Dietrichson, Jens; Kristiansen, Ida Lykke; Nielsen, Bjørn C. V.

Working Paper

Universal preschool programs and long-term child outcomes: A systematic review

Working Paper, No. 2018:19

Provided in Cooperation with:
IFAU - Institute for Evaluation of Labour Market and Education Policy, Uppsala

Suggested Citation: Dietrichson, Jens; Kristiansen, Ida Lykke; Nielsen, Bjørn C. V. (2018) : Universal preschool programs and long-term child outcomes: A systematic review, Working Paper, No. 2018:19, Institute for Evaluation of Labour Market and Education Policy (IFAU), Uppsala

This Version is available at:
https://hdl.handle.net/10419/201457

WORKING PAPER 2018:19

Universal preschool programs and long-term child outcomes:

A systematic review

Jens Dietrichson
Ida Lykke Kristiansen
Bjørn C. V. Nielsen

IFAU
INSTITUTET FÖR ARBETSMARKNADS- OCH UTBILDNINGSPOLITISK UTVÄRDERING

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala.

IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

More information about IFAU and the institute's publications can be found on the website www.ifau.se

ISSN 1651-1166

Universal preschool programs and long-term child outcomes

A systematic review [a]

Jens Dietrichson[b] Ida Lykke Kristiansen[c] and Bjørn C. V. Nielsen[d]

November 23, 2018

Abstract

What are the long-term effects of universal preschool programs on child outcomes? We review 26 studies using natural experiments to estimate the effects of universal preschool programs for children aged 0-6 years on child outcomes measured from third grade to adulthood. Studies comparing preschool with parental, family, or other informal modes of care show mixed effects on test scores, and on measures related to health, well-being, and behavior. However, all estimates for outcomes related to adequate primary and secondary school progression, years of schooling, highest degree completed, employment, and earnings indicate beneficial average effects of universal preschool programs. Three of the included studies calculate benefits-to-costs ratios and find ratios clearly above one. Universal preschool tends to be more beneficial for children with low socioeconomic status, and there are not consistently different effects for boys or girls. Only three studies compare two alternative types of universal preschool programs in terms of long-term outcomes.

Keywords: universal preschool, long-term effects, child outcomes, systematic review

JEL-codes: I00, I20, I24, I38, J24, Z18

---

[a] A previous version of the report has been published as a VIVE Notat. We are thankful to Niels Coley, Mette Deding, Hans Henrik Sievertsen and Miriam Wüst for helpful comments, and to Caroline Westergaard and Bojana Cuzulan for excellent research assistance. Financial support from Innovationsfonden is gratefully acknowledged. Jens Dietrichson also acknowledges financial support from the Swedish Research Council (grant 721-2012-5778) for the project "The impact of Swedish pre-school reforms on student achievements", hosted by the Institute for Evaluation of Labour Market and Education Policy (IFAU), Uppsala, Sweden.

[b] Corresponding author: Jens Dietrichson, VIVE – The Danish Center for Social Science Research, Herluf Trolles gade 11, DK 1052 Copenhagen, Denmark, tel. nr. +45 697 797, *jsd@vive.dk*.

[c] Ida L. Kristiansen, Department of Economics, University of Copenhagen, Øster Farimagsgade 5, building 26, DK 1353 Copenhagen, Denmark, *ilk@econ.ku.dk*.

[d] Bjørn Nielsen, VIVE – The Danish Center for Social Science Research, Herluf Trolles gade 11, DK 1052 Copenhagen, Denmark, *bcn@vive.dk*.

Table of contents

I. Introduction

A large body of literature shows that the early childhood environment has a strong impact on long-term child outcomes, including educational attainment, earnings, health, and well-being (e.g., Almond, Currie, & Duque, 2017; Black et al., 2017). Many children spend a substantial share of their early childhood in preschool programs; that is, they receive formal pre-primary education and care in facilities outside of their homes. The share of children enrolled in preschools has been increasing over the last 50 years in both developed and developing countries, from 43 percent to 79 percent and from 6 to 43 percent respectively (World Bank, 2017; UNESCO, 2018). Public spending on preschools in the OECD countries average just over 0.7 percent of GDP, and private child care expenditures are 15 percent of net family income on average (OECD, 2016, 2017). The importance of the early childhood environment for child development and the resources devoted to preschool make the effect of preschool programs an important issue for families and policy makers alike.

Resource-intensive and high-quality preschool programs targeting highly disadvantaged children and families, such as the Abecedarian and Perry Preschool projects, substantially improve long-term child outcomes (e.g., Campbell et al., 2014; Gertler et al., 2014; Heckman, Pinto, & Savelyev, 2013; Reynolds & Temple, 2008); often with highly beneficial rates of return (e.g. García, Heckman, Leaf, & Prados, 2016; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Reynolds & Ou, 2011). Broader, but still targeted, programs also have long-term beneficial effects (e.g., Carneiro & Ginja, 2014; Currie & Thomas, 1995; Deming, 2009; Kline & Walters, 2016; Ludwig & Miller; McCoy et al., 2017; Rossin-Slater & Wüst, 2017). However, the demand comes from all sorts of families, not just the disadvantaged. Therefore, the results from studies assessing targeted programs are not sufficient to answer the question of whether and in what form – targeted or universal – governments should support preschool programs.

We review the literature on the effects of universal preschool programs on child outcomes from third grade to adulthood. We include studies that compare attendance of universal preschool programs to parental, family, and other informal modes of care, or compare two alternative universal preschool programs to each other, for example in terms

of pedagogical approaches.[1] We investigate heterogeneous effect across family socioeconomic status (SES) and child gender.

We use systematic review methods to maximize our chances of finding all relevant studies and to increase the transparency of our analyses and conclusions. The included studies use natural experiments to obtain a plausible identification of the effects of universal preschool programs. Just including a variable measuring preschool attendance or exposure to a universal preschool program would likely yield biased estimates. Families and children differ in terms of, potentially unobserved, characteristics that influence the attendance decision, where to live, and child outcomes. In successful randomized and natural experiments, the assignment of treatment is unrelated to both observed and unobserved family and child characteristics, and they thus avoid this type of bias. For that reason, we focus on these research designs, but found no randomized experiments. The outcomes we include are only limited with respect to measurement timing (i.e., measurement occurs in third grade or later), and we analyze the following six outcome categories: health, well-being, and behavior; test scores and school grades; primary and secondary school progression; years of schooling and highest grade completed; employment and earnings; and benefits-to-costs.

We find that the average effects are mixed for two out of six categories. The effects on test scores and school grades, and on measures related to health, well-being, and behavior vary between beneficial and harmful across, and sometimes within, studies. The magnitudes of the effects also vary, and most estimates are not statistically significant. On the contrary, all estimates for outcome measures related to adequate primary and secondary school progression, years of schooling and highest degree completed, and earnings and employment indicate beneficial average effects. The magnitudes of these estimates are often substantial, as well as statistically significant. Furthermore, the three included benefits-costs analyses (BCA) indicate benefits-to-costs ratios clearly above one. The majority of studies and estimates therefore indicate that universal preschool programs have beneficial long-term average effects; effects which are found across heterogeneous preschool programs and across countries with very different political and social contexts.

[1] We found however only three studies comparing alternative universal programs, making it difficult to draw any general conclusions except that more studies are needed.

The key message from previous reviews that have included studies of universal preschool is that the evidence is mixed for the general population of children (e.g., Almond et al., 2017; Baker, 2011; Cascio, 2015; Elango et al., 2015; Melhuish et al., 2015; Phillips et al., 2017; Ruhm & Waldfogel, 2012; van Huizen & Plantenga, 2015; Waldfogel, 2015). However, few reviews included more than a handful of studies with adulthood outcomes, and most of them did not provide analyses of separate outcome categories. This review separates itself from the previous by including a broader range of outcomes and countries, and by focusing on long-term outcomes.

Average effects may hide substantial heterogeneity. The effects tend to be more beneficial for low SES children and only one of the four studies finding statistically significant harmful effects do so for low SES children. There is no consistent pattern of different effects for boys and girls. Earlier reviews also find more beneficial effects for low SES students, whereas there are no clear gender differences.[2]

The remainder of the paper is organized as follows. Section II describes the theoretical framework we use to interpret the results. Section III presents the inclusion criteria, and the search, screening and coding methods, as well as the methods used in the analysis. Section IV presents the results of the search and screening process, discusses the risk of bias and quality of inference in the included studies, and describes the results from the included studies. In section V, we discuss our most important findings. Section VI concludes the review.

II. Theoretical framework

This section discusses the theoretical arguments for and against beneficial effects of universal preschool on a variety of skills, and whether we should expect heterogeneity of the effects across SES and gender, as well as over time. We also include a brief discussion of the properties of skill measures.

A. Skill production and the effects of universal preschool

To frame our discussion, we use a simple model based on the model of skill production developed in for example Cunha & Heckman (2007) and Cunha, Heckman, and

[2] Magnuson et al. (2016) also find few gender differences in a review of mainly targeted programs.

Schennach (2010). We are interested in the effect of universal preschool on, e.g., language, math, and social and emotional skills, as well as mental and physical health; concepts that we denote "skills" below to ease the reading. Let the level of a skill obtained after stage *t* be θ_t. The skill level is a function of parental characteristics h, the skill level obtained in the previous stage θ_{t-1}, and parental investments I: $\theta_{t+1} = f_t(h, \theta_t, I_t)$.

We restrict the model to three stages: Stage 1 covers the time occurring before preschool. In stage 2, the parents choose to invest in either a preschool program or in a counterfactual mode of care (e.g., parental care, other informal modes of care or an alternative preschool program), and in stage 3 the skill is measured. A child's skill level in period 3 is either $\theta_3^p = f_2\left(h, \theta_2, \gamma_p I_{p2}\right)$ if parents choose to invest in a universal preschool program, or $\theta_3^c = f_2(h, \theta_2, \gamma_c I_{c2})$ if they invest in the counterfactual mode of care. The skill multipliers $\gamma_m \in (0, \infty), m \in \{p, c\}$ are positive, which implies that investments improve the skill in both modes of care. If $\gamma_p > \gamma_c$, an investment in universal preschool increases skills more – because the program has higher quality than the counterfactual mode of care – and vice versa.

For concreteness, we use a version of the constant elasticity of substitution (CES) model used, e.g., in Cunha et al. (2010) to model the technology of skill formation:

$$\theta_{t+1} = \left(\tau_1 h^{\varphi} + \tau_2 \theta_t^{\varphi} + \tau_3 \gamma_m I_{mt}^{\varphi}\right)^{\frac{1}{\varphi}} \quad (1)$$

where $\sum_{k=1}^{3} \tau_k = 1$ are weights determining the relative importance of the three variables and $\varphi \in (-\infty, 1)$ is a parameter determining the elasticity of substitution; i.e., how easy it is to compensate low levels of skills with future investments. As $\frac{\partial \theta_{t+1}}{\partial \theta_{i,t}} > 0$ and $\frac{\partial^2 \theta_{t+1}}{\partial \theta_{i,t} \partial I_{i,t}} > 0$, skill production exhibits self-productivity – i.e., skills acquired in one period persist into future periods and are self-reinforcing – and dynamic complementarity, i.e., skills produced at one stage of childhood raise the productivity of investment at subsequent stages.

The skill difference at stage 3 between universal preschool and the counterfactual mode of care is $\beta = \theta_3^p - \theta_3^c$. The effect of universal preschool is beneficial if $\beta > 0$ (which happens if $\gamma_p I_{p2} > \gamma_c I_{c2}$) and harmful if $\beta < 0$. The model therefore illustrates a basic point: the effects of universal preschool depend not only on the quality of the

programs, but also the quality of the counterfactual mode of care and on the investment levels in the two modes of care.

What determines the quality of the different modes of care? Enduring forms of interactions with the immediate environment, such as those between children and parents, other adults, and peers, are thought to be the most important influences on child development (e.g., Bronfenbrenner & Morris, 2006). High quality adult-child interactions and caregiving is the strongest predictor of children's skill development (NICHD Early Child Care Research Network, 2002), and is often mentioned as the most important aspect of preschool quality (e.g., Barnett, 2011; Sabol et al., 2013). The connection between the primary caregiver and the young child is assigned an important role in attachment theory (Flaherty & Sadler, 2011). In a preschool setting, multiple caregivers may damage the attachment between the primary caregiver and the child (Belsky, 2001) and may reduce one-to-one adult interaction compared to parental- or family care, which may be harmful, especially at a very young age. However, the counterfactual mode of care is not necessarily one-to-one high quality parental care for all children. Other siblings may compete for attention (Bradley & Corwyn, 2002), and children may attend low quality informal out-of-home care, if universal preschool is not available. Further, insecure attachments to parents may be compensated for by secure attachments to preschool teachers (Goossens & van IJzendoorn, 1990). For these reasons, it is theoretically ambiguous whether the quality of care is higher in universal preschool programs or the counterfactual modes of care.

The model outlined above allows for differential investments in the two modes of care. Preschool programs may give parents, and especially mothers, better labor market opportunities. Because universal programs are often heavily subsidized, they redistribute resources from other tax payers to families with preschool children. In households with preschool children, income may therefore rise if universal preschool is available, and families with more financial resources can invest more in child development (Elango et al., 2015).

B. Heterogeneity

Effects of universal preschool may vary over time. For example, child health is likely to be negatively affected in the short-run by attending preschool due to the increased risk of infection, but the hygiene hypothesis states that such infections may actually strengthen

the immune system and thus have long-run health benefits (Strachan, 1989). Similarly, socializing with other children and adults may be stressful (Vermeer & IJzendooen, 2006) and have short-run harmful effects but may also improve social-emotional skills that are beneficial in the long term (Baker, Gruber, & Milligan, 2008). This reasoning underlines the importance of examining long-term outcomes.

The effects of universal preschool may differ across groups of children. One source of heterogeneity is immediate from Equation (1):[3] the absolute skill difference, $|\beta|$, depends on the child's initial skills in stage 2. To see this, note that

$$\frac{\partial \beta}{\partial \theta_2} = \tau_2 \theta_2^{\varphi-1} \left[\left(\tau_1 h^{\varphi} + \tau_2 \theta_t^{\varphi} + \tau_3 \gamma_p I_{pt}^{\varphi} \right)^{\frac{1}{\varphi}-1} - \left(\tau_1 h^{\varphi} + \tau_2 \theta_t^{\varphi} + \tau_3 \gamma_c I_{ct}^{\varphi} \right)^{\frac{1}{\varphi}-1} \right],$$

which, given our assumptions on the parameters, is only zero if $\gamma_p I_{p2} = \gamma_c I_{c2}$. If $\gamma_p I_{p2} > \gamma_c I_{c2}$, the derivative is positive and implies a larger beneficial effect, while $\gamma_p I_{p2} < \gamma_c I_{c2}$ implies a negative derivative and a larger harmful effect for the group of children with larger initial skills. Reasons for these effects in a preschool context may for example be that a higher level of cognitive skills imply that the child is ready to benefit from pedagogical instruction, a higher level of social-emotional skills may lead to a closer relationship with preschool teachers and less conflicts with peers, or better health implies that the child can attend and benefit from the instruction. As Magnuson et al. (2016) argue, these advantages for relatively high-skilled children are likely to be present also in the counterfactual mode of care. The model predicts that self-productivity and dynamic complementarities amplify effects, regardless of whether they are beneficial or harmful.

However, the initial skill level may not be the only difference between children. The quality of care, γ_p and γ_c, may vary systematically between children in both modes of care, and in ways that are related to skill development before the start of preschool. Parents with higher income or education may be able to give their children a better home environment and may live in neighborhoods that are more conducive to educational achievement and job market success (e.g., Björklund & Salvanes, 2011; Bradley &

[3] The heterogeneity is not confined to the functional form we chose but would arise in any model where $\frac{\partial \beta}{\partial \theta_2} \neq 0$.

Corwyn, 2002; Hart & Risley, 2003). The quality of the same universal program may moreover differ for groups of children. For example, being exposed to high SES peers may have beneficial peer effects for low SES children (Cascio, 2017; Henry & Rickman, 2007).

High SES children are likely to develop a skill advantage early on, if high SES parents provide care of a higher quality. All else equal, the model then predicts larger beneficial effects of universal preschool for this group whenever $\gamma_p I_{p2} > \gamma_c I_{c2}$. However, if the counterfactual mode of care is better for high SES than low SES children, we expect the quality difference between universal preschool and high SES children's counterfactual mode of care to be smaller than for low SES children, or negative. Subsidized preschool may also be more important for low income families, so the investment difference may also be larger. That is, there are two opposing effects, both caused by the more advantageous family environments for high SES children. It is unclear whether the self-productivity and dynamic complementarities are strong enough to offset the low SES children's larger quality and investment differences in the case of beneficial effects for both groups. When $\gamma_p I_{p2} < \gamma_c I_{c2}$ both forces work in the same direction and we expect larger harmful effects for high SES children. The quality difference may also be positive for low SES children and negative for high SES children, in which case the effects will be beneficial for low SES and harmful for high SES children, or vice versa.

A number of studies indicate that girls develop faster than boys in domains like vocabulary and socio-emotional skills. Girls are thus more likely to have an initial skill advantage (see e.g., Magnuson et al., 2016 and the references cited therein), at least in countries and regions without substantial gender bias against girls. The home environment seems less stimulating for boys in the US (Bertrand & Pan, 2013), while there seems to be gender bias in mortality against girls in many low- and middle-income countries, particularly in South Asia, the Middle East and North Africa (Costa, da Silva, & Victora, 2017). Gender heterogeneity is thus likely to be more context dependent than SES differences.

C. Properties of outcome measures

Skills are not directly observable but have to be measured. The properties of outcome measures will therefore affect the estimation and interpretation of the effects.

Most outcome measures in the studies in our review are likely to capture several skills. Standardized test scores, for example, reflect a range of skills, like intelligence, self-discipline, and motivation (Duckworth & Seligman, 2005; Kautz, Heckman, Diris, ter Weel, & Borghans, 2014). However, a test score is still a narrow and noisy measure compared to, say, employment and earnings. Test scores capture fewer skills and are typically measured on one occasion, while employment and earnings are the result of more continuous processes. Preschool may affect a very broad range of skills and broader and less noisy measures ought to capture the effects better.

For outcomes where there is some form of rivalry (like getting into popular college programs, or getting top management positions), heterogeneous effects may be explained by either relative or absolute differences in skill levels between groups. That is, rivalry implies that it is not only the comparison between the treatment and control group that is important for the effects, but also the relative effects within the treated group. Especially in studies where whole areas are treated, there may be within-treatment group competition and the relative skill rank will matter. Rivalry may therefore cause harmful effects of universal preschool for subgroups even when $\gamma_p I_{2p} > \gamma_c I_{2c}$ for all children because the measure captures not only the absolute skill level, but also the relative rank. For non-rival outcomes, where one child's attainment does not decrease other children's attainment, harmful effects instead imply that $\gamma_p I_{2p} < \gamma_c I_{2c}$. From an impartial point of view, harmful effects are less problematic if they occur when all groups improve their skills. Knowledge about rivalry is thus crucial for the interpretation of the effects of universal preschool (or any program).

Ceiling and floor effects can cause heterogeneous effects even when the effects have the same magnitude for all children. For instance, if subgroups of children have different skill levels when they start preschool, the skill improvement may be large enough for one group to attain the skill measure's maximum level, but not for the other group (ceiling effects). Similarly, the improvement may be large enough for one group to surpass the minimum level of a skill measure, but not the other group (floor effects).

III. Method

This section outlines the inclusion criteria and how we located and analyzed relevant studies.

A. Inclusion criteria

We include studies that have at least one estimate of the effect of a program that meet all seven of the following inclusion criteria. To further illustrate how we apply the criteria, we provide examples of included and excluded studies in the supplementary material along with an explanation of why they were included or excluded, see section A2.

Primary empirical studies. We exclude reviews, comments on research, and theoretical papers from the analysis.

Preschool programs. We include studies that examine preschool programs, defined here as formal out-of-home education and care that children attend before they start primary school. In most countries, kindergarten (or preschool class or grade 0) is a part of primary school, and we exclude studies that examine kindergarten in such school systems.

Universal programs. The preschool programs have to be universal; that is, not targeting a specific group of children. Studies of programs targeting selected groups, such as Head Start, are excluded. This criterion does not imply that all or even a large share of children in an area have to attend preschool, only that the program under study should be open to children from the general population.

Long-term child outcomes. We include studies reporting child outcomes in third grade or later. We include all types of long-term child outcome measures, but studies reporting only parental or family (including sibling) outcomes are excluded.

Type of comparisons. We include studies that compare outcomes between children attending or being more exposed to formal preschool programs with children in or being more exposed to modes of family or informal care (e.g., care by parents, relatives, or nannies). We also include studies that compare groups of children receiving care and education in alternative types of preschool programs. Type of preschool could be defined in terms of, for example, the ownership status of preschool or the pedagogical approach. We exclude studies of interventions in existing preschools, where part of a preschool program is changed for some children (e.g., a changed staff-to-child ratio), or where pre-school teachers or managers get professional development.

Country, period, publication status, and language. We do not restrict inclusion by country, time period studied, or publication status of the study. However, we limit the search period backwards in time to 1980 and include only studies written in a language

that at least two members of the research team understand (Danish, English, German, Norwegian, and Swedish). There are no included studies published before 2008 so limiting the search period to after 1980 is unlikely to be restrictive.

Estimation methods. We include studies that estimate the effects of preschool programs by comparing a treatment group to a control group, and where the assignment of treatment was made by randomization or some form of natural experiment. In the latter type of experiment, the assignment of treatment occurs through some form of "natural" (or administrative) process, which is outside the control of researchers. A successful natural experiment mimics the assignment in a randomized experiment, in the sense that the assignment is likely unrelated to observable and unobservable characteristics of the participating children (see e.g., Cascio, 2015; Ruhm & Waldfogel, 2012; van Huizen & Plantenga, 2015, for reviews using a similar criterion).

B. Search strategy, screening and coding

We searched the following electronic databases for relevant studies: EconLit, ERIC, PsycINFO, Academic Search, Teacher Reference Center and SocIndex. All searches were performed in EBSCO-host in November 2017. We present search documentation for all databases in the supplementary material. In addition to the search of electronic databases, we used the reference lists of included studies and the related reviews mentioned in the introduction for citation tracking.

We screened unique identified records from the electronic databases using the title and the abstract to exclude irrelevant records. We piloted the inclusion criteria until we reached at least a 95 percent agreement between all three screeners (the first two authors and a research assistant). We obtained and screened records that we did not exclude in the first level screening in full text. At least two screeners performed both levels of screening for each study independently. In the case of differences in the assessment, a third screener decided. The first and the second author extracted information from included studies about, for instance, the preschool program, the estimation method, and the effect estimates. We resolved discrepancies by discussion, and it was possible to reach a consensus in all cases.

C. Analysis

In the analysis, we use the estimates from the specification designated as the preferred one by the studies, as long as this specification meet our inclusion criteria. If a study does not indicate a preferred specification, we use the one with the lowest risk of bias according to our assessment. If effects are estimated at different ages, we report the estimate for the oldest children. Some included studies examine the same programs and use (partly) overlapping samples. When they also report the same outcome measures, we only include one estimate in the analysis to avoid double-counting. We chose the study that provide the most information (e.g., had a larger sample) or that have the lowest risk of bias in our view. The section A5 in the supplementary material contains a detailed motivation for each of these cases.

In studies that have access to preschool attendance data, we report treatment-on-the-treated (TOT) effects or local average treatment effects (LATE). In studies that do not have access to data for individual preschool attendance, we report intention-to-treat (ITT) estimates of living in an area that was (more) exposed to the universal preschool program. Some of these studies also report TOT estimates, calculated by scaling the ITT estimate with the difference in take-up rates between the treatment and control group. To be unbiased, such TOT estimates require that the scaling-up of preschool programs did not change the type of children attending or the preschool quality, and that there were no spillover effects on children that grew up in a treated area but did not attend preschool (see e.g., Baker, Gruber, & Milligan, 2015 and Havnes & Mogstad, 2011; and van Huizen, Plantenga, & Dumhs, 2017, for discussions). As it is unclear whether these assumptions are met, we report the ITT estimates.

We calculate effect sizes for the studies that contain sufficient information to ease comparisons across studies. Effect sizes are of three types: for continuous outcome measures with an easily interpretable and comparable scale (e.g., years of schooling) we report the given scale. For other continuous measures, such as standardized test scores, we calculate Cohen's *d* by dividing the effect estimate by the standard deviation in the treatment and control groups. For dichotomous outcome measures, we report the absolute effects in percentage points and the relative effects, calculated as the increase or decrease in percent and using the sample mean as the base rate.

We report average effects for the general population of children and, when available, heterogeneity across SES and gender. We use the full sample means as the base rate for the relative effects in the heterogeneity analysis (separate means for high/low SES children, or boys and girls are not reported in most studies). Statistically significant estimates ($p < 0.05$), as reported by the studies, are shown in bold in the tables.

The definitions of outcome variables and the measures used by the studies are different for nearly all outcome types, and there are few studies that report exactly the same outcomes. Furthermore, the included studies examine very different preschool programs in terms of program features, age of attending children, the studied period, and the broader study context. As mentioned, both estimation strategies and estimands differ too. We therefore believe meta-analysis is problematic, as there is little reason to expect similar effects and directly comparable magnitudes. However, not performing a meta-analysis precludes a formal analysis of the consistency of effects across studies. If there is a stochastic component in effect estimates – due to sampling variance, for instance – we should expect estimates to vary across studies, even if the true effects of the evaluated programs were the same. For this reason, evaluating whether the results in the literature are "mixed" or not by counting negative, null, and positive effects may be misleading ("vote counting"; see e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009, for a discussion). Our supplementary material therefore contains an example of a meta-analysis for test scores, which is the most commonly used outcome in our studies. This example illustrates the problems and underlines the heterogeneity of the results. Our synthesis of the results is therefore not based on meta-analyses and should be read with the caveats about vote counting in mind.

We describe the outcome measures we use in more detail for each outcome type in the Results-section below. Note that because of the procedures described above, the effect sizes and relative effects we report may differ from the ones reported in the studies.

IV. Results

This section presents the results of the search and screening process, a discussion of the risk of bias and quality of inference in the included studies, and the analysis of the effects of universal preschool programs on long-term child outcomes. The analysis of effects is divided into six outcome categories: health, well-being, and behavior; test scores and

school grades; primary and secondary school progression; years of schooling and highest grade completed; employment and earnings; and benefits-to-costs The analysis of the three BCAs in the literature provide a natural context for discussing the magnitude of the effects across studies, so we postpone this discussion until that section.

A. Results of the search and screening process

The search of the electronic databases yielded 1,516 unique records, and we found an additional 88 records from the citation tracking. After excluding irrelevant studies based on information in the title and abstract, we screened 147 studies in full text. Of these, 26 met the inclusion criteria. In some cases, we used information from earlier versions of studies, if, for instance, they included outcomes that were not covered in the published/latest version. Figure A 1 in the supplementary materials describes the search and screening process in a flowchart. We also include a detailed description of each study in Table A 1.

The main characteristics of the included studies are summarized in Table 1. Twenty-three studies compare children attending or being more exposed to universal preschool programs to children in modes of family or informal care, and three studies compare preschool types. We analyze the latter three separately by study rather than by outcome in section C7.

Table 1. Descriptive statistics of the 26 included studies

Variables	*N*	%
Country		
Developing	5	19
Developed	21	81
Continent		
Europe	14	54
North America	7	27
South America	4	15
Africa	1	4
Publication status		
Published in scientific journal/books	17	65
Not published in scientific journal/books	9	35
Publication period		
-2012	8	31
2013-2018	18	69
Studied period		
-1960	2	8
1961-1980	6	23
1981-1999	15	58
2000-	3	12
Age of participants		
0-2	9	35
3-6	25	96
Study design		
Difference-in-differences	17	65
Instrumental variables	7	27
Sibling/family fixed effects	2	8
Outcomes		
Health, well-being and behavior	8	31
Test scores and school grades	10	38
Primary and secondary school progression	10	38
Years of schooling and highest grade completed	8	31
Employment and earnings	6	23
Benefits-costs	3	12

Note: Not all categories sum to 26 because some studies covered more than one category, e.g., included both 0-2 and 3-6-year-olds. In these cases, they were counted in all covered categories. Studied period refers to the period in which the preschool program started. The number of studies that examine the effect on the different outcomes corresponds to the number of studies included in the main analysis, hence the three studies that compare different preschool types are not included in this part of the table. The denominator of the percent-column is 26 (23 for outcomes), so it indicates the share of included studies in a certain category.

The studies cover a broad range of countries: there are studies from 12 countries and 4 continents. 21 studies examine programs in developed countries and 5 in developing countries. Most studies have been published in a journal (9 have not) and are relatively new (8 studies are dated before 2013). The studied periods are wide-ranging, but a majority of studies (18) examines a program that children attended during the period 1981-1999. Few studies include very young children: 9 studies include participants that are between 0 and 2 years of age, and 25 include 3-6-year-olds (some thus include both

age categories). The research designs include difference-in-differences (DID; 17 studies), instrumental variables (IV; 7 studies), and sibling fixed effects (2 studies).

B. Risk of bias and the quality of inference in the included studies

All included studies use some form of natural experiment to estimate the effects of universal preschool programs. However, the studies aim to estimate different types of effects and rely on different assumptions, statistical models, and inference techniques in the estimation of causal effects. Table A 1 and Section A4 in the supplementary material includes a brief description of the research designs and a discussion of the main risk of bias in each type of research design and the quality of the statistical inference. In the following, we provide our overall assessment.

The claim that all included studies estimate the causal effects of universal preschool programs has several caveats. These include having few – in extreme cases only one – treated areas in DID designs, using potentially invalid instruments in IV designs, and correlation between parental investments in education and the decision to send one child and not the other(s) to preschool in sibling fixed-effects designs. However, although individual studies may be biased, the estimates are, in our view, not systematically biased toward showing either beneficial or harmful effects.

The included studies seem more likely to overstate the statistical significance of their findings than understate it. The main reasons are that multiple hypothesis testing is rarely adjusted for and that standard errors are difficult to estimate properly when the assignment of treatment is clustered and there are few clusters.

Publication bias is the tendency that statistically significant results are more likely to be published than null findings. This problem pertains to the whole literature rather than the individual studies. To mitigate this risk of bias, we included unpublished reports. However, Franco, Malhotra, and Simonovits (2014) find that publication bias in the social sciences is more driven by researchers not writing up null results than by journals not publishing them. It is difficult to tell whether this is the case for the literature we review, but as we shall see in the following sections, there are plenty of examples of insignificant results in our included studies.

C. Effects of universal preschool programs on long-term child outcomes

1. Health, well-being, and behavior

Table 2 displays the estimated effects of universal preschool programs on measures related to health, well-being, and behavior. We include personality measures, family formation, and crime, as these measures are related to or is a type of behavior. Personality traits and family formation are not clear-cut measures of beneficial or harmful effects, although some are related to other more unambiguous measures (e.g., conscientiousness is often positively associated with earnings and health; Almlund, Duckworth, Heckman, & Kautz, 2011). Five studies report estimates on measures of problem behavior or personality traits, three studies report measures related to health, healthy behaviors, and well-being, and two studies report the effects on outcomes related to crime. The outcomes are measured when the children are between 8 and 39 years old. It is not possible to convert all estimates to a common effect size measure, continuous measures are converted to Cohen's *d* and binary measure are shown in percentages and percent. We present the estimates so that a positive (negative) sign imply an increase (decrease) of the behavior/health aspect/trait measured. For example, a negative sign on an estimate of overall health imply decreased health and therefore a harmful effect.

Table 2. Health, well-being, and behavior

(1) *Study*	(2) *Average effect*	(3) *SES*	(4) *Gender*
Baker et al. (2015): ITT effect of being more exposed on stress, quality of life, and being accused of a crime at age 12-20	Stress: 0.094 Quality of life: **-0.36** Crime: **0.3** (3.7%)	Not reported	Crime *Girls*: **0.17** (2.1%) *Boys*: **0.43** (5.3%)
Berlinski et al. (2009): ITT effect of new preschool places per child on teacher ratings of behavior in 3rd grade	Attention: **12** (14%) Effort: **21** (24%) Discipline: 11 (15%) Participation: **17** (20%)	Not reported	Not reported
Fort et al. (2018): LATE of 1 extra month of preschool on the (log of) openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N) at age 8-14 (mean 10.7)	Effects in %. O: -0.4% C: -0.0% E: -0.6% A: -0.4% N: 0.2%	*Lower income* O: 0.1% C: 0.7% E: -1.1% A: 0.3% N: -0.5% *Higher income* O: **-1.4**% C: -0.1% E: -0.6% A: **-1.2**% N: 0.9%	*Girls* O: -0.5% C: 0.3% E: **-1.2**% A: -0.3% N: 0.0% *Boys* O: -0.3% C: 0.1% E: -0.3% A: -0.4% N: 0.2%
Havnes & Mogstad (2011): ITT effect on being more exposed on the probability of being a parent (P), single without children (S), and single with children (SC) at age 30-39	P: **-1.4** (1.8%) S: **0.62** (4.4%) SC: -0.04 (-0.48%)	*Mother with no high school (HS)* P: **-1.1** (-1.4%) S: **0.56** (4.0%) SC: 0.12 (1.4%) *Mother with HS* P:**-1.3** (-1.6%) S: 0.23 (1.7%) SC: **-0.61** (-7.3%)	*Girls* P: **-2.0** (-2.5%) S: **0.95** (6.8%) SC: -0.31 (-3.7%) *Boys* P: **-0.9** (-1.1%) S: 0.31 (2.2%) SC: 0.21 (2.6%)
Herbst (2017): ITT effect of $100 more in spending on the probability of a work-limiting disability at age 24-39	-0.3 (-4.8%)	Not reported	Not reported
Kühnle & Oberfichtner (2017): LATE effect of one extra year of preschool on Openness (O), Conscientiousness (C), extraversion (E), Agreeableness (A), Neuroticism (N), SDQ - pro-social (PS) and SDQ – peer problems (PP) at age 15.	O: 0.051 C: 0.035 E: 0.010 A: 0.038 N: -0.038 PS: 0.029 PP: 0.011	*Parents' education: Low* O: 0.000 C: 0.044 E: -0.028 A: 0.017 N: -0.103 PS: -0.005 PP: 0.101 *High* O: 0.108 C: 0.066 E: 0.026 A: 0.004 N: -0.005 PS: -0.024 PP: -0.153	*Girls* O: 0.068 C: 0.088 E: 0.057 A: 0.089 N: -0.005 PS: 0.098 PP: -0.025 *Boys* O: 0.055 C: -0.016 E: -0.036 A: -0.005 N: -0.094 PS: -0.033 PP: 0.075

(1) *Study*	(2) *Average effect*	(3) *SES*	(4) *Gender*
Lebihan et al. (2017): ITT effect of being more exposed on Hyperactivity (H), Anxiety (A), Physical aggression (PA) and Indirect aggression (IA) at age 8-9, and health and well-being at age 12-14	*Behavior* H: 0.074 A: **0.21** PA: 0.10 IA: 0.094 *Health and well-being* Overall health: -0.04 Had asthma attack: 0.005 Mental health: -0.037 Belonging: -0.023 Life satisfaction: -0.098 Drank alcohol: -0.018 Doesn't smoke: **0.080**	*Mothers, post-secondary schooling* *Without* H: -0.026 A: **0.30** PA: 0.067 IA: **0.27** *With* H: 0.12 A: **0.19** PA: 0.11 IA: -0.01	Not reported
Smith (2015): ITT effect of being more exposed on the probability of being charged with Felonies (F) and Misdemeanors (M) at age 18-19	*Black* F: -2.8 (-17%) M: -5.7 (-32%) *White* F: -0.6 (-20%) M: 0.9 (18%)	Not reported	Not reported

Note: Whenever there was sufficient information, we calculated either effect sizes in standard deviations or as percentage points and percent, the latter shown as X (Y%). The estimates were calculated so that a positive (negative) sign implied an increase (decrease) of the behavior/health/trait measured. Statistically significant effects ($p < 0.05$), as reported by the studies, are shown in bold. Type of SES heterogeneity is shown in italics in column 5.

For all three subcategories (behavior/personality, health and well-being, and crime) the estimates are mixed. There are examples of beneficial and harmful effects in all subcategories. Furthermore, most estimates are insignificant. Few studies report heterogeneity across SES or gender, and there is no clear pattern in either category.

2. Test scores and school grades

The effect sizes in Table 3 are based on standardized tests of science, mathematics, and literacy, combinations of the latter two subjects, broader tests of cognitive skills and IQ, or school grades. The tests are performed when children are between 8 years and 18-20 years old. Although tests of educational achievement and school grades measure different skills than an IQ test, achievement and IQ tests are significantly correlated (e.g., Borghans, Golsteyn, Heckman, & Humphries, 2016). Furthermore, motivation and incentives to perform well are important for all tests, which is another reason to believe that standardized achievement tests and IQ capture overlapping skills. We therefore analyze these outcomes together.

Table 3. Effects on test scores and school grades

(1) *Study*	(2) *Average effect*	(3) *SES*	(4) *Gender*
Baker et al. (2015): ITT effect of being more exposed on math, reading, and science test scores at age 13-16.	Math PISA: **0.26** SAIP/PCAP: -0.23 Reading PISA: 0.074 SAIP/PCAP: -0.074 Science PISA: -0.032 SAIP/PCAP: -0.042	Not reported.	Not reported.
Berlinski et al. (2009): ITT effect of new preschool places per child on math and Spanish test scores in 3rd grade (age 8).	Math: **0.24** Spanish: **0.23**	*Share living in poverty by municipality* Effects are 0.08 (Math) and 0.16 (Spanish) larger at the 75th percentile than at the median (significance not reported).	*Girls* Math: **0.26** Spanish: **0.27**
Bietenbeck et al. (2017): TOT effect of attending preschool on a composite standardized literacy and numeracy test at age 13-16.	*Kenya*: 13-16: **0.12** *Tanzania*: 13-16: **0.080**	*Household wealth above or below median* No consistent differences (results shown in figure).	Not reported.
Blanden et al. (2016): ITT effect of availability of free preschool places in an area of residence on standardized tests of reading and math at age 11.	Math: -0.002 Reading: **0.006**	*Free school meals eligible* Reading: 0.008 Math: 0.003 *Not eligible* Reading: **0.006** Math: -0.002	*Girls* Reading: **0.007** Math: -0.002 *Boys* Reading: 0.005 Math: -0.001
Cascio and Schanzenbach (2013): ITT effect of being more exposed on math and reading test scores in 8th grade.	*Effects in points* Math: 0.9 (not standardized and average effect for reading is not reported)	*Free/reduced-price lunch eligible* Math: 2.2 Reading: 0.82 *Not eligible* Math: -1.3 Reading: -0.81	Not reported.
Felfe & Lalive (2010): LATE of having spent some time in preschool during 0-3 years of age on school grades at age 9-10.	Significant beneficial effect on grades but the scale of the effect is unclear.	Not reported.	Not reported.
Felfe et al. (2015): ITT effect of being more exposed on PISA scores in math and reading at age 15.	Math: 0.05 Reading: **0.15**	*Parents without secondary school degree* Math: 0.041 Reading: **0.17** *With* Math: 0.025 Reading: 0.11	*Girls* Math: 0.11 Reading: **0.19** *Boys* Math: -0.01 Reading: **0.12**
Fort et al.: (2018): LATE of 1 extra month of preschool on IQ	**-0.045**	*Lower income*: -0.02 *Higher income*: **-0.13**	Girls: **-0.07** Boys: -0.042

(1) *Study*	(2) *Average effect*	(3) *SES*	(4) *Gender*
test score at age 8-14 (mean 10.7).			
Havnes & Mogstad (2015): ITT effect of being more exposed on cognitive skills at age 18-20.	Not reported.	Quantile effects for males reported in figure. Small, insignificant effects in all quantiles.	The sample consists only of boys.
Kühnle & Oberfichtner (2017): LATE effect of one extra year of preschool on language, STEM, overall cognition score, and track choice at age 15.	Language: 0.028 STEM: 0.006 Cognition: -0.000 Academic track: -0.011 Vocational track: 0.005	*Parents' education:* *Low:* Language: 0.037 STEM: 0.019 Cognition: -0.017 *High:* Language: 0.025 STEM: -0.032 Cognition: 0.017	*Girls:* Language: 0.054 STEM: 0.048 Cognition:0.038 Academic track: -0.014 Vocational track: -0.004 *Boys:* Language: 0.013 STEM: -0.028 Cognition: 0.051 Academic track: -0.009 Vocational track: 0.013

Note: Effect sizes measured in standard deviation units unless otherwise mentioned. Statistically significant effects ($p < 0.05$), as reported by the studies, are shown in bold. Type of SES heterogeneity is shown in italics in column 5.

Most studies included in Table 3 report beta-coefficients with the scores standardized to have mean zero and standard deviation equal to one. We report effect sizes calculated in this way whenever possible in the table (some studies lack information), but it should be noted that the standardization procedure differs between studies (e.g., some are standardized by grade, site, or year, and some by the overall standard deviation). As the standardization may affect the effect size, variation in this procedure may be one reason for the variation in effect sizes between studies. Positive estimates imply a beneficial effect.

The effects of universal preschool programs for the general population of children on test scores and school grades are mixed, in the sense that Table 3 contains significant beneficial and harmful effects, as well as insignificant estimates. Both the ITT and TOT effects range from large harmful effect sizes to large beneficial effect sizes. The meta-analysis included in the supplementary material, section A6, indicate substantial heterogeneity, which supports our assessment that the effects are mixed.

Most studies reporting heterogeneity over SES find more beneficial/less harmful effects for children from families with low SES, and no study found a consistent opposite pattern. The absolute magnitude of effects was larger for girls in all studies reporting heterogeneity across gender. However, most gender differences are small.

3. Primary and secondary school progression

Effects of universal preschool programs on outcomes related to primary and secondary school progression are measured by indicators of making age-adequate progress (e.g., being on-grade and probability of not being retained), of having been retained (one exception, Dumas & Lefranc, 2012, use the number of retentions), or by indicators for having graduated/being enrolled, or for having dropped out. We transform measures of making age-adequate progress into measures of grade retention and dropout measures into measures of graduation or being enrolled. Progress and retention rates mirror each other in the sense that the probability of adequate progress equals the probability of never being retained. Graduation and dropout could differ, if there are students who did not graduate on time but had not yet dropped out. However, high school dropout is always measured several years after appropriate graduation in the included studies, making such problems unlikely. The range for child age at measurement is 9-39 years.

In Table 4, we convert all estimates to percentage points and report the relative effects in percent. The negative estimates represent beneficial effects regarding grade retention, while positive estimates represent beneficial effects regarding graduation and being enrolled.

The included studies indicate that universal preschool programs have beneficial effects on measures related to primary and secondary school progression. All estimates of the average effects for the general population indicate beneficial effects of either attending a preschool program (TOT/LATE estimates) or growing up in a more exposed area (ITT estimates), and 7 out of 12 estimates are statistically significant on the five percent level. The effects are larger for children from low SES families in all but two cases in the six studies reporting heterogeneous effects. Two studies report harmful effects for high SES children, but none of the estimates are significant. Effects are less beneficial for girls in three studies and more beneficial in two.

Table 4. Effects on school enrollment, grade retention, being on-grade, and dropout

(1) *Study*	(2) *Average effect*	(3) *SES*	(4) *Gender*
Bastos et al. (2017): ITT estimate of having access to a preschool on probability of primary school enrollment (E) and being retained (R) at age 12.	E: 3.0 (3.5%) R: **-2.4** (-2.7%)	*Share of adults with no education in community* *Low* E: **5.1** (5.9%) R: -**3.6** (-4.1%) *High* E: 0.51 (0.6%) R: -1.1 (-1.2%)	*Girls* E: 2.1 (2.4%) R: **-2.5** (-2.8%) *Boys* E: 4.3 (5.0%) R: **-2.5** (-2.8%)
Berlinski et al. (2008): TOT estimate of attending preschool 1-3 years (Mean = 1.75) compared to 0-1 years on probability of being enrolled at age 15.	**27** (30%)	*Mother's education* *Low*: **27** (30%) *High*: 8.4 (9.2%)	*Girls*: **24** (27%) *Boys*: **36** (40%)
Bietenbeck et al. (2017): TOT effect of attending preschool on probability of being enrolled at age 13-16.	*Kenya*: **2.0** (2.1%) *Tanzania*: **9.0** (10.1%)	Not reported.	Not reported.
Bingley et al. (2018 ITT effect of living in a neighborhood with a preschool when 4 years-old on the probability of obtaining a high school/vocational degree at age 35.	**0.9** (1.2%)	Not reported.	Not reported.
Borraz & Cid (2013): LATE estimate of attending preschool on probability of being retained at age 15.	-4.4 (-15%)	*Mother's education* Less educated: 7.5 (25%)	*Girls*: -2.5 (8.3%) *Boys*: 16 (54%)
Dumas & Lefranc (2012): LATE of attending one more year of preschool on the number of grade repetitions at age 16 and probability of high school graduation.	No. of grade repetitions: -0.076 (-9.4%) High school graduation: **15** (20%)	Not reported (for IV specification).	Not reported (for IV specification).
Felfe et al. (2015): ITT effect of being more exposed on probability of being retained in secondary school.	-3.2 (-10.9%)	*Parents without/with a secondary school degree* *Without*: -3.7 (-12.6%) *With*: -1.9 (-6.5%)	Girls: -4.5 (-15%) *Boys*: -1.9 (-6.5%)
Fitzpatrick (2008): ITT effect of being more exposed on probability of being retained in 4th grade.	-0.7 (-4.5%)	*Free/reduced price lunch White* *Eligible*: **-2.0** (-12.7%) *Not eligible*: 0.1 (0.6%) *Black* *Eligible*: **-2.5** (-15%) *Not eligible*: **-6.0** (-38%)	No differential effects by gender (results mentioned in text).
Havnes & Mogstad (2011): ITT effect on being more exposed on probability of high school graduation at age 30-39.	**1.0** (1.4%)	*Mother's high school degree* *No degree*: **1.3** (1.7%) *Degree*: 0.21 (0.29%)	*Girls*: **0.81** (1.1%) *Boys*: **1.2** (1.7%)
Herbst (2017): ITT effect of $100 more in spending on probability of high school graduation at age 24-39.	**2.1** (2.7%)	Not reported.	Not reported.

Note: Absolute effects are reported in percentage points and relative effects in percent (in parentheses). Positive estimates represent beneficial effects for enrollment and graduation. Negative estimates represent beneficial effects for the number of grade repetitions and probability of being retained. Significant estimates ($p < 0.05$), as reported by the studies, are shown in bold. Type of SES heterogeneity is shown in italics in column 5.

4. Years of schooling and highest grade completed

Table 5 shows the estimates from four studies that report effects on years of schooling and two studies that report effects on the highest grade attained. As the highest grade is typically measured in years, the two outcomes are comparable and we report them in years. Furthermore, the table includes two estimates of the probability of obtaining a college and bachelor's degree. Age at measurement range from 13 to 66.

Six studies report average effects of attending or being more exposed to universal preschool on the years of schooling and highest grade completed. All estimates indicate significant improvements. Six studies examine how the effect differ across SES. Four of these studies find that the effect is largest for low SES children, one study finds that children with highly educated mothers gained the most, and one study find no consistent differences over SES. Four studies report heterogeneous effects across gender. All four studies find beneficial effects for both genders, and the differences between boys and girls are mostly small. One study finds larger effects for boys.

5. Employment and earnings

Table 6 shows the estimates from studies that examine the effect of universal preschool programs on measures related to earnings, employment, and welfare, measured at ages from 23 to 59 years. We report the estimates for earnings in percent and the probability of being employed and on welfare in percentage points and percent. Positive estimates in Table 6 indicate beneficial effects on earnings and employment, while negative estimates indicate beneficial effects on the probability of being on welfare.

Table 5. Effects on years of schooling and highest grade completed

(1) *Study*	(2) *Average effect*	(3) *SES*	(4) *Gender*
Berlinski et al. (2008): TOT effect of attending preschool on years of schooling at age 15.	**0.79**	*Mother's education* *Low*: **0.74** *High*: 0.25	*Girls*: **0.88** *Boys*: **0.89**
Bietenbeck et al. (2017): TOT effect of attending preschool on highest grade completed at age 13-16.	*Kenya*: **0.12** *Tanzania*: **0.11**	*Household wealth above or below median* (Results shown in figure only). *Kenya*: Children below median have insignificantly higher effects. *Tanzania*: Children below median have significantly lower effects.	Not reported.
Bingley et al. (2018): ITT effect of living in a neighborhood with a preschool when 4 years-old on years of schooling and the probability of obtaining a college degree at age 35.	<u>Years of schooling</u> **0.092** <u>College degree</u> Percentage points (%) **0.017** (5.4%)	*Maternal schooling level* <u>Years of schooling</u> *Basic*: 0.021 *High school/vocational training*: **0.064** *College/university*: **0.077**	<u>Years of schooling</u> *Girls*: **0.049** *Boys*: **0.133**
Bingley & Westergård-Nielsen (2012): ITT of being more exposed on years of schooling at age 23-30.	Not reported	Estimates reported for many subgroups: Children of less educated mothers have significantly more years of schooling. Preschool does not have any significant effect on education for children with higher educated mothers or fathers.	No general pattern found.
Havnes & Mogstad (2011, 2015): ITT effect on years of schooling at age 30-39 and the probability of attending college at age 33-42.	<u>Years of schooling</u> **0.074** <u>Attending college</u> Percentage points (%) **1.2** (3.3%)	*Mother high school (HS) education or not and family income* <u>Years of schooling</u> *Low*: **0.24** *Mid*: **0.081** *High*: 0.018 <u>Attending college</u> *No HS*: **1.4** (3.7%) *HS*: 0.33 (0.88%)	<u>Years of schooling</u> *Girls*: **0.066** *Boys*: **0.084** <u>Attending college</u> *Girls*: **1.2** (3.3%) *Boys*: **1.2** (3.2%)
Haimovich Paz (2015): ITT effect of exposure to kindergarten on maximum grade attainment at age 30-66.	**0.18**	*Mother tongue* *Non-English*: **0.29** *English*: **0.14**	The sample consists only of boys.
Herbst (2017): ITT effect of \$100 more in spending on probability of obtaining a bachelor's degree at age 24-39.	Percentage points (%) **1.9** (27%)	Not reported.	Not reported.

Note: Effects measured in years or in percentage points and percent, the latter shown as X (Y%). Statistically significant effects ($p < 0.05$), as reported in the studies, are shown in bold. Type of SES heterogeneity is shown in italics in column 5.

Table 6. Effect on earnings, employment, and being on welfare

(1) *Study*	(2) *Average effect*	(3) *SES*	(4) *Gender*
Bingley et al. (2018): ITT effect of living in a neighborhood with a preschool when 4 years-old on (log) earnings and probability of having no earnings at age 35.	Earnings: **1.2**% No earnings: -0.2 (-1.6%)	*Maternal education* Earnings: *Basic*: 0.00% *High school*: 1.1% *College*: 1.5%	Earnings: *Girls*: 0.1% *Boys*: **2.2**% No earnings: *Girls*: **0.005** (4.1%) *Boys*: -0.008 (-6.6%)
Bingley & Westergård-Nielsen (2012): ITT of being more exposed on (log) earnings at age 23-30.	Not reported	Interaction terms between dummies for parental earnings quartile and preschool density are largely negative for low earnings and positive for high.	No general pattern found.
Havnes & Mogstad (2011, 2015): ITT effect on being more exposed on earnings at age 33-42 and probability of being on welfare at age 30-39.	Earnings: 0.092% Being on welfare: **-0.91** (-5.6%)	*Mother high school (HS) education or not and family income* Earnings: *Low*: **2.9**% *Mid*: -0.50% *High*: -2.0% The quantile treatment effects indicate significant beneficial effects in the lower earnings quantiles. The effects start to turn negative around the 80th quantile and are substantial and significant at the top. Being on welfare: *No HS*: **-0.84** (-5.1%) *HS*: **-1.2** (-7.6%)	Earnings: *Girls*: 0.22% *Boys*: -0.22% Being on welfare: *Girls*: **-1.2** (-7.2%) *Boys*: **-0.63** (-3.9%)
Haimovich Paz (2015): ITT effect of being more exposed on earnings at age 25-45 for white males.		*Mother tongue* *Non-English*: **4**% *English*: 1%	**1.5**% (The sample consists only of boys.)
Herbst (2017): ITT effect of $100 more in spending on ln(earnings), being employed, or receiving public assistance at age 44-59.	Earnings **2.5**% Employed last year: **0.5** (0.6%) Public assistance: **-0.2** (-7.1%)	Earnings: The quantile treatment effects (reported in a figure) are positive for all quintiles, but the magnitudes are larger for the lower quintiles.	Not reported.

Note: Effects on earnings are given in percent and are either calculated by dividing the effect estimate by mean earnings, or from beta-coefficients where the outcome variable is transformed to ln(earnings). Effects on the probability of being employed, being a recipient of welfare benefits, or having no earnings are given in percentage points and percent, shown as X (Y%). Statistically significant effects ($p < 0.05$), as reported in the studies, are shown in bold. Type of SES heterogeneity is shown in italics in column 5.

All estimates for the three outcomes indicate beneficial effects for the general population of children. Most estimates are statistically significant. However, the average effect

contains substantial heterogeneity in some cases: all but two studies find larger effects for low SES children or lower quintiles of the income distribution,[4] and some estimates indicate significantly harmful effects for high SES children or higher quantiles of the income distribution.[5] In contrast, there is no consistent pattern in the studies that examine heterogeneity across gender.

6. Benefit-cost analyses

Universal preschool programs involve a substantial amount of public spending, and one of the most important questions for policy makers is whether the total benefits outweigh the cost of implementation. Three studies in our sample include a BCA: Berlinski, Galiani, and Manacorda (BGM, 2008) examine a program in Uruguay; Cascio and Schanzenbach (CS, 2013) examine programs in the US states of Georgia and Oklahoma; and van Huizen, Plantenga, and Dumhs (vHPD, 2017) use estimates from Felfe, Nollenberger, and Rodríquez-Planas (2015) to analyze a Spanish program.

In all three studies, the estimated benefit-to-cost ratio is clearly above one, meaning that for every dollar the government invest in the universal preschool program society receive more than one dollar in return. The ratios therefore indicate that the examined universal preschool programs were a worthwhile investment. However, the three BCAs build on several assumptions and estimates. We discuss the main assumptions and compare their estimates to other estimates from the included studies below.

The three studies extrapolate child earnings from the effect of universal preschool programs on either test scores (vHPD, CS) or years of schooling (BGM, 2008) when children were around 15 years of age. They assume that that the relationship between test score/years of schooling and earnings is constant over a child's career. BGM use a TOT estimate, while vHPD and CS transform ITT estimates to TOT estimates by dividing the ITT with the differential take-up rate between treatment and control groups. The estimated increase in lifetime earnings are 1.3 (CS), 6.0 (vHPD), and 7.9 percent (BGM). To compare these estimates to the estimates of earnings reported in Table 6, we convert the TOT estimates in CS and vHPD back to ITT estimates. The test scores amount to

[4] Both the exceptions study Danmark (Bingley et al., 2018; Bingley & Westergaard-Nielsen, 2012).

[5] The quantile treatment effects for high and low quantiles estimated in Herbst (2017) and Havnes & Mogstad (2015) are not necessarily comparable to the estimates for high and low SES in other studies, as the quantiles are defined by the post-treatment income distribution and not by (pre- or post-treatment) parental indicators of SES. However, as parental SES and child income tend to be positively correlated, we report and discuss them together.

around 0.03 (CS) and 0.15 standard deviations (vHPD), and the earnings estimate to around 0.3% (CS) and 1.5% (vHPD). The CS estimate are not particularly large compared to our other estimates, while vHPD is among the largest. We have fewer estimates to which we can compare BGM's estimates, but both the years of schooling effect and the earnings estimate seem to be larger than most of our other estimates. The larger effects might by fully reasonable though, given their developing country context.

BGM and CS does not include effects on maternal employment or tax revenues. The program studied by vHPD increased maternal employment, and vHPD include increased earnings for mothers, extrapolated from the employment estimate, as an additional benefit. Furthermore, vHPD include increased tax revenues from the increased child and maternal income, as well as benefits to tax payers from improving graduation and retention rates. The main share of benefits derives from improving child earnings though; tax revenues and maternal earnings make up less than 35 percent of total benefits.

All three studies assume that the estimates extend to all treated children and that any spill-over or general equilibrium effects of the intervention are ignorable. The studies do not include effects on for example welfare dependency, crime, health, and well-being, and there are no estimates of intergenerational effects included in the analyses.[6] These omissions may understate the total benefits of the programs, as the omitted outcomes seem likely to be positively associated with test scores and years of schooling.

All studies include the direct cost of the program for tax payers and parents (net of any decreased costs due to, for instance, out-of-pocket spending on other programs for parents). Only BGM include a cost of children staying in school for more years and a cost of obtaining revenue to finance the program (in their case, the projected interest on a loan). Raising tax revenue to pay for operating costs, or the interest on a loan, may, depending on how the tax is designed, be costly due to deadweight losses. It is not obvious how large such losses would be, but some government guidelines for benefit-cost analyses use 10-20 percent of the costs funded by general taxation (e.g., Finansministeriet, 2017; New Zealand Treasury, 2015).

The discount rate typically has a great impact on the results of BCAs. BGM and vHPD use a 3 percent discount rate for their baseline scenarios, and CS used the 30-year return

[6] Rossin-Slater & Wüst (2017) find beneficial intergenerational effects on educational attainment from a targeted Danish preschool program.

on US Treasury bills, which was 3.4 percent at the time. The benefit-to-cost ratios in the baseline scenarios were 3.2 in CS, 4.3 in vHPD, and 19 in BGM. All ratios are above one, also when substantially higher discount rates are used.

Summing up, both benefits and costs appear to be underestimated in the three studies. The omitted posts on the benefit side are, in our view, potentially more consequential than the omitted posts on the cost side. Increasing program costs by 20 percent to account for deadweight losses of taxation would, for example, not drive the ratios below one in any of the studies. Regarding the extrapolation of test scores and years of schooling to earnings, CS did not stand out in comparison to our other estimates, but still produce a ratio quite far above one. Other universal preschool programs showing beneficial effects may therefore also have benefit-to-cost ratios comfortably over one. In turn, this indicates that the magnitude of the included effects is often substantial; especially considering that many control groups were not no-treatment controls, only not as exposed as the treatment group, which reduces the magnitudes of the effects.

7. Comparison of preschool types

This section describes the three studies that compare long-term outcomes for different types of universal preschool programs. Because the studies compare different programs and are so few, we are unable to draw any general conclusions about preschool types on the long-term effects for children.

Datta Gupta and Simonsen (DGS, 2012, 2016) compare children who attend public center-based care to children who attend family day care, exploiting the variation in the composition of the type of child care Danish municipalities provide. DGS (2012) find that children who attend center-based preschool at age three like school more than children who attend family day care at age 11, but find no significant differences on a number of other outcomes, such as the Strengths and Difficulties Questionnaire, language and cognitive skills, delayed school entry, smoking, alcohol, and petty theft and vandalism. DGS (2016) find that enrollment in center-based care at age two increases enrollment in the academic track in high school at age 17 and the average grade in Danish at age 15. The authors find significant effects for boys on all outcomes, while only the increase in the average grade in Danish is significant for girls. The effects are larger for children of mothers with no more than high school education compared to children of mothers with some higher education.

Biroli et al. (2018) compare the Reggio Emilia approach, originating from the Italian city of the same name, with preschool approaches given to children in the nearby cities of Padova and Parma. They find that the Reggio Emilia approach preschool significantly increased the probability of ever having voted in municipal and Regional elections at age 30 (compared to Padova), but find no significant effects on a number of measures, including IQ, educational attainment, and health. The authors conclude that the differences in quality between the Reggio Emilia approach and the alternative programs were not sufficiently large to show substantial differences in outcomes for the adult population.

V. Discussion

Below we discuss our most important findings, first regarding the average effects for the general population of children, and second regarding the heterogeneity in terms of SES and gender. We then discuss the limitations of the review, and lastly offer some concluding remarks and suggestions for further research.

A. Effects for the general population of children

We have two main findings regarding the average effects: Firstly, the effects on test scores and school grades, and on measures related to health, well-being, and behavior vary across and sometimes within studies. The magnitudes also vary, and most estimates are not statistically significant. Secondly, all estimates for outcome measures related to adequate primary and secondary school progression, years of schooling and highest degree completed, and earnings and employment indicate beneficial average effects. The magnitudes of these estimates are often substantial and statistically significant. Furthermore, the three included BCAs indicate benefits-to-costs ratios clearly above one. While the majority of studies and estimates thus indicate that universal preschool programs have beneficial long-term effects on average, the differences between outcome types are important to understand, and we discuss potential explanations below.

A simple explanation for the differences between outcome types is that some programs are of a low enough quality to be harmful on average. The few studies that include estimates in primary and secondary school as well as in adulthood show consistent

beneficial average effects over time. The studies showing harmful effects have not yet estimated effects in adulthood and harmful effects may be equally persistent.

A different interpretation is that the full effects of universal preschool are better captured by the longer-term measures. Measures like graduation, earnings, and employment are arguably influenced by a broader set of skills than some of the measures for which studies found harmful effects, e.g., test scores. For example, improved personality skills seem to be the best explanation for the patterns in the Perry Preschool program of on the one hand enduring beneficial effects on crime, health, and earnings, and on the other hand short-term but quickly fading effects on cognitive skills (Heckman et al., 2013). However, as the studies we include found some harmful effects on crime, health, and behavioral measures, lasting effects on personality skills cannot explain all the differences found between outcome types.

Another potential explanation is that harmful effects may wane, either because other interventions are later given to children who for example fall behind in school or naturally as children get older. For example, there could be short-term harmful effects on health and socialization from being around other children, but such effects may pass or even turn beneficial over time (e.g., Strachan, 1989; Baker et al., 2008). Although there is some evidence of fadeout of initial harmful effects (see e.g., Lebihan et al., 2017), most of the estimates seem too long-term for waning effects to be a major explanation of the differences between outcomes.

Some outcome measures may be noisier and therefore more likely to produce both harmful and beneficial estimates by chance. The cognitive skills tests in the included studies were often not high stakes for students, and incentives and motivation to perform well matter for test results (e.g., Kautz et al., 2014). If children do not put in a lot of effort, the chance component of test scores may be substantial. However, there are harmful effects on outcomes that are not measured at one test occasion and involve high stakes. The increased crime rates found in Baker et al. (2015) is perhaps the best example.

In our view, the differences between outcome types are not due to upward bias in studies showing beneficial effects. Individual studies may of course be biased, but the risk of upward bias in studies showing beneficial effects did not seem to be higher than the risk of downward bias in studies showing harmful effects. The included studies are more likely to systematically overstate statistical significance, due to for example lack of

proper adjustment for multiple hypothesis testing and clustering. These problems also pertain to studies showing harmful effects and would not change the direction of the effects.

Many universal preschool programs lower the cost of child care for families. Beneficial effects may partly be explained by increased family incomes, as emphasized in the theoretical framework. However, some of our included studies examine programs with clearly positive income effects but still find significant harmful effects (e.g., Baker et al., 2015; Lebihan et al., 2017), which suggests that the quality of universal preschool programs is of first-order importance. This is also in line with results from related literatures where increased income has not reliably produced beneficial effects on child development (e.g., Heckman & Mosso, 2014; but see Black, Devereux, Løken, & Salvanes, 2014 for an example of positive income effects in a preschool context).

In sum, we cannot rule out a combination of other explanations, but the simplest explanation of the differences between outcome types is that they are caused by different universal preschool programs having different quality, and therefore, different effects. Indeed, given the variation in factors related to quality in the studied programs (see Table A 1 in the supplementary material) it would have been surprising if we had not found some differences. It is perhaps more surprising, also in relation to the message from prior reviews, that the results are not more mixed. We return to the causes of quality differences below, where we discuss heterogeneity in terms of SES and gender.

B. Heterogeneity across socioeconomic status and gender

Beneficial effects on average do not imply that universal preschool is good for all children. Our theoretical discussion noted that gender differences may be quite subtly dependent on the features of both the preschool program and the counterfactual mode of care, as well as the initial skill levels of boys and girls. The information needed to tease out the effects of these features is rarely present in the included studies and we do not find a consistent pattern of gender differences in the effects of universal preschool programs. Therefore, we focus the discussion below on the differences between children with high and low SES.

Previous reviews emphasize that the effects of universal preschool programs are more beneficial for disadvantaged, or low SES, children. Our synthesis shows that this tendency is present for many outcomes also in our review. The relatively large beneficial

effects found in studies from developing countries, where more or most children are low SES in comparison with developed countries, is also consistent with the pattern of more beneficial effects for low SES children.

Our theoretical framework shows that a more advantageous home environment could both imply smaller and larger beneficial effects of universal preschool. The effects could be smaller because the difference in quality between preschool and the counterfactual mode of care are smaller and larger because self-productivity and dynamic complementarities amplify initial skill differences. Our findings indicate that the first of these mechanisms is the most important. The effects of self-productivity and dynamic complementarities are typically not strong enough to offset low SES children's larger quality difference between universal preschool and their counterfactual mode of care. Bingley, Jensen, and Sander (2018) is the only counterexample of high SES children having consistently and significantly more beneficial effects than low SES children. The authors explain the findings by high SES children's counterfactual mode of care being low quality informal care and not, like for most low SES children, parental care. That is, it is again quality differences between the preschool program and the counterfactual mode of care that explain the different effects.

In most contexts, universal preschool programs therefore seem to reduce socioeconomic inequalities. This is good news for governments looking for ways to provide equal opportunities to all children, but it matters greatly whether the reduction in inequality is caused by relative improvements of low SES children's skills or by an absolute reduction of high SES children's skills. Our theoretical discussion in Section 2 indicate that harmful effects would be more likely and largest for high SES children because quality differences are more likely to be negative and such negative differences would be amplified by self-productivity and dynamic complementarities. In line with this discussion, nearly all examples of significant harmful effects we find are for children with relatively high SES (Fort et al., 2018; Havnes & Mogstad, 2015; Lebihan et al., 2017) or, in the case of Baker et al. (2015), was most likely driven by this group.[7] We center the further discussion on these studies.

[7] Lebihan et al. (2017) is the only study finding significant harmful effects for low SES children (on measures of anxiety and indirect aggression).

Havnes and Mogstad (2011, 2015) find significant harmful effects on total earnings for the highest quantiles of the earnings distribution and the probability of being a high and top earner for high SES children. The effects on cognitive test scores, years of schooling, the probability of attending college, graduating from high school, being a low earner, and being on welfare indicate beneficial effects for high SES children or the highest quantiles, although some are insignificant.[8] As one would expect more rivalry on local labor markets over (high) earnings than rivalry over the other measures, this pattern of results is consistent with harmful effects being due to larger relative improvements of skills for low and middle SES children. The pattern is, however, also consistent with preschool lowering the absolute level of some skills for high SES children that are important for earnings, but not (as important) for the other measures.

The significant harmful effects found by Fort et al. (2018) in Bologna, Italy, on a cognitive skills test and two personality measures, and by Baker et al. (2015) and Lebihan et al. (2017) in Quebec, Canada, on measures of anxiety, quality of life, and crime are more likely due to an absolute reduction of skills for high SES children. These outcome measures have low or no degree of rivalry and low risk of ceiling and floor effects. The program in Quebec have received low quality assessments, especially at its inception (e.g., Almond et al., 2017; Cascio, 2015; Lebihan et al., 2017; van Huizen & Plantenga, 2015), and had a low staff-to-child ratio (see Table A 1). The program in Bologna is often considered high quality, although it has a relatively low staff-to-child ratio for the 0-2 age group studied by Fort et al. (2018). The samples in Baker et al. (2015) and Lebihan et al. (2017) also include younger children (their age range was 0-4 years), and, as discussed in Section 2, preschool may be more likely to have harmful effects for very young children. Relatively few other included studies examined 0-2-year-olds, and most of those that did also included older children. It is therefore difficult to tell from our sample whether these harmful effects were program or age specific, or a combination of the two.

[8] The effects on cognitive test scores, years of schooling, college attendance, and high school graduation are insignificant. The latter three are measures were ceiling effects are conceivable for high SES children, which may make it more difficult to find significant effects.

VI. Conclusion

We present evidence from a systematic review on the long-term effects of universal preschool programs. Across a wide range of countries and programs, estimates related to adequate primary and secondary school progression, years of schooling and highest degree completed, and earnings and employment, indicate beneficial average effects. However, the effects on test scores and school grades, and health, well-being, and behavior – which are measured earlier in children's lives – are mixed.

Choosing among preschools, rather than choosing whether to put their child in preschool or not, is the choice facing many parents in large parts of the world. However, we found few studies comparing preschool types on long-term outcomes, and we cannot draw any general conclusions other than that more studies are needed.

We find no general gender pattern in the results, whereas children from low SES families seem more likely to benefit from universal preschool programs than high SES children. This creates an opportunity for policy makers to reduce inequality between children from different backgrounds by providing universal preschool for all children, although we want to stress that it matters greatly how inequality is reduced. The few significant harmful effects we find for primarily high SES children should be taken seriously.

An inherent limitation in a review of long-term outcomes is that it is unclear how the universal preschool programs examined in the included studies relate to present day programs. That is, extrapolation of the results to the universal preschool programs of today should be done with caution.

References

Studies included in the synthesis are marked with *.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In *Handbook of the Economics of Education* (Vol. 4, pp. 1-181). Amsterdam: Elsevier.

Almond, D., Currie, J., & Duque, V. (2017). *Childhood circumstances and adult outcomes: Act II* (No. w23017). National Bureau of Economic Research.

Baker, M. (2011). Innis Lecture: Universal early childhood interventions: what is the evidence base?. *Canadian Journal of Economics*, 44(4), 1069–1105.

Baker, M., Gruber, J., & Milligan, K. (2008). Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy*, 116(4), 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* (No. w21571). National Bureau of Economic Research.

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

*Bastos, P., Bottan, N. L., & Cristia, J. (2017). Access to Preprimary Education and Progression in Primary School: Evidence from Rural Guatemala. *Economic Development and Cultural Change*, 65(3), 521-547.

Belsky, J. (2001). Emanuel Miller Lecture: Developmental risks (still) associated with early child care. *Journal of Child Psychology and Psychiatry*, 23, 396-404.

*Berlinski, S., Galiani, S., & Gertler, P. (2009). The effect of pre-primary education of 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.

Bertrand, M., & Pan, J. (2013). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1), 32-64.

*Bietenbeck, J., Ericsson, S., & Wamalva, F. (2017). *Preschool attendance, school progression, and cognitive skills in East Africa* (IZA Discussion Paper Series, DP No. 11212).

*Bingley, P., & Westergaard-Nielsen, N. (2012). Intergenerational transmission and day care. In: Ermisch, J., M. Jäntti, and T. Smeeding (eds). *From Parents to Children: The Intergenerational Transmission of Advantage* (p. 190–204). New York: Russell Sage Foundation.

*Bingley, P., Jensen, V. M., & Sander, S. (2018). *One size fits all? Effects of universal daycare on long-run child and mother outcomes*. Unpublished manuscript.

*Biroli, P., Del Boca, D., Heckman, J. J., Heckman, L. P., Koh, Y. K., Kuperman, S., Moktan, S., Pronzato, C. D. & Ziff, A. L. (2018). Evaluation of the Reggio Approach to early education. *Research in Economics*, 72(1), 1-32

Björklund, A., & Salvanes, K. G. (2011). Education and family background: Mechanisms and policies. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 201–247). Amsterdam, Netherlands: North-Holland.

Black, S. E., Devereux, P. J., Løken, K. V., & Salvanes, K. G. (2014). Care or cash? The effect of child care subsidies on student performance. *Review of Economics and Statistics*, 96(5), 824–837.

Black, M. M., Walker, S. P., Fernald, L. C., Andersen, C. T., DiGirolamo, A. M., Lu, C., ... & Devercelli, A. E. (2017). Early childhood development coming of age: Science through the life course. *Lancet*, 389(10064), 77–90.

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

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons, Ltd.

Borghans, L., Golsteyn, B. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences*, 113(47), 13354–13359.

*Borraz, F., & Cid, A. (2013). Preschool attendance and school-age profiles: A revision. *Children and Youth Services Review*, 35, 816-825.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, 53(1), 371–399.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In W. Damon (Series Ed.) & R. M. Lerner (Vol. Ed.), *Handbook of child psychology: Theoretical models of human development* (pp. 793–828). New York, NY: Wiley.

Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E., & Pan, Y. (2014). Early childhood investments substantially boost adult health. *Science*, 343(6178), 1478–1485.

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. (2015). The promises and pitfalls of universal early education. *IZA World of Labor*, 116, 1–16.

Cascio, E. U. (2017). *Does universal preschool hit the target? Program access and preschool impacts* (No. w23215). National Bureau of Economic Research.

*Cascio, E. U., & Schanzenbach, D. W. (2013). The impacts of expanding access to high-quality preschool education. *Brookings Papers on Economic Activity*, 127–178.

Costa, J. C., da Silva, I. C. M., & Victora, C. G. (2017). Gender bias in under-five mortality in low/middle-income countries. *BMJ Global Health*, 2(2), e000350.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97(2), 31–47.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3), 883-931.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference?. *American Economic Review*, 85(3), 341–364.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, 1(3), 111–134.

Duckworth, A. L., & Seligman, M. E. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological science*, 16(12), 939-944.

*Dumas, C., & Lefranc, A. (2012). Early schooling and later outcomes: Evidence from pre-school extension in France. In: Ermisch, J., M. Jäntti, and T. Smeeding (eds). *From Parents to Children: The Intergenerational Transmission of Advantage* (p. 164–189). New York: Russell Sage Foundation, 2012.

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2015). *Early childhood education* (No. w21766). National Bureau of Economic Research.

*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., Nollenberger, N., & Rodríquez-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.

Finansministeriet (2017). *Vejledning i samfundsøkonomiske konsekvensvurderinger*. Copenhagen: Finansministeriet.

*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 and Policy*, 8(1), Article 46.

Flaherty, S. C., & Sadler, L. S. (2011). A review of attachment theory in the context of adolescent parenting. *Journal of Pediatric Health Care*, *25*(2), 114–121.

*Fort, M., Ichino, A., & Zanella, G. (2018). *The cognitive cost of daycare 0-2 for children in advantaged families.* Unpublished manuscript.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505.

García, J. L., Heckman, J. J., Leaf, D. E., & Prados, M. J. (2016). *The life-cycle benefits of an influential early childhood program* (No. w22993). National Bureau of Economic Research.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., ... & Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*, 344(6187), 998–1001.

Goossens, F. A., & IJzendoorn , M. H. (1990). Quality of infants' attachments to professional caregivers: Relation to infant-parent attachment and day-care characteristics. *Child Development*, 61(3), 832–837.

*Gupta, N. D. & Simonsen, M. (2012). The effects of non-parental child care on pre-teen skill and risky behavior. *Economics Letters*, 116, 622-625.

*Gupta, N. D. & Simonsen, M. (2016). Academic performance and type of early childhood care. *Economics of Education Review*, 53, 217-229.

*Haimovich Paz, F. (2015). The long-term return to early childhood education: Evidence from the first US kindergartens. In Haimovich Paz, F., *Three Essays on the Economics of Education and Early Childhood* (Ch. 1), Dissertation, University of California.

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3), 607–68.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, 27(1), 4–9.

*Havnes, T. & Mogstad, M. (2011). No child left behind: Subsidized child care and children's long-run outcomes. *American Economic Journal: Economic Policy*, 3(2), 97–129.

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

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., & Mosso, S. (2014). The economics of human development and social mobility. *Annual Review of Economics*, 6(1), 689–733.

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6), 2052–2086.

Henry, G. T., & Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review*, 26(1), 100-112.

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

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* (OECD Education Working Papers, No. 110). Paris: OECD Publishing.

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *Quarterly Journal of Economics*, 131(4), 1795–1848.

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

*Lebihan, L., Haeck, C., & Merrigan, P. (2017). Universal childcare and long-term effects on child well-being: Evidence from Canada. *Journal of Human Capital*, forthcoming.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, 122(1), 159–208.

Magnuson, K. A., Kelchen, R., Duncan, G. J., Schindler, H. S., Shager, H., & Yoshikawa, H. (2016). Do the effects of early childhood education programs differ by gender? A meta-analysis. *Early Childhood Research Quarterly*, 36, 521–536.

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-Stevens, K., Petrogiannis, K., Ariescu, A., Penderi, E., Rentzou, K., & Leseman, P. (2015). *A review of research on the effects of early childhood education and care (ECEC) upon child development*. CARE project; Curriculum quality analysis and impact review of European early childhood education and care (ECEC).

New Zealand Treasury (2015). *Guide to social cost benefit analysis*. Wellington: New Zealand Treasury.

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(3), 199–206.

OECD (2016). *Society at a Glance 2016: OECD Social Indicators*. OECD Publishing, Paris.

OECD (2017). *Public spending on childcare and early education*. OECD family database, *http://www.oecd.org/els/family/database.htm. Accessed 2018-04-24*.

Phillips, D., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., & Weiland, C. (2017). *The current state of scientific knowledge on pre-kindergarten effects*. Brookings Institution and the Duke Center for Child and Family Policy.

Reynolds, A. J., & Temple, J. A. (2008). Cost-effective early childhood development programs from preschool to third grade. *Annual Review of Clinical Psychology*, 4, 109–139.

Reynolds, A. J., & Ou, S.-R. (2011). Paths of effects from preschool to adult well-being: A confirmatory analysis of the Child–Parent Center Program. *Child Development*, 82(2), 555–582.

Rossin-Slater, M., & Wüst, M. (2017). *What is the added value of preschool? Long-term impacts and interactions with an infant health intervention* (No. w22700). National Bureau of Economic Research.

Ruhm, C., & Waldfogel, J. (2012). Long-term effects of early childhood care and education. *Nordic Economic Policy Review*, 1(1), 23–51.

Sabol, T. J., Hong, S. S., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-K programs predict children's learning?. *Science*, 341(6148), 845–846.

*Smith, A. (2015). *The long-run effects of universal pre-K on criminal activity*. Unpublished manuscript.

Strachan, D. P. (1989). Hay fever, hygiene, and household size. *BMJ: British Medical Journal*, 299(6710), 1259–1260.

UNESCO (2018). Gross enrolment ratio, pre-primary, both sexes (%). UNESCO Institute for Statistics, *http://data.uis.unesco.org/index.aspx?queryid=142#*. Accessed 2018-04-24.

van Huizen, T., & Plantenga, J. (2015). *Universal child care and children's outcomes – A meta-analysis of evidence from natural experiments* (U.S.E Discussion Paper Series 15-13).

*van Huizen, T., Plantenga, J. & Dumhs, L. (2017). The costs and benefits of investing in universal preschool: Evidence from a Spanish reform. *Child Development*, forthcoming.

Vermeer, H. J., & van IJzendoorn, M. H. (2006). Children's elevated cortisol levels at daycare: A review and meta-analysis. *Early Childhood Research Quarterly*, 21(3), 390-401.

Waldfogel, J. (2015). The role of preschool in reducing inequality. *IZA World of Labor*, 219.

World Bank (2017). *Gross enrolment ratio, pre-primary, both sexes (%)*. UNESCO Institute for Statistics. *https://data.worldbank.org/indicator/SE.PRE.ENRR*. Accessed 2018-04-24.

Appendices

Supplementary material to universal preschool programs and long-term child outcomes: A systematic review (for online publication)

The contents of this supplementary material are as follows: Section A1 describes the included studies and examined preschool programs. Section A2 illustrates how we apply the inclusion criteria by providing examples of studies that we screened in full text but did not include, and some studies that met parts of a criterion and are included. Section A3 contains additional details about the results of the search and screening process. Section A4 provides a discussion about the main risks of bias in the type of research designs used and the quality of inference in the included studies. Section A5 describes the motivation for choosing one estimate over another in the cases where the choice did not obviously follow the principles described in the section IV in the main text. Section A6 describes the results of a meta-analysis of studies using test scores as the outcome variable. Section A7 contains the full search strings used to search the electronic databases.

A1. Information about included studies

Table A 1 describes the included studies in terms of country and region, the preschool program and control condition, staff-to-child ratios, group sizes and staff education, and the natural experiment and estimation strategy used. Studies that examine the same preschool programs are grouped together. When details about a preschool program was not included in a study, we use related information, or, if possible, information from other sources. We use the following acronyms: difference-in-differences (DID), intention-to-treat (ITT), treatment-on-the-treated (TOT), local average treatment effect (LATE), instrumental variable (IV), and regression discontinuity (RD). *N* denotes the number of areas, and *n* the number of child observations. Both numbers refer to the sample sizes used in the estimations of the mean effects. Ranges refer to the minimum and maximum *N* and *n* used in a study.

Table A 1. Characteristics of included studies

Included study	Country/region, period, and sample	Preschool program(s) & control condition	Staff-to-child ratio & staff education in preschool program(s)	Identification and estimation
Baker, Gruber & Milligan (2015) Lebihan, Haeck & Merrigan (2015)	*Country/Region:* Canada *Period*: 1997-2001 Preschool program introduced in 1997 and phased in over a period of four years to 2001. *Sample*: 0-4 years old. The program was open for four-year-olds in 1997 and became available for 0-1 years in 2000-2001. *N* = 10 *n* = 10,857-140, 926 (not reported per specification in Baker et al., 2015, some estimated from PISA/SAIP/PCAP participants from www.cmec.ca)	*Preschool program:* Quebec introduced a subsidy on universal preschool in 1997, making preschool available for everyone for 5 dollars a day. The program was introduced step-wise by age. Preschool under the program was provided in two venues: preschool centers (centres de la petite enfance, CPE) and home-based care. *Control condition*: Children in Quebec shift from informal care into center-based care. The proportion of 0-4-year-olds in care rose by 14 percentage points, or roughly one-third of the baseline rate. There are no substantial changes in the number of children that were cared for in their own home (Baker et al., 2008, Table 2, p. 724), indicating that the introduction of publicly available preschool crowds out informal care arrangements/privately provided child care.	*Staff-to-child ratio*: 0-3-year-olds: 1:8 4-5-year-olds: 1:10 (Baker et al., 2008). *Staff education*: Two-thirds of staff must have a college diploma or university degree in early childhood education (Baker et al., 2008).	*Identification:* Exploits the introduction of the subsidy on universal preschool for children aged 0-4 in Quebec. They use other provinces of Canada as a control group. *Estimation*: Use a DID strategy to estimate an ITT effect of being more exposed to a universal program, as the sample comprises all children and not only those that attend a preschool program.
Bastos, Bottan & Cristia (2016)	*Country/Region:* Guatemala, Rural communities *Period*: 1992-2000	*Preschool program:* Guatemala expanded their provision of public pre-primary schools from 5,300 to 11,500 during the period 1998-2005. The beneficiary communities were selected by the central government with no strict guidelines.	*Staff-to-child ratio*: Not reported. *Staff education:* Teachers must have a pre-primary education qualification; this is obtained	*Identification:* Exploits the large expansion of pre-primary schools and the variation over time and between communities.

	Sample: 4-6-year-olds. *N* = 960 *n* = 8,543	*Control condition:* Mainly parental care, as 0.8-1.2% of the communities had a private preprimary in 2005. Little or no crowding out of informal or private alternatives.	in teacher-training colleges (UNESCO, 2006b; however, this information is from 2006, a few years after the period examined in this article. Staff requirements may have been different in the period examined in the paper.)	*Estimation:* Use a DID strategy with trimming and propensity score re-weighting to estimate an ITT effect. The authors also estimate a TOT effect, but without any data on actual attendance.
Berlinski, Galiani & Manacorda (2008) Borraz & Cid (2013)	*Country/Region:* Uruguay *Period*: 1995-2004 Berlinski et al. (2008): *Sample*: The preschool program comprises 3-5-year-olds. *N*=55 *n*= 23,042 Borraz & Cid (2013) *Sample:* 4-5-year-olds *N* = not reported *n* = 19,732	*Preschool program:* Following a reform in the mid-1990s, the Government of Uruguay made pre-primary education universally available. Enrollment in public preschool rose by 76% over nine years. The expansion attracted mostly children from disadvantaged backgrounds. *Control condition:* Private provision/informal care (not explicitly described). Private fee-based education was common. In Montevideo, around one third of the children in primary education attended a private institution.	*Staff-to-child ratio*: 3-year-olds: groups of 20 4-year-olds: groups of 25 5-year-olds: groups of 35 *Staff education:* Early education teachers study in teacher training colleges to earn a qualification at the non-university tertiary level (UNESCO, 2006c. However, this information is from 2006, a few years after the period examined in this article. Staff requirements and ratios may have been different in the period examined in the paper.)	*Identification:* Exploits the expansion in the provision of public pre-primary education. Main specification in Berlinski et al. (2008) contrasts having attended preschool for 1-3 years (treatment) with the group that attends 0-1 years (control). The main specification in Borraz & Cid (2013) contrast attending preschool with not attending preschool. *Estimation:* Berlinski et al. (2008) use a sibling fixed-effects strategy to estimate a TOT effect. Borraz & Cid (2013) instrument preschool attendance with the mean preschool attendance by child age in each locality.
Berlinski, Galiani & Gertler (2009)	*Country/Region:* Argentina *Period*: 1994-1999 *Sample*: 3-5-year-olds *N* = 417 municipalities; 2,750-3,024 schools	*Preschool program:* Argentina increased the number of preschool classrooms during the period 1993-1999. The increase in pre-primary enrollment varies between provinces. All provinces increased enrollment by at least 10 percentage points. *Control condition:* family care (not explicitly described).	*Staff-to-child ratio*: Class size is 25. *Staff education:* Preschool teachers must be trained at teacher training colleges or at universities (UNESCO, 2006a; however, this information is from 2006, a few years after	*Identification:* Exploits the variation of treatment intensity across regions and cohorts following the expansion of pre-primary school facilities. They are unable to separate one, two or three years of exposure. *Estimation:* Use a DID strategy to estimate an ITT effect. The authors

	n =117,515-145,292		the period examined in this article. Staff requirements may have been different in the period examined in the paper.)	write that they cannot reject that the take-up rate was one, which would result in the estimates being close to a TOT estimate.
Bietenbeck, Ericsson & Wamalva (2017)	*Country/Region:* Kenya and Tanzania *Period*: Kenya: 2000-2013 Tanzania: 2000-2012 *Sample*: Kenya: 3-6-year-olds *N* = not reported *n* = 218,134 Tanzania: 5-6-year-olds *N* = 120 *n* = 288,084	*Preschool program:* There are three types of preschool in Kenya: public preschool, private preschool, and information neighborhood schools. In Tanzania, the vast majority of preschools are public. During the period 1997-2004, preschool enrollment increased from 79-84% in Kenya and from 61-69% in Tanzania. *Control condition:* Family care (not explicitly described).	*Staff-to-child ratio*: Kenya: 1:25-27 Tanzania: increased from 1:45 in 2007 to 1:100 in 2011 in state schools. *Staff education:* Kenya: Primary or secondary education. 41.4% trained teachers Tanzania: Teachers must have completed lower-secondary school. Information for Kenya: (UNESCO, 2005) Information for Tanzania: (World Bank, 2012).	*Identification:* Compare differences between siblings. The authors argue that differences between siblings are due to changes in the local availability of preschool because of an expansion of the pre-primary sector during the studied period. *Estimation:* Use a sibling fixed-effect strategy to estimate a TOT effect.

Bingley, Jensen & Sander (2018) Bingley & Westergård-Nielsen (2012)	*Country/Region:* Denmark Bingley et al. (2018): *Period*: 1967-1979 *Sample*: 3-6-year-olds *N* = 1,098 *n* = 403,241 Bingley & Westergård-Nielsen (2012): *Period:* 1976-1989 *Sample:* 0-6-year-olds *N* = 275 *n* = 531,733	*Preschool program:* A reform from 1964 increased the number of preschool slots. From 1966 to 1979, the number of institutions tripled. From 1976-1989, preschool coverage tripled for the youngest children (age 1-2) and doubled for the oldest (age 3-6). *Control condition:* parental care/informal/private non-maternal care (mainly for high SES mothers with work). Private alternatives to public institutions existed.	*Staff-to-child ratio*: Not reported *Staff education:* Each institution was to have a certain minimum proportion of educated personnel but could also employ untrained helpers. The duration of the education was three years.	*Identification:* Exploit the step-wise roll-out of reforms increasing universal preschool provision in Denmark. Use variation over time and between neighborhoods (Bingley et al., 2018) or municipalities (Bingley & Westergård-Nielsen, 2012). *Estimation:* Use a DID strategy to estimate an ITT effect.
Biroli et al. (2018)	*Country/Region:* Italy, Reggio Emilia, Parma & Padova *Period*: 1954-2000 *Sample*: 0-6-year-olds *N*=3 *n:* Adolescents = 836 Adults 30s = 782 Adults 40s = 791 Adults 50s = 449	*Type comparison:* Reggio Emilia approach is compared to the approaches in the nearby cities of Padova and Parma. *Preschool program:* The Reggio Emilia approach is notable for its investment in staffing, early inclusion of children with disabilities, and high rates of provision of early childhood services.	*Staff-to-child ratio*: 3-year-olds: 1:12-13 *Staff education:* On a biweekly basis, a pedagogista with at least a bachelor's degree in psychology or pedagogy supports the professional development for the educational staff of approximately 4-5 municipal preschools.	*Identification:* Compare children from Reggio Emilia with children from Parma and Radova, who received different kinds of child care approaches. *Estimation:* Use a DID with matching strategy to estimate an ITT effect.

Blanden, Bono McNally & Rabe (2016)	*Country/Region:* England *Period*: 2002-2007 *Sample*: 3-4-year-olds *N* = 888 *n* = 2,900,000	*Preschool program:* England implemented universal part-time preschool for three-year-olds in the early 2000s. The government funded private and voluntary institutions to provide free early education places. The expansion happened entirely in the private sector. *Control condition:* private or parental care. The expansion in preschool mainly crowded out other types of private provision of preschool, as 82% of 3-years-olds already attended some type of preschool education before the reform. The expansion increased the enrollment of three-year-olds by 14.4 percentage points.	*Staff-to-child ratio*: Public sector: 1:13. Private sector: 1:8 if no qualified teacher, 1:13 if qualified teacher. *Staff education:* Public sector: Almost all employed staff hold a degree. Private sector: 10-20% hold a degree.	*Identification:* Exploits the staggered implementation of universal part-time preschool education for 3-year-olds across Local Education Authorities in England. Compare low and high intensity areas. *Estimation:* Use a DID strategy to estimate an ITT effect.
Cascio & Schanzenbach (2013) Fitzpatrick (2008)	*Country/Region:* Cascio & Schanzenbach (2013): US, Georgia & Oklahoma Fitzpatrick (2008): US, Georgia *Period*: Cascio & Schanzenbach (2013): Georgia: 1995-2005 Oklahoma: 1998-2005 Fitzpatrick (2008): 1995-1999	*Preschool program:* Georgia and Oklahoma introduced universal preschool for 4-year-olds in the 1990s. The program in Georgia and Oklahoma increased the likelihood of enrollment in preschool at age four by 19-20 percentage points for low SES children and 11-14 percentage points for high SES children. The enrollment in pre-kindergarten in Georgia increased from 13.9% in 1995 to 53.0% in 1999. *Control condition:* informal/formal care, different for different subgroups of children. High SES children moved from private to public preschool.	*Staff-to-child ratio*: 1:10. *Staff education:* In both states, classroom lead teachers must hold a bachelor degree and participate in annual training.	*Identification:* Cascio & Schanzenback (2013): Compare changes in preschool enrollment in the two states that introduced universal preschool initiatives with the rest of the country over the same period. Fitzpatrick (2008): The article compares children in Georgia that were offered the public pre-kindergarten to children in other states and children before the program was introduced. *Estimation:* Use a DID strategy to estimate an ITT effect. Cascio & Schanzenback (2013) also perform a benefit-cost analysis.

	Sample: 4-year-olds *N* = 50 Cascio & Schanzenbach (2013): *n* = 295-334 state-years Fitzpatrick (2008): *n* = 537,112-1,241,994			
Datta Gupta & Simonsen (2012, 2016)	*Country/Region:* Denmark *Period*: 1996-1997 2012: *Sample*: 3-year-olds *N*= not reported *N* = 2,571-3,784 2016*:* *Sample:* 2-year-olds *N* = 253 *N* = 60,907	*Type comparison:* Compare center-based preschool to non-center-based but municipally-regulated family day care. *Preschool program:* Most children enrolled in family day care eventually enroll in center-based care. The interpretation of the result is an additional 1.5 years of early center-based care. They have data on actual attendance. At age 2 (3), 25% (33%) of enrolled children attend center-based child care arrangements, and 75% (67%) attend family day care.	*Staff-to-child ratio*: Center-based: 1:3.5 Family day care: 1:5 or less *Staff education:* Center-based: Most of core center staff hold a pedagogical degree Family day care: No formal education, but are offered vocational courses.	*Identification:* 19-30% of the municipalities offer guaranteed access to center-based preschool. In the municipalities that offer the guaranteed access, children have a higher probably of getting access to center-based preschool. *Estimation:* Use an IV strategy to estimate LATEs of center-based care relative to family day care for the group of children whose parents choose center-based care when access is guaranteed, but not otherwise.
Dumas & Lefranc (2012)	*Country/Region:* France *Period*: 1952-1983 *Sample*: 2-5-year-olds *N* = 95 *n* = 6,799-21,710	*Preschool program:* During the 1960s and 1970s the enrollment in preschool for 3-year-olds rose from 35% to 90%. The increase varied between regions. *Control condition:* Parental care. The contrast is described as getting one more year of preschool.	*Staff-to-child ratio*: Class size: 25 children. *Staff education:* Preschool teachers have a bachelor's degree.	*Identification:* The authors exploit regional variation in access to preschool. *Estimation:* Use an IV strategy to estimate a LATE.

Felfe & Lalive (2010)	*Country/Region:* Germany, every state except Berlin *Period:* 1996-2000 *Sample:* 0-3-year-olds *N* = Not reported *n* = 850	*Preschool program:* A substantial difference exists in child care offer rates across Germany, due to historical differences in the separated East and West Germany. Child care coverage rates are of the order of 40% in the former East Germany and below 10% in the former West Germany. There is also variation within regions. *Control condition:* Informal/parental care, although informal care arrangements were rarely used.	*Staff-to-child ratio*: East: 1:6.8 West: 1:5.1 *Staff education:* Staff has to undergo special training before being allowed to work in the sector. *Staff with a degree in child care:* East: 90% West: 84%	*Identification:* The authors use the difference in child care offer rates across Germany induced by the former East/West division as an instrument for attending preschool. *Estimation:* Use an IV strategy to estimate the effects of formal care for children of mothers who use formal care because of an increase in the child care offer rate.
Felfe, Nollenberger & Rodrígues-Planas (2015) Van Huizen, Duhms & Plantenga (2017)	*Country/Region:* Spain *Period:* 1991-1996 *Sample:* 3-year-olds. *N* = 15 (treatment: 8, control: 7) *n* = 20,458-40,340	*Preschool program:* Spain expanded their subsidized full-time, high quality universal child care supply in the early 1990s. The enrollment of 3-year-olds in public child care increased from 8.5 to 67.1% from 1990/1991 to 2002/2002. *Control condition:* mainly parental care, but part of the control group might have been in preschool.	*Staff-to-child ratio*: Maximum number of children per class is 20. *Staff education:* Preschool teachers are required to have a college degree in pedagogy.	*Identification:* Exploits the variation in the speed of expansion across states. Divide 15 states into treatment and control based on their increase in public child care enrollment of 3-year-olds. *Estimation:* Use a DID strategy to estimate an ITT effect. Estimates the effect of having a greater opportunity of one year of preschool when the child is three, no data over actual preschool attendance. Van Huizen et al. (2017) use estimates from Felfe et al. (2015) to perform a benefit-cost analysis.
Fort, Ichino & Zanellax (2018)	*Country/Region:* Italy, Bologna *Period*: 2001-2005	*Preschool program:* Parents in Bologna apply for a preferred child care program. Acceptance into a preferred child care program depends on a Family Affluence Index. Less affluent families get offered a spot first. This creates a threshold. On	*Staff-to-child ratio*: 0-year-olds: 1:4 1-2-year-olds: 1:6 *Staff education:* Not reported.	*Identification:* Use the threshold in the admission system that determines whether children are offered a preschool slot as an instrument for attendance.

	Sample: 0-2-year-olds *N* = 1 *n* = 444	average, children that get offered the preferred spot will be in child care for a longer time, compared to children that are not offered their preferred spot. *Control condition:* informal/parental care. Private child care is almost absent; extended family services are the most relevant substitution for child care.		*Estimation:* Use a fuzzy RD strategy to estimate a LATE.
Haimovich Paz (2015)	*Country/Region:* USA *Period:* 1890-1910 *Sample:* White males, 4-6-year-olds *N* = 220 *n* = 20,263-239,390	*Preschool program:* The kindergarten movement provided preschool for children aged 4-6. The increase in enrollment in the years following the incorporation of public kindergartens was rapid in many cities, ranging from 20 to 80 percentage points. *Control condition:* The mothers were most likely the care providers before the kindergarten movement. Some crowding out of private alternatives.	*Staff-to-child ratio*: Not reported. *Staff education:* Most kindergarten teachers were high school graduates with two years of specific training that included child psychology, music, and children's literature.	*Identification:* Exploit geographical variation and variation over time in the number of public kindergartens in cities following the kindergarten movement. *Estimation:* Use a DID strategy to estimate ITT effects.
Havnes & Mogstad (2011, 2015)	*Country/Region:* Norway *Period:* 1976-1979 *Sample:* 3-6-year-olds. *N* = 414 *n* = 499,026 (2011) *n* = 341,170 (2015)	*Preschool program:* A reform from 1975 increased the federal subsidy for child care. The local government was responsible for offering child care. The reform created large variation in the access to child care across municipalities and over time. *Control condition:* The analysis suggests that the new subsidized child care crowded out informal child care arrangements with almost no net increase in total use or maternal labor supply.	*Staff-to-child ratio*: 1:8 with at least one educated preschool teacher per 18 children. *Staff education:* Every formal child care institution had to be run by an educated preschool teacher responsible for day-to-day management. Preschool education is a college degree with supervised practice in a formal preschool institution included.	*Identification:* Compare municipalities with high coverage to municipalities with low coverage (above or below median percentage point increase in preschool coverage rates) *Estimation:* Use a DID strategy to estimate an ITT effect.
Herbst (2017)	*Country/Region:* USA, all states except New	*Preschool program:* During World War 2, The Lanham Act established center-based	*Staff-to-child ratio*: 1:10.	*Identification:* The article exploits the variation between states with

	Mexico, Alaska, and Hawaii. *Period:* 1943-1946 *Sample:* 0-12-year-olds *N* = 47 *n:* age 24-39 = 456,070 age 34-49 = 2,500,553 age 44-59 = 2,481,049	preschool for children aged 0-5 and after-school services for children aged 6-12. The intensity differed between states. *Control condition:* Parental care (not explicitly described).	*Staff education:* Program employed certified school teachers and contracted with universities to establish formal training programs.	low/high spending on the preschool program for children in states with high spending. *Estimation:* Use a DID strategy to estimate an ITT effect.
Kühnle & Oberfichtner (2017)	*Country/Region:* Germany, West Germany and for some measures the regions of Bavaria & Schleswig-Holstein *Period:* 1997-2002 *Sample:* 2-6 years-old *N* = 234 *n* = 7,211-102,523	*Preschool program:* Children usually start in preschool in summer of the calendar year they turn three. This creates a December/January discontinuity. Children born in January on average spend 5 months more in preschool than children born in December. *Control condition:* Parental care or informal preschool settings (not explicitly described)	*Staff-to-child ratio*: 1:7 *Staff education:* Not reported	*Identification:* Exploits the December/January discontinuity to estimate the effect of attend preschool earlier and thereby attending preschool for a longer time. *Estimation:* Use a Fuzzy RD strategy to estimate a LATE and also report the reduced form ITT effect.
Smith (2015)	*Country/Region:* USA, Oklahoma *Period*: 1998-1999 *Sample*: 4-year-olds *N* = 1 *n* = 365	*Preschool program:* Oklahoma introduced universal pre-kindergarten in the 1998-1999 school year. To attend kindergarten, the child had to be five by 1 September. This created a birthday cut-off at the year of the implementation, where children born on or before 1 September are assigned to kindergarten, while children born after 1 September were assigned to pre-	*Staff-to-child ratio*: Maximum 1:10 *Staff education:* Pre-kindergarten teachers are required to be certified in early childhood education.	*Identification:* The author uses the birthday cut-off at the year of the implementation of pre-kindergarten. *Estimation:* Use a combined DID and RD strategy to estimate an ITT effect.

	kindergarten. Around 60% of students offered pre-kindergarten attended. *Control condition:* Formal/private/parental care. The prior conditions were a mix of Head Start, private preschool and no preschool (approximately 20%, 25% and 50%)	

Note: Included studies in alphabetical order, except studies that study the same preschool programs, which are grouped together. When information about, for instance, staff education was not available in an included study, we used, if possible, information from other sources. These are referenced in the table. All other information is taken from the included studies. Acronyms: difference-in-differences (DID), intention-to-treat (ITT), treatment-on-the-treated (TOT), local average treatment effect (LATE), instrumental variable (IV), regression discontinuity (RD). *N* denotes the number of areas included in the estimations, *n* the number of child observations. Both numbers refer to the sample sizes used in the estimations of the mean effects.

A2. Examples of included and excluded studies

To illustrate how we apply the inclusion criteria, we give examples of excluded studies for each criterion below. We also describe a few included studies that meet parts of a criterion. Note that studies could have been excluded by several criteria, but we only mention one below.

Primary Empirical Research. Bradley and Vandell (2007) was excluded because it did not contain primary empirical research, but a review of child care studies on the impact of age at entry and amount, quality, and type of care on children's adaptive functioning.

Preschool Programs. Cascio (2009) study the long-run effect of introducing kindergarten programs as a part of (public) primary school. As these programs are an integrated part of primary school, they do not count as preschool according to our definition, and we excluded the study from the analysis. Haimovich Paz (2015) examine the early kindergarten movement in the US, which operated at a time when kindergarten was not a regular part of primary school and the content of the program was more like contemporary preschool programs than primary school. We included this study for that reason. The participants in the program studied by Herbst (2017) were both preschool and school children. However, the author describes the goal of the paper as to analyze the effect of a child care program. Further, school children were only offered before- and after-school programs and most likely spend most of their time in school. The largest part of the children served are likely to have been preschool children and we included the study.

Universal Programs. Dodge, Bai, Ladd, and Muschkin (2017) studied the long-term effects of North Carolina's Smart Start and More at Four early childhood programs. These programs primarily target disadvantaged children and high-risk children, and the study was therefore excluded.

Long-term Child Outcomes. Baker, Gruber, and Milligan (2008) studied the same introduction of highly subsidized preschool in Quebec as Baker et al. (2015) and Lebihan et al. (2017) but reported outcomes for younger children (primarily 0-4 years) and we excludedthe study was excluded.

Types of Comparisons. Similar to the above-mentioned studies from Canada, Black et al. (2014) used a subsidy scheme to study long-term child outcomes. However, as there were no effects on preschool utilization from a sharp discontinuity in the subsidy scheme

the study did not examine any effects of different types of care and was therefore excluded.

Country, Period, Publication Status, and Language. We did not restrict inclusion by country, time period, or publication status of the study, but included only studies written in a language that at least two members of the research team understand (Danish, English, German, Norwegian, and Swedish). Devaux-Spatarakis (2014) is only in available in French and was therefore excluded.

Methods. Apps, Mendolia, and Walker (2013) used an elaborate matching procedure to control for a very rich set of child and family characteristics and to estimate the impact of preschool on adolescent outcomes. However, they do not use any natural or randomized experiment in the identification and estimation of the effects, and we excluded the study for this reason.

A3. Additional results from the search and screening process

The search of the electronic databases yielded 1,516 unique records (1,861 before duplicates were removed). Table A 2 shows the distribution of records in databases. We identified an additional 88 records from other sources and screened a total of 147 studies in full text. Of these, 26 studies were included: 23 comparing children attending or being more exposed to universal preschool programs with parental, family, and other informal modes and 3 studies comparing alternative preschool types. The full search and screening process is illustrated in Figure A 1 below (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009).

Table A 2. The distribution of records per database

Database	Hits
Academic Search Premier	434
ECONLIT	238
ERIC	381
PsycINFO	694
SocIndex	100
Teacher Reference Center	14
Total	**1,861**

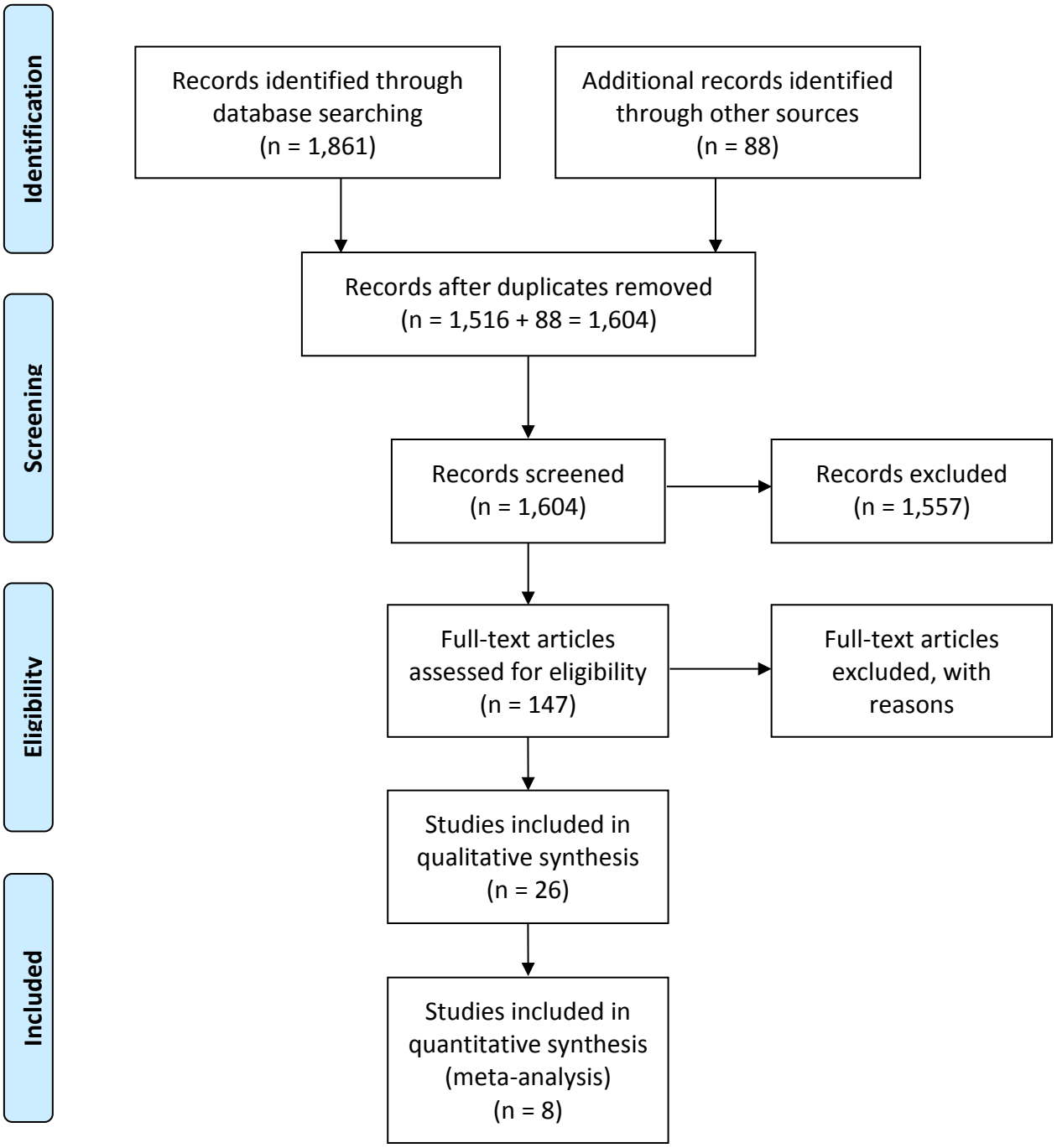

Figure A 1. F Flowchart of the search and screening process

A4. Risk of bias and quality of inference

Risk of bias

The most common research design, used in some form by 17 studies (including one study comparing alternative preschool types), exploited expansions of universal preschool programs that created variation over time (between cohorts of children) and groups (often defined by an area, such as a municipality, state, or city) in how much children were exposed to the programs. The control groups in these studies were often in informal or parental care, but in most cases at least some children in the control group also attended a formal preschool program.

The studies typically lack information on which children attended preschool and used a DID design to estimate an ITT effect of being more exposed to the preschool program. An ITT estimate has the advantage of capturing the full effects of the program, including any peer effects on children in treated areas that did not attend preschool (e.g., Cascio & Schanzenbach, 2013; Havnes & Mogstad, 2011). Because the control group was in many cases not a no-treatment control, only not as exposed, the absolute magnitudes of the estimates are smaller than a contrast between a treatment and no-treatment control group would have been. That is, beneficial effects would be more beneficial and harmful effects more harmful in the latter type of contrast.

The main assumption needed for DID designs to estimate the causal effects is that the trends of the outcome variable would have been parallel, had the treatment group not been more exposed to the preschool program (e.g., Abadie, 2005). The most serious risk of bias in the included studies is that several studies included few areas (seven studies have less than 20 areas, see Table A 1 in the supplementary material). In the most extreme case, only one area was treated. In the case of one treated area, the treatment effect will be confounded by any idiosyncratic trend or shock affecting the outcome variable differently in the treated area compared to the control areas, even if the shock is completely random. This risk of bias decreases, the more treated areas there are, as positive and negative shocks will be more likely to cancel each other out. However, the direction of such bias is difficult to sign and there were both beneficial and harmful effects among the studies with few treated areas (including few areas in the estimation also makes inference more problematic, which we return to below).

Five studies use some form of IV design to estimate the effects of attending preschool (including two studies of alternative preschool types). Just including a variable measuring preschool attendance would likely yield biased estimates as families and children differ in terms of characteristics that influence both the attendance decision and child outcomes. The IV designs attempt to solve this problem by using a two-stage least squares estimation procedure. In the first step, a set of variables predict attendance, at least one of which (the instrument) is assumed to 1) exert a substantial influence on attendance, and 2) only affect child outcomes through its influence on attendance. The included studies use either thresholds in the admission system that determined whether a child was offered a preschool slot or variants of differences in the preschool supply created by for example historical differences, or similar preschool expansions to those used in the DID designs.

The IV studies all have access to data on preschool attendance and estimates variants of a local average treatment effect (LATE). A LATE is the effect for the so-called "compliers"; that is, the children who would not have attended preschool, if they had not been influenced by the instrument (Imbens & Angrist, 1994). One problem with the IV designs is that this group is not readily observable and may not be representative of the larger population of interest. LATE estimates are therefore not easily comparable across studies, as the compliers change from context to context.

The instruments used in the IV designs seem to be strong enough according to the information contained in the studies (i.e., they met condition 1) above). However, it is hard to rule out correlation with child outcomes through other channels than preschool attendance for all instruments used. Historical and geographical differences in the supply of preschools may be correlated with other unobserved determinants of child outcomes (e.g., the value placed by families on having an education or school quality), and admission rules may compare families with different characteristics when samples include children who are not directly at the cut-off created by the rule. Signing this bias across the IV designs is difficult, however, and there are IV designs showing both beneficial and harmful effects.

Two studies employ family or sibling fixed effects, both in the context of expansions of access to universal preschool. The research design uses variation in preschool attendance among siblings to estimate a treatment-on-the-treated (TOT) effect. The sibling fixed effects control for all influences that affect the siblings in the same way, so

if the attendance differences between siblings was only driven by access preschools, for instance, this design may recover the causal effect. A problem is that expanding preschools often means that access increases over time and therefore tends to affect younger rather than older siblings. The effects may therefore be confounded by birth-order effects, which tend to favor older siblings (e.g., Black, Grönqvist, & Öckert, 2017). Both studies control for birth-order effects to mitigate these problems. More generally, parental investments in education and care may be correlated with the decision to send one child and not the other(s) to preschool, and there may also be spillover effects between children. The sign of the bias, if any, is therefore again uncertain. Both studies employing this design show beneficial effects.

Quality of inference

For a number of reasons, the standard errors and *p*-values reported in the included studies are more likely downward than upward biased. Most studies report multiple outcomes but only two adjust for multiple hypothesis testing (Heckman et al., 2017; Lebihan et al., 2017). Treatment is often assigned on the area level, which means that the standard errors need to be adjusted for the clustering of children in areas. However, standard methods for cluster-robust variance estimation often underestimate the standard errors when there are few clusters or the number of children per cluster differs a lot among clusters (e.g., Cameron & Miller, 2015; Mackinnon & Webb, 2017). Few included studies use methods that have been found to work better in these cases (like the wild-cluster bootstrap of Cameron, Gelbach, & Miller, 2008). Furthermore, Mackinnon and Webb (2017) found that even these methods may yield poor results, when the number of treated units is very small. Lastly, Young (2017) find that IV designs tend to produce too small standard errors and *p*-values when standard inference methods are used.

A5. Included estimates

This section provides a motivation of our choice of included estimates in the cases where there were overlapping samples between two studies, or where the choice was not obvious from the principles laid out in the section Analysis in the main text.

Health, well-being, and behavior

Baker et al. (2015) and Lebihan et al. (2017) examine the effects of a preschool reform in Quebec, Canada, use similar estimation methods, and report outcomes from partially

overlapping samples. We include Lebihan et al.'s estimates in the analysis of problem behavior, as they provided separate estimates for children aged 8-9, and in the analysis of health, healthy behaviors, and well-being, as they had access to one more survey wave. Except for life satisfaction/quality of life, where Lebihan et al.'s estimates indicate insignificant beneficial effects and Baker et al. significant harmful effects, the signs of the estimates are the same.

Baker et al. (2015) include estimates on both the probability of being accused and convicted of a crime. The accused in Baker et al. are those charged, plus those dealt with through the use of extrajudicial measures. The latter seemed closer to the measures used by Smith (2015), and we included them. The direction of the results in Baker et al. is similar for both measures.

Test scores and school grades

Fitzpatrick (2008) reports results from the same preschool program and 4th grade tests as Cascio and Shanzenbach (2013). We included the results from Cascio and Shanzenbach (2013), as both Georgia and Oklahoma were in the treatment group in their study, while Fitzpatrick (2008) only included Georgia.

Baker et al. (2015) report two estimates from the PISA tests, one where the 2009 cohort is considered treated and one where this cohort is in the control group, because not all students in this cohort were exposed to treatment. As most estimates in Table 3 are based on contrasts between children who live in areas that were more or less exposed to universal preschool programs, we report the former estimates. Using the latter estimates yields effect sizes indicating less beneficial or more harmful effects.

Primary and secondary school progression

Borraz and Cid (2013) study the same expansion of universal preschool in Uruguay as Berlinski et al. (2008), but use data from only one survey wave. We therefore use the latter study for all estimates of overlapping outcomes.

Years of schooling and highest grade completed

Bingley and Westergård-Nielsen (2012) and Bingley et al. (2018) examine the effect of universal preschool programs in Denmark on years of schooling. The two studies exploit a similar type of expansion/reform, but use non-overlapping samples (cohorts), which is why both studies are included in the analysis. Havnes and Mogstad (2011) and Havnes

and Mogstad (2015) studied the same reform and use an overlapping sample. We used estimates from the latter regarding years of schooling, as they had access to a longer sample in terms of how long they followed the children. Havnes and Mogstad (2015) do not include estimates of the probability of attending college. Consequently, those estimates are taken from their 2011 article.

Employment and earnings

Havnes and Mogstad (2011) and Havnes and Mogstad (2015) study the same reform and use an overlapping sample. However, we include both because Havnes and Mogstad (2011) estimate the effects on the probability of being a top, high, average, and low earner, and Havnes and Mogstad (2015) present quantile treatment effects. As mentioned in the previous section, both Bingley and Westergård-Nielsen (2012) and Bingley et al. (2018) use data from Denmark, but their samples do not overlap. Bingley et al. (2018) provide heterogeneity estimates both across maternal education and earnings quartiles. We report the former, as they are closer to the definition of SES used in most other articles.

A6. Example meta-analysis

This section provides an example of a meta-analysis using the eight studies that use a standardized test scores in math, reading or language arts, science, STEM, cognitive skills or IQ as the outcome (see Table 3 in the main text). We choose this outcome as it is the outcome type used by most studies. The main purpose of the meta-analysis is to investigate whether the observed difference between studies could be due to sampling errors or systematic differences between studies. To make effect sizes as comparable as possible, we transform effect estimates in the following way: First, we divide ITT estimates by the difference in take-up rates between treatment and control groups. Second, in studies estimating the effect of an extra preschool slot, we also divide by the take-up differences. That is, the first and second step transforms ITT estimates to TOT estimates. We do not transform LATE estimates as we lack the necessary information but note that they are typically not the same as a TOT. Third, we divide all effect estimates by the standard deviation of the outcome variable to get an effect size. Fourth, we divide the effect size by the average number of treatment years, which yields an effect size per treatment year. Note that this step assumes that effects can be linearly extra- or interpolated. Fifth, we average the effect sizes and their standard errors by study to avoid

double-counting children that take more than one test. Denote the resulting effect size *d* (after Cohen's *d*). Standard errors are calculated as (Lipsey & Wilson, 2001):

$$SE = [((n_T + n_C)/(n_T\, n_C)) + (d^2)/2(n_T + n_C)]\text{^}0.5$$

Where n_T and n_C is the number of children in the treatment and control group, respectively. These numbers are not included in all papers (some are also missing the total sample size for certain analyses). When the numbers could not be found, we approximate the share in each group by e.g., the share of children in treated areas, and by information from other sources (e.g., *www.oecd.org/pisa* and *www.cme.ca*). Exact calculations for each study is available on request.

It is not obvious how to define treatment and control groups in DID designs where treatment is assigned by area and cohort. We count untreated cohorts in treated areas in the control group, in addition to the cohorts in the control areas. In DID designs, or more generally, when treatment is assigned by area, effect sizes and standard errors should be adjusted using, e.g., information about the intra-cluster correlation (ICC) and cluster sizes (Hedges, 2007). No study included enough information for such adjustments to be possible. The primary consequence of not making this adjustment is that standard errors are underestimated. However, because the ICC is likely to be relatively low when clusters are large, ignoring ICC may be an acceptable approximation.

We use a random effects model for the meta-analysis (Borenstein et al., 2009). Each study is weighted by the inverse variance of the effect size. Studies with more participants are therefore given more weight, all else equal. To assess heterogeneity, we use the chi-squared (or *Q*), the I^2, and τ^2 statistics (Higgins, Thompson, Deeks, & Altman 2003). The statistics provide different perspectives on whether the dispersion of effect sizes is likely to be due to sampling error or systematic differences between studies (Lipsey & Wilson, 2001).

We find a small and insignificant average effect size (d = 0.043, 95% confidence interval = [-0.160, 0.239]),[13] but more importantly, very high levels of heterogeneity. The *Q*-statistic test rejects the null hypothesis of only sampling error with $p < 0.001$, the I^2 is 99.9, which is close to its theoretical maximum value (100), and the τ^2 is very large

[13] Adjusting for cluster-assigned treatments is likely to, in particular, increase standard errors. It is therefore unlikely that such an adjustment would change the qualitative conclusions.

compared to the overall effect size, 0.061. The average effect size is therefore not particularly informative and the effect size differences are highly unlikely due to sampling error alone.

We had to make several strong assumptions and approximations to arrive at comparable effect sizes. There are, as indicated in the main text, many conceptual caveats to this analysis. Nevertheless, it is reassuring that the results are in line with the qualitative analysis: the effects of universal preschool on test scores seem mixed.

A7. Search strings

We searched the following electronic databases for relevant studies: Academic Search Premier, EconLit, ERIC, PsycINFO, SocIndex, and Teacher Reference Center. All searches were performed in EBSCO-host and limited to 1980-2018 in November 2017. The search strings for the six databases follow below.

Academic search premier

Search	Search Terms	
S13	S4 AND S8 AND S12	434
S12	S9 OR S10 OR S11	703,897
S11	SU ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	60,875

S10	AB ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	656,011
S9	TI ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	41,423
S8	S5 OR S6 OR S7	1,821,837
S7	SU (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	108,124
S6	AB (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	1,707,851

S5	TI (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	221,871
S4	S1 OR S2 OR S3	77,686
S3	SU (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	40,738
S2	AB (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	45,548
S1	TI (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	28,484

ECONLIT

Search	Search Terms

S13	S4 AND S8 AND S12	238
S12	S9 OR S10 OR S11	26,243
S11	SU ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	334
S10	AB ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	25,651
S9	TI ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR	1,117

	“IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	
S8	S5 OR S6 OR S7	364,749
S7	SU (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	317,119
S6	AB (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	111,888
S5	TI (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	17,833
S4	S1 OR S2 OR S3	21,408
S3	SU (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	20,451
S2	AB (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	2,163

S1	TI (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	1,129

ERIC

Search	Search Terms	
S13	S4 AND S8 AND S12	381
S12	S9 OR S10 OR S11	29,271
S11	SU ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	729
S10	AB ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR	34,012

	“exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	
S9	TI ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	2,018
S8	S5 OR S6 OR S7	164,085
S7	SU (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	17,024
S6	AB (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	199,972
S5	TI (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	21,740
S4	S1 OR S2 OR S3	65,921
S3	SU (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	71,441

S2	AB (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	43,281
S1	TI (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	19,393

PsycINFO

Search	Search Terms	
S13	S4 AND S8 AND S12	694
S12	S9 OR S10 OR S11	144,671
S11	SU ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	12,166

S10	AB ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	138,237
S9	TI ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	12,278
S8	S5 OR S6 OR S7	455,654
S7	SU (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	60,192
S6	AB (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	494,562
S5	TI (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	42,555
S4	S1 OR S2 OR S3	102,863

S3	SU (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	91,700
S2	AB (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	47,848
S1	TI (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	3,856

SocINDEX

Search	Search Terms	
S13	S4 AND S8 AND S12	100
S12	S9 OR S10 OR S11	25,035
S11	SU ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR	1,776

	"experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	
S10	AB ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instru-ment*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	26,399
S9	TI ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	1,890
S8	S5 OR S6 OR S7	164,819

S7	SU (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	6,353
S6	AB (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	199,604
S5	TI (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	19,668
S4	S1 OR S2 OR S3	20,801
S3	SU (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	15,230
S2	AB (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	14,822
S1	TI (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	5,599

Teacher reference center

Search	Search Terms	
S13	S4 AND S8 AND S12	14
S12	S9 OR S10 OR S11	5,325
S11	SU ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	364
S10	AB ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR randomized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differences”)	4,920
S9	TI ("treatment-control" OR "treatment-comparison" OR "random* control* trial*" OR random-ized field" OR "experiment*" OR "quasi-experiment*" OR "quasi-random* control* trial*" OR "Sibling sample design*" OR "sibling fixed	400

	effect*" OR "family fixed effect*" OR "instrumental variable*" OR "random-assignment design" OR "program effect*" OR "intervention* effect*" OR “instrument*” OR “IV” OR “exogenous variation” OR “evaluate” OR “discontinuity” OR “difference-in-difference*” OR ”with-in household difference” OR ”within household differ-ences”)	
S8	S5 OR S6 OR S7	42,118
S7	SU (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	5,959
S6	AB (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	38,212
S5	TI (universal OR general OR comprehensive OR expan* OR nationwide OR large-scale OR community-wide OR statewide)	
S4	S1 OR S2 OR S3	5,505
S3	SU (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	19,419
S2	AB (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	6,588

S1	TI (preschool* OR “childhood program” OR “child* develop* program*” OR pre-kindergarten OR childcare OR daycare OR "early childhood care" OR "pre-primary education" OR "childhood program*" OR "early education" OR prekindergarten OR "early childhood education" OR Pre-K OR "childhood care" OR "center based day care" OR "family day care" OR "childhood initiative*")	12,046

References to the supplementary material

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72(1), 1-19.

Apps, P., Mendolia, S., & Walker, I. (2013). The impact of pre-school on adolescents' outcomes: Evidence from a recent English cohort. *Economics of Education Review*, *37*, 183–199.

Black, S. E., Grönqvist, E., & Öckert, B. (2017). Born to lead? The effect of birth order on non-cognitive abilities. *Review of Economics and Statistics*, forthcoming.

Bradley, R. H., & Vandell, D. L. (2007). Child care and the well-being of children. *Archives of Pediatrics & Adolescent Medicine*, *161*(7), 669–676.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, *90*(3), 414–427.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

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. 1495).

Devaux-Spatarakis, A. (2014). L'experimentation "telle qu'elle se fait": Lecons de trois experimentations par assignation aléatoire. *Formation Emploi: Revue Francaise de Sciences Sociales*, (126), 17–38.

Dodge, K. A., Bai, Y., Ladd, H. F., & Muschkin, C. G. (2017). Impact of North Carolina's early childhood programs and policies on educational outcomes in elementary school. *Child Development*, *88*(3), 996–1014.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. Journal of Educational and Behavioral Statistics, 32, 341–370.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

MacKinnon, J. G., & Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2), 233–254.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine 6*(7), e1000097.

UNESCO (2005). *Policy review report: Early childhood care and education in Kenya.* Early Childhood and Family Policy Series.

UNESCO (2006a). *Argentina – Early childhood care and education (ECCE) programmes.* UNESCO International Bureau of Education (IBE).

UNESCO (2006b). *Guatemala – Early childhood care and education (ECCE) programmes.* UNESCO International Bureau of Education (IBE).

UNESCO (2006c). *Uruguay – Early childhood care and education (ECCE) programmes.* UNESCO International Bureau of Education (IBE).

World Bank (2012). *Tanzania – Early childhood development*.

Young, A. (2017). *Consistency without inference: Instrumental variables in practical application*. Unpublished manuscript, London: London School of Economics and Political Science.