Universal Child Care and Children's Outcomes: A Meta-Analysis of Evidence from Natural Experiments

Working Paper: October 2015

Thomas van Huizen
Janneke Plantenga

Utrecht University School of Economics

Abstract

This study examines the effects of ECEC on child development and children's later life outcomes. Using meta-analytical techniques, we synthesize the findings from a recent strand of literature, exploiting natural experiments to identify the causal effects of universal ECEC arrangements (IV; DID; RDD). We use 253 estimates from 30 studies conducted between 2005 and 2015. Our meta-regressions include estimates on a wide variety of children's outcomes, ranging from (non-)cognitive development measured during early childhood to educational outcomes during adolescence and labor market performance during adulthood. We classify these diverse outcomes by whether the effect of ECEC on children's outcomes is significantly negative, statistically insignificant or significantly positive and estimate our main meta-analytical models regressions with ordered probit models. Our findings indicate that the evidence on universal ECEC is mixed. Whether the impact is positive or negative cannot be explained by the age of enrollment nor the intensity of the program. Quality, however, matters critically. Furthermore, there is no evidence of fading out and the effects of ECEC appear to be more favorable in the long run (during adolescence/adulthood). The study shows that the gains of ECEC are concentrated within the group of disadvantaged children.

1. Introduction

With the rise in female labor force participation during the past decades, a growing number of children spend a substantial part of their early childhood in Early Childhood Education and Care (ECEC) arrangements.[1] In OECD countries, on average about 70 percent of three year olds are enrolled in some form of ECEC (OECD, 2012). The increase in the use of ECEC arrangements is likely to be continued given a wide range of policy initiatives and objectives. For instance, increasing the enrollment in (high-quality) ECEC arrangements is central in Obama's "Preschool for All" proposal and the EU 2020 benchmarks for education. The policy attention towards public investment in early childhood is fueled by results from a large body of evidence pointing out that the early years matter crucially (Heckman, 2008; Currie and Almond, 2011). Various studies claim that investments in ECEC may improve child development and school readiness, increase children's performance in school and lead to higher levels of intergenerational mobility (Barnett, 2011; OECD, 2012a; 2012b; Duncan and Magnuson, 2013).

Whereas many recent policy aims and proposals concern increasing the coverage of universal child care schemes[2] (i.e. arrangements accessible to all children), the policy debate on universal child care arrangements has been dominated to a large extent by evidence from small scale, targeted interventions (e.g. Perry Preschool, Abecedarian). Although some of these studies provide compelling evidence in favor of ECEC investments (Belfield et al., 2006; Karoly et al., 2005; Heckman et al., 2010), for several reasons the results from these interventions have in fact limited applicability to universal child care arrangements. For instance, these interventions target very disadvantaged children, who are more likely to enter school with low levels of cognitive achievement and therefore have more scope to catch-up. Moreover, the major share of Perry Preschool's social returns is the result of crime reduction (Belfield et al., 2006; Heckman et al., 2010) – which may be unlikely for universal child care arrangements.[3] In addition, the relatively successful programs are small scale 'model' programs, which may be too costly to expand to a larger scale. Given these limitations, the actual impact of universal ECEC arrangements should be assessed by evaluating existing universal arrangements.

---

[1] ECEC may refer to all kind of non-partental child care arrangements before the child enters school/kindergarten. In this study, ECEC includes all formal, center-based child care arrangements – whether or not they have an explicit educational component. We focus on pre-kindergarten arrangements.

[2] We use the term 'universal' to refer to private and public center-based child care arrangements that are non-targeted and are, in principle, accessible to all children (e.g. there are no income criteria). However, this does not always mean that these arrangements are free (or affordable for all) and that the availability is guaranteed.

[3] About 88 percent of the social benefits are due to reductions in crime. However, a substantial share of both the control (55%) and the treatment group was arrested more than 5 times by the age of 40 (Schweinhart et al., 2005), indicating that the target group came from very disadvantaged backgrounds.

Only recently have scholars begun to identify the causal effects of universal ECEC schemes: these studies exploit exogenous variation in the use of child care, for instance as the result of policy reforms, and use Instrumental Variables, Difference-in-Differences, or Regression Discontinuity Design techniques to estimate the causal effect of child care (e.g. Meyer, 1995; Angrist and Pischke, 2009). The overall evidence from these natural experiments, so far, is inconclusive: some studies find that ECEC attendance improves child development (Gormley et al., 2005; Drange and Havnes, 2015), others show that ECEC has no significant impact (Fitzpatrick, 2008) or even worsens children's outcomes (Baker et al., 2008; Herbst, 2013a; Herbst and Tekin, 2015). As a result, universal child care expansions, currently discussed in many countries, may in some cases be considered as promising policy instruments but in others as a form of costly, ineffective (or even counterproductive) public policy.

This study synthesizes this recent (conducted in the period 2005-2015) empirical evidence on universal ECEC arrangements, using meta-analytical techniques. We aim to explain the heterogeneity in estimated effects of universal child care arrangements on children's outcomes: under which conditions are ECEC arrangements likely to have positive effects on child development? Do the effects 'fade out' in the longer run? And which specific groups benefit most from attending childcare? We focus on micro-econometric studies that use data from the US, Canada or Western Europe and apply quasi-experimental techniques exploiting exogenous variation in child care attendance (RDD; IV; DiD) to evaluate the causal effects of universal ECEC arrangements. Our meta-regressions are based on a sample of 253 estimates extracted from 30 studies.

We include estimates on a wide variety of children's 'human capital' outcomes, ranging from (non-)cognitive development measured during early childhood to adolescent educational outcomes and labor market performance during adulthood. Following several recent studies (e.g. Card et al., 2010), we classify the estimates by whether the effect of ECEC on children's outcomes is significantly negative, statistically insignificant or significantly positive. This allows us to compare results from studies that use completely different indicators of children's outcomes and estimation techniques. Our main meta-regressions are estimated with ordered probit models.

Although several reviews and meta-analytical studies on the effects of ECEC exist, this paper contributes to the field by its unique focus on evaluations of *universal* child care arrangements exploiting *natural experiments*. Previous meta-analyses include or focus exclusively on evaluations of targeted interventions (Karoly et al., 2005; Camilli et al., 2010; Nores and Barnett, 2010; Duncan and Magnuson, 2013). Whereas the effectiveness of targeted interventions has been extensively assessed using randomized controlled trials, it is unfeasible (and unethical) to randomly restrict access to universal schemes as result of which these schemes cannot be evaluated by RCTs. An

important estimation issue then arises, as selection into these schemes is voluntary: the decision of parents whether or not to use child care may be related to other (unobserved) factors that are related to child development. Many studies on the relation between ECEC and child outcomes have not explicitly taken this into account (although some include an extensive set of controls). Hence, these estimated relations may be spurious and provide limited insights into the actual causal effects of ECEC arrangements. Controlling for the selection into child care may be crucial: in fact, estimations that do take these selection issues into account may produce completely opposite results compared to estimations that do not (Herbst, 2013: 99). Another innovative element of this study is that the basic classification of sign and significance of the estimated impacts allows us to compare different types of child outcomes, capturing different dimensions (e.g. cognitive and non-cognitive) measured at different ages (from early childhood to adulthood). The meta-analysis therefore goes beyond the short-term cognitive development impact, which has been the focus of most meta-analyses on ECEC.

The remainder of this study is structured as follows. The next section discusses some general lessons from the existing literature: on the basis of this short review, we identify the moderators that will be central in the meta-analyses; section 3 provides a description of the sample of estimates and presents some descriptive evidence; section 4 presents the main results from our meta regression analysis; the final section concludes.

2. Theoretical considerations

There is a growing body of research on the effects of ECEC on child development and later life outcomes such as earnings. This section aims to present the most important lessons, puzzles and remaining questions. On the basis of this short review we will select the moderators that may explain the variation in the estimated effects of child care arrangements.

2.1 ECEC features: starting age, intensity and quality

It is an unsettled question whether the age of enrollment into child care is positively or negatively related to child development.[4] On the one hand, leading scholars in neuroscience have shown that the brain develops rapidly in the early years and that the speed of development slows down with the age of the child. The brain is particularly "malleable" during the early years of life. In line with this argument, early learning is the foundation for further learning according to recent human capital

[4] The relation between the starting age of ECEC is, by definition, related to the total duration of ECEC: total ECEC duration is equal to the school starting age minus the ECEC starting age. The discussion on the optimal starting age is also related to the effects of maternal employment in the early childhood years (and therefore to maternal/parental leave literature).

models; that is, “skill begets skill”. The implies that the returns of human capital investments are higher the earlier in life these investments are made (Cunha and Heckman, 2007; Heckman, 2007; Doyle et al., 2009). A recent meta-analysis (Leak et al., 2010) shows that ECEC programs that start before the age of three have larger positive effects than programs that start later, although this difference is not statistically significant. Others argue that “starting education interventions before age 3 does not appear to be a major contributor to effectiveness” (Barnett, 2011: 977). Given that an early starting age implies separation from the primary caregiver, there are concerns that an early starting age may lead to insecure attachment, generate stress and anxiety and cause negative effects on child development (Bowlby, 1969; Phillips and Lowenstein 2011). This may be especially a problem when children enroll below the age of two and spend long hours in child care (e.g. Haeck et al., 2015). Melhuish et al. (2015: 2) conclude on the basis of an extensive literature review that the results for 0-2 year olds is somewhat mixed, but that “for three years onwards the evidence is consistent that pre-school provision is beneficial to educational and social development for the whole population.” Hence, earlier may not always be better.

Next, the intensity (e.g. part-time/part-day versus full-time/full-day) is another important feature of ECEC programs. There is no consensus in the literature about the relation between the hours spent in child care and the benefits in terms of child development (Melhuish et al., 2015). Some (observational) studies found that children in full-day programs benefit more than those in part-day programs (e.g. Loeb et al., 2007; Robin et al., 2006). Given that the number of ECEC hours is generally determined by the parents, selection into part-time versus full-time schemes is endogenous and the results should be interpreted with caution. Evidence from a natural experiment in Canada in fact demonstrated that full-time child care can be detrimental for child development: more intensive ‘treatment’ does not always produce better outcomes. A full-time program is neither sufficient nor necessary for positive effects: some part-day interventions (e.g. Perry Preschool) show significant improvements in child outcomes.

While the developmental impact of the starting age and the dosage of ECEC is rather ambiguous, there is a growing consensus that the quality of services is crucial: “The positive impact of child care quality on various aspects of children's development is one of the most consistent findings in developmental science” (Melhuish et al., 2015). Various scholars point out that low quality care is a major concern and that the potential benefits can only be realized when the quality is sufficiently high (Haskins and Barnett, 2011; OECD, 2012; Cascio and Schanzenbach, 2013). The issue of quality is relevant in the case of universal child care arrangements, as these arrangements may be mainly targeted towards stimulating parental employment, with less emphasis on child development. Moreover, universal arrangements are also available to parents with higher

income/socio-economic status (SES): if the provided quality by day care centers is low, it is likely that the alternative mode of care (parental care) may be of higher quality (see subsection 2.3). Furthermore, lower SES children are more likely to attend lower quality childcare, as a result of which the gains from the ECEC experience may be limited. Given these arguments, there is more and more emphasis on the quality of ECEC within the policy debate.[5]

2.2 Children's outcomes: timing and type

One of the most controversial issues in the debate on the effects of ECEC is whether the potential developmental gains persist. Indeed, from a policy perspective it is crucial to understand whether ECEC has any long-term (social) benefits. Several targeted ECEC programs (e.g. Perry Preschool, Abecedarian, Head Start) have followed the children for several years - sometimes decades - and generally find that the positive gains in (cognitive) test scores diminish with the time since the end of the program (Leak et al., 2011; Camilli et al., 2010; Heckman et al.,2010). More recently, this pattern of fading out had been documented in the large scale Head Start Impact Study: treatment improves cognitive development significantly in the short-run, but these gains are no longer significant in the first grade (also information on non-cognitive is available, but this is somewhat more mixed) (Barnett, 2011; Puma et al.,2010; 2012). As the findings indicate rather poor results, the funding of the program is under debate (Gibbs et al., 2011).

However, various studies point out that even though the gains in test scores fade out during kindergarten or the first years of school, participation in ECEC may improve longer run outcomes in terms of educational achievements, labor market performance, and crime rates (see Duncan and Magnuson (2013) for a discussion). For instance, Perry Preschool produced limited cognitive achievement gains during childhood but had long-lasting effects on outcomes during adolescence and adulthood. Similarly, evidence indicates that Head Start improves adult outcomes (Garces et al., 2000; Deming, 2009; Ludwig and Miller, 2007). The combined finding of test scores fading out during childhood and significant long-run benefits has also been documented in the STAR project: the best predictor of these long run effects appears to be the short-term test score impact (Chetty et al., 2011).

These findings suggest that there may be "sleeper effects" and that mechanisms producing these long-run gains are rather complex. It is therefore important to assess not only when outcomes are measured, but also what is measured. Recent studies argue that development of non-cognitive skills during the early years may explain a substantial share of the generated adult outcomes (e.g.

[5] In Obama's Preschool for All aims to expand high quality ECEC. In the EU debate, there is a clear paradigm shift from emphasizing child care coverage (Barcelona targets for 2010), towards emphasizing quality ECEC (EU 2020 targets). The OECD as well has a strong focus on ECEC quality: for example, the OECD launched the project "Encouraging Quality in ECEC".

Heckman et al., 2013). These findings suggest that the ECEC impact on child development depends on what is actually measured (e.g. cognitive or non-cognitive skills). Also within the cognitive domain, ECEC may be more effective in improving language than in numeracy skills. This issue is related to the timing of measurement, since data from cognitive achievement tests are generally only available during (early) childhood, whereas school success and labor market outcomes are observed in the longer run.

2.3 Heterogeneous effects

A general finding in the literature is that children from parents with lower socio-economic status (SES) gain more than children from higher SES parents (Cascio, 2015). This is an important reason why some countries such as the US tend to focus on targeted interventions. As universal schemes are also accessible to less disadvantaged children from higher educated parents, the alternative modes of care – parental or informal care – are likely to be of higher quality than the care provided by (at best) mediocre quality ECEC arrangements: child care participation may therefore lead to detrimental effects for higher SES children (Cascio and Schanzenbach, 2013; Havnes and Mogstad, 2015). Clearly, this issue is critical in the discussion on universal child care schemes: if the benefits accrue only to specific groups, the rationale for public investment in ECEC for all is not evident.

In the US literature it has also been examined to what extent the effects of programs differ by race and ethnicity. Results from Head Start for instance suggest that black children benefit more than white children, but also experience a faster fading out of the effect (potentially due to the enrolment into worse quality schools). The cognitive gains of Head Start seem to be more persistent for Hispanics (Bitler et al., 2014). Furthermore, results from a universal US pre-kindergarten show that Hispanic children gain, but that these gains seem to be concentrated within the group of children with non-US born parents (Gormley, 2008).

3. Data

3.1 Included estimates and sample overview

Any meta-analysis starts with an extensive literature search. We performed internet key word searches and used recent reviews and meta-analyses (Ruhm and Waldfolgel, 2011; Brilli, 2014; Duncan and Magnuson, 2013; Meluish et al., 2015). Additional references to studies were obtained using “snowballing techniques”. We applied the following selection criteria:

1) Universal ECEC: studies included in the meta-analysis evaluate child care arrangements that are in principle not restricted in terms of access, i.e. we exclude evidence from targeted interventions.

2) Methodology: studies should exploit natural experiments and apply quasi-experimental techniques, using exogenous variation to correct for selection into these arrangements.[6] Studies use Instrumental Variables (IV), Difference-in Differences (DID) or Regression Discontinuity Designs (RDD).
3) Data: the estimations are based on micro data.
4) Treatment: the study provides an estimate of a treatment effect (formal, center-based childcare) versus a control (parental and/or informal care). The meta-analysis does not include estimates of treatment effects of non-parental care (which includes informal care). The control group should be not participating in any form of ECEC: we exclude studies that compare high-quality versus low-quality childcare, or one specific preschool curriculum versus another.
5) Region: to focus on results from relatively comparable settings, we include only evidence from Western, developed countries (US, Canada and Western Europe).
6) Outcomes: we focus on children's outcomes that capture their human capital, including indicators of cognitive and non-cognitive ability, skills, school performance and labor market outcomes. The study does not include estimates on effects on health, crime or parenting behavior.[7]

Although there are many estimates on the relation between ECEC and child outcomes, only a small share meets criteria 1 and 2: these are the key selection criteria. In fact, while the literature on the topic dates back several decades, these studies are very recent (the first study that meets these criteria was published in 2005). Some studies met all but one criteria: Cascio (2009) for example uses macro data; Berlinski et al. (2008; 2009) provide evidence from Uruguay and Argentina. Moreover, within many studies some or only one of the estimates meet all criteria, but others do not: for instance, IV studies generally also provide OLS estimates (which are not included) and some studies focus on non-parental care but provide also specific estimates on formal care (Bernal and Keane, 2011; Herbst, 2013a). We include only those estimates from the primary studies that meet all selection criteria.

[6] "Good natural experiments are studies in which there is a transparent exogenous source of variation that determine the treatment assignment. A natural experiment induced by policy changes, government randomization, or other events may allow a researcher to obtain exogenous variation in the main explanatory variables." (Meyer, 1995; 151). We therefore excluded estimates based on family/teacher fixed effects or those using Propensity Score Matching (as the latter is based on observed variables).

[7] In contrast to studies on targeted programs, studies on universal childcare arrangements using natural experiments generally do not include health and crime outcomes.

Most of the included studies provide multiple estimates of ECEC impacts, because they use different child outcome indicators (dependent variables), measure outcomes at different points of time (e.g. before and after entering school) or use multiple cohorts in their evaluation. For instance, a single study uses both cognitive and non-cognitive development measures (e.g. Herbst and Tekin, 2015) or multiple estimates on educational and labor market outcomes (e.g. Havnes and Mogstad, 2011). It is generally not clear what is the main result and the estimates provide valuable additional information. Estimates on different child outcomes are therefore included. However, most studies present a battery of robustness tests using different model specifications. These estimates are not included: only the estimates of the base or preferred model are used. Furthermore, if a discussion paper is available in addition to a journal publication, the information from the discussion paper will be used only if it provides additional information (e.g. using a different cohort or different child outcome). For example, Herbst (2013a) provides estimates on cognitive skills, whereas Herbst (2012) also includes estimated effects on motor skills. When estimates use exactly the same outcome indicator, only the final, published estimates are used.

Our final sample consists of 253 estimates obtained from 30 primary studies, published in the period 2005-2015. Various studies contain impact estimates for specific subsamples in addition to the overall (pooled sample). We collected these estimates to test whether the effects of ECEC are heterogeneous. We included impacts on children from low and high SES backgrounds (113 estimates for each group), and for blacks, Hispanics, whites (around 30 estimates for each subgroup). The race/ethnicity specific estimates are all from the US literature.

Although some meta-analyses use only one observation per primary study (e.g. Stanley, 2001), this may lead to a loss of information (Bijmolt and Pieters, 2001). It is not uncommon to use multiple estimates per study: for instance, the sample of Doucouliagos and Stanley (2009) consists of 1474 estimates obtained from 64 studies. Also in meta-analytical studies using a similar estimation method as we apply in our study ('sign and significance' ordered probit models), it is common to use multiple observations per primary study.[8]

Table 3.1 provides an overview of the studies included in the meta-analysis. The studies are ordered by 'cluster': studies from the same author using the same data source (generally the discussion paper and the published paper, e.g. Herbst (2012; 2013a) and Felfe et al. (2012; 2015)). An exception are a series of studies on a Canadian reform (e.g. Baker et al., 2008): all use the same data and exploit the same Quebec reform. The countries covered are US, Canada, Germany, France,

[8] Card et al. (2010): 199 estimates from 97 studies; Butschek and Walter (2014): 99 estimates from 33 studies; Longhi et al. (2008): 1572 estimates from 45 studies.

Norway, Spain. Some study clusters only contribute one relevant observation; most however contribute several observations.

3.2 Extraction of estimates and other information

As outcome measures (the dependent variable(s) used in the primary studies), we collected ECEC effects on cognitive test scores (math/numeracy, language), motor skills, non-cognitive skills, social-emotional development and externalizing problem behavior for different ages. Most of these indicators are measured before the age of 7. Furthermore, educational outcomes during primary and secondary school (e.g. grades, grade repetition, special educational needs) and later life outcomes (e.g. school dropout, completed education, employment state, wages) are included. As the outcomes are very different, there is no common metric. Therefore, the impacts are classified as significantly negative (coded as -1), insignificant (0) and significantly positive (1), using the 10 percent significance level ($p<0,10$).[9]

Following section 2, information on the following moderators is obtained:

- Starting age of child care: we measure whether the arrangement is used by children below the age of 3 or whether children start at the age of 3 or later. While preschools or pre-K programs generally start at the age of 4, in other programs a substantial number of babies and toddlers are enrolled.
- Intensity of the program: is the program offered on a full-time basis or on a part-time basis? We code this variable as 1 if it concerns a full-time program and 0 otherwise (part-time or the intensity varies).
- Quality: quality of ECEC is difficult to measure in general and (even more) difficult to compare across different settings. We therefore use two important quality indicators that are to a large extent comparable: educational levels of child care staff and staff-to-child ratios. We rely on the information about quality provided by the study, generally discussed in the "institutional background" section of the paper. In addition to information provided by the primary study, we use external reports to assess the quality of the specific ECEC arrangement (e.g. from the OECD or independent research institutes). Both the ratio and education dimension are scored on a 3-point scale: low (0), medium (1) or high (2): see Appendix Table A1 for the scoring scheme and Table A2 for the documentation on how we have derived these scores. The quality indicator we use in our main analyses is the sum of these scores. In addition, we complement this with a more qualitative assessment: is the arrangement considered as high-quality (by the authors or external evaluators)? Do the centers generally comply with the regulations? The qualitative

[9] We performed sensitivity tests using the 5 percent significance level instead: see section 4.3.

assessment is generally consistent with the derived quality scores except for the Canadian (Quebec) case, where quality regulations seem relatively strict but noncompliance was a major issue. In some cases it is not entirely clear which score to assign. In our main analyses we use the lower bound scores, but we test the sensitivity of the results using the upper bound.

- Timing of the effects: we distinguish in the analysis between four time periods in the child's life: immediate (during or directly after the program; i.e. including measurements at the start of kindergarten); short-term (during kindergarten, generally a year after the program); medium term (elementary and secondary school); long term (adolescent educational outcomes, wages). As an alternative, we calculate the gap in years between the age of measurement and school entry. Furthermore, in additional analyses we also distinguish between cognitive and non-cognitive outcomes, and between language and numeracy skills.
- Additional variables: we extracted information about the estimation method, whether the study is published or not and the number of observations used in the primary study.

3.2 Descriptive analysis

Table 3.2 provides the descriptive statistics of the sample. Most of the estimates are from the US and are derived from studies on ECEC arrangements for children aged 3 and older. About one third of the estimates refer to full-time programs (which are more frequently observed in the US/Canada). Their appears to be quite some variation in quality: the quality scores from Europe are higher on average. The majority of the estimates concern immediate effects, measured during or directly after the ECEC program.

Table 3.3 presents the distribution of the estimated effects. A striking feature is that the evidence on universal ECEC seems rather mixed. Although about a third of all the estimates indicates positive impacts on children´s outcomes, many estimates are insignificant and over 20 percent of the estimates is significantly negative. The impacts from Europe tend to be more favorable than those estimated for the US and Canada. Consistent with previous literature surveys, negative ECEC effects seem to be concentrated within the younger age groups. For children enrolling at age 3 or later, most effects are positive and only a small fraction is negative. Furthermore, medium term effects tend to be insignificant, whereas long term effects are generally positive. Comparing the results for the different subgroups, the more favorable outcomes are concentrated within the group of lower SES children. Most of the estimates for the subsamples of higher SES children are insignificant and negative outcomes are as likely as positive. Hispanics and children from migrant parents also gain more from ECEC than the other groups.

To further examine the overall evidence we calculated for each of the (clusters of) studies an average score: the estimates are coded as -1 (negative and significant), 0 (insignificant) or 1 (positive

and significant). The score for each study is calculated as the average of these outcomes and therefore can vary, theoretically, between -1 (only significantly negative estimates) and 1 (only significantly positive estimates). Figure 1 presents the distribution of the study average scores: although it is clear that there are more positive than negative scores, the figure also points out that there is substantial heterogeneity in the causal effect estimates. About a quarter of the scores are negative and 60 percent of the scores are positive. However, not all of the studies with positive scores provide compelling evidence: about half of the positive scores are equal or below 0,5, indicating that at least half of the estimates within the study are insignificant (or that some estimates are significantly negative).

4. Multivariate analysis

4.1 Meta-analytic model

We estimate the relation between outcomes (significantly positive, insignificant, significantly) and the relevant variables using an ordered probit model. Given that we use multiple estimates per study, we use sampling weights. Although we could use a simple weighting scheme, where the weight is equal to one divided by the number of estimates provided by the study (e.g. Horváthová, 2010), this approach seems not appropriate in our study. For instance, if both study A and B provide one estimate on short-term cognitive development but study B also provides several estimates on non-cognitive development and later school outcomes, the weight attached to the cognitive development estimate may be substantially larger for A than for B. We therefore use the following equation to calculate the weight of each estimate (see Appendix Table A3 for information on the child outcome domains):

$$\frac{1}{\text{number of study estimates within child outcome domain}} \times c$$

where *c* represents a correction factor which is equal to 1 divided by the number of studies using the same data and child outcome indicator. In addition, when subsamples are used, the factor corrects for the sample size.[10] Whereas the number of estimates per study varies between 1 and 63, the total weight per study varies between 1 and 5. This approach adjusts for the number of estimates per study, while at the same time more weight is assigned to studies that provide a more comprehensive evaluation (in terms of variety in child outcomes). We test the sensitivity of our results using alternative weighting methods.

[10] For instance, Felfe et al. (2015) provide estimates for separate cohorts and for a pooled sample: the sum of weights for the cohort-specific estimates is equal to the pooled sample estimate (N subsample/N complete sample). Similarly, Dustmann et al. (2012) provide separate estimates for native and immigrant samples.

4.2 Main results

The main estimation results from the meta-regression analyses are presented in Table 4.1. We start by separately analyzing the various dimensions of heterogeneity: ECEC features, measured outcomes (type and timing) and study features. Column 1 shows the results when three central ECEC elements are included: starting age, intensity and quality. Column 4 and 5 present models that include all dimensions simultaneously. Starting below the age of three does not lead to more favorable results: in fact, the coefficient is negative (though insignificant). Also when more controls are included (column 4 and 5) the coefficient remains negative and insignificant. Overall, there is no clear relation between ECEC enrollment age and outcomes in terms of child development. Next, the coefficient on the intensity of the arrangement (full-time or part-time) is positive but insignificant. This finding is consistent across specifications. Hence, there is no clear evidence that full-time programs lead to more positive results than part-time programs. While our findings on starting age and intensity are rather ambiguous, the results on ECEC quality are positive and highly significant ($p<0.01$). The findings presented in Table 4.1 are consistent with previous qualitative literature reviews (e.g. Melhuish et al., 2015), pointing out that the evidence on the impact of the starting age and intensity is inconsistent and that quality is a crucial determinant of the positive child care effects. A very robust finding is that quality matters: we explore the quality dimension more extensively in section 4.4.

The second dimension concerns the timing (and type) of outcome. The type of outcome is highly correlated to the timing. For example, estimates on cognitive and non-cognitive development are only available in the short-run. The models presented in Table 4.1 therefore include only timing indicators. Compared to immediate outcomes (measured while in child care), short-term, medium and long-term effects are more positive. The coefficients on the short-term and medium outcomes are significant in the model presented in column 4, but are insignificant in most other specifications. However, the coefficient indicating long-term outcomes is consistently positive. Instead of the three timing dummies, we have used the gap in years between the start of school (end of child care) and the age the outcome is measured. The results from these alternative models show that this time gap is significantly positively related to the ECEC outcomes: outcomes that are measured later in life are more favorable. The estimates in Table 4.1 point out that this is driven by the more positive long-term effects. The findings indicate that there may be fading out of the effects after entering school (though we find no evidence of this), but that there may be sleeper effects and the ECEC impact may be more positive in the long run. It should however be noted that only few studies provide information on long-term effects. Next, additional analyses show that in terms of different development domains there are little differences: for instance, it appears that effects in the non-

cognitive domain are not significantly more positive (or negative) than in the cognitive domain. However, we do find that ECEC has relatively large positive effects on the child´s language skills.

The third cluster of moderators we explore are the study features. Concerning the estimation method, IV and DiD do not lead significantly different results in terms of child outcomes. The results of RDD estimations are significantly more positive. However, all of the RDD estimates are from relatively high quality pre-kindergarten programs (most of the estimates are from the Tulsa pre-K program). The final column shows there is no clear indication of publication bias. In addition, the number of observations used in the study is not significantly related to the child outcome. If sample size matters, we may expect larger positive t-statistics and larger negative t-statistics. In that case, significantly positive and significantly negative outcome are more likely. However, the predicted relation between the sample size and the outcome variable in ordered probit models is not clear. We discuss this issue in more detail in the next subsection.

Overall, the findings from the meta-regressions suggest that starting age and intensity do not matter significantly, but that the quality is a crucial determinant of the effects of ECEC. Moreover, effects tend to be more favorable when long run outcomes are measured.

4.3 Sensitivity checks

We performed a series of tests to examine the robustness of our main results. First, we applied alternative weighting schemes: each estimate is weighted equally (i.e. no weighting) or each study is weighted equally (estimate weights are defined as 1 divided by the number of estimates provided by the primary study). Second, we used the 5 percent rather than the 10 percent level to determine whether an estimate from a primary study is significantly positive or significantly negative. This of course changes the distribution of the outcome variable.[11] Appendix Table A4 presents the results of these sensitivity analyses. The estimates are generally consistent with the main results: quality matters and long run outcomes are relatively favorable. The exception seems to be the coefficient on the publication status, which is negative and significant. Furthermore, when clustering the estimates we allowed for alternative clusters (for instance, the studies evaluating the Quebec child care reform have been classified as a single cluster). This did not significantly affect the results.

Next, we examined to what extent country (institutional) effects drive our results. An important source of variation is due to different institutional settings. Because this is also to variation we aim exploit, we did not include country or region controls in our main meta-regression models. However, if we include a set of country/region dummies, we obtain similar results (see Appendix Table A5, column 1 and 2). As an additional test, we performed the analyses on the US

[11] 51 percent of the outcomes is insignificant, while 19 and 30 percent are significantly negative and positive respectively.

sample only (Appendix Table A5, column 3 and 4). The results using the US subsample show that quality is important and that long term effects are relatively more positive. However, most of the results are stronger. There is, for instance, some indication that fulltime arrangements are more positive. The most interesting result concerns the estimates on timing. The results show again that longer term effects are more favorable. However, the short term effects are more negative than the immediate effect, indicating that positive effects may fade out shortly after entering school. This is consistent with the results from Perry Preschool and recent evidence from the Head Start Impact Study. Our results suggest that a negative evaluation of Head Start - based on the finding that effects fade out during the first grades - is rather premature (see Gibbs et al., 2011).

Following Card et al. (2010), we evaluate our meta-regression model (ordered probit model) by estimating probit models: see Appendix Table A6. The dependent variable of these models is a dummy indicating a significantly negative (column 1 and 2) or a significantly positive (column 3 and 4). If the ordered model is accurate, the estimated coefficients on the negative outcomes should have the opposite sign as the coefficients from the ordered probit model, while the coefficients on the positive outcomes should have the same sign as the coefficients from the ordered probit model. Although there are some inconsistencies, overall the results from the probit models are in line with our main results. The main difference is the significantly negative coefficient of "medium term" in the negative outcome model. This discrepancy is probably due to the fact that this model cannot fully control for estimation method (due to multicollinearity).[12] Another test of the ordered probit model is based on evaluating the coefficients on the (square root of the) sample size in the probit models: they appear to be small and insignificant. This indicates the mechanical effect (a larger sample size translates into larger negative and positive t-statistics) is offset by research design: when larger datasets are available, scholars tend to use more complicated estimation techniques, which generally decrease the chance of finding a significant effect. The finding that the sample size is not related to finding a positive or negative effect indicates that our ordered probit model is correct (see Card et al. (2010) for a more extensive discussion on this test).

4.4 Analyzing the role of quality

The analyses presented so far consistently point out that ECEC quality is important. In this subsection we explore the effects of quality in more detail: see Table 4.2 for the estimation results. First, in panel A the results are shown from re-estimated models using a dummy indicating high-quality child care (scoring 3 or 4). The coefficient of this dummy is consistently positive and

[12] More specifically, results from RDD studies are never negative and only estimate ECEC effects on immediate children's development.

significant. Second, we explore the role of staff-child ratios and educational requirements separately (panel B). The coefficients of both quality dimensions are positive across specifications, but the coefficients on the staff educational requirements score are positive in the three estimated models. This suggest that educational standards are a more important confounding factor for the effects of ECEC than staff-child ratios. Third, in panel C we distinguish between three different high-quality arrangements: those with high staff-child ratios standards but medium educational requirements (total score 3); with medium staff-child ratios standards, but high educational requirements (total score 3); and arrangements that score high on both dimensions (score 4). The coefficients of the two score 3 arrangements are positive but not consistently significant in all specifications (although they are always jointly significant). It is not clear from these estimations that educational standards are more relevant than regulations on staff-child ratios: they both seem to matter. However, a very robust finding is that arrangements that score high on both dimensions lead to more favorable child outcomes.

4.5 Heterogeneous effects

Previous literature suggests that the effects of ECEC may be heterogeneous. In some cases, gains are concentrated within specific group, without affecting other groups. In other cases, some groups may gain and others lose (Havnes and Mogstad, 2015). We explore this issue here, focusing on subsample estimates from different socio-economic groups and different racial and ethnic groups: see Table 4.3. Socio-economic status is frequently measured using the education of the mother (e.g. high school completed or not) or the income level of the father (e.g. above or below the median level). The upper part of table presents three different models comparing the effects on children from different socio-economic backgrounds: the estimates in column 1 and 2 are based on the subset of primary studies that provide SES-specific estimates, while all primary studies are used to estimate the results presented in column 3. In all three specifications, the impact of ECEC is less favorable for higher SES groups (i.e. they are less likely to gain or more likely to lose from ECEC).

The lower panel of Table 4.3 Panel B shows the estimation results using race and ethnicity specific subsamples (black, Hispanic, white). This type of estimates is available for only a small subset of our studies: all of these are on rather high quality prekindergarten programs. In column 1 we included only this subset (using the “white” subgroup as the reference category), in column 2 we added estimates from other US prekindergarten programs and in column 3 all estimated effects are included. The results indicate that Hispanic children are more likely to benefit for child care arrangements than the other groups. This may be the results of a language development effect

(Hispanic children may have Spanish speaking, non-US born parents).[13] This is consistent with our finding (section 4.2) that child care is relatively positive for language development (compared to other skills). Overall, we interpret the results on the different subsamples as evidence that the benefits of ECEC are concentrated within more disadvantages children and Hispanics. Children from higher socio-economic status parents are less likely to gain from ECEC.

4.6 Perry Preschool as a benchmark

The promising results from Perry Preschool have to a large extent influenced the policy debate on Early Childhood Education and Care. But how do the results of this small scale, 1960's model program compare to the evidence on universal child care schemes? To address this question, we extracted 132 Perry effect estimates from two recent re-evaluations (Heckman et al., 2010;2011).
We focus on the development and other human capital indicators (i.e. excluding the results on crime and family formation): the two re-evaluations include a large variety of outcomes, from age 4 IQ to age 40 wages. About 30 percent of these estimates are significantly positive; the other estimates are insignificant. Given the small sample size, these results are remarkable (for instance, the results for males are based on 72, with less than half receiving the Perry treatment).

We again estimate the ordered probit models for sign and significance, using the same weighting scheme for the Perry estimates as for the estimates on universal schemes (Table A3). The results from Perry serve as a benchmark (reference category). Table 4.4 presents the results: column 1 and 3 include only the subset of US estimates, while column 2 and 4 include estimates from other countries as well. Considering the results using our main (pooled) sample (column 1 and 2), there is no indication that Perry significantly outperforms universal schemes. In fact, results from high quality programs are significantly more favorable than the results from Perry. Comparing the results of Perry with the estimates on the low SES group (the group that is generally targeted by ECEC interventions), there again appears to be no significant difference on average (panel A). The results are highly dependent on the quality of the universal arrangement (panel B): low quality universal schemes produce significantly worse outcomes than Perry, whereas high quality universal arrangements seem to perform better in terms of children's outcomes. These findings point out that quality matters significantly, also for the outcomes of lower SES children.

[13] In addition, we performed analyses using (the small number of) available estimates on the migrant subsamples. Although these results should be interpreted cautiously, the estimations suggest that migrants benefit more from ECEC than natives.

5. Conclusion

This study synthesizes the main lessons from the recent 'natural experiment' research on Early Childhood Education and Care, using meta-analytical techniques. Although it is frequently claimed that child care arrangements improve child development and lead to positive outcomes in the long run, the overall evidence on universal ECEC is quite mixed: about a third of all the estimates indicates positive impacts on children´s outcomes, many estimates are insignificant and about one out of five estimates is significantly negative. This study examined what explains the heterogeneity in results.

Results from our meta-regressions showed that both the age of enrollment and the intensity (full-time versus part-time) are not important confounding factors. One of the most robust findings of this study is that quality matters significantly: across many different specifications, using different samples and controls, and measuring child care quality in different ways, high quality ECEC arrangements consistently produce more favorable outcomes. Furthermore, there is no evidence that the effects of ECEC fade out. Interestingly, our results from the US show that the effects may fade out within the first years of school, but that the long run effects (in terms of completed education and labor market success during adulthood) are more favorable than the immediate effects (measured during early childhood). The results show that the gains of ECEC are concentrated within the group of disadvantaged children. Descriptive analysis suggest that for higher SES children, ECEC is about as likely to be beneficial as it is to be harmful.

The results have important policy implications. Given that a large share of the effects are insignificant and some are negative, and that this is to a significant degree driven by low to mediocre quality, it is imperative not to compromise on quality. It is unlikely that the gains from child care investments can be realized when quality levels are insufficient: child care may even be harmful. This may be especially be a risk when child care arrangements are used mainly as a labor market policy to raise (female) employment levels. The relevance of quality does fit well with the current policy consensus: Obama's preschool for all is about expanding access to high-quality ECEC and at the EU more attention is shifted from increasing coverage - to increase female labor force participation and gender equality - towards improving quality levels – to improve child development and well-being. Our results suggest that a partial child care investment strategy, increasing coverage without setting high quality standards, may be ineffective or even counterproductive. If policy makers do not prefer to spend substantial resources on high quality child care, it may be more effective to focus on targeted interventions.

References

Baker, M., Gruber, J., and Milligan, K. (2008). Universal Child Care, Maternal Labor Supply, and Family Well -Being. Journal of Political Economy, 116(4), 709-745.

Bernal, R., and Keane, P. (2011). Child Care Choices and Children's Cognitive Achievement: The Case of Single Mothers. Journal of Labor Economics, 29(3), 459-512.

Blanden, J., Del Bono, E., McNally, S., and Rabe, B. (2015). Universal Pre-School Education: The Case of Public Funding With Private Provision. CEP Discussion Paper No 1352, London.

Dumas, C., and Lefranc, A. (2010). Early schooling and later outcomes : Evidence from pre-school extension in France. Thema Working Paper No 2010-07, Cergy Pontoise.

Dustmann, C., Raute, A., and Schönberg, U. (2012). Does Universal Child Care Matter? Evidence from a Large Expansion in Pre-School Education. Working paper.

Drange, N., and Havnes, T. (2015). Child Care Before Age Two and the Development of Language and Numeracy: Evidence from a Lottery. IZA Discussion Paper No. 8904, Bonn.

Felfe, C., and Lalive, R. (2010). How does Early Child Care affect Child Development? Learning from the Children of German Unification. Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Economics of Child Care and Child Development, No. B11-V2

Felfe, C., and Lalive, R. (2013). Early Child Care and Child Development: For Whom it Works and Why. SOEPpapers on Multidisciplinary Panel Data Research No. 536.

Felfe, C., and Lalive, R. (2014). Does Early Child Care Help or Hurt Children's Development? IZA Discussion Paper No. 8484, Bonn.

Felfe, C., Nollenberger, N., and Rodríguez-Planas, N. (2012). Can't Buy Mommy's Love? Universal Childcare and Children's Long-Term Cognitive Development. IZA Discussion Paper No. 7053, Bonn.

Felfe, C., Nollenberger, N., and Rodríguez-Planas, N. (2015). Can't Buy Mommy's Love? Universal Childcare and Children's Long-Term Cognitive Development. Journal of Population Economics, 28, 393–422.

Fitzpatrick, M.D. (2008). Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement. The B.E. Journal of Economic Analysis & Policy Advances, 8(1), 1-38.

Gormley, W.T., and Gayer, T. (2005). Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program. Journal of Human Resources, 40(3), 533-558.

Gormley, W.T., Gayer, T., Phillips, D., and Dawson, B. (2005). The Effects of Universal Pre-K on Cognitive Development. Developmental Psychology, 41(6), 872–884.

Gormley, W.T., and Phillips, D. (2005). The Effects of Universal Pre-K in Oklahoma: Research Highlights and Policy Implications. Policy Studies Journal, 33(1), 65-82.

Gormley, W.T. (2008). The Effects of Oklahoma's Pre-K Program on Hispanic Children. Social Science Quarterly, 89(4), 916-936.

Haeck, C., Lefebvre, P., and Merrigan, P. (2013). Canadian Evidence on Ten Years of Universal Preschool Policies: the Good and the Bad. Working Paper 13-34, Montréal.

Haeck, C., Lefebvre, P., and Merrigan, P. (2015). Canadian Evidence on Ten Years of Universal Preschool Policies: the Good and the Bad. Labour Economics, forthcoming.

Havnes, T., and Mogstad, M. (2010). Is Universal Child Care Leveling the Playing Field? Evidence from Non-Linear Difference-in-Differences. IZA Discussion Paper No. 4978, Bonn.

Havnes, T., and Mogstad, M. (2011). No Child Left Behind: Subsidized Child Care and Children's Long-Run Outcomes. American Economic Journal: Economic Policy, 3(2), 97-129.

Havnes, T., and Mogstad, M. (2015). Is universal child care leveling the playing field? Journal of Public Economics, 127, 100-114.

Herbst, C.M. (2012). The Impact of Non-Parental Child Care on Child Development: Evidence from the Summer Participation "Dip". IZA Discussion Paper No. 7039, Bonn.

Herbst, C.M. (2013a). The Impact of Non-Parental Child Care on Child Development: Evidence from the Summer Participation "Dip". Journal of Public Economics, 105, 86–105.

Herbst, C.M. (2013b). Universal Child Care, Maternal Employment, and Children's Long-Run Outcomes: Evidence from the U.S. Lanham Act of 1940. IZA Discussion Paper No. 7846, Bonn.

Herbst, C.M., Tekin, E. (2010a). Child care subsidies and child development. Economics of Education Review, 29, 618–638.

Herbst, C.M., Tekin, E. (2010b). The Impact of Child Care Subsidies on Child Well-Being: Evidence from Geographic Variation in the Distance to Social Service Agencies. IZA Discussion Paper No. 5102, Bonn.

Herbst, C.M., Tekin, E. (2015). The Impact of Child Care Subsidies on Child Well-Being: Evidence from Geographic Variation in the Distance to Social Service Agencies. Journal of Policy Analysis and Management, forthcoming.

Kottelenberg, M.J., and F. Lehrer, S,F, (2013). New Evidence on the Impacts of Access to and Attending Universal Child-Care in Canada. Canadian Public Policy, 39(2), 263- 285.

Lefebvre, P., Merrigan, P., and Verstraete, M. (2008). Childcare Policy and Cognitive Outcomes of Children: Results from a Large Scale Quasi-Experiment on Universal Childcare in Canada. CIRPEE Working Paper No 08-23, Montréal.

Magnuson, K.A., Ruhm, C., and Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? Economics of Education Review, 26, 33-51.

Peisner-Feinberg, E. S., Schaaf, J. M., LaForett, D. R., Hildebrandt, L.M., and Sideris, J.(2014). Effects of Georgia's PreK Program on children's school readiness skills: Findings from the 2012 –2013 evaluation study. Chapel Hill: The University of North Carolina, FPG Child Development Institute.

Wong, V.C., Cook, T.D., Barnett, W.S., and Jung, K. (2008). An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs. Journal of Policy Analysis and Management, 27(1), 122–154.

Table 3.1 – Overview of studies

Study	Country	Data	Outcome measure	Est. method	# estimates [weight]	Average score
Baker et al. (2008); Haeck et al. (2013; 2015); Lefebvre et al. (2008); Kottelenberg & Lehrer (2013)	Canada	NLSCY	1,2,3	DID; IV (K&L)	44 [3]	-0,27
Bernal & Keane (2011)	US	NLSY79	1	IV	1 [1]	0
Blanden et al. (2015)	England	NPD	1,2,3,4,6	DID	15 [5]	0,47
Drange & Havnes (2015)	Norway	Admin. Oslo/ Statistics Norway	1	IV	5 [1]	1
Dumas & Lefranc (2010)	France	DEPP & FQP	6,7	IV	3 [2]	0,67
Dustmann et al. (2012)	Germany	Admin. SEE	1,3	IV	10 [2]	0,3
Felfe & Lalive (2010; 2013)	Germany	GSOEP & GCP	1,2,3,5,6	IV	33 [5]	0,42
Felfe & Lalive (2014)	Germany	SEE data	1,2,3	IV	4 [3]	0,5
Felfe et al. (2012; 2015)	Spain	PISA	4 [x2],6,7	DID	8 [4]	0,5
Fitzpatrick (2008)	US (Georgia)	NAEP	4,6	DID	3 [2]	0
Gormley et al. (2005)	US (Tulsa, OK)	TPS (2003)	1	RDD	3 [1]	1
Gormley & Gayer (2005); Gormley & Phillips (2005)	US (Tulsa, OK)	TPS (2001)	1,2,3	RDD	5 [3]	0,8
Gormley (2008)	US (Tulsa, OK)	TPS (2006)	1	RDD	3 [1]	1
Havnes & Mogstad (2010; 2011; 2015)	Norway	Statistics Norway	7,8,9	DID	10 [2]	0,4
Herbst (2013b)	US	Census ('70-'90)	7,8,9	DID	21 [3]	0,81
Herbst (2012; 2013a)	US	ECLS-B	1,3	IV	2 [2]	-1
Herbst & Tekin (2010a; 2010b; 2015)	US	ECLS-K	1,2,3,4,5	IV	63 [5]	-0,51
Magnuson et al. (2007)	US	ECLS-K	1,2	IV	4 [2]	0
Peisner-Feinberg et al. (2014)	US (Georgia)	Collected– Georgia Pre-K evaluation	1,2,3	RDD	10 [3]	0,8
Wong et al. (2008)	US (OK; WV)	Collected– NIEER (04)	1(x2 states)	RDD	6 [2]	0,5

Table 3.2 - Sample descriptive statistics

		All	US/Canada	Europe
1.	Number of estimates	253	165	88
2.	Age enrolment			
	Age 3+	64.82	71.52	52.27
	Age below 3	35.18	28.48	47.73
3.	Intensity			
	Part-time/varies	66.01	54.55	87.50
	Full-time	33.99	45.45	12.50
4.	Quality			
	Low quality (score<3)	54.55	81.82	3.41
	High quality (score≥3)	45.45	18.18	96.59
5.	Timing measurement			
	Immediate	52.96	58.79	42.05
	Short term	15.81	13.94	19.32
	Medium term	17.79	14.55	23.86
	Long term	13.44	12.73	14.77
6.	Estimation method			
	DiD	37.55	37.58	37.50
	IV	54.15	49.70	62.50
	RDD	8.30	12.73	-
7.	Published	45.45	59.39	19.32
	Unpublished	54.55	40.61	80.68

Table 3.3 - Estimated effects of ECEC

			Significantly negative	Insignificant	Significantly positive
		N			
1.	All	253	21.74	43.48	34.78
2.	Countries				
	US	121	28.93	38.84	32.23
	Canada (Quebec)	44	36.36	54.55	9.09
	Germany	47	4.26	51.06	44.68
	Other EU	41	4.88	36.59	58.54
3.	Age enrolment				
	Age 3+	164	21.95	40.24	37.80
	Age below 3	89	21.35	49.44	29.21
4.	Intensity				
	Part-time/varies	167	23.35	44.91	31.74
	Full-time	86	18.60	40.70	40.70
5.	Quality				
	Low quality (<3)	135	36.96	45.65	17.39
	High quality ($\geq$3)	118	3.48	40.87	55.65
6.	Timing				
	Immediate	134	27.61	41.79	30.60
	Short term	40	32.50	35.00	32.50
	Medium term	45	6.67	73.33	20.00
	Long term	34	5.88	20.59	73.53
7.	Estimation method				
	DiD	95	13.68	46.32	40.00
	IV	137	30.66	45.99	23.36
	RDD	21	0	14.29	85.71
8.	Publication status				
	Unpublished	138	12.32	41.30	46.38
	Published	115	33.04	46.09	20.87

Figure 1 – Distribution of mean study outcomes

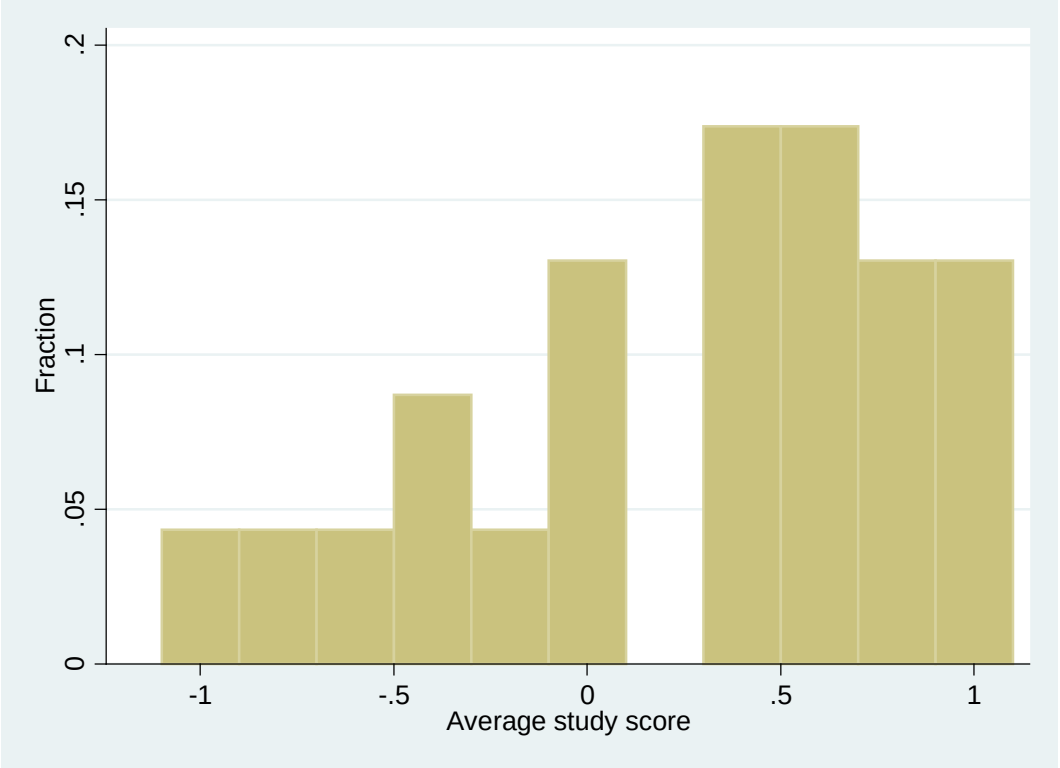

Table 4.1 - Main results

	(1)	(2)	(3)	(4)	(5)
ECEC FEATURES					
Age below 3	-0.468			-0.148	-0.230
Ref: age 3+	(0.300)			(0.400)	(0.429)
Fulltime	0.558*			0.258	0.126
Ref: Part-time/varies	(0.330)			(0.273)	(0.427)
Quality score (0-4)	0.436***			0.380***	0.340***
	(0.126)			(0.119)	(0.112)
TIMING EFFECTS					
Short term		0.706		1.019**	0.795
		(0.442)		(0.490)	(0.484)
Medium term		0.212		0.558*	0.517
		(0.309)		(0.297)	(0.348)
Long term		0.989**		1.511**	1.392**
Ref: Immediate		(0.408)		(0.692)	(0.614)
STUDY FEATURES					
Estimation method: IV			-0.583	0.104	0.0127
			(0.366)	(0.262)	(0.530)
Estimation method: RDD			0.639**	1.178***	1.164**
			(0.282)	(0.356)	(0.561)
Ref: DiD					
Published					-0.476
					(0.404)
Sample size (square root)					9.56e-07
					(0.000598)
Pseudo R2	0.154	0.0487	0.0653	0.226	0.240
Log likelihood	-44.17	-49.64	-48.78	-40.40	-39.68

Clustered standard errors in parentheses. The estimated are based on 253 estimates from 30 studies (23 clusters of studies).
*** p<0.01, ** p<0.05, * p<0.1

Table 4.2 - The role of quality

		(1)	(2)	(3)
A.	High quality	0.872**	1.050***	0.792**
	(dummy: score≥3)	(0.389)	(0.318)	(0.377)
B.	Staff: child ratio score	0.135	0.326	0.275
		(0.221)	(0.277)	(0.311)
	Educ. requirements score	0.715***	0.603*	0.477*
		(0.217)	(0.338)	(0.289)
C.	Ratio[High]xEduc[Medium]	0.562	1.012**	0.657
		(0.438)	(0.463)	(0.432)
	Ratio[Medium]xEduc[High]	0.913**	0.567	0.684
		(0.395)	(0.558)	(0.443)
	Ratio[High]xEduc[High]	1.379***	1.684***	1.644**
		(0.424)	(0.463)	(0.660)
	Ref: quality scores<3			
	Controls: features and timing	NO	YES	YES
	Controls: study characteristics	NO	NO	YES

Clustered standard errors in parentheses. The estimated are based on 253 estimates from 30 studies (23 clusters of studies).
*** p<0.01, ** p<0.05, * p<0.1

Table 4.3 - Effects for different subsamples

		(1)	(2)	(3)
A.	High SES	-0.569**	-0.687***	-0.679***
		(0.270)	(0.221)	(0.252)
	Low SES	-	-0.0139	0.0240
		-	(0.225)	(0.237)
	Nr of estimates	226	365	479
	Nr of clusters	12	12	23
	Pseudo R2	0.0346	0.215	0.207
	Log likelihood	-17.42	-34.95	-56.73
	Reference group	Low SES	Pooled	Pooled
	Controls	NO	YES	YES
	Studies included	SES subs. available	SES subs. available	All
B.	Black	0.806***	0.191	0.124
		(0.0391)	(0.388)	(0.322)
	Hispanic	1.597***	0.959*	0.721*
		(0.372)	(0.523)	(0.397)
	White	-	-0.584	-0.383
		-	(0.379)	(0.355)
	Nr of estimates	93	120	343
	Nr of clusters	4	6	23
	Pseudo R2	0.169	0.0683	0.235
	Log likelihood	-4.678	-12.31	-44.44
	Reference group	White	Pooled	Pooled
	Controls	NO	NO	YES
	Studies included	Race/ethnic. subs. available	US Pre-K	All

Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 4.4 - Perry Preschool versus universal arrangements

		Main sample		Low SES sample	
		(1)	(2)	(3)	(4)
A.	Universal arrangement	-0.0770	0.108	-0.0631	0.338
		(0.354)	(0.201)	(0.312)	(0.250)
	Ref: Perry Preschool				
B.	Universal – low quality	-0.700	-0.695	-1.978***	-1.739***
		(0.609)	(0.445)	(0.385)	(0.295)
	Universal – high quality	0.620**	0.559***	0.505***	0.458***
		(0.267)	(0.123)	(0.0658)	(0.0704)
	Ref: Perry Preschool				
	Nr of estimates	253	385	175	245
	Nr of clusters	13	25	7	14
	Sample	US	All	US	All

Clustered standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix

Table A1 – Quality scores

Score	Staff-child ratio	Educational requirement staff
0 [Low]	Lower requirements / substantial variation	Lower requirements / substantial variation
1 [Medium]	Age≥3: 1:11 – 1:15 or better; Age<3: 1:10 or better	At least vocational education in ECE or substantial share of programs with Bachelor degree teachers
2 [High]	Age≥3: 1:10 or better; Age<3: 1:8 or better	Bachelor degree required

Table A2 - An assessment of ECEC quality

Study	Country (region), type of ECEC	Criteria	Evaluation	Score [upper bound]
Baker et al. (2008); Haeck et al. (2013; 2015); Lefebvre et al. (2008); Kottelenberg & Lehrer (2013)	Canada (Quebec), child care	Staff-child ratio	• 1:10 (age 4-5) • Problems of non-compliance (see Qualitative assessment)	(2-1=)1
		Educ. requirement staff	• Not-for-profit centers: 2/3 qualified staff • For-profit centers: 1/3 qualified staff • Qualified defined as: university diploma; university level training in ECE; post-secondary "non-university" degree; secondary school diploma; vocational training in ECE • Problems of non-compliance (see Qualitative assessment)	(1-1=)0
		Qualitative assessment	Overall: low-medium quality (estimated score<3) • "The transition to the new system created frictions… The Family Policy also emphasized an increase in the quality of care, in part through regulatory changes. Formal qualifications were raised for both CPE and home-based caregivers... An audit study based on detailed site visits by Japel, Tremblay, and Côté (2005) indicates that quality in the public child care centers exceeded quality in private options, but still only 61 percent of the public facilities met their criteria for minimum quality." (Baker et al., 2008) • "in 2001: 42% of not-for-profit centers do not respect the ratio of two out of three 'qualified' educators; 25% of for-profit centers do not respect the very less stringent ratio of 1/3; overall, 40% of educators have no specific qualification in ECEC" (Lefebvre et al., 2008) • "...Japel, Tremblay, and Côté (2005) found that in Quebec after the policy was implemented, only 5 percent of programs were high quality and the majority of the arrangements scored below the mid-range with slightly over 10 percent classified at below minimum quality" (Kottelenberg & Lehrer, 2013) • "two major studies on the first years of the program (ISQ, 2004; Japel et al., 2005) showed that the average quality in Québec's subsidised daycare network was at best satisfactory and in many cases low or not acceptable, particularly for children in lower income families... The audit [conducted by the Office of the Auditor General of Québec (VGQ, 2011)] also finds that the percentage of subsidised centres not respecting the maximum ratio of number of children per qualified educator was 42% during 2008–2009 and 54% during 2009–2010...The lack of supply […] eventually forced the government to create spaces at a very rapid rate, with an evident lack of trained personnel." (Haeck et al., 2015)	
		Total		1
Bernal & Keane (2011); Herbst (2012; 2013a); Herbst & Tekin (2010; 2015)	US, child care	Staff-child ratio	Varies: see Qualitative assessment	0
		Educ. requirement staff	Varies: see Qualitative assessment	0
		Qualitative assessment	Overall: large variation, mediocre (estimated score<2) • "Recent studies suggest, however, that average quality in U.S. child care settings tends to be mediocre and highly variable (Helburn et al., 1995; NICHD, 2000a; Phillips & Adams, 2001; Vandell & Wolfe, 2000)." (Herbst & Tekin, 2010)" (Herbst & Tekin, 2010) • "Studies by the NICHD Early Child Care Research Network (2000a) estimate that 42 percent of preschool child care settings are "poor" or "fair" quality, and that positive caregiving is "highly characteristic" for only 12	

			percent of children. A review of child care settings by the National Research Council appears to corroborate this, finding that 10 percent to 20 percent of early care and education environments are "inadequate" and pose serious risks to child development (National Research Council & Institute of Medicine, 2000)." (Herbst & Tekin, 2010) • "a key feature of the CCDF is the principle of "parental choice…Although this design feature increases flexibility for working parents, it also means that children may be exposed to low-quality care." (Herbst & Tekin, 2015) • "by mandating only minimum quality standards, the CCDF reduces the incentive for providers to invest in costly quality improvements that promote child development." (Herbst & Tekin, 2015)	
		Total		0
Blanden et al. (2015)	England, child care	Staff-child ratio	Age 3 and older: • Nursery schools/public provision arrangements: 1:13 • Private sector: 1:13 or 1:8 if no qualified teacher is present	1 1-2
		Educ. requirement staff	• Nursery schools/public provision arrangements: qualified teacher • Private sector: managers and supervisory staff require level 3 ECE qualification; at least 50% of the other staff needs to hold at least a level 2 qualification; in about 40% of the cases a teacher is present (see Qualitative assessment)	2 1-2
		Qualitative assessment	Overall: medium-high quality (estimated score 2-3) Quality assessment nursery schools/public provision: 3 Quality assessment private sector: if no qualified teacher is present: 3 (2+1); if qualified teacher is present: 3 (1+2) • Results are driven by expansion in the private sector: "A unique feature of the free entitlement in England is that the expansion relied exclusively on private settings to provide the new places." (Blanden et al., 2015) • Notes on educational requirements: o "40% of children in private nurseries have a teacher present, compared to 100% in public nurseries (Gambaro et al., 2013)" (Blanden et al., 2015) o "Level 2 and 3 qualifications are achieved after 1 or 2 years of post-compulsory school training, which often can be on-the-job training, and attract those with the poorest academic records (Nutbrown, 2012)" (Blanden et al., 2015) • "Public nurseries also have higher quality based on detailed observation of classroom practice and adult-child interactions (Sylva et al., 2004)." (Blanden et al., 2015) • "Both complier and non-complier LEAs might have experienced improvements in childcare quality at the same time, as providers registered for funding status and had to comply with quality regulations." (Blanden et al., 2015) • Ofsted performs inspections and enforces standards, however "[t]here has been some criticism of the regulator focus on health, safety and environment rather than pedagogical quality (National Audit Office, 2004; Mathers et al., 2012)." (Blanden et al., 2015)	
		Total		3
Drange & Havnes (2015)	Norway (Oslo), child care (before age 2), 2005-2010	Staff-child ratio	age 0-3: 1:3 (staff) 1:10 (teacher)	2
		Educ. requirement staff	• pre-school teacher: college degree (one year, including supervised practice in a formal child care institution) • assistant teacher: no educational requirements	2
		Qualitative assessment	Overall: relatively high quality (estimated score 3-4) • "Child care in Norway is heavily regulated, with provisions on staff qualifications, number of children per adult and per teacher, size of play area, and educational orientation… In Oslo, about 60 % of child care institutions are public… Both public and private institutions require municipal approval and supervision…" (Drange & Havnes, 2015) • Average ratios in child care institutions: teacher/children=0,077; staff/children=0,298 (Drange & Havnes, 2015)	

			• "For children 0-3 years, the ratio is 7-9 children per trained pre-school pedagogue when children attend more than six hours per day." (OECD, 2006) • "Heads and pedagogues in ECEC have 3-year tertiary level training at one of the state university colleges or private colleges. There is no formal qualification requirement for assistants, who make up the bulk of the staff in direct contact with children; although an increasing number holds either secondary vocational or tertiary diplomas." (OECD, 2006)	
		Total		4
Dumas & Lefranc (2010)	France, preschool (école maternelle), 1972-1980	Staff-child ratio	In the case of école maternelle: No national regulations on staff-child ratios for école maternelle. Average class size: 25 (Dumas & Lefranc, 2010) – 30 in 1980 (OECD, 2006). In addition, there is generally a teacher assistant present, suggesting a ratio of 1:12,5 – 1:15	0-1
		Educ. requirement staff	Teachers hold a bachelor's degree	2
		Qualitative assessment	Overall: medium-high quality (estimated score 2-3) • "Preschool in France is centrally administered by the ministry of Education and is to a very large extent offered within public schools. The stated objective of école maternelle is to help children reach autonomy and acquire knowledge and skills in order to promote their readiness for elementary school. To reach these goals, preschool follows a standardized and integrated curriculum, for a duration of three years." (Dumas & Lefranc, 2010) • "Preschool teachers are national civil servants and receive the same level of training as primary school teachers, typically a bachelor's degree level." (Dumas & Lefranc, 2010)	
		Total		2 [3]
Dustmann et al. (2012) Felfe & Lalive (2010; 2013); Felfe & Lalive (2014)	Germany, child care	Staff-child ratio	Requirements vary and are set at the state (Bundesländer) level: Average staff-child ratios: • Age 0-3: 1:5,13 (West), 1:6,76 (East) (Felfe & Lalive, 2010) • In West German states: 1:2,8-5,1 (Felfe & Lalive, 2013) • Age 0-2: 1:6,4; Age 3-5: 1:10 (CESifo DICE report, 2010)	2
		Educ. requirement staff	Requirements vary and are set at the state (Bundesländer) level: West: 84 percent of the staff in the East has a specialized degree (Felfe & Lalive, 2010) East: 90 percent of the staff in the East has a specialized degree (Felfe & Lalive, 2010)	1
		Qualitative assessment	Overall: varies, but relatively high quality (estimated score 3-4) • "Regions in the West and regions in the East of Germany are on average quite similar with respect to formal criteria such as the tightness of their regulations concerning maximal size of the group of children, staff-child ratios, availability, further education, size of the child care center, etc." Felfe & Lalive, 2010) • "States (Bundesländer) are in charge of regulating the quality of center-based care in Germany…centers have a clear educational mission… State regulations regarding the share of qualified staff are rather lax. Yet, across all West German states, the majority of the pedagogical staff possesses of an educational degree [varies between 80-95%]." (Felfe & Lalive, 2013) • "Playgroups can have at most ten children and need to be supervised by at least one certified education specialist and one or two assistants. The degree required to work as a group leader in a child care center requires two years of theoretical training and at least two years of practice in a child care center. Care centers comply with these regulations: in 2006 and 2007 groups accommodated on average 10.2 children, 62.3% of the employed staff had a degree in early childhood education, and the ratio of children to staff was about 3:1." (Felfe & Lalive, 2014) • "centers offer high-quality and low-cost care for young children…The German context is also unusual in maintaining a fairly high level of quality while expanding." (Felfe & Lalive, 2014) • "a program that is aimed at 3- to 6-year-olds for the case of Germany, where the quality of care is high, fairly homogenous and almost exclusively public… In Germany, formal child care provision of children aged between 3 and 6 is mostly public. Child care provision is further characterized by strict country-wide quality standards: in	

			all child care institutions, the teacher-child ratio must not exceed 2 teachers for 25 children. Moreover, teachers must have completed a three-year vocational programme, and are certified by the state. Further regulations exist regarding the space provided by the institution for each child. As a consequence, the quality of child care is relatively homogeneous... In terms of educational content, Germany follows the social pedagogy tradition..." (Dustmann et al., 2012) • 64% of the child care staff are Erzieherinnen ("kindergarten pedagogues"), who have received three years of vocational training (including a one year internship) (OECD, 2006)	
		Total		3
Felfe et al. (2012; 2015)	Spain, child care	Staff-child ratio	National standards: 1:8 (age below 1); 1:13 (age 1); 1:20 (age 2); 1:20 (age 3); 1:20 (age 4-6) (Moss, 2000; Felfe et al., 2015) Actual levels: 1:13,9 (OECD, 2010); 1:10 (CESifo, 2010)	1
		Educ. requirement staff	Preschool teachers are required to hold a college degree in pedagogy	2
		Qualitative assessment	Overall: relatively high quality (estimated score: 3-4) • "This reform implied universal access to high-quality public childcare for all 3-year-olds...LOGSE also provided federal provisions for the first time in Spain regarding educational content, group size, and staff skill composition for children 3 to 5 years old, regardless of ownership status." (Felfe et al., 2015) • Psycho-educational theories served as guidelines for the design of the curriculum. (Felfe et al., 2015) • "LOGSE established the maximum number of students per class to be 20 for 3-year-olds and 25 for 4- and 5-year-olds...classes are grouped based on the year in which children were born and thus are not mixed in ages." (Felfe et al., 2015) • "The post-LOGSE early childhood teacher has a three year post-18 training to degree level in a university-based teacher training institute; the training, pay and status is at the same level as for primary school teachers." (Moss, 2000)	
		Total		3
Fitzpatrick (2008)	US (Georgia), prekindergarten 1994-2005	Staff-child ratio	1:10	2
		Educ. requirement staff	• a diploma in early childhood education or an associate's degree • however 80 percent of the teachers hold a bachelor's degrees (see Qualitative assessment)	1-2
		Qualitative assessment	Overall: relatively high quality (estimated score 3-4) • Score≤Score Tulsa/Oklahoma programs (see Gormley & Gayer (2005) and others/Wong et al. (2008)) • See Peisner-Feinberg et al. (2014) for a more recent study of the Georgia prekindergarten program • Average annual spending per child: $4,010 (2007) • Maximum group size: 20 • Instruction is based on a state approved curriculum • Funding increases with the lead teacher's educational level (Henry et al., 2004) • "At a minimum, Pre-K teachers in Georgia are required to have a technical diploma related to early childhood education or an associate's degree, which approximately 20 percent do. Most, however, have bachelor's degrees in related fields but are not certified to teach (20 percent) and about 60 percent are certified teachers." (Henry et al., 2004) • Most classes are offered by private-for-profit and local public school systems (Henry et al., 2004) • 2003: the Georgia preschool program met 7/10 NIEER benchmark requirements for high quality; 13/43 state preschool programs score 8 or higher (Barnett et al., 2003; State of Preschool 2003)	
		Total		3[4]
Gormley & Gayer (2005); Gormley & Phillips (2005);	US (Tulsa, OK), prekindergarten, 2000-	Staff-child ratio	1:10	2
		Educ. requirement staff	college degree for preschool teacher; no requirements for teacher assistants	2

Gormley (2008)		Qualitative assessment	Overall: relatively high quality (estimated score 4) • preschool teacher receive the same compensation as elementary schools teachers • Maximum group size: 20 • No specific curriculum mandated • "…,Oklahoma's program is a very high-quality program" (Gormley, 2008) • "Oklahoma made a commitment to providing high quality early education from the program's inception and has sustained this commitment through the transition to universality … Whereas 100% of Oklahoma's pre-K teachers have a college degree, fewer than 20% of non-home-based day care providers nationwide have a college degree" (Gormley & Phillips, 2005) • "Tulsa's pre-K programs … exceed their counterparts in other states in terms of both the amount and quality of instruction to which young children are exposed." (Phillips et al. 2009) • 2003: the Oklahoma preschool program met 8/10 NIEER benchmark requirements for high quality (Barnett et al., 2003; State of Preschool 2003); 5/43 state preschool programs score 8 or higher	
		Total		4
Havnes & Mogstad (2010; 2011; 2015)	Norway, child care (age 3-6), 1976-1979	Staff-child ratio	1:8 (1:16 for teachers)	2
		Educ. requirement staff	college degree for preschool teacher; no requirements for teacher assistants	2
		Qualitative assessment	Overall: relatively high quality (score 3-4) • Educational content: a social pedagogy tradition • Average annual spending per child: $6,600, which is relatively high compared to other universal programs: "The high expenditure levels were mirrored in fairly extensive requirements to qualifications of child care staff and physical environment, as well as a relatively low number of children per staff" (Havnes & Mogstad, 2011)	
		Total		4
Herbst (2013b)	US, child care (age 0-5)	Staff-child ratio	Varies. However, 1:10 was recommended	0-1
		Educ. requirement staff	Varies substantially: see 'Qualitative assessment'	0
		Qualitative assessment	Overall: varies, not consistently high quality (score<3) • "Lanham Act centers operated for long hours… It was common for preschoolaged children to spend at least 12 hours per day in the center, usually on a 6am to 6pm schedule …The recommended teacher-child ratio in Lanham Act centers was 10-to-1, and many centers abided by the recommendation…California centers were among the highest-quality: they had an explicit nutrition focus; children were given a medical exam; parents completed a developmental history; and teachers were provided with in-service training and college credit (Koshuk, 1947). On the other hand, low-quality was pervasive in other areas… variation in quality across states and localities was likely to be substantial." (Herbst, 2013)	
		Total		0[1]
Magnuson et al. (2007)	US, prekindergarten, 1997	Staff-child ratio	Varies by state: 1:10 or better in 30/43 Pre-K programs (Barnett et al., 2003)	1
		Educ. requirement staff	Varies by state: • 21/43 Pre-K programs require a bachelor's degree, 29 require specialized training in ECE (Barnett et al., 2003) • Almost all states require at least a teacher certification in ECE, several states require a bachelor's degree (Ripple et al., 2003)	1
		Qualitative assessment	Overall: prekindergarten programs offer on average higher quality ECEC than other non-prekindergarten child care arrangements in the US, but there is substantial variation between states in requirements and quality (estimated score 2). • Score>general child care programs (see Bernal & Keane (2011) and others); • Score<relatively high-quality state programs (e.g. Georgia, Oklahoma, see Peisner-Feinberg et al. (2014); Gormley & Gayer (2005) and others; Fitzpatrick (2008); Wong et al. (2008))	

			• "typical preschool or prekindergarten programs, which vary in the extent to which they offer high quality early learning environments… data on [child–staff ratios, class sizes, and caregiver education and pay] suggest that school-based prekindergarten is of relatively high quality…some evidence indicating that prekindergarten programs located in public schools may be of relatively high quality" (Magnuson et al., 2007) • 86% of school-based prekindergarten teachers have a college degree (Smith et al., 2003) • Educational levels and wages are higher for staff in publicly-operated programs than privately operated programs (Bellm et al., 2002) • "All states needed to improve their quality standards. State quality standards varied widely. Most states did not meet a majority of our research-based benchmarks for minimum state standards…Low state quality standards and funding levels in many states raise serious concerns about state commitment to providing a good education to our young children." (Barnett et al., 2003)	
		Total		2
Peisner-Feinberg et al. (2014)	US (Georgia), prekindergarten, 2012	Staff-child ratio	1:11	1
		Educ. requirement staff	Teacher: bachelor's degree in ECE; Assistances: Child Development Associate	2
		Qualitative assessment	Overall: relatively high quality (estimated score 3-4) • Maximum class size: 20-22 • Comprehensive curriculum standard • Minimum salary requirements for teachers and assistants • Before 2011-2012 reform: o Staff-child ratio: 1:10 o Teacher degree requirement: Associate • Meets 8/10 NIEER benchmark requirements for high quality (Barnett et al., 2012)	
		Total		3
Wong et al. (2008)	US (Oklahoma), prekindergarten, 2004	Staff-child ratio	1:10	2
		Educ. requirement staff	Bachelor's degree with training in early education	2
		Qualitative assessment	Overall: relatively high quality (estimated score 4) • Maximum class size: 20 • Average amount of state spending per child: $6,167 • Comprehensive curriculum standard • Meets 8/10 NIEER benchmark requirements for high quality (Barnett et al., 2005)	
		Total		4
Wong et al. (2008)	US (West Virginia), prekindergarten, 2004	Staff-child ratio	1:10	2
		Educ. requirement staff	Bachelor's or Associate degree with ECE training	1
		Qualitative assessment	Overall: relatively high quality, but lower than Oklahoma's program (estimated score 3) • Maximum class size: 20 • Average amount of state spending per child: $6,829 • Comprehensive curriculum standard (since 2004) • Meets 6/10 NIEER benchmark requirements for high quality (Barnett et al., 2005)	
		Total		3

Table A3 - Estimation weights

	Development domain	Timing measurement	Examples	N
1	Cognitive skills	Immediate / Short term	Math test scores; reading test scores	81
2	Non-cognitive skills	Immediate / Short term	Hyperactivity-inattention; social skills	43
3	Other/general indicators	Immediate / Short term	Motor skills; everyday skills	50
4	Cognitive skills	Medium term	Math test scores ; reading test scores	19
5	Non-cognitive skills	Medium term	Social skills; concentration problems	20
6	Academic performance	Medium term	Grades; falling behind prim. school	6
7	Education	Long term	Years of schooling; grade repetition second. school	13
8	Employment	Long term	Employment position; on welfare	13
9	Earnings	Long term	Annual earnings	8

Table A4 - Robustness tests: weights and significance levels

	No weights		Weight=1/n		α=0.05	
	(1)	(2)	(3)	(4)	(5)	(6)
ECEC FEATURES						
Age below 3	-0.00306	-0.119	-0.248	-0.324	-0.0254	-0.0979
Ref: age 3+	(0.236)	(0.300)	(0.356)	(0.394)	(0.345)	(0.353)
Fulltime	0.446	0.389	0.166	-0.101	0.0475	-0.0676
Ref: Part-time/varies	(0.455)	(0.473)	(0.436)	(0.508)	(0.295)	(0.414)
Quality score (0-4)	0.422***	0.379***	0.381***	0.321**	0.348***	0.320***
	(0.0805)	(0.0893)	(0.134)	(0.142)	(0.128)	(0.108)
TIMING EFFECTS						
Short term	0.363	0.294	0.549	0.288	0.626**	0.367
	(0.363)	(0.374)	(0.500)	(0.499)	(0.271)	(0.255)
Medium term	0.643*	0.716*	0.289	0.366	0.714***	0.676**
	(0.337)	(0.378)	(0.295)	(0.356)	(0.233)	(0.281)
Long term	1.687**	1.568**	1.459*	1.429**	1.856***	1.728***
Ref: Immediate	(0.784)	(0.693)	(0.759)	(0.708)	(0.584)	(0.526)
STUDY FEATURES						
Estimation method: IV	0.0832	0.119	0.297	0.136	0.278	0.243
	(0.381)	(0.498)	(0.332)	(0.416)	(0.238)	(0.442)
Estimation method: RDD	1.452***	1.582***	1.715***	1.690***	1.755***	1.792***
	(0.354)	(0.545)	(0.449)	(0.539)	(0.295)	(0.417)
Ref: DiD						
Published		-0.523**		-0.574**		-0.490
		(0.255)		(0.272)		(0.358)
Sample size (square root)		2.47e-05		-0.000244		7.58e-05
		(0.000414		(0.000438		(0.000547
Pseudo R2	0.243	0.259	0.274	0.288	0.236	0.252
Log likelihood	-203.2	-198.9	-17.35	-17.01	-38.45	-37.61

Clustered standard errors in parentheses. The estimated are based on 253 estimates from 30 studies (23 clusters of studies).
*** p<0.01, ** p<0.05, * p<0.1

Table A5 - Robustness tests: country effects

	Complete sample		US sample	
	(1)	(2)	(3)	(4)
COUNTRIES				
Canada (Quebec)	-1.703**	-1.669**		
	(0.799)	(0.795)		
Germany	-0.197	-0.976		
	(0.664)	(0.715)		
Other Europe	-0.295	-0.559		
	(0.539)	(0.540)		
ECEC FEATURES				
Age below 3	0.202	0.260	-0.251	-0.328
Ref: age 3+	(0.557)	(0.547)	(0.898)	(0.891)
Fulltime	0.690**	0.542	0.398	1.115***
Ref: Part-time/varies	(0.273)	(0.397)	(0.408)	(0.431)
Quality score (0-4)	0.415***	0.497***	0.806***	0.773***
	(0.151)	(0.134)	(0.183)	(0.183)
TIMING EFFECTS				
Short term	0.914*	0.735	-0.343***	-0.448***
	(0.501)	(0.459)	(0.0538)	(0.0643)
Medium term	0.290	0.343	1.201***	1.104***
	(0.403)	(0.418)	(0.166)	(0.164)
Long term	1.064*	1.093*	5.727***	6.385***
Ref: Immediate	(0.588)	(0.593)	(0.957)	(1.121)
STUDY FEATURES				
Estimation method: IV	-0.279	-0.215	2.100***	1.507**
	(0.368)	(0.502)	(0.409)	(0.602)
Estimation method: RDD	0.589	0.373	2.574***	2.006***
	(0.579)	(0.713)	(0.492)	(0.668)
Ref: DiD				
Published		-0.634		0.760***
		(0.418)		(0.0660)
Sample size (square root)		-4.47e-05		-0.000941
		(0.000541)		(0.000773)
Observations	253	253	121	121
Nr of cluster	23	23	11	11
Pseudo R2	0.248	0.262	0.441	0.444
Log likelihood	-39.27	-38.53	-14.52	-14.44

Clustered standard errors in parentheses. The estimated are based on 253 estimates from 30 studies (23 clusters of studies).
*** p<0.01, ** p<0.05, * p<0.1

Table A6 - Robustness tests: probit models

	Significantly negative estimate		Significantly positive estimate	
	(1)	(2)	(3)	(4)
ECEC FEATURES				
Age below 3	0.416	0.462	0.0585	0.0478
Ref: age 3+	(0.472)	(0.518)	(0.445)	(0.486)
Fulltime	-0.127	0.100	0.371	0.450
Ref: Part-time/varies	(0.792)	(0.662)	(0.255)	(0.516)
Quality score (0-4)	-0.498***	-0.412***	0.338**	0.365**
	(0.125)	(0.137)	(0.135)	(0.145)
TIMING EFFECTS				
Short term	-0.314	0.0136	1.018**	0.854
	(0.408)	(0.354)	(0.516)	(0.572)
Medium term	-1.523***	-1.480***	0.152	0.0838
	(0.217)	(0.198)	(0.445)	(0.472)
Long term	-0.838	-0.334	1.347**	1.195*
Ref: Immediate	(0.914)	(0.875)	(0.675)	(0.630)
STUDY FEATURES				
Estimation method: IV	0.431	0.480	-0.00864	0.138
	(0.485)	(0.404)	(0.319)	(0.641)
Estimation method: RDD			1.031**	1.151*
			(0.443)	(0.652)
Ref: DiD				
Published		0.739*		-0.188
		(0.396)		(0.481)
Sample size (square root)		-0.000190		0.000305
		(0.000555)		(0.000733)
Pseudo R2	0.415	0.440	0.228	0.239
Log likelihood	-11.80	-11.30	-28.21	-27.84

Clustered standard errors in parentheses. The estimated are based on 253 estimates from 30 studies (23 clusters of studies).
*** p<0.01, ** p<0.05, * p<0.1