provided at PDS sites. However, each pre-service teacher, regardless of the type of programme, participated in either two seven-week field experiences, or one 16-week field experience (for the study, PDS and non-PDS pre-service teachers were matched based on which field experience they undertook).

Pre-service teacher data was collected through the Student Teaching Experience Survey which contained 103 Likert-style questions. Descriptive

statistics and MANOVA were used to analyse the differences between the PDS and non-PDS group. Measurements from this study were used in the meta-analyses of Urban Environments and Diversity, Self-Efficacy, and Reflective Educator. As is common, with each Weight of Evidence measurement rated as “high,” this study was determined to contain clear purpose and details, appropriate methodology, and relevant findings.

McKinney, Robinson, & Spooner (2004)

WoE-A	WoE-B	WoE-C	WoE-D
High	Medium	High	Medium

Students who participated in this study consisted of those enrolled in a PDS ITE programme and those enrolled in a non-PDS programme. The nature of the PDS and non-PDS programmes discussed in this study is the same as is discussed by McKinney, et al. (2008). Since both studies are conducted by the same author they are likely based on, or at a minimum, related to, the same PDS sites.

Each participant was administered the Urban Teacher Selection Interview before and after their student teaching experience to see the development on 10 effective urban teacher characteristics (these are mentioned when discussing McKinney et al. (2008) as both papers employed the same scored interview). Scores from each interview were analysed with descriptive statistics and ANCOVA. Frequency Distributions were also used in which the number of students achieving a high score, average, and low score on various characteristics was reported. This paper included measurements that were used in the Urban Environments and Diversity section.

Overall, this study was given a Weight of Evidence rating of “high.” However, WoE-B received a rating of “medium” due to the use of frequency distributions, which were not useful for the quantitative meta-analysis, relegating their use to discussion points during the non-quantitative section of the results. It also appears (based on another study conducted by this lead author, McKinney, et al. (2008)) that the reported means may be rounded to the closest integer, but

this was determined to be irrelevant for the WoE ratings, as the data was still useful and understandable.

Sandholtz & Wasserman (2001)

WoE-A	WoE-B	WoE-C	WoE-D
Medium	Medium	High	Medium

Both PDS and non-PDS programmes in this study were fifth-year programmes completed post-graduation from a bachelor’s degree programme. The PDS programme included a placement at a PDS site which lasted one year. In addition, PDS pre-service teachers participated in activities typically reserved for in-service teachers at the school site and attended weekly seminars led by school teachers. University supervisors worked from school sites and their classes were also held at those same sites. Feedback is delivered through collaborative and reflective assessments which includes supervisors, school faculty, and pre-service teachers. Students in the non-PDS programme complete six less weeks of student-teaching. The collaborative feedback, integration into school faculty activities, and coursework conducted at the school site are pieces which are absent from the non-PDS programme.

At the end of the school year student teachers and cooperating teachers were given questionnaires, were interviewed, and were observed in order to study comparisons between a PDS and non-PDS programme. Surveys completed by cooperating teachers included sections on problems, benefits, changes in teaching, evaluation of the programme, and demographic information. Student-teacher surveys included sections on developmental concerns (which consisted of a checklist created to determine a pre-service teacher’s development), teaching experiences, programme evaluation, and demographic information. Interviews were conducted as part of the programmes and provided more information about experience and evaluation of programmes. Observations provided corroborating data. Descriptive statistics and ANOVA were used to evaluate survey responses with significance set at the .05 level. Qualitative data was analysed by coding and annotating the data, seeking corroborative evidence, and identifying themes and patterns. Measurements from this study were used in results for Self-Efficacy, Classroom Management and Organisation, and Lesson Planning.

Overall, this study was given a WoE-D rating of “medium,” due to its “medium” ratings on WoE-A and WoE-B measures. A “medium” rating was given for WoE-A due to a lack of detail as standard deviation was not reported for any study. The usefulness of the study’s measurements was also impacted by the lack of standard deviation, hence the “medium” rating given for WoE-B. Issues with standard deviation and solutions are discussed in each respective results category.

Sharpe, Lounsbery, Golden, & Deibler (1999)

WoE-A	WoE-B	WoE-C	WoE-D
High	Medium	High	High

Two undergraduate cohorts enrolled in Physical Education Teacher Education (PETE) programmes were sorted into two groups: an “involved” group which participated collaborative activities in a PDS environment and a “non-involved” which was not involved in collaborative activities, nor held at a PDS site. The PDS environment included supervision by co-operating teachers and university faculty, in-service teachers teaching undergraduate teacher education coursework, university faculty teaching in the in-service teachers’ gymnasiums, and programme development worked on collaboratively with in-service teachers and university faculty. Research and development projects related to the improvement of public school and university programmes were also conducted. The non-PDS, non-involved group included none of these collaborative activities.

Observations were conducted to study the time spent on instruction, management, interpersonal activities, Academic Learning Time in Physical Education (ALT-PE), off-task, and organisational activities. Open-ended interviews were also conducted at the end of each semester over the 5-year study period to gather participant perspectives on challenges and successes on various activities, professional commitment, ideas for what makes an effective PE teacher, and what was most and least beneficial in their programme. Unique among the included studies, graphs were prominent in this study, displaying trends over time regarding time spent on the above-listed activities. Measurements in these studies were included in analysis for Classroom Management and Organisation.

This study was classified as “high” on an overall WoE-D measurement, despite being rated as “medium” for WoE-B. The “medium” rating for WoE-B was due to the display of graphs, which were useful for trends, but did not provide any numerical data aside from approximate percentages of class time spent on specific activities.

Yerian & Grossman (1997)

WoE-A	WoE-B	WoE-C	WoE-D
High	High	High	High

Two groups (one PDS and one non-PDS) were involved in this study, which took place over three years. PDS students had a core seminar on teaching and learning conducted jointly by university faculty and school staff, with a total of 3-5 instructors who represented areas such as curriculum and instruction, educational psychology, educational leadership, and special education. Each field experience was aligned with this seminar and took place at a PDS site; students progressively attained more classroom responsibility through the year. Supervision and evaluation of PDS pre-service teachers was conducted by in-service teachers who were each responsible for a group of 2-4 pre-service teachers. The programme also focused on middle-level education. The non-PDS programme had four courses, rather than one seminar, which were more disconnected to the field experience. The process of the field experience for non-PDS students was also progressive, and culminated in a 10-week student teaching experience. Rather than consistent evaluation by an in-service teacher, university supervisors observed up to 15 students 2-3 times during the field experience. Moreover, the non-PDS programme was not specific to middle-level education.

Questionnaires consisting of Likert scale and open-ended questions were given to pre-service teachers, who were also interviewed. These measurements were used to study the differences in programme perceptions, personal knowledge perceptions, and beliefs toward students with disabilities. ANOVA was used to analyse the Likert-scale questions, and responses to open-ended questions were grouped according to emerging themes; these emerging themes were used to analyse interview responses. Measurements from this paper were included in analysis for the categories of Urban Environments and Diversity and

Self-Efficacy. Each individual Weight of Evidence measurement was rated as “high,” indicating this study contains clarity of purpose and details, appropriate methodology, and relevant findings.

Mantle-Bromley, Gould, McWhorter, & Whaley (2000)

WoE-A	WoE-B	WoE-C	WoE-D
High	High	High	High

This paper studied three different programmes: A PDS-based programme, a more traditional ITE programme, and a 10-month graduate programme meant for mid-career professionals who wanted to transition into the teaching profession. Measurements that were included in the meta-analysis portion were comparisons between PDS and the more traditional programme, only. The PDS programme was a one semester experience at the end of ITE. Not only did pre-service teachers attend the school site for their 15-week field experience, but also for their general methods course and a seminar. Twice weekly, they attended a university course co-taught by university faculty and high school staff, worked with their mentor teacher in a classroom for 90 minutes, and then attended a noon-hour seminar on various topics of mutual interests to the pre-service teachers, in-service teachers, and university faculty. Further, they taught a minimum of three university-supervised lessons, observed/discussed lessons conducted by their peers, performed managerial duties, and participated in school-wide activities outside of the classroom. The non-PDS group also completed a 15-week field experience, had at least 100 hours of pre-student-teaching experiences, and engaged in a micro-teaching lab where they taught lessons with their peers as the students. However, courses and seminars taken at the school site, a 90-minute block with which to work with their mentor teacher, and participation in school-wide activities were not part of the non-PDS programme.

An overall ANOVA for differences between all three programmes was discussed in the Lesson Planning category analysis. A survey of 36 questions was designed to find information on job searches, employment status, future plans, factors in attrition, and programme satisfaction. Descriptive statistics were used along with chi-square analysis and ANOVA to test significant differences between programmes. Likert scales were used for programme

satisfaction as it related to preparation for classroom management, working with diversity, lesson planning, technology use, and teaching strategies. Data from this paper were used in the meta-analysis for Urban Environments and Diversity, Classroom Management and Organisation, and Lesson Planning (the overall ANOVA between all three programmes discussed in the Lesson Planning section). The overall Weight of Evidence was rated as “high,” indicating this study contains clarity of purpose and details, appropriate methodology, and relevant findings.

Because of the high marks of most of the articles, we can be confident in the validity of the results shown. Where any article caused a hesitation in confidently interpreting the displayed results, the reasons were discussed. Two studies, McKinney, et al. (2004) and Sandholtz & Wasserman (2001), were discussed at certain points due to issues with data reporting.

3.2 Aggregated Results for Indicators of Teacher Preparedness

Before showing the results of this study, clarification should be made regarding the studies included in each meta-analysis. Multiple papers included in this assessment sometimes reported on multiple measures of teacher preparedness. For instance, Holbein, et al. (2016) reported on a measure of motivation to teach in urban environments, as well as a measure of self-efficacy for pre-service teachers. Because these separate measurements are each relevant to a different meta-analysis category, Holbein, et al. (2016) is listed in more than one meta-analysis. Specifically, the measure of motivation to teach in urban environments is included in the Urban Environments and Diversity meta-analysis, while the measure of self-efficacy for pre-service teachers is included in the Self-Efficacy meta-analysis. This also explains why the same cited paper may display different effect sizes in different meta-analysis categories.

In cases where one category had multiple meta-analyses, this was due to a possible outlier in the Urban Environments and Diversity category (where the possible outlier was included in the first meta-analysis, and excluded in the second meta-analysis; this way, effect sizes with and without the outlier could be seen) or assumed standard deviations in the Self-Efficacy and Lesson

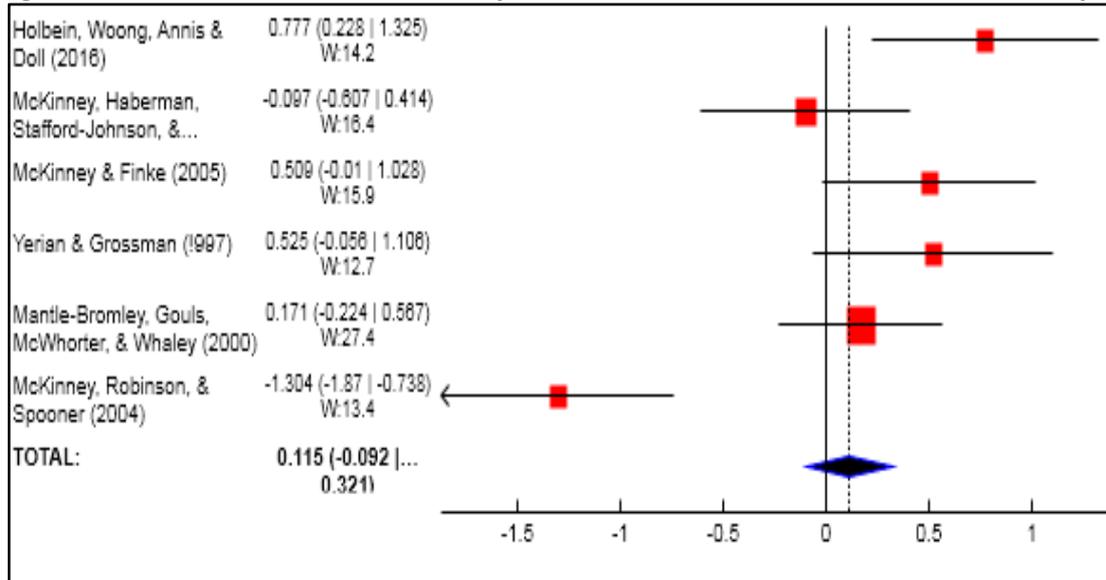
Planning categories (where one meta-analysis included higher assumed standard deviations, one meta-analysis included lower assumed standard deviations). This process was discussed in the Data Synthesis section. In both cases, the unaffected studies in each meta-analysis are the exact same. For example, a measurement by Yerian & Grossman (1997) appears in the Urban Environments and Diversity meta-analysis. Because the only difference between the two meta-analyses in this category is the inclusion or exclusion of the possible outlier (McKinney, et al. (2004)), each other study remained the same in both meta-analyses for this category. This was done so the effects of the possible outlier (or different assumed standard deviations in the other sub-categories) could be determined with everything else remaining the same.

Urban Environments and Diversity

Measurements included in this meta-analysis relate to a pre-service teacher's ability to effectively teach in urban and diverse environments. Quantified indicators such as attitude and approach to students in these environments are contained in the analysis of these six studies. Additionally, two studies, McKinney, et al. (2004) and McKinney, et al. (2008), included in this analysis employed a scored interview intended to measure ability in urban environments. One qualitative piece will be used to analyse the intervention in this category as well. Figures 3.1 and 3.2 include the same studies.

It should be noted that the study conducted by McKinney, Robinson, and Spooner (2004), reported a standard deviation of -1. After further review, a typing mistake was determined to be the most reasonable conclusion, and the standard deviation was changed to 1 and placed into the MetaLight meta-analysis tool as such.

Figure 3.1: Fixed Effects Meta-Analysis for Urban Environments and Diversity¹²

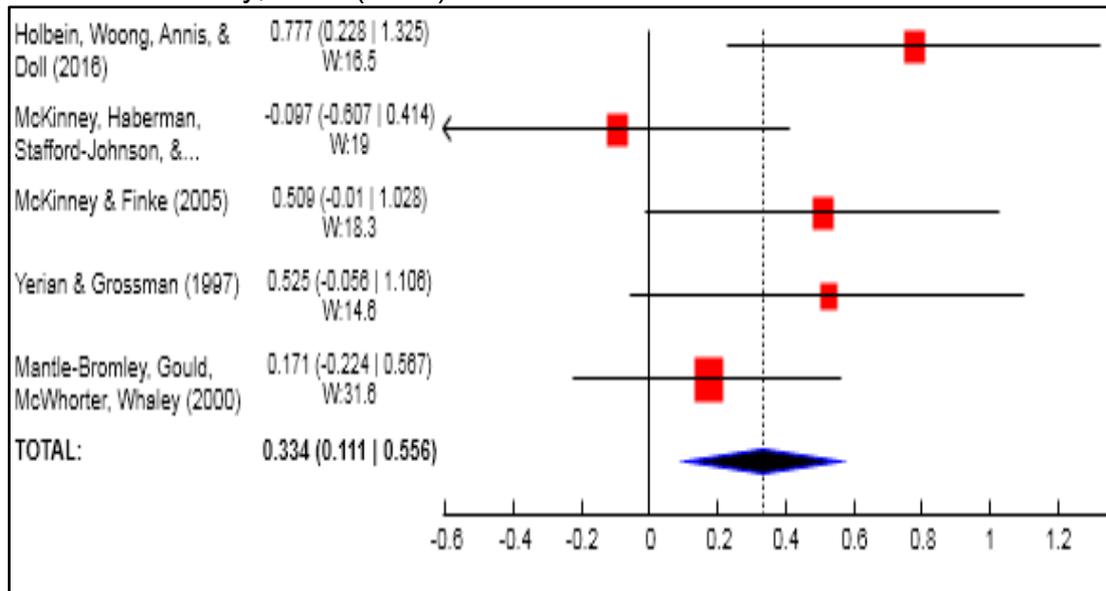


As displayed in Figure 3.1, the fixed effect model demonstrated a very small effect size of 0.115 favouring PDS ITE. When a random effects model was employed, an effect size of 0.099 was displayed, favouring the PDS intervention. However, both pooled effect sizes had a confidence interval which crossed the null effect line, so it cannot be confidently determined that the intervention caused a positive result when all data is aggregated. Further, the I^2 and Q measures of heterogeneity are 85.6% and 34.6, respectively. While these are both high measures of heterogeneity, the final study (McKinney, et al. (2004)), appearing to be an outlier, provides a possible explanation for this result. This study is the only one with a confidence interval that lay completely to the left side, always favouring the control group. The effect size of this study is also much more to the left than any other study, providing further indication it may be an outlier. When this study is taken out,¹³ the pooled estimate effect size changes in important ways, as displayed in Figure 3.2.

¹² In order from top to bottom, each measure in Figures 3.1 indicates (as named by the original study) motivation to teach in urban environments, a scored interview (Urban Teacher Selection Interview) that determines one's ability in a specifically urban environment, how well pre-service teachers approach urban/at-risk students, pre-service teachers' confidence in working with diversity, pre-service teachers' satisfaction with preparation for diverse settings, and the same scored interview used earlier (Urban Teacher Selection Interview).

¹³ The studies displayed in Figure 3.2 do not include the final study listed in Figure 3.1, but otherwise are the exact same measurements presented in the exact same order that appear in Figure 3.1.

Figure 3.2: Fixed Effects Meta-Analysis for Urban Environments and Diversity without McKinney, et al. (2004)



When that study is taken out the pooled effect size has increased to 0.334 for the fixed effects model. The random effects model also saw an increase in effect size, to 0.353, while both confidence intervals now lay completely to the right of the null effect line. These new confidence intervals indicate a positive intervention effect with 95% certainty. It ought to be noted that, despite the increased effect size, they still may be categorised as small effect sizes. We also now have a lower I^2 value of 40.6%, while Q takes on a value of 6.73. These levels are much lower than their initial measurements, and it appears the study by McKinney, et al. (2004) may be largely responsible for the original, high heterogeneity values.

In addition to the meta-analysis, McKinney, et al. (2004) reports on another measurement, ability to “work with at-risk students,” which is relevant to this sub-category and presented in the form of frequency distributions. The frequency distributions are reported as the number of students who attained a high, average, or low score in the pre-test and post-test. For the specific measurement ability to “work with at-risk students,” three PDS students achieved a high score, 13 students achieved an average score, and 13 more students attained a low score. On the post-test (conducted after field experiences), only two students scored highly, five achieved an average score, and 22 students attained a low score. This is compared with the pre-test for the

non-PDS group in which two students achieved a high score, six achieved an average score, and 22 achieved a low score. The non-PDS post-test displayed two students achieving a high score, two students achieving an average score, while 26 students attained a low score.

These results showed a downward trend for both PDS ITE students and non-PDS ITE students. Interestingly, this means students actually decreased in ability to work with at-risk students after more exposure to urban environments during clinical field experiences. However, because these measures are also from McKinney, et al. (2004) (which appeared to be an outlier), and included oddities in reported data such as a negative standard deviation and numerical data which appeared to be rounded to the closest integer (rather than using decimal points), perhaps we ought to be careful not to rely too much when drawing conclusions from this data.

On the other hand, there is the possibility that this study provides a different insight – the possibility of self-perception bias. If true, this means that, on self-reported surveys/questionnaires, pre-service teachers may have perceived (and rated) themselves better than their performance warrants. This possibility is indicated by the two studies which showed negative effect sizes in this meta-analysis, McKinney, et al. (2008) and McKinney, et al. (2004). Both of these studies used an interview in which university-based interviewers scored student responses. The other studies in this meta-analysis, which all displayed a positive effect size, used student self-reporting questionnaires/surveys to gather their data. If PDS students were aware of the claimed benefits of PDS sites, they may have felt they were better prepared simply because they attended a PDS ITE programme. As a result, they may have perceived themselves to be better prepared than their non-PDS counterparts, even if there was little to no difference. The possibility of self-perception could be a problem with much larger implications not limited to this meta-analysis.

McKinney, Robinson, and Spooner (2004) appear to hold reservations about their results and say while a link could not be established between the intervention and competence growth when working in urban school environments, it does not mean PDS interventions provide no benefits; in fact, they note how other studies attest to benefits of PDS sites. However, they

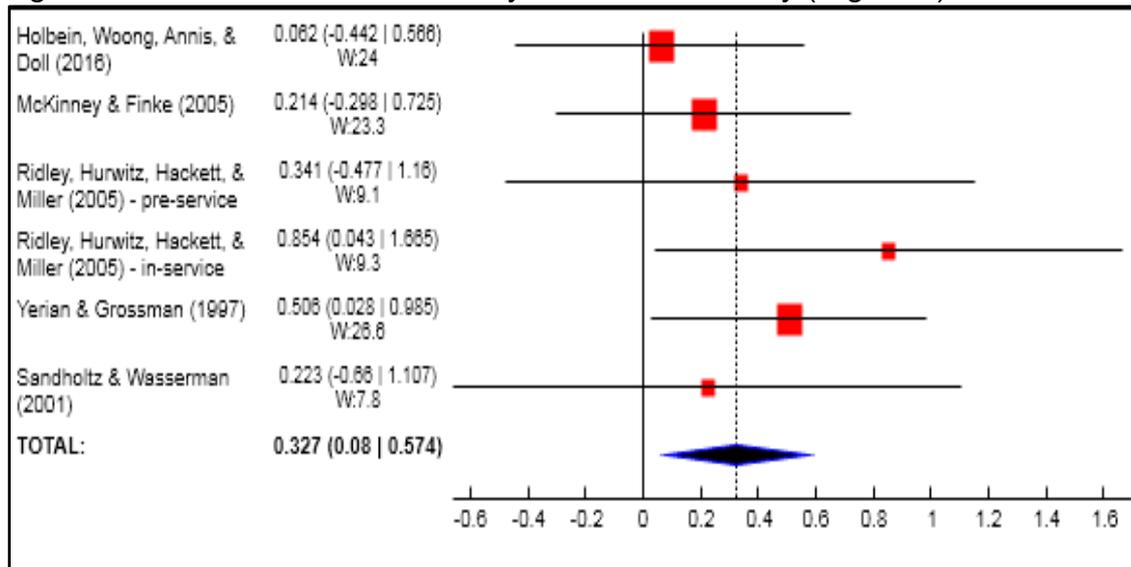
specifically question how teacher education prepares students for urban environments, stating problems such as the length of exposure during ITE (which may not be long enough) and the possibility that pre-dispositions gathered through one's life may affect teaching ability in urban environments. They propose that schools of education should try to attract more individuals with pre-dispositions which would allow them to be successful in urban environments.

Self-Efficacy

Studies included under this category include quantified measures of self-efficacy and belief in affecting student achievement. These measures were determined by surveys and questionnaires given to pre-service teachers. One report Ridley, et al. (2005) measures the effectiveness of first-year in-service teachers, with one group which was prepared at a PDS site, and one group which was prepared at a more traditional site.

Before moving on it should be noted that two studies, Sandholtz & Wasserman (2001) and Yerian & Grossman (1997) had Likert scale measurements by which a lower mean indicated a more positive result, opposite of every other measurement in this meta-analysis. In this case, the means on the studies were transformed in accordance with the process and explanation in section 2.4. Additionally, the measurement by Sandholtz & Wasserman (2001) did not include a standard deviation. This was solved by performing two different meta-analysis in this category: the first meta-analysis (Figure 3.3) used the assumption that the missing standard deviations would equal the highest reported standard deviations in the other studies, and the second meta-analysis (Figures 3.4) used the assumption the missing standard deviations would equal the lowest reported standard deviations of the other studies. The reasoning for such assumptions is also stated in section 2.4. Figures 3.3 and 3.4 use the same measurements.

Figure 3.3: Fixed Effects Meta-Analysis for Self-Efficacy (High SD)¹⁴



As can be seen by Figure 3.3, the fixed effects meta-analysis shows a total effect of 0.327 with a confidence interval that lay completely to the right of the line of null effect. When viewing a random-effects model, the same effect size and confidence interval were shown. In this case, with 95% confidence, we have a small, positive effect size favouring PDS ITE. Interestingly, measures for heterogeneity were noticeably low, with $Q = 3.47$ and $I^2 = 0\%$, indicating very little heterogeneity between studies.

To reiterate, this meta-analysis includes a study which did not report standard deviation, and we assumed the standard deviation for this report was equal to the highest reported standard deviation of the group. The following figure (3.4) represents the assumption that the non-reported standard deviation is equal to the lowest standard deviation reported in the other studies.

¹⁴ In order from top to bottom, each measure in Figure 3.3 is a measure of (as named by the original paper) self-efficacy, self-efficacy, belief in teaching effectiveness for pre-service teachers, belief in teaching effectiveness for in-service teachers, confidence to affect student achievement, confidence in self to be an effective educator. The studies presented in Figures 3.3 and 3.4 are the same and appear in the same order.

Figure 3.4: Fixed Effects Meta-Analysis for Self-Efficacy (Low SD)

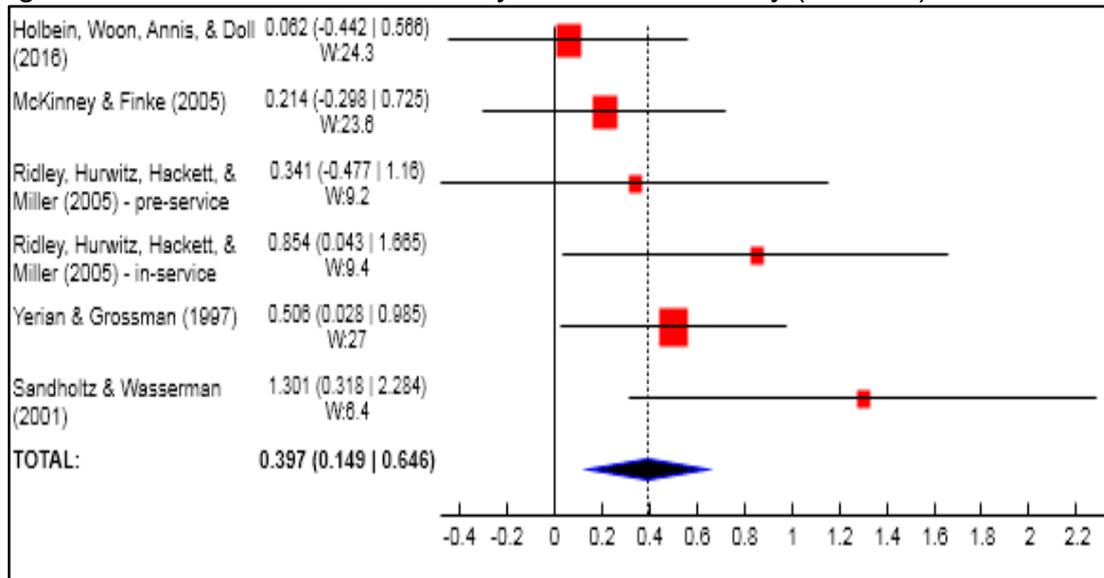


Figure 3.4 shows a more positive effect (compared with Figure 3.3) for PDS ITE on self-efficacy, with a total effect size of 0.397. For the random effects model, this effect size was even larger at 0.426. Both had confidence intervals which lay to the right of the null effect line, indicating a statistically significant result. These larger effect sizes indicate an approach toward an overall more moderate, positive effect. This result, plus the shift to the right for the confidence intervals, are due to the assumed lower standard deviation that was used for the measurement taken from Sandholtz & Wasserman (2001), as there were no other changes between the first and second meta-analysis for this category.

However, measures for heterogeneity have increased when the lower standard deviation value was assumed. Q now equals 6.88 and I^2 stands at 27.3%. While neither value should cause for serious alarm with regard to study heterogeneity, it is interesting that a change in one study's (assumed) standard deviation from high to low affected these levels to this magnitude.

This meta-analysis included six studies, the most in any one meta-analysis (the first meta-analysis for Urban Environments and Diversity also included six studies), of which none reported a negative effect. Again, of the final 10 papers included in the final results and data analysis, many reported on multiple measures within one paper. Any studies which were used here and in the Urban Environments and Diversity sub-category (or used in any combination of

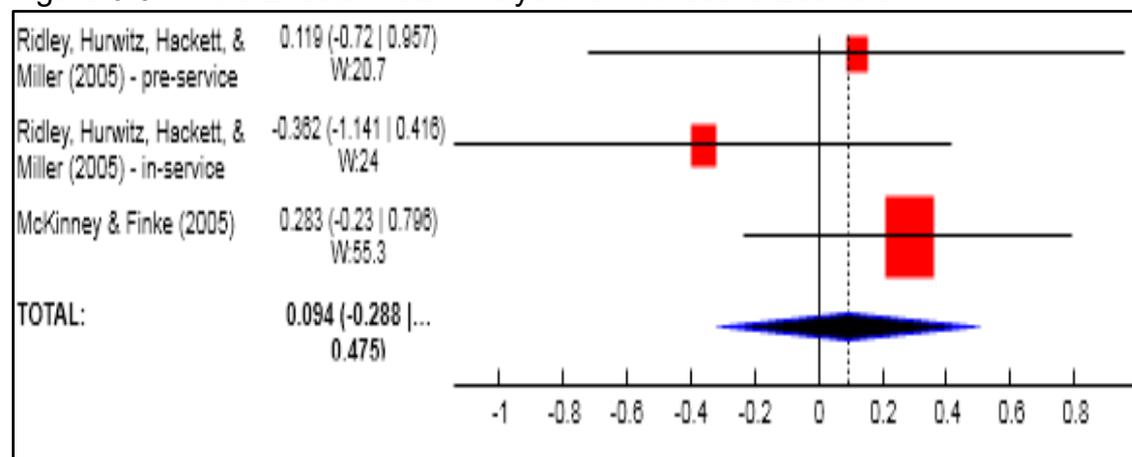
multiple sub-categories) reported on different measurements that were relevant to each sub-category, as previously explained.

While the significance of the meta-analysis effect size for Urban Environments and Diversity was dependent on the possibility that one study may be an outlier, the effect sizes for Self-Efficacy were positive and significant regardless of the different assumptions in each meta-analysis. These quantitative results provide us with a confident conclusion that the PDS ITE had a positive effect on students' self-efficacy, possibly due to the longer exposure to a classroom setting provided by PDS models of ITE.

Reflective Educator

The studies in this meta-analysis category include both pre-service and first-year in-service teachers. Measurements are centred around an ability to reflect on classroom practice as part of a continual learning process for improvement. There are three quantitative studies and one qualitative study for this section. Figure 3.5 reflects the quantitative meta-analysis.

Figure 3.5: Fixed Effects Meta-Analysis for Reflective Educator¹⁵



Both the fixed effects and random effects meta-analysis provide the same, very small, effect size of 0.094, and a confidence interval which crosses the line of null effect. In this case, we cannot be certain that the aggregate effect of PDS

¹⁵ In respective order from top to bottom, the studies in Figure 3.5 are measures (as named by their original papers) of post-lesson reflection for pre-service teachers, post-lesson reflection for in-service teachers, and reflective practice on the part of a pre-service teacher.

intervention produces a positive effect, even though the mean effect size is positive. With $Q = 1.84$ and $I^2 = 0\%$, heterogeneity issues are not a concern.

Interestingly, the one negative result of the meta-analysis was of in-service teachers. This could possibly indicate that reflection is a skill which can be learnt within a requisite amount of time spent teaching. Because PDS pre-service teachers have more exposure to classrooms than non-PDS pre-service teachers, this could explain why PDS pre-service teachers appear to be better at reflective practice. However, as non-PDS in-service teachers reach that requisite exposure (which PDS students may have reached during their clinical field experience), they may catch up (and move ahead, in this case) to effectively practicing reflection. Of course, this would require that once the requisite amount of exposure is reached, there is little benefit of more exposure as it relates to reflection (otherwise, PDS students would always be ahead of non-PDS students because their total exposure time would be greater).

While, by itself, this meta-analysis may not be sufficient because there are only three studies, there is a qualitative measure that may be able to provide more insight into the intervention effect on reflection. In their study, Castle, Fox, & Fuhrman (2009) also evaluated a theme of “Reflections connected to practice” versus “Reflections not connected to practice” in audiotapes and portfolios of PDS and non-PDS programme graduates. Qualitative measures were coded as PDS or non-PDS, and audiotapes were transcribed by graduate research assistants with no connection to the ITE programmes. It is possible, because each portfolio and audiotape was coded as PDS or non-PDS, that those who graded and took note of emergent themes were aware of the group to which a student belonged. In each of the three groups evaluated, a larger proportion of PDS students were labelled as having their reflections and practice connected than were labelled otherwise. The opposite is true for non-PDS ITE students, a majority of which were labelled as not having connections between their reflections and practice in each study. Overall, more PDS ITE students appeared to include deep and integrated discussion within their course reflections with regard to connecting their reflections of class time to their classroom practice.

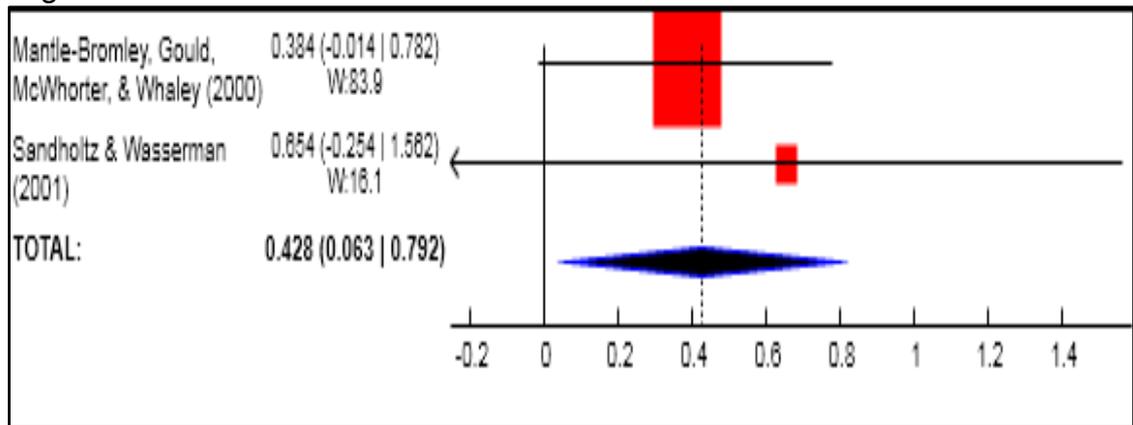
Added to our quantitative analysis is this evaluation that appears to say students trained at PDS-sites are better at reflection and connecting it to their teaching practice, compared to non-PDS trained students. It is also important to note that, because the one negative result in the meta-analysis was of in-service teachers, measurements of pre-service teachers all displayed positive effects. While the overall results were insignificant, we may be tentatively confident that pre-service teachers gain positive benefits from a PDS intervention (and, ergo, possibly better preparation), although the differences may subside during in-service teaching.

Classroom Management and Organization

This meta-analysis consisted of measurements designed to capture classroom management, the effective handling of student problems, and the effective handling of student behaviour. There were three quantitative studies and two qualitative studies used in this section. Figure 3.6 represents the meta-analysis for the quantitative studies.

The measurement taken from Sandholtz & Wasserman (2001) for this meta-analysis used a Likert scale in which a lower number indicated a more positive response. The results were adjusted according to the process outlined in section 2.4. The same study also did not report any standard deviation for their measurements. Because there was only one other study in this meta-analysis, the assumed standard deviation for the Sandholtz & Wasserman (2001) measurement was equal to that of Mantle-Bromley et al. (2000).

Figure 3.6: Fixed Effects Meta-Analysis for Classroom Management and Organization¹⁶



As displayed in Figure 3.6, the pooled effect size stood at 0.428, a small-approaching-moderate effect. The random-effect meta-analysis produced the same effect size. Furthermore, in both cases the confidence interval lay completely to the right of the line of null effect, allowing us to state with 95% confidence that there is a positive effect due to the PDS intervention. Measures of heterogeneity I^2 and Q are 0% and 0.28, respectively, showing low levels of heterogeneity. These overall effect sizes and heterogeneity measurements are most likely affected by the inclusion of only two studies in this meta-analysis. While positive effects on the intervention's results are shown, such a limited number of studies can bring concerns about properly interpreting the effect size, and whether one study exerts an non-proportional amount of influence on the effect size, causing the results to be skewed

As with the results for Reflective Educator, a qualitative study may be able to help provide more insight. For this category, a study by Sharpe, Lounsbery, Golden, and Deibler (1999) can provide more insight. While studying physical education ITE, evaluators observed pre-service teachers five times each while practice teaching (experience prior to a clinical field experience) and student teaching (teaching during a clinical field experience). Also observed five times during a semester were in-service teachers. The observers for each programme went through the same workshops to ensure familiarity and agreement on all evaluation protocol. Only observers who had been involved with the PDS site

¹⁶ In respective order from top to bottom, the studies in Figure 3.6 are measures (as named in their original papers) of effective classroom management and effectively handling student behavioural issues.

group observed the PDS group, and only observers who were not involved with the PDS group observed the non-PDS group. No one who worked with the PDS-based group observed the non-PDS group, and vice-versa. Data was reported as a percentage of class time spent on certain activities over these five observations as displayed on a graph, but no further quantitative analysis was done. Since such data could not be included in the MetaLight meta-analysis, it makes most sense to report the results as a discussion of trends.

Over the course of the observations, the time spent on classroom management showed a steady decrease for PDS pre-service teachers, and was at a relatively low level (less than 10% of class time) for PDS-trained in-service teachers. For PDS pre-service teachers, similar decreasing trends were found for class time spent being off-task, and class time spent on organisation. For both of these measures, PDS-trained in-service teachers also decreased the time spent on these tasks, using approximately 5% of class time on each of these tasks by the final observation. Further observations showed that the time spent on class instruction and student learning trended upward over the course of the semester for PDS-based pre-service teachers and PDS-trained in-service teachers (besides the occasional decreasing dip). Instruction increased by nearly 20% across all groups (practice teaching, student teaching, and in-service teaching) and time spent by student's learning increased similarly by the end of the fifth observation.

On the other hand, non-involved pre-service teachers and teachers who were not PDS-trained showed almost no trends, and therefore no improvement, in the amount of time spent on classroom management. These percentages appear to hover between 15% and 35%, a much larger proportion of class time than PDS-based students and teachers. Similar issues were seen when observing time spent on organisation and being off-task. For those non-involved and not PDS-trained, the percentages of time spent on organisation and being off-task did not decrease, and in one case even increased by 5%, between the first and fifth observation. Additionally, time spent on instruction by non-PDS in-service and pre-service teachers showed no consistent increasing trend, and actually decreased by approximately 10% for in-service teachers not trained at a PDS site. With regards to student learning, there was no consistent increasing

trend, but rather a consistent level showing no improvement over the course of the observations. The observations in this study demonstrate a positive effect for PDS-trained teachers and current PDS ITE students. Time spent on organisation, management and being off-task decreased for PDS students and teachers. No decrease, and sometimes an increase, was seen for those who were not involved in PDS ITE.

While the meta-analysis was comprised of only two studies, the additional study included for augmentation suggested positive effects of the PDS intervention. More studies would be able to provide a more robust analysis, but, because all studies included show a positive effect (including data that strongly supports PDS in Sharpe, et al. (1999)), we may be cautiously confident that PDS ITE may provide some positive benefits of classroom management and organisation to pre-service teachers.

Lesson Planning

Included in the meta-analysis for lesson planning are measures of effective lesson plans, teaching strategies/methods, and instructional materials. There are four quantitative studies, one qualitative study, and a discussion of quantitative data which was not able to be displayed in MetaLight due to reasons discussed below. Figures 3.7 and 3.8 all include the same studies. As in previous meta-analyses, Ridely, et al. (2005) includes a measurement for in-service teachers, which is the second measurement listed from that paper.

Again, a measurement from Sandholtz & Wasserman (2001) in this meta-analysis used a Likert scale in which a lower number indicated a more positive response and did not report standard deviation. Adjustments were made based on the processes outlined earlier, and two meta-analyses were conducted – one assuming the highest standard deviation based on the other studies, and the second assuming the lowest standard deviation based on the other studies.

Figure 3.7: Fixed Effects Meta-Analysis for Lesson Planning (High SD)¹⁷

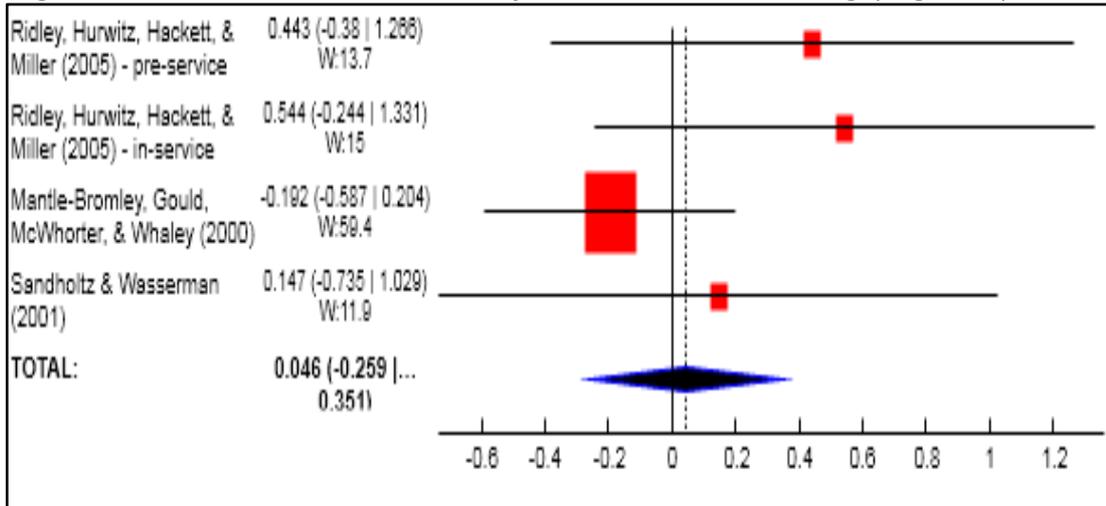
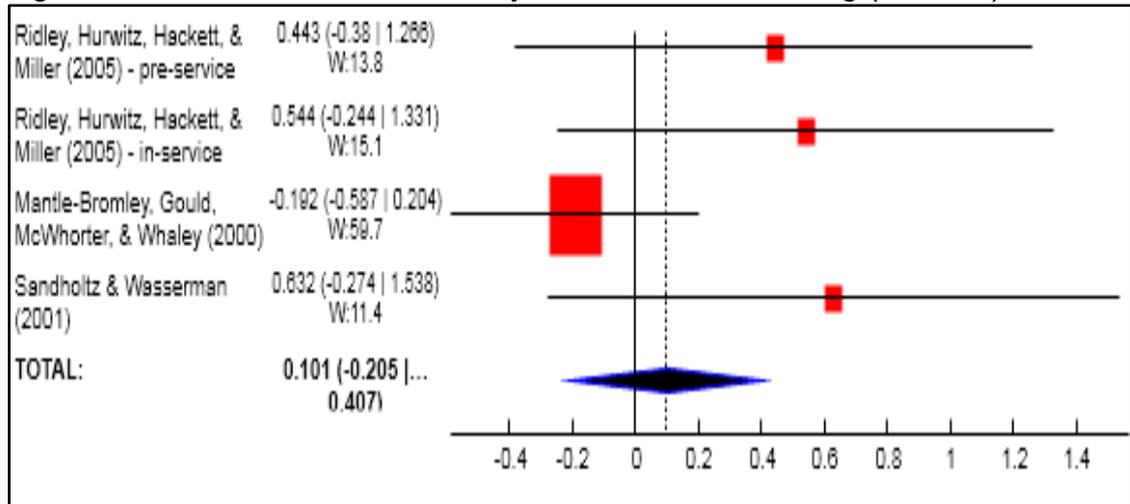


Figure 3.7 displays a very small, but positive, pooled effect size of .046 for the intervention. The random effects model showed a slightly larger pooled effect size of 0.109. However, the pooled confidence interval in both the fixed and random-effect meta-analyses crosses the line of null effect, indicating we cannot be sure there is a positive effect, despite the positive nature of the pooled effect size. Regarding heterogeneity, $Q = 3.87$ and $I^2 = 22.4\%$, both of which indicate relatively low heterogeneity.

Figure 3.8: Fixed Effects Meta-Analysis for Lesson Planning (Low SD)



¹⁷ In respective order from top to bottom, the studies included in Figure 3.7 are measures (as named in their original papers) of effective lesson planning of pre-service, effective lesson planning of in-service teachers, quality of teaching strategies, and confidence in using a variety of instructional materials. These same measurements are presented in the same order in Figure 3.8.

As we can see in Figure 3.8 above, the lower assumed standard deviation did not change the effect sizes or confidence intervals in any significant way. While the displayed fixed effects model shows an effect size of 0.101, the random effects model effect size increased to 0.241. Both confidence intervals still cross the null effect line. Differences in our assumed standard deviation for the Sandholtz & Wasserman (2001) measurement did not significantly alter the results. As with the Self-Efficacy meta-analysis, heterogeneity levels increased with a lower standard deviation, with Q now equal to 5.3, and I^2 now equal to 43.4%. While these are not high levels of heterogeneity, they did increase with a lower assumed standard deviation.

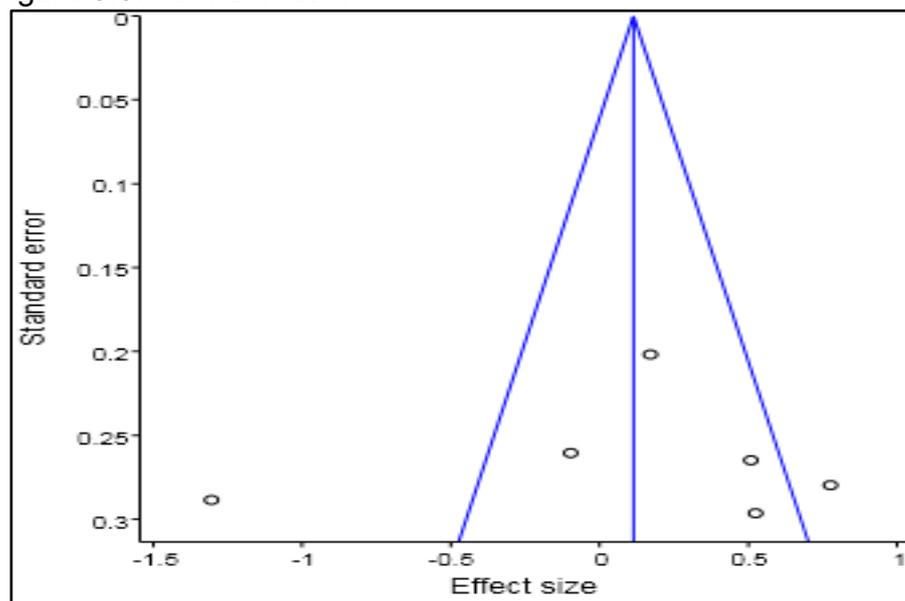
While Mantle-Bromley et al. (2000) is included in the meta-analysis, another measurement of lesson planning in that study is of interest, but was not able to be included in the meta-analysis because it was not broken down by programme route. Rather, the measurement was a one-way ANOVA which was conducted to evaluate differences in programme satisfaction regarding training in lesson planning. The ANOVA displayed a measurement of difference between three groups: a PDS ITE programme, a more traditional non-PDS ITE programme, and a third, specialised programme intended for mid-career professionals who wanted to become teachers. This result is of interest to us because there was no significant difference found between any of the programmes for satisfaction in relation to preparation of lesson planning. Each programme's students felt satisfied with the preparation they received in terms of lesson planning.

While one qualitative study and three quantitative studies included in the meta-analysis demonstrated positive effects (including a more positive effect for in-service teachers), this indicator displays insignificant and inconclusive results overall for both the quantitative and qualitative studies. More studies may be able to better determine what the actual effect of PDS interventions are on this indicator (assuming there is an effect), but the measurements included in this indicator are not persuasive either way.

3.3 Funnel Plots

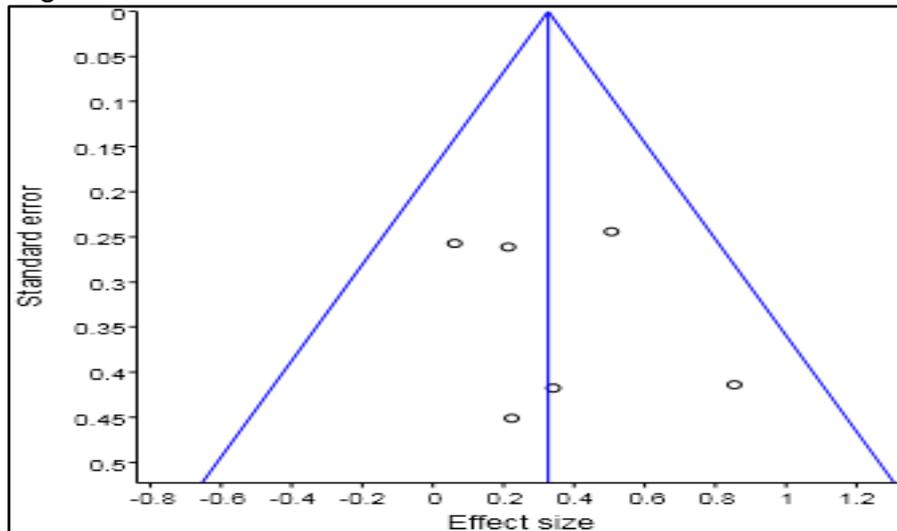
Below, two funnel plots are displayed as Figures 3.9 and 3.10. Since funnel plots are used to determine publication bias, there is no need to display a funnel plot for every category; as long as each study appears in a funnel plot, we can interpret the results as they relate to publication bias. Because there was not a single meta-analysis which included every study, two funnel plots of different meta-analyses – the funnel plots for Urban Environments and Diversity (Figure 3.9, Funnel Plot A) and Self-Efficacy (Figure 3.10, Funnel Plot B) – are displayed. These two funnel plots were chosen because each study included throughout the various meta-analyses is included within one or both of these studies.

Figure 3.9: Funnel Plot A



In Figure 3.9, we see the studies do not appear to be distributed equally on each side of the funnel. Four out of six studies fall on the right side of the funnel, and are more closely grouped together than the two studies on the left side. The farthest study is the previously mentioned McKinney, et al. (2004) which was identified as a possible outlier in the Urban Environment and Diversity analysis.

Figure 3.10: Funnel Plot B



In Figure 3.10, the studies appear to be more closely grouped and evenly distributed on each side. One could argue that because the bottom left portion of the funnel plot (negative effect size and higher standard error) is empty, this funnel plot is not evenly distributed. While true, these studies are more closely grouped when compared to Figure 3.9.

Taking both funnel plots into account, there may be slight bias in the selected studies, as indicated by the uneven distribution of studies. This may possibly be due to the simplified search string (discussed earlier) and a filter for “scholarly journals” (discussed earlier) during the search process. Such a search string and a filter may reduce the number of articles returned, and cause this slight bias that appears possible in the articles gathered. The results of this funnel plot should be kept in mind when interpreting and discussing these results.

4. Discussion

Overall, the various meta-analyses had mixed results. Although each meta-analysis had a positive pooled effect size, some had confidence intervals which crossed the line of null effect, allowing no confident conclusion to be made regarding the effect size. Out of these insignificant results (the first meta-analysis for Urban Environment and Diversity, the meta-analysis for Reflective Educator, and both meta-analyses for Lesson Planning), all models provided a pooled effect size below Cohen's threshold for small effect (0.2) except one – the random effects model for Lesson Planning when the assumed standard deviation was low turned out to be 0.241.

We should also keep in mind a point brought up when discussing the possible reasons why McKinney, et al. (2004) displayed very different results compared to other studies: the issue of self-perception. Each study used in these quantitative meta-analyses is one of self-perception (with the exceptions of McKinney, et al. (2004) and McKinney, et al. (2008)). While few would argue that teachers should not feel confident in their ability to be effective across these various indicators, it is possible that pre-service teachers may believe they are more effective than they are. However, because we do not have measurements of how effective these pre-service teachers were once they became in-service classroom teachers, there is no way to be certain whether this bias exists or not. Although the effect of PDS ITE on students' confidence in their ability to perform a variety of functions seems to be generally more positive than that of more traditional programme students, many results are insignificant nonetheless.

When looking at each meta-analysis together, it seems that there is no definitive answer as to the effects of PDS intervention across-the-board. While acknowledging the time-constrained nature of this paper, this evidence assessment points to some possible discrepancies between what PDSs are intended to accomplish versus what they have accomplished. It calls into question whether the principles enshrined by The Holmes Group (2007) and the benefits espoused by many education researchers (such as Darling-Hammond (1994, 1996, 2007), Kochan (1999), and Abdal-Haqq (1991)) are supported by

empirical evidence. It is also possible that this discrepancy is due to incomplete integration of the PDS programme. If an ITE programme is not fully integrated with a PDS site, one could argue it does not properly constitute the ideal of a PDS ITE programme as envisioned by The Holmes Group (2007). However, to say this research did not provide any important information would be untrue. Some measures of PDS interventions showed positive and significant effects, such as Self-Efficacy, Classroom Management and Organisation, and the second meta-analysis for Urban Environments and Diversity. In areas where the meta-analysis was overall insignificant, other studies provided insight which favoured PDS interventions such as the practice of reflection and classroom management. With this research in hand, the question now turns to how this can inform reforms for both education research and ITE programmes.

4.1 Limitations of this Study

Before moving onto the implications of this research, we should acknowledge the limitations of this study. The time-constraint of this study only allowed for a shortened systematic review – a systematic rapid evidence assessment. Because of this, I do not purport this review to have accumulated all of the available data on this topic and indicators. It may be the case that much more research can be gathered, and a longer-term systematic review may be conducted.

As discussed previously, it is also possible there is self-perception bias through some of these studies. Without more data to confirm or rule this out, we should be careful about interpreting these results as definitive measurements on ability; it is possible they may represent something more like confidence in performing a specific type of task.

Because most data used in this study is from a group of ten papers, it may be expected that the results would be somewhat homogenous across categories. Similar to the limitations due to the time constraint, it is possible there is important data not included in this study which could shed more light on the effects of PDS interventions. As stated before, the search string, although providing a limited number of studies, provided outstanding specificity. Although

some data may have been missed due the limited number of articles, the simplified search string was viewed as a strength due to the time constraint.

In addition, while this paper will make recommendations based upon the findings of this study, these recommendations are not meant to be exhaustive. For instance, recommendations regarding standardised processes for ITE will be mentioned, but I do not advocate in this paper for any specific type of standardised assessment, whether test, exam, or otherwise.

Additionally, this study chose to focus on ITE programmes which described themselves as PDSs by having PDS sites for field experiences. It is possible that some PDSs studied were not yet fully integrated as ITE programmes, and were at a developmental stage. This study did not break results down by any differentiation in stages of PDS development. It is always possible that different results and interpretations could have been provided had results been broken down by stage of development. However, citing the previously mentioned constraints, it was determined this would not be done.

4.2 Implications for Research

Overall, it is common that both ITE programmes and education researchers use self-reported surveys to gauge programme effectiveness. However, researchers and ITE programmes should be cautious when interpreting results from such surveys, and not treat them as fool-proof indicators of effectiveness. As reported by Donaldson and Grant-Vallone (2002) self-reported surveys are common data gathering tools because they are easy to obtain. It seems clear that there can be serious trade-offs for this ease of access. They report that research participants often respond in ways which they believe will make them look better. While some of this may be out of fear that potential employers will see their data, there is also the possibility that they believe they are better/more effective than they are in practice. Despite the fact that sole reliance on self-reported data for research has been criticized and even called unacceptable, these methods persist and have been commonly seen throughout this review process. While some papers included in the final results did include other sources of data, many relied solely on self-reported surveys. The authors further argue that because of the multi-dimensional aspects that cause bias in

self-reported data (such as participant level of sensitivity to personal ability and their perceived risks and benefits from such reports), it is unlikely that any single method will be able to confidently check the self-reported data for bias. Donaldson and Grant-Vallone (2002) point out that it is likely at least two other measures would be needed to provide a confident check on the results of self-reported data. Many studies found through this review process did not include multiple checks on self-reports, and, as stated previously, some relied solely on these types of data gathering methods.

The bias of self-reporting and seeming reliance on it for educational research is something that ought to be addressed. While understanding the benefits of self-reported data (ease of access, ensuring student satisfaction, etc.) it is something that often may not provide a sound methodological basis for programme effectiveness conclusions. While we cannot know for sure whether the differences between findings of self-reported studies and those which were not self-reported (McKinney et al. (2004) and McKinney, et al. (2008), which were scored interviews) are due to the methods of data collection, it does raise the question of whether education research methods should focus on non-self-reported data, both quantitatively and qualitatively. If policy reforms are to be based on research, it seems obvious that we would want the most robust and methodologically sound research methods to be used.

Further, it seems that there is little common terminology and definitions when it comes to evaluating ITE programmes. For instance, what exactly classifies a programme as effective or ineffective? What does a category such as lesson planning actually entail? What characteristics are needed to teach in specific contexts? These questions point not only to a need to develop common terminology with common definitions, but also a need for common assessment tools to determine competence and effectiveness within these categories. The studies which involved McKinney as the lead author are continually mentioned, but it is because they use a tool unlike the other studies: a specific quantitative assessment which has been tested and researched over decades to determine effectiveness in a specific context (McKinney et al., 2004). Castle et al., (2009) also uses a method where students do not self-report (reports and end-of-term portfolios are graded), although these types of methods must employ

scorers/ graders who have no incentive to provide biased information. Such objective measures ought to be more common in education, and could even be developed to enhance examinations for entrance to ITE programmes and subsequent certifications. However, such developments require common terminology, definitions, and a general acceptance of what makes a teacher effective – something that does not appear to be concretely defined at the moment.

As a final note on future research, the review process revealed very few studies which combined data on pre-service and in-service teachers. Teacher education (and many other aspects of education) could benefit from studies which follow pre-service teachers (broken down by ITE programme type, if desired) through their coursework, field experience(s), certification exams, and post-certification in-service teaching, with data also collected on their students' achievement levels. It is likely that only such long-term studies (including longer, more thorough systematic reviews) can provide a better understanding of the advantages and disadvantages of certain ITE preparation, and in which specific contexts they work.

4.3 Implications for Policy

Research displayed in this paper (both in the meta-analysis and outside such analyses) showed that PDS interventions increased self-efficacy, as well as classroom management and organisation, the ability to be a reflective practitioner, and possible improvement in urban teaching effectiveness. The earlier literature review also points to the importance of longer and more extensive teacher education when compared with shorter training programmes. Other cited research has mentioned the possibilities which PDSs provide to drastically reform teacher education, and the entirety of the education system. However, we cannot forget about the concerns brought about by PDSs, which may shed light on the (at times) mediocre effects demonstrated by PDS ITE interventions.

Darling-Hammond (1994) notes a large obstacle to the realisation of PDS ideals is the structure of the current education system. New developments such as PDSs inherently require collaboration, an understanding of mutual and

individual benefits, and increased respect toward professional development. However, current structures keep teachers in their own individual classroom focused on individual work rather than collaboration on interdisciplinary projects. Further, current standardised assessments and evaluations continue to incentivise teachers to focus their energy on ensuring students pass a test, rather than promoting critical thinking, discourse, debate, and in-depth study in a collaborative environment.

While the current structures encourage non-collaborative and surface-level learning, the low-value typically placed on university schools of education leads to under-funding, putting further stress on teacher education. Without appropriate changes collaboration, integration of the school and university, and improved teacher education will be much more difficult to achieve. It may be due to these challenges that PDSs have not noticeably (in this study) improved the overall outcomes of ITE.

Built on the idea that teacher education cannot improve in isolation from the rest of the education system (The Holmes Group, 2007), PDSs show strong potential benefits. Previously cited research has shown the benefits of full certification and extensive preparation, and this review has provided more insight into potential positives for PDS models of ITE. Since the PDS model appears to provide benefits in some areas of ITE (although more thorough research is necessary), moving more of the teaching force to a TFA or Teach First-style alternative programme seems both unrealistic and unlikely to be particularly beneficial in the long-term. As noted by McKinney et al. (2004), it is possible that field experiences for pre-service teachers ought to be longer, even more than the already-elongated PDS field experience. A longer experience would allow more preparation, more exposure to a classroom environment, and more experience on which to reflect. Longer exposure and partnerships in various contexts and low-income areas provided by the PDS model would also allow the structure of ITE, schools, and universities over time (Abdal-Haqq, 1991; Darling-Hammond, 1994, 1996). While Byrd and McIntyre (1999) acknowledge the small research base, they maintain that PDSs show great potential to realise the goals of improved teacher education if higher workloads

can be handled, education hierarchy can be restructured to allow for more innovation and university faculty choose to engage in school sites.

The model proposed by Darling-Hammond (2007) also may guide us in forming new systems of teacher education. Through story, she proposes a system where one declares a non-education major during university (presumably the subject which the student plans to teach) and after a couple of years is admitted to a selective teacher education programme. During this programme, the student has field experiences in classrooms and community centres/organisations, takes courses on learning and development, and appears to complete their university major within the typical four-year period from entrance at university (in the United States), while then moving onto a fifth-year master's degree programme. That fifth year also includes a year-long internship at a PDS site. A process like this combines intensive education work, preparation, and internships at a PDS site with the firmly established U.S. university track.

In certain ways, proposals like the one above have similarities to common teacher education tracks in the United Kingdom. An undergraduate degree can be completed, followed by a year studying for a PGCE, where most time is spent in a classroom setting. Some coursework is taken during this year as well in order to continue building on content-area knowledge for the pre-service teacher. With a pre-requisite level of subject area coursework needed for PGCE admission, this pathway has the benefits of assuring teachers are competent in their content areas, while also placing them in a nearly year-long field experience, similar to Darling-Hammond's proposed "fifth year."

The Holmes Group, throughout its various reports, also spoke about the professionalisation of teaching, making it more like other professions (it regularly equated PDS sites to teaching hospitals). Such professionalisation could occur via undergraduate tracks such as those that are seen in medicine and dentistry (pre-med coursework and pre-dental coursework, respectively). A potential track may include introduction to non-content area coursework important for education (such as psychology of learning, development, sociology, and courses on inclusion and diversity). This would be able to ensure those admitted to teacher education programmes have a sufficient level of

knowledge to be successful during intensive coursework and field experiences. As per the earlier discussion regarding the need for more common standards and terminology in education, holistic entrance exams (which may include interview components such as that which is used in McKinney et al. (2004) and McKinney, et al. (2008)) tailored to determine an applicant's potential for and dedication to the teaching profession could also be developed as a requisite for entry into ITE programmes. As a note, this does not necessarily mean a pen-and-paper standardised assessment would be necessary, but debate on the virtues, or lack thereof, of standardised testing is outside the purview of this paper.

Combining these different aspects may result in teacher education programmes that look similar to Darling-Hammond's proposal with some adjustments: entrance at a university with a plan to complete a subject area major and pre-ITE coursework. After the first two or three years, an application would be submitted for the university's teacher education programme, including scores on the holistic ITE exam and/or interview. Once admitted, students would not only finish their major coursework and begin more in-depth study in subjects such as learning, development, and sociology, but be constantly immersed in community-based and school-based field experiences. As Darling-Hammond proposes, this would culminate in a "fifth year" where the ITE student is in a year-long internship while bulking up on subject-specific and general education coursework that may lead to a master's degree or a professional certificate – one may consider this year to be similar to a PGCE year at a U.K. university. Schools of education providing this intensive teacher preparation may be granted a name such as "academy," but that is a secondary issue related more to style than substance.

Such an institution keeps the important structures of PDSs, while adding standardisation into the transition process from undergraduate coursework to teacher training, and through graduation and certification. An ITE programme like this would provide more context exposure and longer field experiences for students. Field experiences which were too short were noted as a possible reason why students appeared ill-prepared for teaching in McKinney et al. (2004). The pre-requisite coursework would ensure each ITE student has a

base level of knowledge upon which to build. Its attachment to a university allows interdisciplinary research to be undertaken among faculty, staff, and students (who could even be required to complete a research project), and would also allow for a variety of perspectives to be heard. As Darling-Hammond (2007) notes, this structure would combine what we know works for teacher education. It further allows for the professionalisation and standardisation of the teaching profession, supported by The Holmes Group (2007), to occur. Perhaps most importantly, this structure does not isolate one idea and implement it into current teacher education programmes. Rather, it is a structural shift that integrates multiple research-based ideas. As argued by The Holmes Group, teacher education cannot improve in isolation. It is my belief, based on the evidence, that such changes would improve the integration, collaboration, professionalisation, and quality of graduates produced by ITE programmes, and be more in alignment with the original vision for Professional Development Schools.

4.4 Conclusion

This analysis, while time constrained, has provided a base for future research and information for ITE reforms. Keeping in mind the limitations of this study, the data indicate notes of positivity for PDS interventions in ITE. Clearly, more research should be gathered, but it is notable that not one meta-analysis displayed a pooled negative effect of PDS interventions (insignificance, notwithstanding). Moreover, different PDSs may have been at different developmental stages, leaving the possibility open that more developed PDSs would show larger effects than novice PDS sites. Despite these limitations, I believe the implications for research and policy are sound based on the evidence provided, and would provide a better system of ITE and education research.

References

- Abdal-Haqq, I. (1991). Professional Development Schools and Educational Reform: Concepts and Concerns. ERIC Digest 91-2.
- Boe, E. E., Cook, L. H., & Shin, S. (2007). Does Teacher Preparation Matter for Beginning Teachers in Either Special or General Education? *The Journal of Special Education, 41*(3), 158-170.
- Byrd, D. M., & McIntyre, D. J. (1999). Research on Professional Development Schools. In D. M. Byrd & D. J. McIntyre (Eds.), *Teacher Education Yearbook* (Vol. VII). Thousand Oaks, California: Corwin Press, Inc.
- Castle, S., Fox, R. K., & Fuhrman, C. (2009). Does Professional Development School Preparation Make a Difference? A Comparison of Three Teacher Candidate Studies. *School-University Partnerships, 3*(2), 58-68.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Vol. 2). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Coleman, J. S., & Others. (1966). *Equality of Educational Opportunity* (OE-38001). Retrieved from <http://files.eric.ed.gov/fulltext/ED012275.pdf>
- Darling-Hammond, L. (1994). *Professional Development Schools: Schools for Developing a Profession* (L. Darling-Hammond Ed.). New York: Teacher's College Press.
- Darling-Hammond, L. (1996). The Quiet Revolution: Rethinking Teacher Development. *Educational Leadership, 53*(6), Improving Professional Practice, 4-10.
- Darling-Hammond, L. (2000a). How Teacher Education Matters. *Journal of Teacher Education, 51*(3), 166-173. doi:10.1177/0022487100051003002
- Darling-Hammond, L. (2000b). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives, 8*(1), 1-44.
- Darling-Hammond, L. (2007). The Story of Gloria Is a Future Vision of the New Teacher. *Journal of Staff Development, 28*(3), 25-26.
- Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does Teacher Certification Matter? Evaluating the Evidence. *Educational Evaluation and Policy Analysis, 23*(1), 57-77.
- Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness. *Education Policy Analysis Archives, 13*(42), 51.

- Donaldson, M. L., & Moore Johnson, S. (2011). TFA Teachers: How Long Do They Teach? Why Do They Leave? *Education Week*.
- Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding Self-Report Bias in Organizational Behavior Research. *Journal of Business and Psychology, 17*(2), 245-260.
- EPPI-Centre. (2003). Review Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research. Version 0.9.7. London: EPPI-Centre, Social Science Research Unit.
- Goldhaber, D. (2002). The Mystery of Good Teaching. *Education Next, 2*(1).
- Goldhaber, D. D., & Brewer, D. J. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement *Educational Evaluation and Policy Analysis, 22*(2), 129-145.
- Gough, D. (2004). Systematic Research Synthesis to Inform the Development of Policy and Practice in Education. In G. Thomas & R. Pring (Eds.), *Evidence-based Practice* (pp. 44-62). Buckingham: Open University Press.
- Gough, D. (2007). Weight of Evidence: A Framework for the Appraisal of the Quality and Relevance of Evidence. *Research Papers in Education, 22*(2), 213-228.
- Haberman, M. (1986). Licensing Teachers: Lessons from Other Professions. *Phi Delta Kappan, 67*(10), 719-722.
- Haberman, M. (1987). Recruiting and Selecting Teachers for Urban Schools. ERIC/CUE Urban Diversity Series, Number 95 (pp. 76).
- Haberman, M. (1995). The Dimensions of Excellence in Programs Preparing Teachers for Urban Poverty Schools. *Peabody Journal of Education, 70*(2), 24-43.
- Hanushek, E. A. (1971). Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data. *The American Economic Review, 61*(2), 280-288.
- Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature, 24*(3), 1141-1177.
- Hanushek, E. A. (2011). The Economic Value of Higher Teacher Quality. *Economics of Education Review, 30*(3), 466-479.
doi:10.1016/j.econedurev.2010.12.006
- Hawk, P. P., Coble, C. R., & Swanson, M. (1985). Certification: It Does Matter. *Journal of Teacher Education, 36*(3), 13-15.

- Hemphill, F. C., & Vanneman, A. (2011). *Achievement Gaps: How Hispanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress* (NCES 2011-459). Retrieved from Washington, D.C. : <https://nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf>
- Holbein, M., Woong, L., Annis, K., & Doll, V. (2016). The Impact of an Urban Professional Development School Model on Attitudes and Self-Efficacy of Developing Preservice Teachers. *School-University Partnerships*, 9(1), 64-70.
- Ingersoll, R. M. (2002). The Teacher Shortage: A Case of Wrong Diagnosis and Wrong Prescription. *NASSP Bulletin*, 86(631), 16-31. doi:10.1177/019263650208663103
- Innovation, U. S. D. o. E. O. o. P. P. a. (2003). *Meeting the Highly Qualified Teachers Challenge: The Secretary's Second Annual Report*. Retrieved from Washington, D.C.: <https://www2.ed.gov/about/reports/annual/teachprep/2003title-ii-report.pdf>
- Kochan, F. K. (1999). Professional Development Schools: A Comprehensive View. In D. M. Byrd & D. J. McIntyre (Eds.), *Research on Professional Development Schools* (Vol. VII). Thousand Oaks, California: Corwin Press, Inc.
- Mantle-Bromley, C., Gould, L. M., McWhorter, B. A., & Whaley, D. C. (2000). The Effect of Program Structure on New Teachers' Employment and Program Satisfaction Patterns. *Action in Teacher Education*, 22(1), 1-14.
- McKinney, S. E., & Finke, J. A. (2005). A Comparison of the Internship Experience for Student Interns Placed in Different Urban School Environments. *Professional Educator*, 27(1), 7.
- McKinney, S. E., Haberman, M., Stafford-Johnson, D., & Robinson, J. (2008). Developing Teachers for High-Poverty Schools: The Role of the Internship Experience. *Urban Education*, 43(1), 68-82.
- McKinney, S. E., Robinson, J., & Spooner, M. (2004). A Comparison of Urban Teacher Characteristics for Student Interns Placed in Different Urban School Settings. *Professional Educator*, 26(2), 17-30.
- Monk, D. H. (1994). Subject Area Preparation of Secondary Mathematics and Science Teachers and Student Achievement. *Economics of Education Review*, 13(2), 125-145.
- Obama, B. (2011). Remarks by the President in State of the Union Address [Press release]. Retrieved from <https://obamawhitehouse.archives.gov/the-press-office/2011/01/25/remarks-president-state-union-address>

- Olson, L. (1989). 'Clinical Schools': Theory Meets Practice on the Training Ground. *Education Week*, 8, 1.
- Partnership, T. H. (2007). *The Holmes Partnership Trilogy: Tomorrow's Teachers, Tomorrow's Schools, Tomorrow's Schools of Education*. New York, NY: Peter Lang Publishing, Inc.
- Pawson, R., Boaz, A., Grayson, L., Long, A., & Barnes, C. (2003). Types and Quality of Social Care Knowledge. London: Social Care Institute for Excellence.
- Reardon, S. F. (2012). The Widening Academic Achievement Gap between the Rich and the Poor. *Community Investments*, 24(2), 19-39.
- Reinhartz, J., & Stetson, R. (1999). Teachers as Leaders: A Question or an Expectation? In D. M. Byrd & D. J. McIntyre (Eds.), *Research on Professional Development Schools* (Vol. VII). Thousand Oaks, California: Corwin Press, Inc.
- Ridley, S. D., Hurwitz, S., Hackett, M. R. D., & Miller, K. K. (2005). Comparing PDS and Campus-Based Preservice Teacher Preparation: Is PDS-Based Preparation Really Better? *Journal of Teacher Education*, 56(1), 46-56.
- Rockoff, J. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*, 94(2), 247-252.
- Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement*. Retrieved from Knoxville, Tennessee:
[http://www.stcloudstate.edu/tpi/initiative/documents/preparation/Cumulative and Residual Effects of Teachers.pdf](http://www.stcloudstate.edu/tpi/initiative/documents/preparation/Cumulative%20and%20Residual%20Effects%20of%20Teachers.pdf)
- Sandholtz, J. H., & Wasserman, K. (2001). Student and Cooperating Teachers: Contrasting Experiences in Teacher Preparation Programs. *Action in Teacher Education*, 23(3), 54-65.
- Sharpe, T., Lounsbery, M. F., Golden, C., & Deibler, C. (1999). Analysis of an Ongoing, District-Wide Collaborative Approach to Teacher Education. *Journal of Teaching in Physical Education*, 19(1), 79-96.
- Sykes, G., Bird, T., & Kennedy, M. (2010). Teacher Education: Its Problems and Some Prospects. *Journal of Teacher Education*, 61(5), 464-476. doi:10.1177/0022487110375804
- Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). *Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress*. Retrieved from Washington, D.C.:
<https://nces.ed.gov/nationsreportcard/pdf/studies/2009455.pdf>

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education*, 11, 57-67. doi:10.1023/A:1007999204543

Yerian, S., & Grossman, P. L. (1997). Preservice Teachers' Perceptions of Their Middle Level Teacher Education Experience: A Comparison of a Traditional and a PDS Model. *Teacher Education Quarterly*, 24(4), 85-101.

Appendix 1: Data Extraction Tool

Guidelines for extracting data and quality assessing primary studies in educational research Version 0.9.7

Section A: Administrative details

These guidelines are adapted from the following source:

EPPI-Centre (2003) Review Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research. Version 0.9.7. London: EPPI-Centre, Social Science Research Unit.

A.1 Name of the reviewer	A.1.1 Details
A.2 Date of the review	A.2.1 Details
A.3 Please enter the details of the paper being reviewed <i>A paper can be a journal article, a book, or chapter in a book, or an unpublished report.</i>	A.3.1 Type of Paper: A.3.2 Unique Identifier: A.3.3 Authors: A.3.4 Title: A.3.5 Source: A.3.6 Publication Date: A.3.7 Language: A.3.8 Identification Number: A.3.9 Access Date(s):

Section B: Study aim(s) and rationale

B.1 What are the study research questions and hypotheses? <i>Please write in authors' states question if there is one. Elaborate if necessary, but indicate which aspects are reviewers' interpretation. Other, more specific questions about the research questions and hypotheses are asked later.</i>	B.1.1 Explicitly stated (please specify) B.1.2 Implicit (please specify) B.1.3 Not stated/unclear (please specify)
---	--

<p>B.2 What is/are the explanation(s) given for the potential success of PDS or the specific PDS model being studied in bringing about improved outcomes ?</p> <p><i>Please write in authors' explanation if there is one. Elaborate if necessary, but indicate which aspects are reviewers' interpretation.</i></p>	<p>B.2.1 Explicitly stated (please specify)</p> <p>B.2.2 Implicit (please specify)</p> <p>B.2.3 Not stated/unclear (please specify)</p>
<p>B.3 Which of the following groups were consulted in working out the aims of the study, or issues to be addressed in the study?</p> <p><i>Please write in authors' description if there is one. Elaborate if necessary, but indicate which aspects are reviewers' interpretation. Please cover details of how and why people were consulted and how they influenced the aims/issues to be addressed.</i></p>	<p>B.3.1 Researcher(s) (please specify)</p> <p>B.3.2 Funder(s) (please specify)</p> <p>B.3.3 Education Practitioner(s) (please specify)</p> <p>B.3.4 Other(s) (please specify)</p> <p>B.3.5 None/Not stated</p> <p>B.3.6 Coding is based on: Authors' description</p> <p>B.3.7 Coding is based on: Reviewers' inference</p>
<p>B.4 When was the study carried out?</p> <p><i>If the authors give a year, or range of years, then put that in. If not, give a 'not later than' date by looking for a date of first submission to the journal, or for clues like the publication dates of other reports from the study.</i></p>	<p>B.4.1 Explicitly stated (please specify)</p> <p>B.4.2 Implicit (please specify)</p> <p>B.4.3 Not stated/unclear (please specify)</p>

Section C: Study research question(s) and its policy or practice focus

The first eight questions come from Keywording – please note if there is disagreement with previous coding

Note: If there was a PDS comparison group, questions C.1 through C.8 were repeated for the PDS comparison group

<p>C.1 What is/are the population focus/foci of the PDS group?</p>	<p>C.1.1 Learners</p> <p>C.1.2 Head teacher / Senior management</p> <p>C.1.3 Teaching staff</p> <p>C.1.4 Teachers as learners</p> <p>C.1.5 Pre-service teachers</p> <p>C.1.6 Other education practitioners</p> <p>C.1.7 Government</p> <p>C.1.8 Local education authority officers</p> <p>C.1.9 Parents</p>
---	---

	<p>C.1.10 Governors</p> <p>C.1.11 Other</p> <p>C.1.12 Coding is based on: Authors' description</p> <p>C.1.13 Coding is based on: Reviewers' inference</p>
<p>C.2 What is the relevant university level of the group?</p>	<p>C.2.1 Undergraduate</p> <p>C.2.2 Graduate</p> <p>C.2.3 Other (please specify)</p> <p>C.2.4 Unstated</p>
<p>C.3 What is the sex of the teachers-in-training?</p>	<p>C.3.1 Female only</p> <p>C.3.2 Male only</p> <p>C.3.3 Mixed sex</p> <p>C.3.4 Not stated/unclear</p> <p>C.3.6 Coding is based on: Authors' description</p> <p>C.3.7 Coding is based on: Reviewers' inference</p>
<p>C.4 In which country or countries was the study carried out? <i>Region/locality should be specified if possible</i></p>	<p>C.4.1 Explicitly stated (please specify)</p> <p>C.4.2 Not stated/unclear (please specify)</p>
<p>C.5 At what type of school are the partnership Professional Development Schools? <i>Schools in different nations may have different classifications. This ought to be explained further if necessary</i></p>	<p>C.5.1 Traditional Public School (U.S.); State School (U.K.)</p> <p>C.5.2 Private School (U.S.); Public School (U.K.)</p> <p>C.5.3 Charter (U.S.); Academy (U.K.)</p> <p>C.5.4 Independent school</p> <p>C.5.5 Other educational setting (please specify)</p> <p>C.5.6 Multiple Educational Settings (please specify)</p> <p>C.5.6 Not stated/Unclear (specific if necessary)</p>
	<p>C.6.1 Pre-primary</p> <p>C.6.2 Primary</p>

<p>C.6 At what level of school are the partnership Professional Development Schools? <i>Schools in different nations may have different classifications. This ought to be explained further if necessary</i></p>	<p>C.6.3 Middle/Junior High (U.S. only) C.6.4 Secondary C.6.5 Multiple (please specify)5 elemen C.6.6 Other (please specify) C.6.7 Not Stated C.6.8</p>
<p>C.7 In which location type did the partnership Professional Development School reside?</p>	<p>C.7.1 Urban C.7.2 Suburban C.7.3 Rural C.7.4 Other (please specify) C.7.5 Not Stated C.7.6 Coding is based on: Authors' description C.7.7 Coding is based on: Reviewers' inference</p>
<p>C.8 Are there other locational details are relevant to this study?</p>	<p>C.8.1 Explicitly stated (please specify) C.8.2 Not stated/unclear (please specify)</p>
<p>C.9 Is there a non-PDS comparison group?</p>	<p>C.9.1 Yes C.9.2 No (If no, X-Y not applicable)</p>
<p>C.10 What is/are the population focus/foci of the non-PDS comparison group?</p>	<p>C.10.1 Learners C.10.2 Head teacher / Senior management C.10.3 Teaching staff C.10.4 Teachers as learners C.10.5 Non-teaching staff C.10.6 Other education practitioners C.10.7 Government C.10.8 Local education authority officers C.10.9 Parents C.10.10 Governors C.10.11 Other C.10.12 Coding is based on: Authors' description</p>

	C.10.13 Coding is based on: Reviewers' inference
C.11 What is the relevant university level of the group?	C.11.1 Undergraduate C.11.2 Graduate C.11.3 Other (please specify) C.11.4 Not stated
C.12 What is the sex of the teachers-in-training?	C.12.1 Female only C.12.2 Male only C.12.3 Mixed sex C.12.4 Not stated/unclear C.12.6 Coding is based on: Authors' description C.12.7 Coding is based on: Reviewers' inference
C.13 In which country or countries was the study carried out? <i>Region/locality should be specified if possible</i>	C.13.1 Explicitly stated (please specify) C.13.2 Not stated/unclear (please specify)
C.14 At what type of school is the comparison group? <i>Schools in different nations may have different classifications. This ought to be explained further if necessary</i>	C.14.1 Traditional Public School (U.S.); State School (U.K.) C.14.2 Private School (U.S.); Public School (U.K.) C.14.3 Charter (U.S.); Academy (U.K.) C.14.4 Independent school C.14.5 Other educational setting (please specify) C.14.6 Coding is based on: Authors' description C.14.7 Coding is based on: Reviewers' inference
C.15 At what level of school are the non-comparison group (if a clinical field experience is undertaken)Not stat? <i>Schools in different nations may have different classifications. This</i>	C.15.1Pre-Primary C.15.2 Primary C.15.3 Middle/Junior High (U.S. only) C.15.4 Secondary C.15.5 Other (please specify) C.15.6 Not Stated

<i>ought to be explained further if necessary</i>	C.15.7 Coding is based on: Authors' description C.15.8 Coding is based on: Reviewers' inference
C.16 In which location type did the non-comparison group partake in a clinical field experience?	C.16.1 Urban C.16.2 Suburban C.16.3 Rural C.16.4 Other (please specify) C.16.5 Not Stated C.16.6 Coding is based on: Authors' description C.16.7 Coding is based on: Reviewers' inference
C.17 Are there other locational details are relevant to this study?	C.17.1 Explicitly stated (please specify) C.17.2 Not stated/unclear (please specify)

Section D: Methods - Design

The EPPI Centre team and the Review Groups are working on more detailed questions that are appropriate to different types of studies. In the meantime, please add extra comments on the methods if you think they are relevant, or suggest questions.

D.1 Which type(s) of study does this report describe? <i>Multiple codings accepted, but please take care to code in relation to the main focus of this data extraction.</i> <i>Note this has already been asked in keyword Q10.</i> <i>(Studies that look at the development of methodology, or that review primary research, are not addressed by data extraction.)</i>	D.1.1 Description D.1.2 Evaluation of Relationship(s) D.1.3 Evaluation of Relationship(s) with Comparison D.1.4 Review of Primary Research
D.2 Which variables or concepts, if any, does the study aim to measure or examine?	D.2.1 Explicitly stated (please specify) D.2.2 Implicit (please specify) D.2.3 Not stated/unclear (please specify)
D.3 If the study is an evaluation, when were measurements of the	D.3.1 Not applicable (not an evaluation)

<p>variable(s) used for outcomes made, in relation to the intervention?</p> <p><i>If at least one of the outcome variables is measured both before and after the intervention the before and after category was used.</i></p>	<p>D.3.2 Before and after</p> <p>D.3.3 Only after</p> <p>D.3.4 During (specify if necessary)</p> <p>D.3.5 Other (please specify)</p> <p>D.3.6 Not stated/unclear (please specify)</p>
---	---

Section E: Methods - Groups

<p>E.1 If comparisons are being made between PDS and non-PDS groups, please specify the basis of any differences in these programmes.</p>	<p>E.1.1 Not applicable (not more than one group)</p> <p>E.1.2 Explicitly Stated (please specify)</p> <p>E.1.4 Implicit (please specify)</p> <p>E.1.5 Not stated/unclear (please specify)</p>
<p>E.2 Do the groups differ in any way aside from the type of programme (PDS or non-PDS) in which they are enrolled?</p>	<p>E.2.1 Not applicable (not more than one group)</p> <p>E.2.2 Explicitly stated (please specify)</p> <p>E.2.3 Implicit (please specify)</p> <p>E.2.4 Not stated/unclear (please specify)</p>
<p>E.3 Number of groups</p> <p><i>The reviewer should specify the precise number of PDS and non-PDS groups</i></p>	<p>E.3.1 One (please specify)</p> <p>E.3.2 Two (please specify)</p> <p>E.3.3 Three (please specify)</p> <p>E.3.4 Four or more (please specify)</p> <p>E.3.5 Other/unclear (please specify)</p>
<p>E.4 Study design summary</p> <p><i>In addition to answering the questions in this section, describe the study design in your own words. You may want to draw upon and elaborate the answers you have already given.</i></p>	<p>E.4.1 Details</p>

Section F: Methods - Actual sample

<p>F.1 What was the total number of participants in the study (the actual sample)? <i>If more than one group is being compared, please give numbers for each group.</i></p>	<p>F.1.1 Not applicable (e.g. study of policies, documents etc) F.1.2 Explicitly stated (please specify) F.1.3 Implicit (please specify) F.1.4 Not stated/unclear (please specify)</p>
<p>F.2 What is the proportion of those selected for the study who actually participated in the study? <i>Please specify numbers and percentages if possible.</i></p>	<p>F.2.1 Not applicable (e.g. study of policies, documents etc) F.2.2 Explicitly stated (please specify) F.2.3 Implicit (please specify) F.2.4 Not stated/unclear (please specify)</p>
<p>F.3 Which country/countries are the individuals in the actual sample from? <i>If UK, please distinguish between England, Scotland, N. Ireland and Wales, if possible. If from different countries, please give numbers for each.</i></p> <p><i>If more than one group is being compared, please describe for each group.</i></p>	<p>F.3.1 Not applicable (e.g. study of policies, documents etc) F.3.2 Explicitly stated (please specify) F.3.3 Implicit (please specify) F.3.4 Not stated/unclear (please specify)</p>
<p>F.4 What ages are covered by the actual sample? <i>Please give the numbers of the sample that fall within each of the given categories. If necessary refer to a page number in the report (e.g. for a useful table).</i></p> <p><i>If more than one group is being compared, please describe for each group.</i></p> <p><i>If follow-up study, age at entry to the study.</i></p>	<p>F.4.1 Not applicable (e.g. study of policies, documents etc) F.4.2 0 to 4 F.4.3 5 to 10 F.4.4 11 to 16 F.4.5 17 to 20 F.4.6 21 and over F.4.7 Not stated/unclear (please specify) F.4.8 Coding is based on: Authors' description F.4.9 Coding is based on: Reviewers' inference</p>

<p>F.5 What is the sex of the individuals in the actual sample? <i>Please give the numbers of the sample that fall within each of the given categories. If necessary refer to a page number in the report (e.g. for a useful table).</i></p> <p><i>If more than one group is being compared, please describe for each group.</i></p>	<p>F.5.1 Not applicable (e.g. study of policies, documents etc)</p> <p>F.5.2 Single sex (please specify)</p> <p>F.5.3 Mixed sex (please specify)</p> <p>F.5.4 Not stated/unclear (please specify)</p> <p>F.5.5 Coding is based on: Authors' description</p> <p>F.5.6 Coding is based on: Reviewers' inference</p>
<p>F.6 What is the socio-economic status of the individuals within the actual sample? <i>If more than one group is being compared, please describe for each group</i></p>	<p>F.6.1 Not applicable (e.g. study of policies, documents etc)</p> <p>F.6.2 Explicitly stated (please specify)</p> <p>F.6.3 Implicit (please specify)</p> <p>F.6.4 Not stated/unclear (please specify)</p>
<p>F.7 What is the ethnicity of the individuals within the actual sample? <i>If more than one group is being compared, please describe for each group.</i></p>	<p>F.7.1 Not applicable (e.g. study of policies, documents etc)</p> <p>F.7.2 Explicitly stated (please specify)</p> <p>F.7.3 Implicit (please specify)</p> <p>F.7.4 Not stated/unclear (please specify)</p>
<p>F.8 Is there any other useful information about the study participants?</p>	<p>F.8.1 Not applicable (e.g. study of policies, documents etc)</p> <p>F.8.2 Explicitly stated (please specify)</p> <p>F.8.3 Implicit (please specify)</p> <p>F.8.4 Not stated/unclear (please specify)</p>
<p>F.9 What proportion of the sample that dropped out/provided invalid data over the course of the study?</p>	<p>F.9.1 Not applicable (e.g. study of policies, documents etc)</p> <p>F.9.2 Not applicable (no drop outs)</p>

	<p>F.9.3 Explicitly Stated (please specify)</p> <p>F.9.4 Implicit (please specify)</p> <p>F.9.5 Unclear (please specify)</p>
<p>F.10 Do authors provide any information on whether and /or how those who dropped out/provided invalid data differ from those who remained in the study?</p>	<p>F.10.1 Not applicable (e.g. study of policies, documents etc)</p> <p>F.10.2 Not applicable (no drop outs)</p> <p>F.10.3 Yes (please specify)</p> <p>F.10.4 No</p>

Section G: Methods - Data collection

<p>G.1 Which methods were used to collect the data? <i>Please indicate all that apply and give further detail where possible.</i></p>	<p>G.1.1 Curriculum-based assessment</p> <p>G.1.2 Focus group</p> <p>G.1.3 Group interview</p> <p>G.1.4 One to one interview (face to face or by phone)</p> <p>G.1.5 Observation</p> <p>G.1.6 Self-completion questionnaire/survey</p> <p>G.1.7 Self-completion report or diary</p> <p>G.1.8 Exams</p> <p>G.1.9 Clinical test</p> <p>G.1.10 Practical test</p> <p>G.1.11 Psychological test</p> <p>G.1.12 Hypothetical scenario including vignettes</p> <p>G.1.13 School/college records (e.g. attendance records etc)</p> <p>G.1.14 Secondary data such as publicly available statistics</p> <p>G.1.15 Other documentation</p> <p>G.1.16 Not stated/unclear (please specify)</p> <p>G.1.17 Coding is based on: Authors' description</p>
---	--

	G.1.18 Coding is based on: Reviewers' inference
<p>G.2 Details of data collection methods or tool(s). <i>Please provide details including names for all tools used to collect data, and examples of any questions/items given. Also, please state whether source is cited in the report.</i></p>	<p>G.2.1 Explicitly stated (please specify)</p> <p>G.2.2 Implicit (please specify)</p> <p>G.2.3 Not stated/unclear (please specify)</p>
<p>G.3 Who collected the data? <i>Please indicate all that apply and give further detail where possible.</i></p>	<p>G.3.1 Researcher</p> <p>G.3.2 Head teacher/Senior management</p> <p>G.3.3 Teaching or other staff</p> <p>G.3.4 Parents</p> <p>G.3.5 Pupils/students</p> <p>G.3.6 Governors</p> <p>G.3.7 LEA/Government officials</p> <p>G.3.8 Other education practitioner</p> <p>G.3.9 Other (please specify)</p> <p>I.4.10 Not stated/unclear</p> <p>G.3.11 Coding is based on: Authors' description</p> <p>G.3.12 Coding is based on: Reviewers' inference</p>
<p>G.4 Do the authors describe any ways they addressed the <u>reliability</u> of their data collection tools/methods? <i>e.g. test - re-test methods</i></p> <p><i>(Where more than one tool was employed, please provide details for each.)</i></p>	<p>G.4.1 Details</p>
<p>G.5 Do the authors describe any ways they have addressed the <u>validity</u> of their data collection tools/methods? <i>e.g. mention previous validation of tools, published version of tools, involvement of target population in development of tools.</i></p> <p><i>(Where more than one tool was employed, please provide details for each.)</i></p>	<p>G.5.1 Details</p>

<p>G.6 Where were the data collected? <i>e.g. school, home.</i></p>	<p>G.6.1 Explicitly stated (please specify) G.6.2 Implicit (please specify) G.6.3 Unclear/not stated (please specify)</p>
<p>G.7 Are there any other important features of data collection? <i>e.g. use of video or audio tape; ethical issues such as confidentiality etc.</i></p>	<p>G.7.1 Details</p>

Section H: Methods - Data analysis

A section on statistical analysis is being developed for use where relevant. In the meantime please add any comments about statistical analysis.

<p>H.1 Which methods were used to analyse the data? <i>Please give details eg. for in-depth interviews, how were the data handled? Details of statistical analysis can be given next.</i></p>	<p>H.1.1 Explicitly stated (please specify) H.1.2 Implicit (please specify) H.1.3 Not stated/unclear (please specify)</p>
<p>H.2 Which statistical methods, if any, were used in the analysis?</p>	<p>H.2.1 Details</p>
<p>H.3 What rationale do the authors give for the methods of analysis for the study? <i>e.g. for their methods of sampling, data collection or analysis.</i></p>	<p>H.3.1 Details</p>
<p>H.4 Do the authors discuss how they treated missing data and data from drop outs?</p>	<p>H.4.1 Not applicable (e.g. study of policies, documents etc) H.4.2 No applicable (no missing data or drop outs) H.4.3 Yes (please specify) H.4.4 No</p>
<p>H.5 Do the authors describe any ways they have addressed the <u>reliability</u> of data analysis? <i>e.g. using more than one researcher to analyse data, looking for negative cases.</i></p>	<p>H.5.1 Details</p>
<p>H.6 Do the authors describe any ways they have addressed the <u>validity</u> of data analysis?</p>	<p>H.6.1 Details</p>

<i>e.g. internal or external consistency; checking results with participants.</i>	
H.7 Do the authors describe strategies used in the analysis to control for bias from confounding variables?	H.7.1 Details
H.8 Please describe any other important features of the analysis.	H.8.1 Details

Section I: Results and Conclusions

I.1 How are the results of the study presented? <i>e.g. as quotations/figures within text, in tables, appendices.</i>	I.1.1 Details
I.2 What are the results of the study as reported by authors? <i>Please give details and refer to page numbers in the report(s) of the study, where necessary (e.g. for key tables).</i>	I.2.1 Details
I.3 Are there any obvious shortcomings in the reporting of the data?	I.3.1 Yes (please specify) I.3.2 No
I.4 Do the authors report on all variables they aimed to study as specified in their aims/research questions? <i>This excludes variables just used to describe the sample.</i>	I.4.1 Yes (please specify) I.4.2 No
I.5 Do the authors state where the full, original data are stored?	I.5.1 Yes (please specify) I.5.2 No
I.6 What do the author(s) conclude about the findings of the study? <i>Please give details and refer to page numbers in the report of the study, where necessary.</i>	I.6.1 Details

Section J: Quality of the study - Reporting

<p>J.1 Is the context of the study adequately described? <i>Consider your answer to questions: Why was this study done at this point in time, in those contexts and with those people or institutions? Was the study informed by, or linked to an existing body of empirical and/or theoretical research?</i></p> <p><i>Which of the following groups were consulted in working out the aims to be addressed in the study? Do the authors report how the study was funded? When was the study carried out?</i></p>	<p>J.1.1 Yes (please specify) J.1.2 No (please specify)</p>
<p>J.2 Are the aims of the study clearly reported? <i>Consider your answer to questions: What are the broad aims of the study? What are the study research questions and/or hypotheses?</i></p>	<p>J.2.1 Yes (please specify) J.2.2 No (please specify)</p>
<p>J.3 Is there an adequate description of the sample used in the study and how the sample was identified and recruited? <i>Consider your answer to all questions in Methods on 'Sampling Strategy', 'Recruitment and Consent', and 'Actual Sample'.</i></p>	<p>J.3.1 Yes (please specify) J.3.2 No (please specify)</p>
<p>J.4 Is there an adequate description of the methods used in the study to collect data? <i>Consider your answer to the following questions: Which methods were used to collect the data? Details of data collection methods or tools Who collected the data? Do the authors describe the setting where the data were collected? Are there other important features of the data collection procedures?</i></p>	<p>J.4.1 Yes (please specify) J.4.2 No (please specify)</p>
<p>J.5 Is there an adequate description of the methods of data analysis? <i>Consider your answer to the following</i></p>	<p>J.5.1 Yes (please specify) J.5.2 No (please specify)</p>

<p>questions in: Which methods were used to analyse the data? What statistical methods if any, were used in the analysis? Who carried out the data analysis?</p>	
J.6 Is the study replicable from this report?	J.6.1 Yes (please specify) J.6.2 No (please specify)
J.7 Do the authors avoid selective reporting bias? (e.g. do they report on all variables they aimed to study as specified in their aims/research questions?)	J.7.1 Yes (please specify) J.7.2 No (please specify)

Section K: Quality of the study

K.1 Are there ethical concerns about the way the study was done? <i>Consider consent, funding, privacy, etc.</i>	K.1.1 Yes, some concerns (please specify) K.1.2 No concerns
K.2 Who was involved in the study design?	K.2.1 Explicitly stated (please specify) K.2.2 Implicitly stated (please specify) K.2.3 Not Applicable (Explain)
K.3 Was the choice of research design appropriate for addressing the research question(s) posed?	K.3.1 Yes (please specify) K.3.2 No (please specify)
K.4 Have sufficient attempts been made to establish the <u>reliability</u> of data collection methods and tools? <i>Consider your answer to the following question: Do the authors describe any ways they have addressed the reliability of their data collection tools/methods?</i>	K.4.1 Yes, good (please specify) K.4.2 Yes, some attempt (please specify) K.4.3 No, none (please specify)
K.5 Have sufficient attempts been made to establish the <u>validity</u> of data collection tools and methods?	K.5.1 Yes, good (please specify) K.5.2 Yes, some attempt (please specify)

<p>Consider your answer to the following question in: Do the authors describe any ways they have addressed the validity of their data collection tools/methods?</p>	<p>K.5.3 No, none (please specify)</p>
<p>K.6 Have sufficient attempts been made to establish the <u>reliability</u> of data analysis? Consider your answer to the following question in: Do the authors describe any ways they have addressed the reliability of data analysis?</p>	<p>K.6.1 Yes (please specify) K.6.2 No (please specify)</p>
<p>K.7 To what extent are the research design and methods employed able to rule out any other sources of error/bias which would lead to alternative explanations for the findings of the study? e.g. (1) In an evaluation, was the process by which participants were allocated to, or otherwise received the factor being evaluated, concealed and not predictable in advance? If not, were sufficient substitute procedures employed with adequate rigour to rule out any alternative explanations of the findings which arise as a result? e.g. (2) Was the attrition rate low and, if applicable, similar between different groups?</p>	<p>K.7.1 A lot (please specify) K.7.2 A little (please specify) K.7.3 Not at all (please specify)</p>
<p>K.8 Weight of evidence: Taking account of all quality assessment issues, can the study findings be trusted in answering the study question(s)? In some studies it is difficult to distinguish between the findings of the study and the conclusions. In those cases, please code the trustworthiness of this combined results/conclusion.</p>	<p>K.8.1 High trustworthiness (please specify) K.8.2 Medium trustworthiness (please specify) K.8.3 Low trustworthiness (please specify)</p>
<p>K.9 Have sufficient attempts been made to justify the conclusions drawn from the findings so that the conclusions are trustworthy?</p>	<p>K.9.1 Not applicable (results and conclusions inseparable) K.9.2 High trustworthiness K.9.3 Medium trustworthiness K.9.4 Low trustworthiness</p>
<p>K.10 In light of the above, do the reviewers differ from the authors over the findings or conclusions of the study? Please state what any difference is.</p>	<p>K.10.1 Not applicable (no difference in conclusions) K.10.2 Yes (please specify)</p>

Appendix 2: Weight of Evidence Rubric

Criteria for Weight-of-Evidence Judgments

	Weight of Evidence Section*		
Judgment	A	B	C
High	Purpose of research is explicit; methods used and data are detailed, explicit, and clearly justify conclusions; report is easily understandable	Extensive use of thorough questionnaires and/or interviews and/or evaluations; reported in useful manner	Sample of at least one PDS cohort at a specific sites (multiple PDS groups/cohorts at the same or different schools and/or non-PDS sample groups may be included as a comparison); clear focus on teacher preparedness outcomes
Medium	Purpose of research understood but not explicit; methods used and data could be more detailed, understood but may not be explicit, and partially justify conclusions; report difficult to understand at times	Some use of questionnaires and/or interviews and/or evaluations; reported in manner which is moderately useful	Sample of at least one PDS cohort at a specific sites (multiple PDS groups/cohorts at the same or different schools and/or non-PDS sample groups may be included as a comparison); clear focus on teacher preparedness outcomes
Low	Purpose of research not clear, methods used and data are not detailed nor explicit and do not justify conclusions; report not easily understandable	No use of questionnaires and/or interviews and/or evaluations; reported in a non-useful way	Inadequate/no sample of PDS students; no relevant focus on teacher preparedness outcomes

*Section D is not included in this rubric because it is the aggregate judgment of Sections A, B, and C

The Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) is part of the Social Science Research Unit (SSRU), UCL Institute of Education, University College London.

Since 1993, we have been at the forefront of carrying out systematic reviews and developing review methods in social science and public policy. We are dedicated to making reliable research findings accessible to the people who need them, whether they are making policy, practice or personal decisions. We engage health and education policy makers, practitioners and service users in discussions about how researchers can make their work more relevant and how to use research findings.

Founded in 1990, the Social Science Research Unit (SSRU) is based at the UCL Institute of Education, University College London. Our mission is to engage in and otherwise promote rigorous, ethical and participative social research as well as to support evidence-informed public policy and practice across a range of domains including education, health and welfare, guided by a concern for human rights, social justice and the development of human potential.

The report was first published in 2017 by:

Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)
Social Science Research Unit
Institute of Education, University of London
18 Woburn Square
London WC1H 0NR

Tel: +44 (0)20 7612 6397
<http://eppi.ioe.ac.uk>
<http://www.ucl.ac.uk/ioe>

Cover image ©

This document is available in a range of accessible formats including large print. Please contact the UCL Institute of Education for assistance:
telephone: +44 (0)20 7947 9556 www.ucl.ac.uk