Contents lists available at ScienceDirect

Economics of Education Review

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

Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments

Thomas van Huizen*, Janneke Plantenga

Utrecht School of Economics, Utrecht University, Postbus 80125, Utrecht, 3508 TC, The Netherlands

ARTICLE INFO

Keywords:
Preschool
Child care
Universal ECEC
Child development
Meta-analysis

JEL classification:
H40
I26
I28
J13

ABSTRACT

This study examines the effects of universal Early Childhood Education and Care (ECEC) on child development and children's later life outcomes. Using meta-analytical techniques, we synthesize the findings from recent studies that exploit natural experiments to identify the causal effects of universal ECEC arrangements. We use 250 estimates from 30 studies conducted between 2005 and 2017. 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 and earnings in adulthood. Overall, the evidence on universal ECEC is mixed. Age of enrollment is not a major factor in explaining the impact. Some evidence indicates that more intensive programs produce more favorable outcomes. Program quality matters critically: high quality arrangements consistently generate positive child outcomes. Publicly provided programs produce more favorable effects than privately provided (and mixed) programs. There is no evidence of fading out. Furthermore, the gains of ECEC are concentrated within children from lower socioeconomic families.

1. Introduction

A growing number of children spend a substantial part of their early childhood in Early Childhood Education and Care (ECEC).[1] In OECD countries, on average about one out of three children aged 0–2 and over 80% of children aged 3–5 participate in ECEC (OECD, 2016). Various European countries (e.g. France, Germany, Norway and Spain) offer universal ECEC programs, accessible to all children that meet age-eligibility criteria, as a result of which the coverage rate between age 3 and mandatory school age is close to 100%. Some US states (e.g. Georgia, Oklahoma and Florida) also provide universal programs; a nationwide universal system does not exist, however, and enrollment rates are considerably lower.[2] The policy attention towards public investment in early childhood is fueled by results from a large body of evidence pointing out that experiences in early childhood matter crucially for later life outcomes (Currie & Rossin-Slater, 2015; Elango, Garcia, Heckman, & Hojman, 2015).

While many recent reforms and proposals concern implementing or expanding universal child care and preschool programs, the policy debate has been dominated to a large extent by evidence from targeted interventions (e.g. Perry Preschool, Abecedarian, Head Start). Although some studies indeed provide compelling evidence in favor of targeted ECEC interventions (Barnett, 2011; Barnett & Masse, 2007; Carneiro & Ginja, 2014; Elango, Garcia, Heckman, & Hojman, 2015; Heckman et al., 2010), the results from these studies have limited applicability to universal ECEC programs. In fact, the research estimating the causal effects of universal programs is far from conclusive: some studies find that participation in ECEC improves child development (Drange & Havnes, 2015; Gormley, Gayer, Phillips, & Dawson, 2005), while others show that ECEC has no significant impact (Blanden, Del Bono, Hansen, & Rabe, 2017; Fitzpatrick, 2008) or may produce adverse effects on children's outcomes (Baker, Gruber, & Milligan, 2008; 2015). As societal returns depend critically on the effects on children's outcomes (e.g. van Huizen, Dumhs, & Plantenga, 2018), universal child care and preschool expansions may in some cases be considered as a promising but in other cases as a costly and ineffective policy strategy.

This study synthesizes the recent but growing body of international evidence on the effects of universal ECEC programs on children's outcomes, 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 universal ECEC

* Corresponding author.
E-mail address: t.m.vanhuizen@uu.nl (T. van Huizen).

[1] ECEC may refer to all kinds of non-parental child care and preschool arrangements before the child enters school/kindergarten. Here we focus on formal, center-based child care and preschool arrangements.

[2] The aim of the "Preschool for All" proposal of the Obama administration was also to provide universal (high-quality) prekindergarten.

https://doi.org/10.1016/j.econedurev.2018.08.001
Received 18 July 2016; Received in revised form 7 August 2018; Accepted 7 August 2018
Available online 16 August 2018
0272-7757/

arrangements likely to generate positive effects on child development? Do the effects fade out in the longer run? And which specific groups benefit most from attending ECEC? We focus on studies that exploit exogenous variation to evaluate the causal effects of universal ECEC programs implemented in Western, developed countries.

Our meta-regressions are based on a sample of 250 estimates extracted from 30 studies conducted between 2005 and 2017. The study contributes to the field by focusing on the international evidence of the effects of *universal* ECEC programs exploiting *natural experiments*, taking into account a *large range of children's outcomes*. Existing meta-analyses generally focus on the US, are (almost) completely based on targeted programs, use different (less strict) criteria concerning methodology for the selection of studies and tend to concentrate on short-term cognitive outcomes. Because our analytical sample hardly overlaps with the analytical samples of previous meta-analyses, the present study complements the literature in three ways.

First, by focusing on universal programs, the study provides insights into the mechanisms driving variation in the estimated effects of this type of program. Previous meta-analyses include mostly or exclusively estimates from evaluations of targeted interventions (Camilli, Vargas, Ryan, & Barnett, 2010; Duncan & Magnuson, 2013; Karoly, Kilburn, & Cannon, 2005; McCoy et al., 2017; Shager et al., 2013). Given the differences in program eligibility criteria,[3] the population enrolled in universal programs (a more general population of children) differs substantially from the population enrolled in targeted programs (children from disadvantaged families). The latter may be more likely to gain from participating in ECEC, for instance because these children may have a lower quality home environment (e.g. Cascio, 2015). In addition, some relatively successful programs are small-scale model programs, which may be too costly to expand to a larger scale. Hence, results from meta-analyses including (only) estimates from targeted programs are largely (fully) driven by these estimates and cannot be extrapolated to universal programs.[4]

Second, we focus on evidence from natural experiments because these studies account for selection into ECEC by exploiting a source of exogenous variation. One of the central challenges in the literature on the effectiveness of ECEC is to account for non-random selection: the decision of parents whether or not to enroll their child in ECEC may be related to (unobserved) factors that are related to child development. Whereas the effectiveness of targeted interventions has been extensively evaluated using randomized controlled trials (RCTs), it is unfeasible (and unethical) to randomly restrict access to universal programs. Hence, RCTs of universal programs do not exist. As an alternative strategy, scholars have exploited natural experiments to account for selection bias. These studies use instrumental variables (e.g. Cornelissen, Dustmann, Raute, & Schönberg, 2018; Drange & Havnes, 2015), difference-in-differences (e.g. Baker et al., 2008; Havnes & Mogstad, 2011), or regression discontinuity design (e.g. Blanden et al., 2017; Gormley & Gayer, 2005) techniques to identify the causal effects of universal ECEC programs. Covariate-adjusted associations, reported in a vast number of studies on ECEC and child development, are prone to bias. In fact, estimations that do not account for endogenous selection into ECEC may produce completely opposite results, even when using the same sample (Dearing & Zachrisson, 2017; Herbst, 2013). We argue that studies which are highly susceptible to selection bias should be excluded from the study sample as fundamental errors in primary studies will be carried over to the meta-analysis.[5] By focusing on evidence from natural experiments, we aim to synthesize the recent literature that makes a relatively strong claim of causality.

Third, we include estimates on a wide variety of children's outcomes, ranging from (non-)cognitive development measured during early childhood to longer-term outcomes such as adolescent educational outcomes and labor market outcomes during adulthood. Following several recent studies (e.g. Card, Kluve, & Weber, 2010; Groot, Poot, & Smith, 2016), we classify the estimates by whether the effect of ECEC on children's outcomes is significantly negative, statistically insignificant or significantly positive. Our main meta-regressions are estimated with ordered probit models. Given that the evidence on the causal effects of universal programs is still relatively scarce, an important advantage of this approach is that we can include a larger number of studies in our meta-regressions. Moreover, this approach allows us to compare results from studies that measure completely different outcomes. An innovative element of this study, therefore, is that we take into account different types of child outcomes, capturing different dimensions measured at different points in life (from early childhood to adulthood). The meta-analysis therefore goes beyond the short-term cognitive development impact, which has been the focus of almost all existing meta-analyses on the effectiveness of 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; Section 4 presents some descriptive evidence; Section 5 presents the results from our meta-regression analysis; the final section concludes.

2. Theoretical considerations

2.1. ECEC features: starting age, intensity and quality

It is an unsettled question whether the age of enrollment into child care or preschool is positively or negatively related to child development. 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 (Doyle, Harmon, Heckman, & Tremblay, 2009). According to recent models on the technology of skill formation, early learning is the foundation for further learning and therefore “skills beget skills.” This implies that the returns of human capital investments are higher the earlier in life these investments are made (Cunha & Heckman, 2007; Heckman & Masterov, 2007). Hence, theoretically one may expect greater effects when children start ECEC at an earlier age.

However, the empirical evidence on this issue is inconclusive. Barnett (2011: p. 977) argues that “starting education interventions before age 3 does not appear to be a major contributor to effectiveness”. Results from a meta-analysis show that ECEC programs which start before the age of 3 provide larger positive estimates than programs that start later, but this difference is not statistically significant (Leak et al., 2010). Given that an early starting age implies separation from the primary caregiver, studies have raised concerns that an early starting age may lead to insecure attachment, generate stress and anxiety and cause negative effects on child development (Bowlby, 1969; Jacob, 2009). This may especially be a problem when children enroll below the age of two and spend long hours in child care (e.g. Haeck, Lefebvre, & Merrigan, 2015). However, the more rigorous evidence on this issue is mixed (Dearing & Zachrisson, 2017). Melhuish et al. (2015: 2) conclude on the basis of an extensive literature review that the evidence for

[3] Eligibility for targeted programs depends on income or other socio-economic conditions of the family.

[4] Note that most ECEC evaluations concern targeted interventions and evidence on universal programs is still relatively scarce. This means that if the sample strategy of a meta-analysis is open to both targeted and universal programs, the results from the meta-regressions will be mainly driven by the targeted programs sample. For instance, this holds for the meta-analysis by Duncan and Magnuson (2013): almost all included estimates are from targeted program evaluations.

[5] This methodological problem is often referred to as ‘garbage in, garbage out’ (Borenstein, Hedges, Higgins, & Rothstein, 2009: p. 380).

0–2 year olds is rather ambiguous, 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 or dosage (part-time versus full-time) of ECEC programs may determine program effectiveness. 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, Bridges, Bassok, Fuller, & Rumberger, 2007; Robin, Frede, & Barnett, 2006). Given that the number of ECEC hours is to some degree determined by the parents, selection into part-time versus full-time arrangements is endogenous and these results should therefore be interpreted with caution. Evidence from a natural experiment in Quebec, Canada, in fact demonstrates that full-time child care can be detrimental for child development (Baker et al., 2008; 2015) – more intensive ‘treatment’ does not always produce better outcomes. Felfe and Zierow (2017) also report that a German expansion of full-day slots in child care improved school readiness but negatively affected socio-emotional development among immigrant children. The evidence thus seems to suggest that the returns to ECEC are non-linear. Overall, a full-time program is neither sufficient nor necessary for positive effects: some part-time programs (e.g. Perry Preschool) show significant improvements in child outcomes.

While the empirical evidence of the developmental impact of the starting age and intensity of ECEC is rather ambiguous, there is a growing consensus that the quality of services is crucial. A recent meta-analysis of European longitudinal studies, for instance, reports that quality of ECEC is positively related to children's cognitive test scores (Ulferts & Anders, 2016). Based on a qualitative review of the literature, Melhuish et al. (2015, p. 45) also conclude that “[t]he positive impact of child care quality on various aspects of children’s development is one of the most consistent findings in developmental science”. 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 (OECD, 2012; Cascio & Schanzenbach, 2013).

2.2. Children's outcomes: timing and domain

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 generates long-term (social) benefits. Several studies on targeted ECEC programs (e.g. Perry Preschool, Abecedarian, Head Start) have followed treatment and control children for several years – sometimes decades – and generally find that the positive gains in test scores diminish with the time elapsed since the end of the program (Elango et al., 2015; Camilli et al., 2010; Leak et al., 2010). For instance, results from the Head Start Impact Study indicate that significant short-run gains were are no longer significant in first grade (Puma, Bell, Cook, & Heid, 2010; 2012). In the evaluation of the Tennessee Voluntary Prekindergarten, some effects even became negative as children moved through primary school (Lipsey, Weiland, Yoshikawa, Wilson, & Hofer, 2015).

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-term outcomes with regard to educational achievements, labor market performance, and crime rates.[6] For instance, the results from Perry Preschool show limited cognitive achievement gains during childhood but do indicate long-lasting effects on outcomes during adolescence and adulthood (Heckman, Pinto, & Savelyev, 2013). Similarly, evidence indicates that Head Start improves long-run outcomes (Carneiro & Ginja, 2014; Deming, 2009; Ludwig & Miller, 2007). A recent meta-analysis of 22 US studies confirms that participation in ECEC improves medium- and long-term educational outcomes such as high school graduation rates (McCoy et al., 2017). These findings suggest that there may be “sleeper effects” and that mechanisms producing these long-run gains are rather complex (Barnett, 2011).

Not only when, but also what is measured matters. Recent studies argue that long-run gains can be attributed to the development of non-cognitive skills during the early years (e.g. Heckman et al., 2013; Kautz, Heckman, Diris, Ter Weel, & Borghans, 2014). Furthermore, within the cognitive domain ECEC may for instance be more effective in improving language than 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. Counterfactual mode of care

Another theoretical consideration is that the effectiveness of an ECEC program crucially depends on the alternative mode of care that the program is substituting for (Cascio, 2015). As Elango et al. (2015: p. 16) stress, we should distinguish between the following questions: “What is the causal effect of an early childhood education program relative to a particular childcare alternative, where one of these alternatives might be no treatment at all?” and “What is the causal effect of adding a program to the available choice set?”. Many ECEC evaluations – both those based on RCTs and those exploiting natural experiments – examine the second question.

The issue of control contamination, i.e. a significant share of the control group is exposed to alternative ECEC services, is important when interpreting the evidence. In general, estimated (intention-to-treat) effects will be smaller when ECEC experiences of the control group are more similar to those of the treatment group. The results from the Head Start Impact Study[7] clearly demonstrate this issue: while effects are weak when not accounting for control contamination effects, effects are relatively strong when comparing Head Start to home care (Feller, Grindal, Miratrix, & Page, 2016). Control contamination provides an important explanation for the finding that effect sizes seem to decline over time. Children seem to gain less from participating in a (specific) ECEC program in more recent times because substantially more children in study control groups participate in some alternative form of ECEC (Duncan & Magnuson, 2013).

The evidence discussed above mainly refers to targeted programs. Various studies also indicate that the introduction or expansion of universal ECEC programs may crowd out existing ECEC services (Bassok, Fitzpatrick, & Loeb, 2014; Blanden, Del Bono, McNally, & Rabe, 2016; Cascio, 2017; Cascio & Schanzenbach, 2013). Because children from medium and high socio-economic status (SES) families also enroll in universal programs, the average quality of the home learning environment is likely to be higher (see Section 2.4) and the role of control contamination may be less relevant.

2.4. Heterogeneous effects

Finally, it has to be taken into account that ECEC produces heterogeneous effects. A general finding in the literature is that children from parents with lower socio-economic status (SES) gain more from participation in ECEC than children from higher SES families (Cascio, 2015; Cascio & Schanzenbach, 2013; 2014; Drange & Havnes,

[6] See Duncan and Magnuson (2013) for a more extensive discussion of the puzzling pattern of short-run fadeout (in test scores) and long-run gains in adulthood.

[7] In this recent RCT study on the impact of Head Start, about 60% of the control group in the HSIS is enrolled in an alternative child care or preschool program.

2015; Felfe, Nollenberger & Rodríguez-Planas, 2015). The explanation for this finding is highly related to the issue discussed in Section 2.3. Parental SES is positively related to the learning environment provided by the counterfactual (parental or informal care by relatives). Children from higher SES families may therefore have less to gain or even lose from participating in ECEC programs (e.g. Havnes & Mogstad, 2015; Herbst, 2013). The finding that more recent programs appear to generate smaller effect sizes can be partly attributed to the striking increase in the educational level of mothers during the past decades, corresponding to an improvement in the quality of the home learning environment (Duncan & Magnuson, 2013).

Furthermore, empirical evidence indicates that children from immigrant families – another indicator of disadvantage – are more likely to benefit from participating in ECEC than children from native families (Cornelissen et al., 2017; Gormley, 2008). Program effects may also depend on 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. The cognitive gains of Head Start seem to be more persistent for Hispanics (Bitler, Hoynes, & Domina, 2014).

Clearly, to what extent effects are heterogeneous is a critical issue in the discussion on universal ECEC: if the benefits accrue only to specific groups, the rationale for public investment in universal ECEC is not evident from a child development perspective.

3. Data

3.1. Included estimates and sample overview

Any meta-analysis starts with an extensive literature search. To identify relevant studies, we performed internet key word searches[8] and used recent reviews and meta-analyses (Duncan & Magnuson, 2013; Elango et al., 2015; Leak et al., 2010; Melhuish et al., 2015; Ruhm & Waldfogel, 2011). Additional references to studies were obtained using snowballing techniques; see Appendix Table B1 for details. We applied the following study selection criteria:

1) Universal ECEC: studies included in the meta-analysis evaluate universal ECEC programs. These programs are accessible to all age-eligible children in a country, state or more local setting. Hence, evaluations of targeted programs, which base eligibility on a measure of disadvantage, are excluded.[9]
2) Methodology: as selection into ECEC is a major concern in the identification of the program effects, we selected studies that exploit a source of exogenous variation to account for this identification problem. Theoretically, our sampling strategy is open to experiments (RCTs) and natural experiments. However, since RCTs of universal programs do not exist, we focus on studies exploiting natural experiments: the effects are estimated using instrumental variables (IV), a difference-in-differences (DID) approach or a regression discontinuity design (RDD).[10]
3) Starting age: we included estimates of ECEC programs that start below the age of 5.
4) Treatment and comparison: the treatment condition concerns (more) participation in ECEC. This implies that treatment refers to participation in, or a higher probability of enrollment in,[11] or an earlier start of ECEC (e.g. age three versus age four). We did not include estimates of effects of overall non-parental care (i.e. a mixture of formal ECEC and informal care by relatives). However, control contamination – the counterfactual is a mixture of parental, informal and formal care – is a feature of most studies in this field; we include only studies in which the comparison group participates significantly less in ECEC.[12] This is for instance relevant in DID studies that exploit regional variation in the timing of expansion of ECEC services. In that case, in order to be included the examined reform should lead to an expansion of ECEC places and not fully crowd out existing formal ECEC arrangements. We do not include studies that define the counterfactual mode of care as an alternative ECEC program: direct comparisons between full-time versus part-time, high-quality versus low-quality ECEC, center-based ECEC versus (subsidized) family day care or one specific preschool curriculum versus another are excluded.
5) Regions: in order to focus on results from relatively comparable settings, we include only evidence from Western, developed countries: the US, Canada, Australia, New Zealand and Western Europe (EU-15 plus Norway and Switzerland).
6) Treatment period: we exclude evaluations of programs from before 1960.
7) Data: the estimations are based on micro (child-level) data.
8) Population: as we are primarily interested in the effects of ECEC for the general population, we exclude studies that focus exclusively on specific subgroups (e.g. children in single-mother households, specific ethnic minorities).
9) Outcomes: we focus on children's outcomes that capture human capital, including indicators of cognitive and non-cognitive skills, school performance and labor market outcomes. The study does not include estimates of effects on health, crime or parenting behavior.[13]
10) Publication and language: we include peer-reviewed journal publications, official reports and discussion/working papers published by academic/research institutes (e.g. NBER, CEPR, IZA). Estimates reported in draft papers are not included. We exclude studies published before 2000 and results reported in non-English studies.

Although there are many estimates on the relation between ECEC and child outcomes, only a small share of studies meet the first two criteria: these are the key selection criteria. Criterion 1 implies that results from targeted programs (e.g. Head Start, Perry Preschool, targeted prekindergarten programs) are excluded from our analytical sample. Estimates of ECEC programs that represent a mixture of targeted and universal programs are also excluded (e.g. Magnuson, Ruhm, & Waldfogel, 2007). Moreover, many studies on universal programs report covariate-adjusted correlations (OLS) or apply propensity score matching (e.g. Gormley et al., 2011): as these studies do not exploit an exogenous source of variation, they are not included.[14] While our

[8] The last literature search was done on 7th December 2017.

[9] This distinction between targeted and universal programs is in line with the general literature. For instance, Elango, Garcia, Heckman, and Hojman (2015: p. 58) define a program as ‘universal’ if access to the program is not means-tested and “is available to a general population of children in a local setting (e.g., county, state, country) when the only eligibility requirement is age.” On the other hand, “programs with eligibility criteria based on income, socio-economic status, or other measures of disadvantage” are referred to as means-tested or targeted programs (Elango et al., 2015: p. 7).

[10] “Good natural experiments are studies in which there is a transparent exogenous source of variation that determines 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).

[11] For example, intention-to-treat effects when exploiting regional variation in a DID approach (Havnes & Mogstad, 2011).

[12] In our meta-regressions, we test the relevance of control contamination as a moderator.

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

[14] Although it is common in meta-analysis to include only “high-quality” studies, several meta-analyses on ECEC and child development also include studies using propensity score matching, sibling fixed effects and models that take into account observed baseline differences (e.g. Duncan & Magnuson,

selection criteria are more restrictive with respect to the type of program and methodology, our approach is less restrictive concerning regions (i.e. we not focus exclusively on the US) and children's outcomes (i.e. we include not only estimates of short-term language or numeracy effects). As a result, our analytical sample differs almost completely from existing meta-analyses.

Several studies meet the two key criteria, but do not meet at least one of the other criteria. For instance, Cascio (2009) evaluates a kindergarten program (criterion 3); Gupta and Simonsen (2010) compare preschool with subsidized family day care (criterion 4); Berlinski et al. (2008; 2009) provide evidence from Uruguay and Argentina (criterion 5); Herbst (2017) evaluates a 1940's program (criterion 6); Cascio & Schanzenbach (2013) use aggregated, state-level data (criterion 7); Gormley (2008) examines the effects of universal prekindergarten for Hispanic children (criterion 8). Moreover, within several studies some of the estimates meet all criteria, but others do not. For instance, IV studies generally also provide OLS estimates (the latter are not included). We include only those estimates from the relevant primary studies that meet all selection criteria.

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, some studies use both cognitive and non-cognitive development measures (e.g. Blanden et al., 2016) or multiple estimates on educational and labor market outcomes (e.g. Havnes & Mogstad, 2011). Estimates on different child outcomes provide valuable additional information and 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 if it provides relevant additional information (e.g. using a different cohort or different child outcome). When estimates concern exactly the same outcome measure, only the estimates published in the journal publication are included.

Our analytical sample contains multiple estimates per primary study. Since using only a single estimate per primary study may lead to a loss of information (e.g. Bijmolt & Pieters, 2001), this is common in the meta-analysis literature: for instance, the sample of Doucouliagos and Stanley (2009) consists of 1474 estimates obtained from 64 studies. Also in meta-analytical studies applying a similar estimation method as we do ('sign and significance' ordered probit models), multiple observations per primary study are used.[15]

The final sample consists of 250 estimates obtained from 30 primary studies, conducted in the period 2005–2017.[16] Table 1 provides an overview of the studies included in our analytical sample. The studies are ordered by cluster, consisting of studies using the same data source (27 clusters in total). The countries covered are the US, Canada, Australia, France, Germany, Norway, Spain and the UK. Authors explicitly refer to the ECEC program as universal or to the system that is (moving towards) a universal ECEC system in all except one study.[17] It should be noted that even when all children within a specific age range are eligible, universal ECEC does not imply universal ECEC coverage at the time of evaluation. In fact, various studies exploit the (staggered) expansion of ECEC services while moving towards a universal ECEC system (e.g. Cornelissen et al., 2017; Felfe, Nollenberger, & Rodríguez-Planas, 2015; Havnes & Mogstad, 2011). All programs are (highly) publicly-subsidized and public subsidies cover all or most of the costs of ECEC. However, not all programs are publicly provided (see below).

3.2. Extraction of estimates and moderator data

3.2.1. Estimates of ECEC effects

As outcome measures (the dependent variable(s) used in the primary studies), we collected estimates of 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. Given the differences in the outcomes, there is no common metric. Therefore, the estimates are classified as significantly negative (coded as -1), insignificant (0) and significantly positive (1), using the 10% significance level ($p < 0.10$).[18]

3.2.2. Methodology

We also extracted information about the estimation method: DID, IV or RDD. DID studies generally exploit regional variation in the timing of an ECEC expansion following a reform, comparing changes in children's outcomes between treatment and comparison regions (e.g. Blanden et al., 2016; Felfe et al., 2015; Havnes & Mogstad, 2011). IV studies also often exploit regional variation, for instance instrumenting ECEC attendance by an indicator of local ECEC supply (e.g. Cornelissen et al., 2017). All included RDD studies use an age-cutoff. For instance, Gormley and Gayer (2005) evaluate universal prekindergarten in Tulsa (Oklahoma) and exploit the fact that children are eligible if they are age 4 on September 1 of a given year. They compare test scores of children who started kindergarten and were just eligible for prekindergarten in the year before the assessment to test scores of children who started prekindergarten and were just ineligible for prekindergarten in the year before the assessment.

Although RDD has been frequently applied in evaluations of US prekindergarten programs and is often considered a strong design, the specific implementation in US evaluations has also been subject to criticism.[19] Lipsey et al. (2015) provide an extensive discussion on threats to the validity of these RDD studies. They argue that the likely direction and magnitude of the potential bias is unclear and depends on the specific application. However, when lower performing children are less likely to enroll in the program or are more likely to drop out, differential attrition would result in an upward bias of the treatment effects. In our meta-regressions, we test whether the estimates from these RDD studies systematically differ from studies applying alternative methods.

3.2.3. ECEC features: starting age, intensity, quality and type of provision

Following the literature reviewed in Section 2, we distinguish between programs that start before the age of 3, and those that start at age 3 or later. We also test the results when using an additional category (below age 3; age 3; age 4). We code the variable indicating the intensity of the program as 1 if it concerns a full-time program and 0 otherwise (part-time program or program with varying intensity).

While it is relatively straightforward to measure starting age and

(footnote continued)
2013; McCoy et al., 2017).

[15] For instance, Card, Kluve, and Weber (2010) use 199 estimates from 97 studies and Butschek and Walter (2014) use 99 estimates from 33 studies.

[16] The first study that meets the selection criteria was published in 2005.

[17] The exception is the study by Weiland and Yoshikawa (2013: p. 2115), but they note that the evaluated program is accessible to all age-eligible children (i.e. not means-tested): "[a]ny child within the city of Boston who turned 4 by September 1 could apply for the program; … children's access was not limited by their family income or other restrictions." Elango et al. (2015: p. 68) also state that "Weiland and Yoshikawa (2013) evaluate a universal preschool program…".

[18] We performed sensitivity tests using the 5% significance level instead.

[19] For instance, Elango et al. (2015: p. 68) are "…skeptical about the interpretation of the estimates…".

Table 1
Overview of studies.

	Study (cluster)	Country (region)	Data	Period in ECEC[a]	Outcome (see Table A2)	Estimation method	# estimates [weight]	Average score
1	Baker et al. (2008; 2015); Haeck et al. (2013; 2015); Kottelenberg and Lehrer (2013; 2014); Lefebvre et al. (2008)	Canada (Quebec)	NLSCY	1998–2009	1,2,3,5	DID; IV (K&L)	69 [4]	−0.43
2	Baker et al. (2015)	Canada (Quebec)	SAIP/PCAP	1997–2000	7	DID	3 [1]	−0.33
3	Baker et al. (2015)	Canada (Quebec)	PISA	1997–2001	7	DID	6 [1]	0
4	Bartik, Gormley, and Adelstein (2012); Gormley (2008)	US (Tulsa, OK)	TPS (2006)	2005–2006	1	RDD	4 [1]	1
5	Blanden et al. (2016)	England	National Pupil Database (NPD)	2002–2007	1,2,3,4,6	DID	15 [5]	0.33
6	Blanden et al. (2017)	England	National Pupil Database (NPD)	2006–2010	1,2,3	RDD	5 [3]	0
7	Cascio (2017)	US	ECLS-B	2005–2006	1,2	DID	12 [2]	0.17
8	Chor, Andresen, and Kalil (2016)	Australia (Queensland)	LSAC	2004–2008	1,2,3	DID	3 [3]	0.33
9	Cornelissen et al. (2018); Dustmann, Raute, and Schönberg (2012)	Germany (Weser-Ems)	Admin. records/SEE	1991–2003	1,3	IV	16 [2]	0.25
10	Dearing, Zachrisson, and Nærde (2015)	Norway (southeast)	BONDS	2007–2009	2	IV	1 [1]	0
11	Drange and Havnes (2015)	Norway (Oslo)	Admin. data/ Statistics Norway	2005–2007	1	IV	5 [1]	1
12	Dumas and Lefranc (2010)	France	DEPP	1971–1980	6,7	IV	3 [2]	0.67
13	Felfe and Lalive (2010)	Germany	German Child Panel	1997–1999	5,6	IV	6 [2]	0.67
14	Felfe and Lalive (2010; 2013)	Germany (West)	GSOEP	2003–2009	1,2,3	IV	27 [3]	0.37
15	Felfe and Lalive (2014)	Germany (Schleswig-Holstein)	Admin. records/SEE	2006–2007	1,2,3	IV	4 [3]	0.5
16	Felfe et al. (2012; 2015)	Spain	PISA	1993–1996	6,7	DID	8 [2]	0.5
17	Fitzpatrick (2008)	US (Georgia)	NAEP	1995–2000	4,6	DID	3 [2]	0
18	Gormley et al. (2005)	US (Tulsa, OK)	TPS (2003)	2002–2003	1	RDD	3 [1]	1
19	Gormley and Gayer (2005); Gormley and Phillips (2005)	US (Tulsa, OK)	TPS (2001)	2000–2001	1,2,3	RDD	5 [3]	0.8
20	Havnes and Mogstad (2010; 2011; 2015)	Norway	Statistics Norway	1976–1979	7,8,9	DID	10 [2]	0.4
21	Kuehnle and Oberfichtner (2017)	Germany (West)	NEPS (SC4)	1997–2003	7	RDD	5 [1]	0
22	Kühnle and Oberfichtner (2017)	Germany (Bavaria)	Bavarian school census	1998–1999	7	RDD	1 [1]	0
23	Kühnle and Oberfichtner (2017)	Germany (Schleswig-Holstein)	Admin. records / SEE	1997–2003	1,2.3	RDD	8 [3]	0.25
24	Peisner-Feinberg, Schaaf, LaForett, Hildebrandt, and Sideris (2014)	US (Georgia)	Collected– Georgia Pre-K evaluation	2011–2012	1,2,3	RDD	10 [3]	0.8
25	Weiland and Yoshikawa (2013)	US (Boston)	BPS (2009)	2008–2009	1,2,3	RDD	12 [3]	0.75
26	Wong, Cook, Barnett, and Jung (2008)	US (OK)	Collected – NIEER	2003–2004	1	RDD	3 [1]	0.67
27	Wong et al. (2008)	US (WV)	Collected – NIEER	2003–2004	1	RDD	3 [1]	0.33

[a] For DID studies, this refers to the relevant post reform period.

intensity, ECEC quality is difficult to measure in general and even more difficult to compare across different settings. We use two quality indicators: educational levels of ECEC staff and staff-to-child ratios. We focus on these indictors for several reasons. First, they are to a large extent comparable between different institutional settings. Second, it is likely that these structural features are important determinants of the quality of center-based facilities (Blau & Curie, 2006; Mashburn et al., 2008; NICHD Early Child Care Research Network, 2002). Third, these dimensions are important for policy makers, as they are clearly related to the costs of the programs and can be influenced by public policy.

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 Appendix Table B2 for the documentation on how we have derived the scores for the specific studies. The quality indicator we use in our main analyses is the sum of these scores. In order to determine the scores, we use the information about quality provided by the primary study, generally discussed in the "institutional background" section of the paper. Moreover, we rely on other academic papers and external reports that evaluated the quality of the specific ECEC arrangement (e.g. OECD and NIEER reports). In addition to the official quality standards, we take into account results from 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 assessment is generally consistent with the derived quality scores. An exception is the Canadian (Quebec) case, where quality regulations seem relatively strict but noncompliance was a major issue during the expansion period. As we aim to capture the actual quality level, we adjust the score if there is a discrepancy.

In addition to this quality indicator, we coded the type of provision of ECEC: public versus private or mixed.[20] Although it is not clear a priori whether public or private provision is more beneficial in terms of children's outcomes, quality may depend on the type of provision. Quality may also be more homogenous in the case of public provision. In mixed markets, quality differences between publicly and privately provided studies have been examined. Results for the UK, for example, show that process quality is lower in private nurseries than in public nurseries (Blanden et al., 2017; Sylva, Melhuish, Sammons, Siraj-Blatchford, & Taggart, 2004).

3.2.4. Children's outcomes: timing and domain

In order to test whether effects diminish as children age, we measure the difference in years between the end of treatment year and the year of measurement. Alternatively, 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, up to 10 years after treatment), and long-term (during adolescence/adulthood, more than 10 years after treatment). Furthermore, we coded whether cognitive or non-cognitive outcomes are measured (outcome type 1 and 4 versus 2 and 5; see Table A2), and within the cognitive domain whether language and numeracy skills are measured.

3.2.5. Counterfactual mode of care

Whether the comparison group has access to alternative ECEC services may be a relevant moderator. We code the 'access to alternative ECEC' as 1 if a study:

- Follows a DID approach and the ECEC expansion resulted in a significant crowding out of existing ECEC services (e.g. Blanden et al., 2016).
- Uses an IV approach and participation in a specific ECEC program is instrumented while a significant share of children in the comparison group participates in an alternative form of ECEC.
- Follows a RDD strategy and a substantial share of the comparison group is enrolled in an alternative form of ECEC.

Furthermore, the time period during which children were enrolled in the program is likely to be related to the degree of control contamination: in more recent times, more alternative ECEC services are available (Duncan & Magnuson, 2013). We therefore also code the year in which the children were participating in ECEC. When a study covers multiple years, we use the average year.

3.2.6. Heterogeneous effects

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 were able to extract a substantial number of estimates for children from low and high SES backgrounds (129 estimates for each subgroup), where SES is generally 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).[21]

4. Descriptive analysis

Table 2 provides the descriptive statistics of the sample. Concerning region (Europe/non-Europe), starting age, intensity, quality and the type of provision, the analytical sample appears to be more or less equally distributed between the different categories. Estimates on full-time programs are, however, relatively more common in the non-European subsample. 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). The quality scores from European cases are higher on average. A larger share of the European estimates also concern publicly provided programs. The majority of the estimates concern immediate effects, measured during or directly after the ECEC program. Most evaluations are on recent programs, especially in the non-European subsample.

Table 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 indicate significantly positive impacts on children's outcomes, half of the estimates are insignificant and 16% of the estimates are significantly negative. The estimates obtained from European studies tend to be more favorable than those estimated in non-European studies. However, this difference is mainly due to the Canadian case (study cluster 1; see Table 1), which generally indicates negative effects. The Canadian case can also explain the rather large share of negative estimates for programs with a starting age below 3 and for full-time programs. Estimates for high-quality programs are relatively favorable, with the majority of estimates being significantly positive and almost no estimates being significantly negative. Studies on publicly provided programs report more significantly positive effects than those on private and mixed programs. Surprisingly, children are more likely to benefit from treatment if the comparison group had access to alternative ECEC services (or significant crowding out of existing ECEC services took place) – although the number of these estimates is relatively low. Furthermore, effects are more favorable when cognitive outcomes are measured; the negative effects appear to be concentrated within the cluster of estimates on non-cognitive outcomes. Impact estimates do not seem to vary considerably with the timing of

[20] Most cases concern either public provision or mixed (private/public) provision. We therefore merged the latter category with private provision.

[21] We also extracted separate estimates for blacks, Hispanics, and whites (around 30 estimates for each subgroup). The race/ethnicity specific estimates are all from US studies.

Table 2
Sample descriptive statistics.

		All	Europe	Non-Europe
1.	Number (%) of estimates	250	114 (45.60)	136 (54.40)
2.	Age enrollment			
	Age 3+	51.60	62.28	42.65
	Age below 3	48.40	37.72	57.35
3.	Intensity			
	Part-time/varies	55.20	89.47	26.47
	Full-time	44.80	10.53	73.53
4.	Quality			
	Low quality (<3)	45.60	28.95	59.56
	High quality (≥3)	54.40	71.05	40.44
5.	Provision			
	Private/mixed	54.80	27.19	77.94
	Public	45.20	72.81	22.06
6.	Outcome domain			
	Cognitive	34.40	21.93	44.85
	Non-cognitive	22.80	16.67	27.94
	Other	42.80	61.40	27.21
7.	Timing measurement			
	Immediate	64.40	45.61	80.15
	Short term	12.40	19.30	6.62
	Medium term	9.60	13.16	6.62
	Long term	13.60	21.93	6.62
8.	Counterfactual			
	No ECEC	71.20	82.46	61.76
	Includes ECEC	28.80	17.54	38.24
9.	Treatment year			
	Before 2000	28.80	50.00	11.03
	After 2000	71.20	50.00	88.97
10.	Estimation method			
	DID	49.20	28.95	66.18
	IV	27.20	54.39	4.41
	RDD	23.60	16.67	29.41
11.	Publication status			
	Working paper/report	57.20	65.79	50.00
	Published	42.80	34.21	50.00
12.	Sample size			
	Small sample (<7441)	50.00	35.96	61.76
	Large sample (≥7441)	50.00	64.04	38.24

The European subsample includes France, Germany, Norway, Spain and the UK; the non-European subsample includes US, Canada, Australia.

Table 3
Distribution of estimated effects of ECEC.

		N	Significantly negative (%)	Insignificant (%)	Significantly positive (%)
1.	All	250	16.00	50.00	34.00
2.	Regions/ countries				
	Europe	114	3.51	56.14	40.35
	Germany	67	2.99	61.19	35.82
	Other Europe	47	4.26	48.94	46.81
	Non-Europe	136	26.47	44.85	28.68
	US	55	–	40.00	60.00
	Other non-Europe	81	44.44	48.15	7.41
3.	Age enrollment				
	Age 3+	129	2.33	52.71	44.96
	Age below 3	121	30.58	47.11	22.31
4.	Intensity				
	Part-time/varies	138	2.90	55.80	41.30
	Full-time	112	32.14	42.86	25.00
5.	Quality				
	Low quality (< 3)	114	32.46	54.39	13.16
	High quality (≥ 3)	136	2.21	46.32	51.47
6.	Provision				
	Private/mixed	137	27.74	52.55	19.71
	Public	113	1.77	46.90	51.33
7.	Outcome domain				
	Cognitive	86	5.81	47.67	46.51
	Non-cognitive	57	35.09	45.61	19.30
	Other	107	14.02	54.21	31.78
8.	Timing measurement				
	Immediate	161	16.77	48.45	34.78
	Short term	31	12.90	51.61	35.48
	Medium term	24	20.83	54.17	25.00
	Long term	34	11.76	52.94	35.29
9.	Counterfactual				
	No ECEC	178	22.47	50.56	26.97
	Includes ECEC	72	–	48.61	51.39
10.	Treatment year				
	Before 2000	72	6.94	59.72	33.33
	After 2000	178	19.66	46.07	34.27
11.	Estimation method				
	DID	123	26.83	54.47	18.70
	IV	68	10.29	47.06	42.65
	RDD	59		44.07	55.93
12.	Publication status				
	Working paper/ report	143	13.29	55.24	31.47
	Published	107	19.63	42.99	37.38
13.	Sample size				
	Small sample (<7441)	125	14.40	42.40	43.20
	Large sample (≥7441)	125	17.60	57.60	24.80
14.	Subgroup estimates				
	Low SES	129	7.75	52.71	39.53
	High SES	129	4.65	79.07	16.28

measurement, treatment year, publication status and sample size. Studies that are based on larger samples are not more likely to report significant effects: in fact, they appear to be somewhat more likely to report insignificant effects.[22] Finally, comparing the results for the different subgroups, the more favorable outcomes are concentrated within the group of lower SES children. Almost 80% of the estimates for the subsamples of higher SES children are insignificant.

To further examine the overall evidence, we calculated for each of the study clusters an average score (labelled 'average study score'). Given that the estimates are coded as −1 (negative and significant), 0 (insignificant) or 1 (positive and significant), the average score for each study can vary theoretically between −1 (only significantly negative estimates) and 1 (only significantly positive estimates). Fig. 1 presents the distribution of the average study scores. The figure points out that there is substantial variation between studies in the estimated effects. About 30% of the study scores are non-positive and 70% of the scores are positive. However, a substantial part of the studies with a positive average score do not consistently provide evidence of positive ECEC effects: half of the studies with positive scores have a score of 0.5 or lower, indicating that at least half of the estimates obtained from the study are insignificant (or that some estimates are significantly negative).

We also plotted the average study scores against the three non-binary moderators (Figs. 2–4). Fig. 2 plots the average study score against the quality scores. It is striking that studies evaluating programs with high-quality scores (3 or 4) generally report positive impacts, whereas the results of low-quality programs appear to be rather mixed. Furthermore, Fig. 3 shows to what extent the effects of ECEC fade out as the child ages. The x-axis indicates the (natural log transformation of the) difference in years between the end of treatment year and the year in which the outcome is measured. The figure does not provide support for the hypothesis that ECEC fade out: medium- and long-term effects are not systematically lower than the immediate and short-term effects. Next, as in Duncan and Magnuson (2013), we plotted our indicator of program effectiveness against the (average) year children were enrolled

[22] In general, most studies seem to have sufficient power as sample sizes are generally large (between 686 children and over 3 million children).

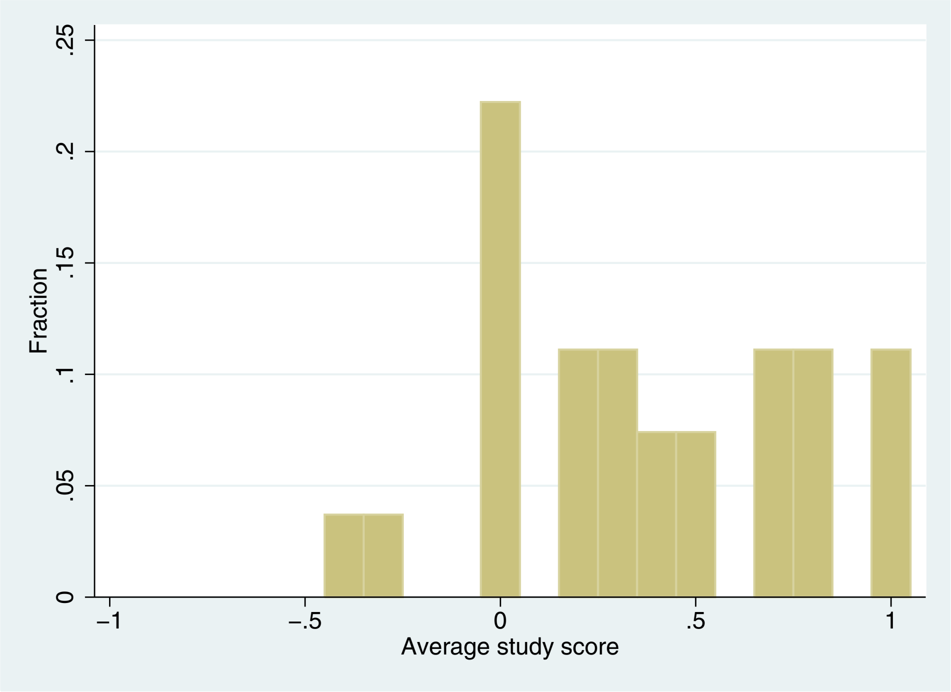

Fig. 1. Distribution of average study outcomes.

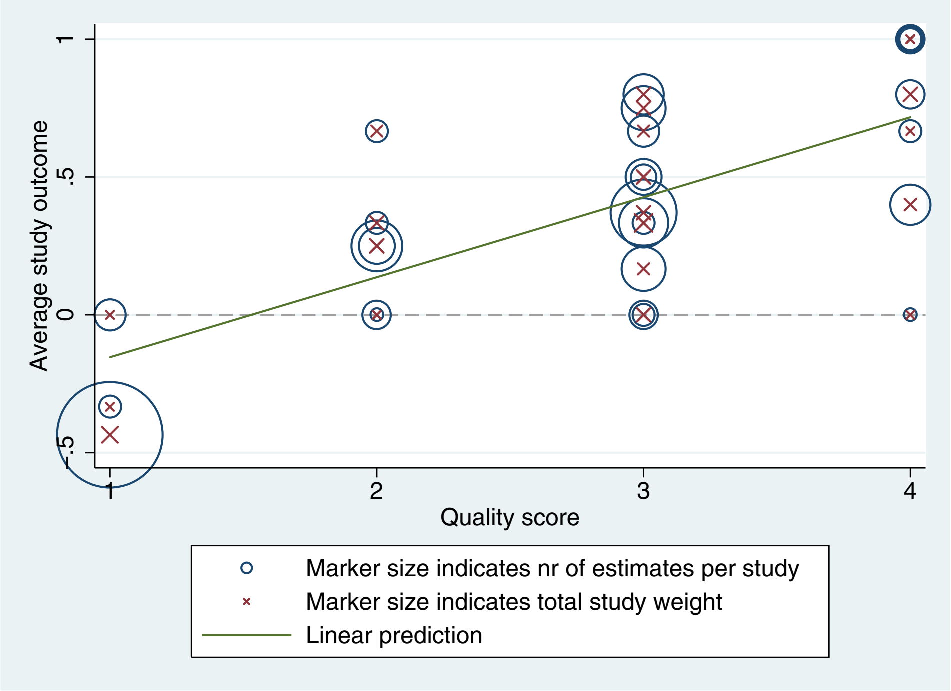

Fig. 2. Average study score and ECEC quality.

in the program (Fig. 4). The figure shows no clear declining profile over time, but our analytical sample consists almost completely of estimates from recent (post-1990) programs.

5. Multivariate analysis

5.1. Meta-analytic model

Following Card et al. (2010), we estimate the relation between outcomes (significantly positive, insignificant, significantly negative) and the relevant moderators using an ordered probit model. The application of more conventional effect size meta-analytical models is unfeasible with our analytical sample for two main reasons. First, there is no common metric: the included studies use very different outcome measures as dependent variables, ranging from early childhood cognitive test scores to earnings in adulthood. Second, the studies apply different econometric techniques and typically do not report the same type of estimate. For instance, in the typical DID study, the estimates should be interpreted as intention-to-treat whereas in most IV studies the local average treatment effects are estimated. The reported

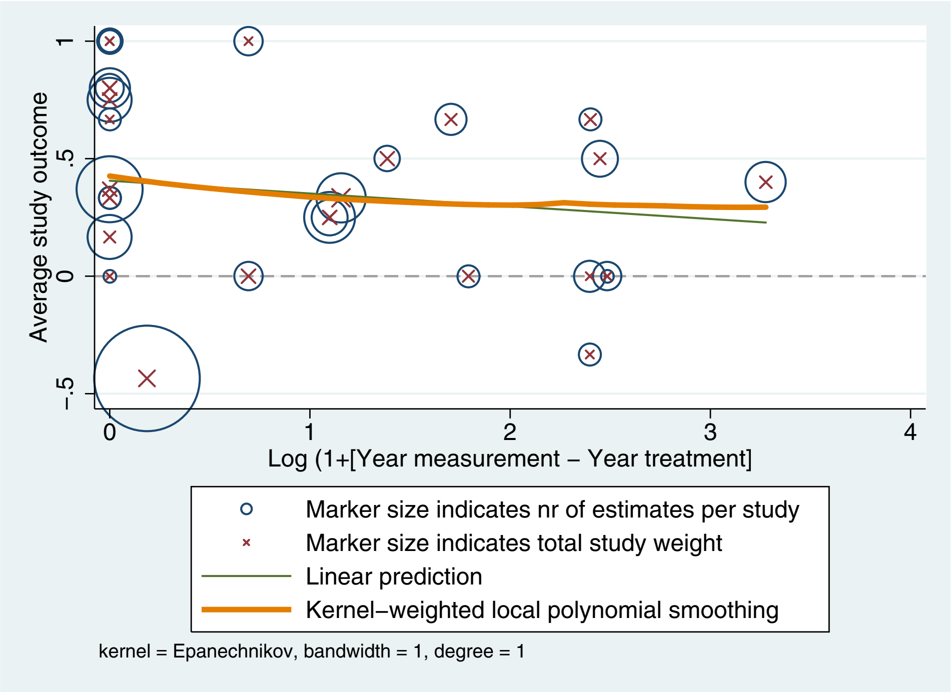

Fig. 3. Average study score and timing of outcome measurement.

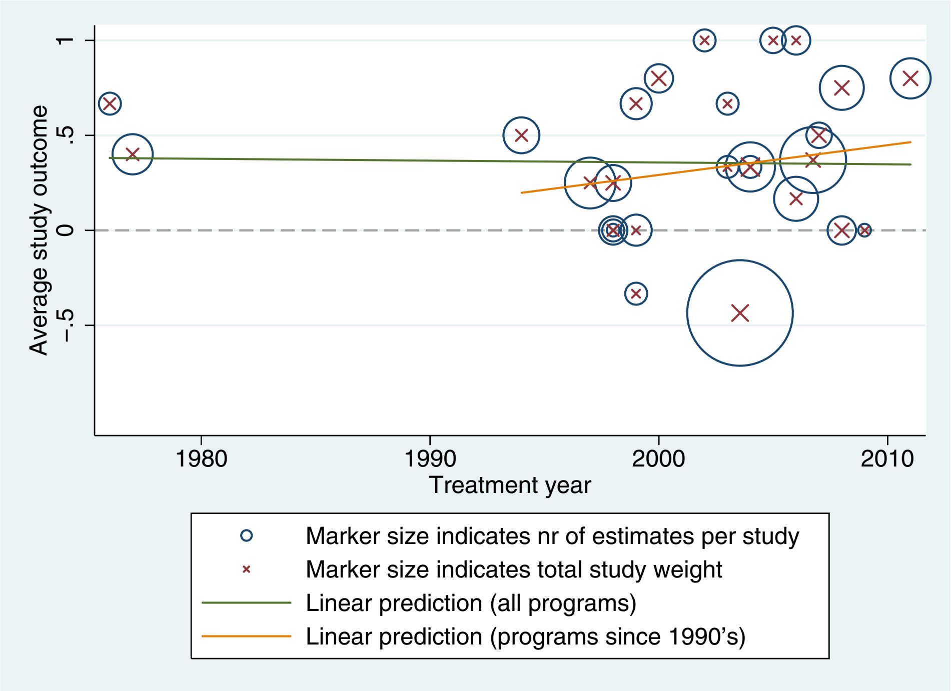

Fig. 4. Average study score and treatment year.

estimates in US RDD studies are neither pure intention-to-treat nor pure treatment-on-the-treated (Lipsey et al., 2015). The sizes of these different estimates are likely to vary even when they concern the same program. The advantage of the ordered probit model approach is that it allows us to compare studies using a large diversity in outcomes and different econometric techniques.[23] However, a limitation is that this approach only allows us to draw relative and not absolute conclusions about the effectiveness of ECEC. Section 5.3 provides a test of our meta-analytical model.

Given that we use multiple estimates per primary 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), we argue that this approach is not appropriate for our analysis. For instance, consider the case where both study A and B provide one estimate on short-term cognitive development but study B also provides several estimates on

[23] If we were to restrict our sample to the most common metric (during/end of program language scores) and estimation method, we would keep only 17 of the 250 estimates (all on US prekindergarten programs).

non-cognitive development and later school outcomes. Applying a simple weighting rule, the weight of the cognitive development estimate from study A will be substantially larger than that of the estimate reported in study B. We therefore use the following equation to calculate the weight of each estimate (see Appendix Table A2 for information on the child outcome domains):

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

If the study provides no pooled sample estimate but provides estimates on subgroups that jointly form the general population, we multiply this weight by an additional correction factor (given by $N_{\text{subsample}}/N_{\text{total sample}}$) to derive the final estimate weight.[24] We follow a similar strategy when estimates for multiple cohorts are provided.

Whereas the number of estimates per study varies between 1 and 69, the total weight per study varies between 1 and 5 (i.e. in our sample, up to five different outcome domains are assessed within a study). 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 in Section 5.4. Furthermore, as it is likely that there are statistical dependencies among estimates from the same study, we cluster the standard errors by study or data source: when different studies use the same data source, the estimates are likely to be correlated and are considered as a single cluster (see Table 1).

5.2. Main results

The main estimation results from the meta-regression analyses are presented in Table 4. 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 four central ECEC features are included: starting age, intensity, quality and type of provision. Columns (2)–(5) present the (separate) results for the timing of outcome measurement, the counterfactual, the time period the children were enrolled in the ECEC program and several study features. Columns (6)–(8) present models that include various dimensions simultaneously.

5.2.1. ECEC features: starting age, intensity, quality and type of provision

Starting below the age of 3 does not lead to more favorable results: in fact, the coefficient in most specifications is negative (though insignificant). We obtain similar results when we introduce an additional dummy indicating starting at age 3 (reference: starting at age 4). Overall, there is no clear relation between ECEC enrollment age and outcomes in terms of child development. Next, the coefficient indicating program intensity (full-time versus part-time/intensity varies) is positive and significant in the more complete specifications. Hence, the results provide some support for the hypothesis that full-time programs lead to more positive results than part-time programs. While our findings on starting age and intensity are somewhat ambiguous, the coefficients of our ECEC quality indicator are positive and highly significant ($p < 0.01$) in all specifications. This result is consistent with previous qualitative literature reviews (e.g. Melhuish et al., 2015) and meta-analyses (e.g. Ulferts & Anders, 2016), pointing out that quality is a crucial determinant of the positive ECEC effects. Interestingly, publicly provided ECEC programs consistently appear to generate more favorable effects than private and mixed programs. A likely explanation is that this variable captures variation in dimensions of ECEC quality that are not measured by our quality indicator.

As the results point out that the ECEC quality score is an important moderator for program effectiveness, we provide a more extensive assessment of quality using different indicators: see Table 5 for the estimation results s–. First, in panel A the results are shown from models using a dummy indicating high-quality services (scoring 3 or 4). The coefficient of this dummy is consistently positive and significant. Second, we explore the role of staff-child ratios and educational requirements separately (panel B). The coefficients of both quality dimensions are positive and significant across specifications, but the results indicate that educational standards are somewhat more important in determining ECEC effects than staff-child ratios. This finding is consistent with the results of a recent meta-analysis of longitudinal studies on European programs (Ulferts & Anders, 2016). 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). Again, these results indicate that ECEC quality plays a critical role. Furthermore, it should be noted that the dummy indicating public provision is significantly positively associated with ECEC effects in all specifications presented in Table 5.

5.2.2. Children's outcomes: timing and domain

The second category of moderators concerns the timing of outcome measurement and the outcome domain. The models presented in Table 4 demonstrate that timing of measurement does not explain variation in ECEC effectiveness. None of the specifications shows that effects significantly decline with the years since treatment. We also used a linear variable (the number rather than the log of years between time of measurement and the end of treatment) and examined potential non-linear effects, distinguishing between four evaluation periods (during or directly after treatment; short-term; medium-term; long-term[25]). In the alternative specifications, the coefficients of these dummies are insignificant and do not indicate a consistent pattern. Hence, the results from our meta-regressions do not provide evidence that suggests initial ECEC effects fading out.

Furthermore, we tested whether children are more likely to gain in specific development domains. We distinguish between cognitive, non-cognitive and other outcomes.[26] The results from all specifications presented in Table 4 indicate that children are less likely to benefit from ECEC in terms of non-cognitive outcomes than in cognitive and other outcomes. This result is in sharp contrast to the argument made in studies on targeted interventions, that children's non-cognitive development is (mainly) positively affected and that this is an important channel for persistent gains in later life (Heckman et al., 2013). It could be that ECEC effects are especially heterogeneous with respect to the non-cognitive development domain: low SES children may benefit from ECEC while the non-cognitive skills of high SES children may be negatively affected. In that case, effects of universal programs are somewhat ambiguous given that a more general population of children participates in these programs. Another potential explanation is that most of the estimates on the non-cognitive domain are from programs for children who start at a relatively young age. As discussed in Section 2.1, the evidence for these programs is rather mixed.

5.2.3. Counterfactual mode of care

The results show that program effectiveness cannot be explained by an indicator of control contamination (whether the control group had

[24] Cascio (2017) for instance provides estimates for 'low income' and 'not low income' children, but not for the pooled sample.

[25] Because only a few studies provide information on long-term effects, we also run specifications where we merged the medium-term and long-term categories. These timing dummies are also not significantly associated with the program effectiveness.

[26] We also examined differences between language and math outcomes, but did not find any systematic differences between these two domains.

Table 4
Main results.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ECEC FEATURES								
Age below 3	−0.253				−0.103	0.0964	−0.0600	−0.0337
	(0.232)				(0.230)	(0.363)	(0.432)	(0.352)
Fulltime	0.408				0.458*	0.489*	0.533**	0.609**
(ref = Part-time/varies)	(0.258)				(0.263)	(0.266)	(0.269)	(0.243)
Quality score (0–4)	0.882***				0.958***	0.925***	0.933***	1.045***
	(0.155)				(0.143)	(0.136)	(0.138)	(0.137)
Public provision	0.996***				1.019***	1.141***	1.158***	1.277***
(ref = Private/mixed)	(0.228)				(0.224)	(0.228)	(0.253)	(0.292)
MEASUREMENT: DOMAIN/TIMING								
Domain[a] (ref = Cognitive)								
Non-cognitive		−0.803***			−0.870**	−0.865**	−0.862**	−0.890**
		(0.303)			(0.401)	(0.405)	(0.403)	(0.410)
Other		−0.137			−0.00556	0.000562	−0.000993	0.0170
		(0.251)			(0.257)	(0.255)	(0.253)	(0.249)
Measurement-treatment gap[b]		−0.115			−0.0694	0.0476	0.0520	−0.00196
		(0.188)			(0.125)	(0.142)	(0.144)	(0.183)
COUNTERFACTUAL/ECEC PERIOD								
Comparison incl. alternative ECEC			0.517			0.379	0.373	0.616
(ref = No alternative ECEC)			(0.371)			(0.370)	(0.369)	(0.382)
Treatment year (coef. × 100)			−1.503			0.488	1.007	0.199
			(1.556)			(1.753)	(1.720)	(1.440)
STUDY FEATURES								
Estimation method (ref = DID)								
Estimation method: IV				0.590**			0.0724	−0.240
				(0.281)			(0.294)	(0.269)
Estimation method: RDD				0.663*			−0.175	−0.490*
				(0.343)			(0.265)	(0.286)
Published								−0.521*
								(0.269)
Sq. root sample size (coef. × 100)								−0.00831
								(0.0353)
Pseudo R2	0.252	0.0418	0.0229	0.0389	0.294	0.299	0.300	0.311
Log likelihood	−37.80	−48.40	−49.35	−48.54	−35.67	−35.42	−35.34	−34.78

Entries represent coefficients of ordered probit models (clustered standard errors in parentheses). The estimated are based on 250 estimates from 27 study clusters (30 studies).

[a] The cognitive domain refers to outcome domain 1 and 4, the non-cognitive domain refers to outcome domain 2 and 5 (see Appendix Table A2).

[b] Measurement-treatment gap = ln (1 + [measurement year] − [end of treatment year]).

*** $p < 0.01$.

** $p < 0.05$.

* $p < 0.1$.

Table 5
Results: quality dimensions.

		(1)	(2)	(3)	(4)
A.	High quality	1.536***	1.661***	1.673***	1.657***
	(dummy: score ≥ 3)	(0.223)	(0.251)	(0.259)	(0.259)
B.	Staff: child ratio score	0.593**	0.631***	0.639***	0.643***
		(0.231)	(0.222)	(0.228)	(0.232)
	Educ. requirements score	1.101***	1.202***	1.148***	1.200***
		(0.208)	(0.213)	(0.223)	(0.232)
C.	Ratio[High]&Educ[Medium]	1.351***	1.454***	1.441***	1.399***
		(0.294)	(0.327)	(0.333)	(0.336)
	Ratio[Medium]&Educ[High]	1.535***	1.679***	1.835***	1.786***
		(0.308)	(0.325)	(0.418)	(0.425)
	Ratio[High]&Educ[High]	1.759***	1.881***	1.761***	1.781***
		(0.308)	(0.266)	(0.297)	(0.308)
	Ref: quality scores < 3				
	Controls:				
	ECEC features	YES	YES	YES	YES
	Domain and timing	NO	YES	YES	YES
	Counterf. and period	NO	NO	YES	YES
	Study features	NO	NO	NO	YES

Entries represent coefficients of ordered probit models (clustered standard errors in parentheses). The estimated are based on 250 estimates from 27 study clusters (30 studies).

*** $p < 0.01$.

** $p < 0.05$.

access to alternative ECEC or whether a preschool reform significantly crowded out existing ECEC programs). This holds across all estimated specifications (also when no other controls are included). This seems to be in contrast with evidence from targeted programs. A plausible explanation is that the relevance of the counterfactual probably depends on the differences between the treatment and comparison group in terms of the quality of the learning environment. In the case of targeted programs, the quality of the home environment is relatively low: effects may be substantial if treatment is compared to the home situation, but small if compared to a comparison group of children who have access to alternative ECEC services. In the case of universal programs, the quality of the home learning environment is likely to be higher on average and therefore whether the counterfactual group has access to alternative ECEC services may be a weaker determinant of the estimated program effectiveness. For children with a high quality home learning environment, substituting parental care by ECEC may not be beneficial, while positive effects may be expected from substituting low quality by high quality ECEC. An alternative explanation is that counterfactual conditions are difficult to measure (especially in natural experiments), and that our measure is not sufficiently precise to capture variation in the counterfactual conditions.

Moreover, the findings do not show that more recent programs are significantly less (or more) effective than older programs. It should be noted though that our analytical sample consists primarily of evaluations of recent (post-1990) ECEC programs. ECEC programs and their counterfactuals probably did not change much during this relatively short period of time.[27]

5.2.4. Study features

The final set of moderators we explore are the study features. Concerning the estimation method, IV and DID do not produce consistently different results in terms of child outcomes. The final column shows that results published in academic journals are somewhat less favorable, though this result appears to be rather sensitive to the specification of the model. In addition, the number of observations used in the study is not significantly related to the child outcome in our ordered probit model.

Next, we tested whether the RDD studies of US prekindergarten programs report systematically different results, given the criticism these studies received (see Section 3.2). It appears that estimates from these studies are significantly more positive when no other moderators are included, but when other moderators are included this no longer is the case. The results from our meta-regression therefore do not indicate that these studies systematically over- or underestimate program effects.

5.2.5. Heterogeneous effects

We analyzed whether ECEC effects are heterogeneous, focusing on subsample estimates from different socio-economic groups. Panel A of Table 6 presents the results when using the estimates for low and high SES subsamples in addition to our main analytical sample, whereas Panel B shows the results based on the estimates from only those studies that provide SES-specific estimates. The coefficients for the low SES groups are positive in the more extensive specifications but the results indicate no significant difference with the estimates from the overall sample. In all specifications, the impact of ECEC is less favorable for higher SES groups: children from more advantaged families are less likely to benefit from ECEC.[28] Overall, the findings consistently point out that the benefits of ECEC are concentrated within more disadvantaged children.

5.3. Testing the meta-analysis model

Card et al. (2010) demonstrate that the ordered probit model approach is valid when the ‘effective’ sample size, a combination of sample size and study design complexity, is constant. Larger (negative and positive) t-statistics may be expected when the sample size is larger. However, this mechanical effect of the actual sample size may be offset by the use of more complex and demanding research designs that can be applied when larger samples are available. Card et al. (2010) test the validity of their ordered probit model using simple probit models that estimate the probability of significantly negative or significantly positive effects and include the square root of the sample size as an additional right-hand-side variable. If larger sample sizes are systematically related to the ‘effective’ sample size (i.e. they are not offset by more complex research designs), one can expect that the sample size increases the likelihood of finding significantly negative and significantly positive effects.

We follow this approach to test the validity of our model: see Table 7. If the ordered model is accurate, the estimated coefficients on the negative outcomes should have the opposite sign to 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. Despite some inconsistencies in the models predicting negative outcomes, the results from the probit models are overall in line with our main results.[29] The main test of the ordered probit model is based on evaluating the coefficient on the (square root of the) sample size in the probit models, which appears to be small and insignificant. The finding that the sample size is not related to finding a positive or negative effect indicates that our ordered probit model is valid (see Card et al., 2010 for a more extensive discussion on this test).

Table 6
Results: heterogeneous effects.

		(1)	(2)	(3)	(4)
A.	Low SES	−0.0877 (0.172)	−0.0554 (0.182)	0.00134 (0.190)	0.0394 (0.173)
	High SES	−0.544** (0.219)	−0.569** (0.224)	−0.614*** (0.217)	−0.633*** (0.208)
	Reference group	Pooled	Pooled	Pooled	Pooled
	Nr of estimates	496	496	496	496
	Nr of clusters	27	27	27	27
B.	High SES	−0.553*** (0.207)	−0.649*** (0.181)	−0.855*** (0.193)	−0.837*** (0.190)
	Reference group	Low SES	Low SES	Low SES	Low SES
	Nr of estimates	258	258	258	258
	Nr of clusters	15	15	15	15
	Controls:				
	ECEC features	YES	YES	YES	YES
	Domain and timing	NO	YES	YES	YES
	Counterf. and period	NO	NO	YES	YES
	Study features	NO	NO	NO	YES

Entries represent coefficients of ordered probit models (clustered standard errors in parentheses).

*** $p < 0.01$.

** $p < 0.05$. * $p < 0.1$.

5.4. Sensitivity tests

We performed a series of tests to examine the robustness of our main results (see Appendix Tables B3–B9). First, we used the 5% rather than the 10% 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: 54% of the estimates are insignificant, while 11% and 35% of the estimates are significantly negative and positive, respectively. Except for the coefficients for program intensity and the non-cognitive domain, which are insignificant in (almost) all specifications, the results are qualitatively similar to our main results.

Second, 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 study cluster). The estimates are generally consistent with our main findings. However, the coefficients for ‘other outcome domains’ are more precisely estimated in the unweighted models. This suggests that the effects on cognitive skills are more positive compared not only to non-cognitive skills but also to other child outcomes.

Third, we tested whether the results are sensitive to whether and how standard errors are clustered. If we do not cluster the standard errors or use alternative study clustering methods (for instance, clustering at the level of actual study titles), the results overall do not change substantially. Nevertheless, when we do not cluster-adjust the standard errors, the standard errors are somewhat larger. The results on program intensity are weaker than in our main specifications, indicating no significant differences between full-time and part-time/mixed programs.

[27] Duncan and Magnuson (2013) present evidence indicating a declining profile over time, but their meta-analysis includes many studies on programs beginning in the 1960s and 1970s.

[28] The results from additional analyses for the US sample indicate that Hispanic children are more likely to benefit from ECEC than the other groups.

[29] The discrepancies are probably due to the fact that this model cannot fully control for estimation method (due to multicollinearity). More specifically, negative outcomes are relatively uncommon in our sample and results from RDD studies, for instance, are never negative.

Table 7
Results: probit models.

	Significantly negative estimate		Significantly postive estimate	
	(1)	(2)	(3)	(4)
ECEC FEATURES				
Age below 3	1.080***	2.258***	0.279	0.606*
	(0.344)	(0.544)	(0.270)	(0.357)
Fulltime	−0.409	−3.774***	0.644**	0.730***
(ref = Part-time/varies)	(0.389)	(1.268)	(0.289)	(0.274)
Quality score (0–4)	−0.704**	−1.795***	0.885***	0.960***
	(0.283)	(0.494)	(0.139)	(0.140)
Public provision	−1.124**	−3.159***	0.892***	1.129***
(ref = Private/mixed)	(0.474)	(0.579)	(0.242)	(0.359)
MEASUREMENT: DOMAIN/TIMING				
Domain[a] (ref = Cognitive)				
Non-cognitive	0.690	1.005	−1.116*	−1.104*
	(0.545)	(0.679)	(0.592)	(0.577)
Other	0.702	1.119	0.117	0.103
	(0.500)	(0.747)	(0.274)	(0.272)
Measurement-treatment gap[b]	−0.158	−0.776***	−0.132	−0.160
	(0.227)	(0.249)	(0.126)	(0.245)
COUNTERFACTUAL/ ECEC PERIOD				
Comparison incl. alternative ECEC		–		0.570
(ref = No alternative ECEC)				(0.491)
Treatment year (coef. × 100)		−17.22***		−1.805
		(4.538)		(1.853)
STUDY FEATURES				
Estimation method (ref = DID)				
Estimation method: IV		1.048***		−0.440
		(0.216)		(0.386)
Estimation method: RDD		–		−0.454
				(0.329)
Published		−0.256		−0.381
		(0.241)		(0.403)
Sq. root sample size (coef. × 100)		−0.0421		−0.00837
		(0.108)		(0.0451)
Pseudo R2	0.526	0.604	0.279	0.291
Log likelihood	−6.566	−5.490	−27.85	−27.38

Entries represent coefficients of ordered probit models (clustered standard errors in parentheses). The estimated are based on 250 estimates from 27 study clusters (30 studies).

[a] The cognitive domain refers to outcome domain 1 and 4, the non-cognitive domain refers to outcome domain 2 and 5 (see Appendix Table A2).

[b] Measurement-treatment gap = ln (1 + [measurement year] − [end of treatment year]).

*** $p < 0.01$.

** $p < 0.05$.

* $p < 0.1$.

Fourth, we tested whether the results are sensitive to excluding specific cases. For instance, in our main sample estimates from working/discussion paper versions of published articles are included if they provide additional information (e.g. on a different child outcome domain). Excluding these studies (25 estimates; Felfe, Nollenberger, & Rodríguez-Planas, 2012; Haeck, Lefebvre, & Merrigan, 2013; Havnes & Mogstad, 2010) has almost no impact on our estimates. Finally, we examined whether the results change if the Canadian case, which contributes by far the most estimates to our analytical sample, is excluded. While our sample size declines substantially (to 172 estimates), the results remain qualitatively similar to our main results. However, the results on age of entry are stronger, indicating more favorable results when children enter ECEC before the age of three.

6. 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 participation in child care and preschool improves child development and leads to positive outcomes in the long run, the overall evidence on universal ECEC is somewhat mixed: About a third of the estimates indicates positive impacts on children´s outcomes, half of the estimates are insignificant and about one out of six estimates is significantly negative. This study examines what explains the heterogeneity in effects: When do children benefit from ECEC?

Results from our meta-regressions do not indicate that the age of enrollment is an important factors explaining the variation in program effectiveness. There is some evidence indicating that full-time programs are more likely to produce significant gains than part-time programs. One of the most robust findings of this study is that quality matters: across many different specifications, using different samples and controls and measuring quality in different ways, high quality ECEC arrangements consistently produce more favorable outcomes. Publicly provided programs also appear to generate more positive effects than privately provided (and mixed) programs. Our results also indicate that children benefit more in the cognitive than in the non-cognitive development domain. Furthermore, the evidence does not provide support for the idea that effects of ECEC fade out. In contrast to evidence from targeted programs, whether the comparison group has access to alternative ECEC services appears not to be a relevant moderator for the effectiveness of the program. However, the direction of the effect of the counterfactual condition is more ambiguous in the case of universal programs, since children enrolled in universal programs have a higher quality home learning environment on average. Finally, the evidence clearly indicates that the gains of ECEC are concentrated within the group of disadvantaged children.

The results have important policy implications. First, given the importance of the early years in shaping later life outcomes, compromising on quality may reduce short-run costs but also long-run benefits (or lead to long-run costs). This highlights the role of public policy. The relevance of quality does fit well with the current policy consensus: the attention is shifted from expanding coverage – to increase female labor force participation and gender equality – towards improving quality levels – to improve child development and well-being. Although these policy objectives are supported by the results from our meta-analysis, increasing coverage levels and investing in quality are both policy strategies that require substantial amounts of public spending. Given that the gains from investments in quality ECEC materialize in the long run, it is important that policies focusing on increasing coverage take these longer-term consequences of quality changes into account. Because high quality typically implies high cost and program benefits for higher SES children are not evident, there is an economic rationale for providing universal ECEC with a sliding fee scale (where higher income families pay more for the services) instead of free ECEC for all.

The evidence on universal programs presented in this study is not always consistent with evidence from existing (meta-analytic) studies on targeted programs. While various scholars claim that there is compelling evidence in favor of targeted ECEC interventions, we would conclude that the effects of universal programs are rather ambiguous given that the majority of estimates is non-positive. Nevertheless, a shared finding in the literature on universal and targeted ECEC is that disadvantaged children are likely to gain from participating in (high-quality) ECEC. As our findings show that children from higher SES families are unlikely to benefit significantly from ECEC, both targeted and universal programs have the potential to narrow SES skill gaps. A key policy question is whether universal programs generate larger benefits for disadvantaged children than targeted programs. Although a recent study suggests that disadvantaged children benefit more from universal than from targeted programs in the short-term (Cascio, 2017), evidence on the differential impact of these programs is almost nonexistent. Testing the differences between program effects as well as the mechanisms explaining these differences are important directions for future research.

Acknowledgments

We would like to extend thanks to the editor, two anonymous referees, Yusuf Emre Akgündüz, Henrik Daae Zachrisson and participants of the InGRID Expert workshop, APPAM Research Conference, SOLE Annual Meeting, CARE Conference, ECCESS Seminar, Trinity College Dublin seminar, CBRT Policy Evaluation Workshop, ESPE Annual Conference for helpful comments on prior drafts.

Appendix A

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; Age < 3:1:9–1:10	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
Estimation weights.

	Development domain	Timing measurement	Example of measures/outcomes	*N*	%
1	Cognitive skills	Immediate/short term	Math test scores; reading test scores	79	31.60
2	Non-cognitive skills	Immediate/short term	Hyperactivity-inattention; social skills	46	18.40
3	Other/general indicators	Immediate/short term	Motor skills; everyday skills	67	26.80
4	Cognitive skills	Medium term	Math test scores; reading test scores	7	2.80
5	Non-cognitive skills	Medium term	Social skills; concentration problems	11	4.40
6	Academic performance	Medium term	Grades; falling behind prim. school	6	2.40
7	Education	Long term	Years of schooling; grade repetition second. school	28	11.20
8	Employment	Long term	Earnings; employment position; on welfare	6	2.40

Immediate refers to outcomes measured during or directly after treatment; short term refers to outcomes measured up to one year after treatment; medium term refers to less than 10 years after treatment; long term refers to ten years or more after treatment.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2018.08.001.

References

Baker, M., Gruber, J., & Milligan, K. (2008). Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy, 116*, 709–745.

Baker, M., Gruber, J., & Milligan, K. (2015). Non-cognitive deficits and young adult outcomes: the long-run impacts of a universal child care program. NBER Working Paper No. 21571.

Bartik, T. J., Gormley, W., & Adelstein, S. (2012). Earnings benefits of Tulsa's pre-K program for different income groups. *Economics of Education Review, 31*(6), 1143–1161.

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science, 333*, 975–978.

Barnett, W. S., & Masse, L. N. (2007). Comparative benefit–cost analysis of the

Abecedarian program and its policy implications. *Economics of Education Review, 26*(1), 113–125.

Bassok, D., Fitzpatrick, M., & Loeb, S. (2014). Does state preschool crowd-out private provision? The impact of universal preschool on the childcare sector in Oklahoma and Georgia. *Journal of Urban Economics, 83*, 18–33.

Berlinski, S., Galiani, S., & Gertler, P. (2009). The effect of pre-primary education on primary school performance. *Journal of Public Economics, 93*, 219–234.

Berlinski, S., Galiani, S., & Manacorda, M. (2008). Giving children a better start: preschool attendance and school-age profiles. *Journal of Public Economics, 92*, 1416–1440.

Bijmolt, T. H. A., & Pieters, R. G. M. (2001). Meta-analysis in marketing when studies contain multiple measurements. *Marketing Letters, 12*(2), 157–169.

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). Experimental evidence on distributional effects of head start. NBER Working Paper No. 20434.

Blanden, J., Del Bono, E., McNally, S., & Rabe, B. (2016). Universal pre-school education: the case of public funding with private provision. *The Economic Journal, 126*, 682–723.

Blanden, J., Del Bono, E., Hansen, K., & Rabe, B. (2017). The Impact of free early childhood education and care on educational achievement: a discontinuity approach investigating both quantity and quality of provision. Discussion Paper No. 0617, School of Economics, University of Surrey.

Blau, D., & Currie, J. (2006). Preschool, daycare, and afterschool care: Who's minding the kids? In E. A. Hanushek, & F. Welch (Eds.). *Handbook of the economics of education, volume 2 of handbooks in economics* (pp. 1163–1278). Amsterdam: Elsevier Chapter 20.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, UK: Wiley.

Bowlby, J. (1969). Separation anxiety. *International Journal of Psycho-Analysis, 41*, 69–113 1960.

Butschek, S., & Walter, T. (2014). What active labour market programmes work for immigrants in Europe? A meta-analysis of the evaluation literature. *IZA Journal of Migration, 3*, 1–18.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *The Teachers College Record, 112*, 579–620.

Card, D., Kluve, J., & Weber, A. (2010). Active labour market policy evaluations:a meta-analysis. *Economic Journal, 120*, F452–F477.

Carneiro, P., & Ginja, R. (2014). Long-term impacts of compensatory preschool on health and behavior: evidence from head start. *American Economic Journal: Economic Policy, 6*(4), 135–173.

Cascio, E. U. (2009). Do investments in universal early education pay off? Long-term effects of introducing kindergartens into public schools. NBER Working Paper No. 14951.

Cascio, E. U. (2015). The promises and pitfalls of universal early education. IZA World of Labor.

Cascio, E. U. (2017). Does universal preschool hit the target? Program access and preschool impacts. NBER Working Paper 23215.

Cascio, E. U., & Schanzenbach, D. W. (2013). The impacts of expanding access to high-quality preschool education. *Brookings Papers on Economic Activity, 2*, 1–54.

Cascio, E. U., & Schanzenbach, D. W. (2014). Proposal 1: expanding preschool access for disadvantaged children. In M. S. Kearney, & B. H. Harris (Eds.). *Policies to address poverty in America* (pp. 19–28). Washington, DC: Hamilton Project 2014.

Chor, E., Andresen, M. E., & Kalil, A. (2016). The impact of universal prekindergarten on family behavior and child outcomes. *Economics of Education Review, 55*, 168–181.

Cornelissen, T., Dustmann, C., Raute, A., & Schönberg, U. (2018). Who benefits from universal child care? Estimating marginal returns to early child care attendance. *Journal of Political Economy* (forthcoming, doi:10.1086/699979).

Cunha, F., & Heckman, J. (2007). The technology of skill formation. NBER Working Paper No. 12840.

Currie, J., & Rossin-Slater, M. (2015). Early-life origins of life-cycle well-being: Research and policy implications. *Journal of Policy Analysis and Management, 34*, 208–242.

Dearing, E., Zachrisson, H. D., & Nærde, A. (2015). Age of entry into early childhood education and care as a predictor of aggression: faint and fading associations for young Norwegian children. *Psychological Science, 26*, 1595–1607.

Dearing, E., & Zachrisson, H. D. (2017). Concern over internal, external, and incidence validity in studies of child-care quantity and externalizing behavior problems. *Child Development Perspectives, 11*, 133–138.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics, 3*, 111–134.

Doucouliagos, H., & Stanley, T. D. (2009). Publication selection bias in minimum wage research? A meta-regression analysis. *British Journal of Industrial Relations, 47*, 406–428.

Doyle, O., Harmon, C. P., Heckman, J. J., & Tremblay, R. E. (2009). Investing in early human development: timing and economic efficiency. *Economics & Human Biology, 7*(1), 1–6.

Drange, N., & Havnes, T. (2015). Child care before age two and the development of language and numeracy: evidence from a lottery. IZA Discussion Paper No. 8904.

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

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*, 109–131.

Dustmann, C., Raute, A., & Schönberg, U. (2012). Does universal child care matter? Evidence from a large expansion in pre-school education. Working paper, University College London.

Elango, S., Garcia, J. L., Heckman, J. J., & Hojman, A. (2015). Early childhood education. NBER Working Paper No. 21766.

Felfe, C., & 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., & Lalive, R. (2013). Early child care and child development: For whom it works and why. SOEP Papers on Multidisciplinary Panel Data Research No. 536.

Felfe, C., & Lalive, R. (2014). Does early child care help or hurt children's development? IZA Discussion Paper No. 8484, Bonn.

Felfe, C., Nollenberger, N., & Rodríguez-Planas, N. (2012). Can't buy mommy's love? Universal childcare and children's long-term cognitive development. IZA Discussion Paper No. 7053.

Felfe, C., Nollenberger, N., & 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.

Felfe, C., & Zierow, L. (2017). From dawn till dusk: implications of full-day care for children's development. CESifo Working Paper Series No. 6490.

Feller, A., Grindal, T., Miratrix, L., & Page, L. C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics, 10*, 1245–1285.

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*, 1–38.

Gormley, W. T., & Gayer, T. (2005). Promoting school readiness in Oklahoma: an evaluation of Tulsa's pre-K program. *Journal of Human Resources, 40*, 533–558.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal Pre-K on cognitive development. *Developmental Psychology, 41*, 872–884.

Gormley, W. T., & Phillips, D. (2005). The effects of universal Pre-K in Oklahoma: Research highlights and policy implications. *Policy Studies Journal, 33*, 65–82.

Gormley, W. T. (2008). The effects of Oklahoma's Pre-K program on hispanic children. *Social Science Quarterly, 89*, 916–936.

Gormley, W. T., Jr., Phillips, D. A., Newmark, K., Welti, K., & Adelstein, S. (2011). Social-emotional effects of early childhood education programs in Tulsa. *Child Development, 82*(6), 2095–2109.

Groot, H. L. F., Poot, J., & Smith, M. J. (2016). Which agglomeration externalities matter most and why? *Journal of Economic Surveys, forthcoming*.

Gupta, N. D., & Simonsen, M. (2010). Non-cognitive child outcomes and universal high quality child care. *Journal of Public Economics, 94*, 30–43.

Haeck, C., Lefebvre, P., & Merrigan, P. (2013). Canadian evidence on ten years of universal preschool policies: the good and the bad. CIRPÉE Working Paper 13-34.

Haeck, C., Lefebvre, P., & Merrigan, P. (2015). Canadian evidence on ten years of universal preschool policies: the good and the bad. *Labour Economics, 36*, 137–157.

Havnes, T., & Mogstad, M. (2010). Is universal child care leveling the playing field? Evidence from non-linear difference-in-differences. IZA Discussion Paper No. 4978.

Havnes, T., & Mogstad, M. (2011). No child left behind: Subsidized child care and children's long-run outcomes. *American Economic Journal: Economic Policy, 3*, 97–129.

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

Heckman, J. J., & Masterov, D. V. (2007). The productivity argument for investing in young children. *Applied Economic Perspectives and Policy, 29*, 446–493.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of public Economics, 94*(1–2), 114–128.

Heckman, J. J., Pinto, R., & Savelyev, P. A. (2013). Understanding the mechanisms through which an inuential early childhood program boosted adult outcomes. *American Economic Review, 103*, 2052–2086.

Herbst, C. M. (2013). 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. (2017). Universal child care, maternal employment, and children's long-run outcomes: Evidence from the U.S. Lanham Act of 1940. *Journal of Labor Economics, 35*(2), 519–564.

Horváthová, E. (2010). Does environmental performance affect financial performance? A meta-analysis. *Ecological Economics, 70*, 52–59.

Jacob, J. I. (2009). The socio-emotional effects of non-maternal childcare on children in the USA: a critical review of recent studies. *Early Child Development and Care, 179*, 559–570.

Karoly, L. A., Kilburn, M. R., & Cannon, J. S. (2005). *Early childhood interventions: Proven results, future promise.* Santa Monica, CA: RAND Corporation.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: improving cognitive and non-cognitive skills to promote lifetime success. IZA Discussion Paper No. 8696.

Kottelenberg, M. J., & Lehrer, S. F. (2013). New evidence on the impacts of access to and attending universal child-care in Canada. *Canadian Public Policy, 39*, 263–285.

Kottelenberg, M. J., & Lehrer, S. F. (2014). Do the perils of universal childcare depend on the child's age? *CESifo Economic Studies, 60*, 338–365.

Kühnle, D., & Oberfichtner, M. (2017). Does early child care attendance influence children's cognitive and non-cognitive skill development? IZA Discussion Paper No. 10661.

Leak, J., Duncan, G. J., Li, W., Magnuson, K., Schindler, H., & Yoshikawa, H. (2010). Is timing everything? How early childhood education program impacts vary by starting age, program duration and time since the end of the program. Paper prepared for the 2010 APAM meeting, Boston.

Lefebvre, P., Merrigan, P., & 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.

Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., & Hofer, K. G. (2015). The prekindergarten age-cutoff regression-discontinuity design: methodological issues and implications for application. *Educational Evaluation and Policy Analysis, 37*, 296–313.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review, 26*, 52–66.

Ludwig, J., & Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics, 122*, 159–208.

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

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732–749.

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., & Shonkoff, J. P. (2017). Impacts of early childhood education on medium-and long-term educational outcomes. *Educational Researcher, 46*(8), 474–487.

Melhuish, E., Ereky, S., Petrogiannis, K., Ariescu, A., Penderi, E., Rentzou, K., Tawell, A., Leseman, P., & Broekhuisen, P. (2015). A review of research on the effects of early childhood education and care (ECEC) on child development, CARE publication.

Meyer, B. D. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business and Economic Statistics, 13*, 151–161.

NICHD Early Child Care Research Network. (2002). Child-care structure→Process→Outcome: Direct and indirect effects of child-care quality on young children's development. *Psychological Science, 13*, 199–206.

OECD (2016). Enrolment in childcare and pre-school. OECD Family database.

OECD. (2012). *Starting strong III: A quality toolbox for early childhood education and care.* OECD Publishing.

Peisner-Feinberg, E. S., Schaaf, J. M., LaForett, D. R., Hildebrandt, L. M., & Sideris, J. (2014). *Effects of Georgia's pre-K 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.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start Impact Study. Final Report. U.S. Department of Health and Human services.* Washington DC: Administration for Children and Family.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A., & Downer, J. (2012). Third grade follow-up to the head start impact study: final report. OPRE Report 2012–45 Administration for Children & Families, Washington DC.

Robin, K. B., Frede, E. C., & Barnett, W. S. (2006). *Is more better? The effects of full-day vs half-day preschool on early school achievement.* Rutgers, National Institute for Early Education Research.

Ruhm, C., & Waldfogel, J. (2011). Long-term effects of early childhood care and education. IZA Discussion Paper No. 6149, Bonn.

Shager, H. M., Schindler, H. S., Magnuson, K. A., Duncan, G. J., Yoshikawa, H., & Hart, C. M. D. (2013). Can research design explain variation in head start research results? a meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis, 35*, 76–95.

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2004). *The effective provision of pre-school education (EPPE) project: Findings from pre-school to end of key stage 1.* Nottingham, United Kingdom: Department for Education and Skills.

Ulferts, H., & Anders, Y. (2016). Effects of ECEC on academic outcomes in literacy and mathematics: meta-analysis of European longitudinal studies. EU CARE Project Report.

van Huizen, T., Dumhs, L., & Plantenga, J. (2018). The costs and benefits of investing in universal preschool: evidence from a Spanish reform. *Child Development* (forthcoming, doi:10.1111/cdev.12993.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*, 2112–2130.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27*, 122–154.