The Effects of a Structured Curriculum on Preschool Effectiveness: A Field Experiment

Mari Rege
Ingunn Størksen
Ingeborg F. Solli
Ariel Kalil
Megan M. McClelland
Dieuwer ten Braak
Ragnhild Lenes
Svanaug Lunde
Svanhild Breive
Martin Carlsen
Ingvald Erfjord
Per Sigurd Hundeland

Abstract: This study tests an intervention that introduces a structured curriculum for five-year-olds into the universal preschool context of Norway, where the business as usual is an unstructured curriculum. We conduct a field experiment with 691 five-year-olds in 71 preschools and measure treatment impacts on children's development in mathematics, language, and executive functioning. The nine-month intervention has effects on child development at post-intervention and the effects persist one year following the end of the treatment. The effects are mainly driven by the preschools identified as low-quality at baseline, indicating that a structured curriculum can reduce inequality in early childhood learning environments.

Keywords: universal preschool, preschool quality, intervention, early childhood investments, randomized controlled trial, field experiment, childcare, child development, curriculum.

JEL: I20, H42

Acknowledgments: Mari Rege is a professor of economics at the University of Stavanger. Ingunn Størksen is a professor of pedagogical psychology at the University of Stavanger. Ingeborg F. Solli is an associate professor of economics at the University of Stavanger. Ariel Kalil is a professor of public policy at the University of Chicago. Megan McClelland is the Katherine E. Smith Healthy Children and Families Professor at Oregon State University. Dieuwer ten Braak is an associate professor of psychology at the University of Stavanger. Ragnhild Lenes is an associate professor of education science at the University of Stavanger. Svanaug Lunde is a university lecturer of education science at the University of Stavanger. Svanhild Breive is an associate professor of mathematics education at the University of Agder. Martin Carlsen is a professor of mathematics education at University of Agder. Ingvald Erfjord is an associate professor of mathematics education at University of Agder. Per Sigurd Hundeland is an associate professor of mathematics education at University of Agder. Rege and Størksen share first authorship. Corresponding author can be reached at mari.rege@uis.no. We are grateful to project coordinator Åse Lea who has juggled all the different logistics in this field experiment, to our trained testers who participated in the three waves of assessments, and the preschool teachers and children who participated in the project. We are grateful for comments from participants at seminars and the CESifo Area Conference on Economics of Education, and from Roberta Golinkoff, James Heckman and Eric Bettinger. Thank you to Roberta M. Golinkoff, Greg Duncan, Clancy Blair, Douglas Clements, Adele Diamond, Christina Weiland, Pamela Morris and Terri Sabol who all provided us with advice on the curriculum design. We acknowledge funding from The Research Council of Norway (237973, 270703 & 318626), The Sørlandet Knowledge Foundations and The Agder County administrations. This study is pre-registered in the registry of the American Economic Association (AEARCTR-0002241). The data used in the analyses are subject to strict regulation by the Norwegian Data Protection Authority. The authors are willing to advise other scholars regarding procedures to apply for access. Supplementary materials are freely available online.

doi:10.3368/jhr.0220-10749R3

I. Introduction

Many European countries, including the U.K., France, Germany, and all Nordic countries, invest heavily in universal preschool programs. Moreover, universal programs are available in several U.S. states and in Quebec, Canada. The investments in preschool are largely motivated by research demonstrating that preschool programs can boost child development and have long-term impacts on school achievement and adult labor market participation (e.g. Heckman et al. 2010, Melhuish 2011). However, despite an enormous policy interest in universal preschool, we have limited understanding of the conditions under which preschool is effective. Moreover, large variation in center quality has given rise to widespread scientific and policy concern (Bennett and Tayler 2006).

Several studies in economics investigate effects of preschool participation on child development, but the evidence is far from unified (Cornelissen et al. 2018). Some studies demonstrate that universal preschool participation might, especially for disadvantaged children, have positive and lasting effects on child development (Havnes and Mogstad 2011, Cornelissen et al. 2018, Berlinski, Galiani, and Manacorda 2008, Berlinski, Galiani, and Gertler 2009, Felfe, Nollenberger, and Rodríguez-Planas 2015). However, there are also studies showing that preschool participation can be detrimental to child development (Baker, Gruber, and Milligan 2008) or have no effects at all (Gupta and Simonsen 2010). It is hard to understand why the different studies yield different results, because the studies are from different countries and the preschool programs (e.g. the curriculum, teacher education, and child-staff ratios), the counterfactual to preschool (e.g. parent care, grandparent care, or other center-based care), the children's age and the populations differ across the studies. Moreover, in order to identify causal relationships, some of these studies utilize policy reforms that led to rapid preschool expansions in certain geographic locations in difference-in-differences type approaches (Baker, Gruber, and Milligan 2008). The quality of care during and immediately

after a rapid expansion is not necessarily representative of the quality of care in centers with established routines and experienced caregivers.

There is a need for research that can help us better understand the conditions under which preschool is effective. One dimension along which preschool programs across the world differ substantially, is the intentionality and the systematicity with which they promote children's cognitive and socioemotional skills, such as numeracy, literacy, and executive functioning. The social pedagogical tradition in Scandinavia and Germany represents one end of this spectrum with its limited curricular focus and its emphasis on free play in mixed-age groups, humanistic values, and the affective qualities of the teacher-child relationship. There is no detailed and structured curriculum, instead teachers facilitate learning through spontaneous engagement, interaction and play, and through crafts projects and story-time (Bennett and Tayler 2006, Engel et al. 2015). A major concern with this approach is that it gives the preschool centers a large degree of freedom with respect to pedagogical content, which can give rise to large differences in learning across centers (Bennett and Tayler 2006, Engel et al. 2015).

The "school readiness tradition" in contrast, takes an instrumental approach to strengthening children's skills in specific areas of development that have been linked to success at the start of formal schooling and to longer-run academic achievement and social adjustment (Bennett and Tayler 2006). In this tradition preschool teachers work intentionally and systematically to stimulate children's skill development using play-based approaches. The school readiness tradition is prevalent in English-speaking countries (Bennett and Tayler 2006). However, unstructured curricula that are modeled after northern European philosophies are increasingly popular, and there is an ongoing debate about their effectiveness compared to more structured and academically-focused curricula (Montie, Xiang, and Schweinhart, 2006; Weiland and Yoshikawa, 2013; Lonigan et al., 2015).

The present study tests an intervention that introduced a structured curriculum for five-year-olds into the universal preschool context of Norway. As the current curriculum is very non-specific and unstructured, and there are large differences in learning across centers (Rege et al. 2018), Norway provides an excellent platform for investigating the effects of a structured curriculum on children's skills. The intervention consisted of a nine-month long comprehensive curriculum with age-appropriate intentional skill-building activities in mathematics, language, social skills, and executive functioning and accompanying teacher training. Our field experiment had 691 five-year-olds in 71 preschool centers. We randomly split the centers between a control and a treatment group using block randomization. Treated centers implemented the comprehensive structured curriculum for the five-year-olds in addition to receiving teacher training, while the control group continued with business-as-usual. We assessed the children's skills in language, mathematics, and executive functioning at baseline, post-intervention, and in a one-year follow-up.[i] At all assessments, the testers were trained, certified, and blind to treatment status.

The structured curriculum intervention had a positive effect on a summary score of children's skills in math, language, and executive functioning at post-intervention. A treatment impact, with a magnitude of about 13 percent of a standard deviation ($p < 0.10$), persisted one year following the end of the treatment. The treatment effect was particularly pronounced in mathematics. The effect on mathematics skills in the one-year follow-up was quite large, 23 percent of a standard deviation ($p < 0.01$). By way of comparison, it takes on average about five months of learning and development at this age to improve children's mathematics skills by this magnitude. The treatment effect on mathematics also represents about two-thirds of the difference in average mathematics skills between children of mothers with and without a college degree in our sample.

Since large differences in learning across centers is a major concern in preschool systems with unstructured curricula (Bennett and Tayler 2006, Engel et al. 2015), we investigated differential treatment effects across centers identified as low- and high-quality centers at baseline. We utilized center fixed effects at baseline as a proxy for quality, i.e. the center mean difference between observed and predicted test scores, given child observables, and defined low- and high-quality by a median split. Since most of the children have been in the same preschool since age one, limited emphasis on pedagogical content in certain centers is presumed to contribute to a lower quality score at baseline in these centers. This is supported by survey data demonstrating that, prior to the intervention, the five-year-olds in low-quality centers have fewer hours with structured pedagogical activities compared to the high-quality centers. As such, our structured curriculum intervention should be particularly effective for the low-quality centers, because for these centers there is a larger contrast in practice between treatment and control. Consistent with this conjecture, our analyses demonstrate that the treatment effect was mainly driven by centers defined as low-quality at baseline. In these centers, the treatment effect on the sum score was 22 percent of a standard deviation ($p < 0.01$) in the one-year follow-up, which was 19 percent of a standard deviation ($p < 0.10$) larger than in the high-quality centers. This suggests that a structured curriculum can reduce inequality in early childhood learning environments by raising center quality at the bottom of the distribution. However, our results must be interpreted with caution as our treatment design, constrained by field practicalities, was bundled. As a consequence, we cannot separate the effects of the structured curriculum from the possible effects of other mechanisms. These mechanisms include more teacher resources, more time with same-age peers, and a shift in children's focus towards mathematics.

We also investigated differential treatment effects across child skills at baseline and across parental education. We expected the treatment to be particularly beneficial for children with fewer learning opportunities at home versus a more limited impact for children with highly-educated parents who provide on average more stimulating home learning environments (Kalil, Ryan and Corey 2012; Kalil et al. 2016). However, our analyses suggest that children in our intervention benefited equally from the treatment regardless of their initial skill level or their parents' education.

Our field experiment makes several important contributions. As noted above, conflictive evidence in the extensive literature investigating effects of preschool participation on child development (Cornelissen et al. 2018) call for research that helps us better understand the conditions under which universal preschool is effective. In our field experiment all children participated in preschool. This allows us to provide important new insight on the importance of a comprehensive structured curriculum for preschool effectiveness.

Our work relates to an emerging economic literature investigating the impact of observable quality indicators, often referred to as structural quality (Blau and Currie 2006), such as teacher education, child-staff ratios, teacher and management experience, and class size, on child development (e.g. Bauchmüller, Gørtz, and Rasmussen 2014, Blau 1999, Currie and Neidell 2007, Drange and Rønning 2020). In general, the evidence from these studies seems to mimic evidence from similar studies in schools, which suggests limited potential for facilitating child development by merely investing in structural quality, and points instead to the need for better understanding the role of preschool process quality for children's skill development (Jackson, Rockoff, and Staiger 2014, Blau and Currie 2006). Developmental psychologists have also made this point (e.g. Sabol et al. 2013). Process quality represents the direct experiences for children and includes factors such as the sensitivity and responsiveness

of caregivers, the pedagogical approaches, and curriculum and materials available for learning (OECD 2015).

One paper closely related to ours is Araujo et al. (2016) who studies kindergarten classrooms in Ecuador. Based on videos from each classroom, the study measures three aspects of teacher practice: instructional support, emotional support and, classroom organization. These three aspects of teacher practice are often invoked in discussions of classroom quality. By using random classroom assignment for identification, Araujo et al. (2016) provide convincing evidence that these teacher practices are important predictors of child development. Yet, as noted by the authors, the measures may be correlated with other unmeasured teacher attributes, which could themselves affect child learning. Moreover, the study does not test tools to enhance these dimensions of teacher practice. Our field experiment complements this work by investigating whether a structured curriculum and accompanying teacher training are important for preschool effectiveness.

The present study also builds on work in psychology and education investigating how structured curricula affect child development (e.g. Dillon et al. 2017, Clements and Sarama 2011, Weiland and Yoshikawa 2013, Schmitt et al. 2015, Diamond et al. 2007). This literature suggests that detailed age-appropriate curricular foci that intentionally and systematically target school readiness skills through play-based activities, along with teacher training, are key determinants of child development in preschool (Burchinal 2018). The closest study to ours is Weiland and Yoshikawa (2013), which offered a comprehensive preschool curriculum with a teacher coaching system to students in the Boston Public School system. This study, however, was conducted in a large urban school district with a predominantly low-income population (e.g. 69% of their sampled children were eligible for free or reduced-price lunch). Weiland and Yoshikawa also employ a quasi-experimental regression-discontinuity design as compared to our true experimental one. Other experimental studies in this vein are also

conducted with children from low-income families or in developing country contexts. Moreover, the curricula in related studies often target one specific skill domain (e.g., Clements and Sarama 2011). We know of no field experiment investigating effects of a comprehensive structured preschool curriculum in the context of a universal preschool program, despite the strong policy interest in such programs.

II. The Norwegian Context and Preschool System

We conducted our field experiment in the universal preschool context of Norway. Norway has a strong welfare system with generous social security and family policies facilitating both child well-being and a strong labor market attachment for parents of young children. After childbirth or adoption, parents have the right to twelve months of parental leave with wage compensation and job security. Thereafter, all children ages one to five have the right to publicly regulated and subsidized preschool. The preschool utilization is very high with an uptake of 98 percent among five-year-olds, which is the age of the targeted children in our experiment. Compulsory primary school starts at age six and more than 95 percent attend public schools.

Norwegian preschool centers typically organize children in mixed-age groups, with one- and two-year-olds and three- to five-year-olds together. The adult-child ratio is regulated so that the child groups with the youngest children have at least one preschool teacher per 7-9 children, whereas the groups with the older children have at least one teacher per 14-18 children. A preschool teacher has a bachelor's degree in early childhood education and care. In addition to the teacher, each child group has two assistants. Many of the assistants have a relevant certificate of apprenticeship. However, there are no formal qualification requirements for the assistants; it is not even required that they have completed high school.

The Norwegian preschool system was established in the 1970s as a response to the need for high-quality care as mothers entered the labor market. The idea that these centers had an important job in preparing children for school was not prevalent. Despite the educational and developmental purpose invoked in contemporary discussions of preschool, the Norwegian program remains dominated by the social pedagogical tradition seen in the Nordic countries and Germany, as opposed to the school readiness approach seen in many English-speaking countries (Bennett and Tayler 2006). In general, free play and children's natural curiosity are highly valued and encouraged in the social pedagogical tradition. Moreover, there is no detailed and structured curriculum, instead teachers facilitate learning through spontaneous engagement, interaction and play, and through crafts projects and story-time (Bennett and Tayler 2006, Engel et al. 2015).

The Norwegian preschool centers' pedagogical content is regulated by the National Framework Plan for Content and Tasks of Kindergartens (Ministry of Education and Research 2017), which defines seven learning areas. These learning areas are loosely described and are the same for all children ages one to five. There are no specific guidelines on how preschool centers should implement the learning areas and the curriculum provides no learning benchmarks.

As noted above, the five-year-olds are typically in mixed-age groups together with three- and four-year-olds. However, since age five is the last year before starting mandatory schooling, some centers choose to provide some planned and structured pedagogical activities exclusively for the five-year-olds (as we do in our treatment) even if this is not required by the Framework Plan. In a survey of our participating centers conducted the year prior to the field experiment implementation (52 of 71 centers responded), only 21 centers reported that they offered at least one hour per week of planned and structured pedagogical activities exclusively for the five-year-olds. Among these 21 centers the average number of hours per week was

close to three. In contrast, our treatment gives the five-year-olds at least eight hours per week with planned and structured curricular focus exclusively for the five-year-olds.

III. Structured Curriculum Intervention: Structured Preschool Curriculum and Accompanying Teacher Training

Our structured curriculum intervention consisted of a comprehensive curriculum with age-appropriate intentional skill-building activities in mathematics, language and executive functioning, and accompanying teacher training. The teachers committed to spending at least eight hours a week for nine months (almost the full preschool year) engaging the five-year-olds in the curriculum, separate from five-year-olds' larger classrooms that also included three- and four-year-olds. The curriculum has 130 learning activities, which we developed in collaboration with Norwegian preschool teachers (Størksen et al. 2018). The learning activities are inspired by existing U.S. curricula with evidence of positive effects in targeted programs, such as I Can Problem Solve (Shure 1992), Interactive Book Reading (Mol, Bus, and de Jong 2009), Building Blocks (Clements and Sarama 2011), California Preschool Curriculum Framework (California Department of Education 2016), Tools of the Mind (Bodrova and Leong 2007), and Red Light, Purple Light (Schmitt et al. 2015). A playful learning approach permeates all the activities, in that the activities were interactive, engaging, and meaningful (Weisberg, Hirsh-Pasek, and Golinkoff 2013), and the curriculum emphasizes a warm and responsive child-teacher relationship (Pianta 1999). Importantly, the curriculum is not a scripted program intended to dictate teacher practice on a daily basis. Instead, teachers are encouraged to integrate the curriculum in their own approach and to augment it with their own ideas. Additionally, the activities are flexible in terms of challenge and complexity, allowing teachers to match their practice to children's skill levels. The activities are organized in a book with suggested schedules for how to structure the activities by day, month, and year.

In mathematics, the curriculum engages children in activities stimulating numbers and quantitative thinking, in addition to measurement, geometry, and statistics. To stimulate early literacy, the children participate in interactive book reading (Mol, Bus, and de Jong 2009) and language games related to letter and sound recognition. The games stimulating executive functioning, in terms of working memory, inhibitory control, and flexible attention (Best and Miller 2010), challenge children to memorize and follow rules that require inhibitory control and doing the opposite to instructions. Social skills were stimulated in games and activities constructed to enhance cooperation, assertion, responsibility, empathy, and self-control in social settings (Gresham and Elliot 1990).

It is generally not enough to simply hand teachers a curriculum; how it is supported matters. A consensus statement on the U.S. preschool evidence (Yoshikawa et al, 2013) highlights that it is not coaching or curricula alone, but the combination that appears particularly effective. As such, since working with a structured curriculum was completely new for most of the teachers in the Norwegian preschool context, our structured curriculum intervention had to be combined with accompanying teacher training. The teacher training consisted of a credit-based university class prior to the year of curriculum implementation and coaching during the year of implementation. For preschool centers with more than 18 five-year-olds, two teachers participated in the training. During the training, the teachers learned about the theoretical and empirical research foundation for the curriculum. Moreover, as part of the class, the teachers practiced the learning activities in the preschool curriculum with their current five-year-olds and provided us with feedback. We revised the activities in the curriculum based on the feedback. This feedback process gave the teachers a sense of ownership of the curriculum and helped us adapt the curriculum to the Norwegian preschool context, both of which are critical for implementation quality and high treatment compliance (Størksen, Ertesvåg and Rege 2021). The coaching during the year of implementation consisted of two gatherings with all

preschool teachers and their coaches and four scheduled one-to-one phone meetings. Teachers could schedule additional phone meetings with their coaches to address any immediate questions or concerns.

The intervention gave particular weight to mathematics. The curriculum had more scripted activities for mathematics and the teacher training had more lectures on mathematics compared to the other skill domains targeted by the intervention. Moreover, half of the scheduled phone meetings for coaching were devoted to mathematics exclusively. Finally, we advised the trained teachers to implement the mathematics activities and the assistants to implement the language activities under the guidance of the trained teacher. We emphasized mathematics because our pre-intervention assessment among teachers revealed that mathematics had low priority in the centers compared to other developmental areas. Furthermore, during the training, teachers reported that the mathematics activities were more novel and challenging compared to the other activities.

IV. Experimental Design and Empirical Strategy

Figure 1 illustrates the experimental design. We randomly split the 71 participating centers between a control and a treatment group using block randomization (see Procedures for details). During the preschool year 2015/2016 the teachers in treated centers attended the teacher training, and, as a part of the training, provided extensive feedback and helped us revise the curriculum. Thereafter, the trained teachers implemented the curriculum intervention with the five-year-olds in their center during the preschool year 2016/2017. The preschool centers in the control group continued with business-as-usual, but teachers received the teacher training and intervention material in 2017/2018, when the children in the control group had left preschool and started 1st grade in school. We assessed the children's skills in

language, mathematics, and executive functioning in August 2016 (baseline, T1), June 2017 (post-intervention, T2), and March 2018 when the children were in 1st grade in school (follow-up, T3). Additionally, we conducted multiple surveys among the preschool teachers to assess teacher compliance and their perceived relevance, importance, and benefit of the training and intervention. We pre-registered the research design, dependent variables, in addition to important sub-samples for investigation of differential treatment effects, in the registry of the American Economic Association (AEARCTR-0002241).

We estimate effect sizes and statistical significance utilizing the following OLS model:

$$Y_{i,c}^{m} = \alpha + \gamma T_c + \mathbf{\beta X_i} + \varepsilon_{i,c} \ ,$$

where $Y_{i,c}^{m}$ is the test score on outcome measure *m* for child *i* in preschool center *c*. $\mathbf{X_i}$ is a vector of child and parent characteristics, including all baseline test scores (T1), α is the constant term, and $\varepsilon_{i,c}$ is the error term. T_c is an indicator for the treatment status of the child's preschool center *c*, and γ is the estimated treatment effect. We estimate the model separately for T2 and T3 outcomes. Given random assignment to treatment, controlling for child and parent characteristics should only to a limited degree affect the treatment estimate. However, we expect increased precision of the treatment estimate, in particular when controlling for baseline test scores. In all models, we include a vector of fixed effects for randomization block, and we cluster on randomization block to adjust for correlated error terms within blocks.[ii] Furthermore, due to the small number of clusters in our sample, we estimate the wild bootstrap p-value for the treatment estimate in order to investigate the sensitivity to the number of clusters.[iii]

We investigate differential treatment effects across child skills at baseline, parental education, and center quality at baseline. Specifically, for all outcomes we estimate the following model:

$$Y_{i,c}^{m} = \alpha + \gamma T_c + \delta T_c \cdot High_i + \theta High_i + \mathbf{\beta X_i} + \varepsilon_{i,c}$$

where $High_i$ is an indicator for high baseline skills, high parental education, or high-quality center[iv]. Apart from this interaction term, the model specification is identical to our main model.

We define high baseline skills as scoring above the median of the relevant T1 score. We measure parental education as the mean number of education years of the mother and father; the high/low split is at the median of parental education. As an indicator for center quality, we use the center mean difference between observed and predicted test scores. Specifically, we follow (Rege et al. 2018) and estimate the following model using T1 assessment data:

$$Y_{ic} = \alpha_c + \mathbf{\beta X_i} + \varepsilon_{i,c}$$

where Y_{ic} is a collapsed test score across all assessments for child *i* in center *c* at baseline, and X_i is a vector of child and parent characteristics (gender, birth month, parent education, earnings, and immigrant status). α_c is the center fixed effects and constitutes our quality measure. Because the center fixed effects are particularly sensitive to outliers in very small centers, we exclude centers with fewer than five children in this specific analysis. In order to define high- and low-quality centers, we split the sample at median value of the center fixed effects.

Notably, the investigation of differential treatment effects across centers identified as low- and high-quality centers at baseline is not described in our pre-registration. We added this after analyzing T1 assessment data and detecting large differences in learning across centers (Rege et al. 2018). This variation motivated us to investigate if our structured curriculum intervention reduced the inequality in early childhood learning by raising center quality at the bottom of the distribution.

V. Procedures

In Norway, our field experiment is referred to as the Agder project, as it was conducted in the Agder counties of southern Norway. In February 2015, we organized informational meetings for all municipalities and preschool centers in the Agder region. Among the 30 municipalities in the region, 15 signed up for the project. Within these municipalities, preschool directors decided themselves if they wanted their preschool center to participate. Among the 190 preschool centers in these municipalities, 72 signed up for the project. Participating municipalities, preschools, and teachers had to sign written agreements that detailed the expected activities and obligations to the project. Prior to the intervention, one center in the control group withdrew from the project, leaving us with 71 participating centers.

We conducted block randomization of the preschool centers into treatment and control. Blocks were constructed based on location and center size (number of children), which was the only available information about centers at the time of randomization.[v] We have a total of 15 blocks, consisting of four to six centers and 29 to 92 children. Our power calculation in the pre-registration (AEARCTR-0002241) demonstrated that 71 centers gave us 80 percent power to identify effect sizes of 0.25 standard deviation. When controlling for pre-intervention test scores, in addition to controls for characteristics of preschool centers, children, and parents, statistical power is likely to be improved.

We collected parental consent in spring 2015 when the children were three to four years old, prior to randomizing preschool centers into treatment and control. However, due to the extensive timeframe between collection of parental consents and curriculum implementation (more than a year), we allowed for additional (late) parental consents after the randomization of centers. In total, we received parental consent for 701 children, which constitute 90 percent of the children in the 71 preschool centers. Among these are 132 late consents. As expected, the majority (80 percent) of the late consents are from treated centers. The preschool teachers were in charge of collecting the parental consents, and teachers in the treated group were

likely more engaged in the project and worked harder to get the remaining parental consents. In order to maintain a large sample size, we include children with late consent in our main analyses, while adding an indicator as control for late consent. In a robustness check (Table A1, Panel A) we investigate robustness to excluding children with late consent from the analyses.

During the preschool year 2015/2016 the teacher responsible for the five-year-olds in treated centers participated in the teacher training, a credit-based university class. In centers with more than 18 five-year-olds, two teachers participated in the training. To make it possible for the headteachers to participate in the training during work hours, the centers received funding for a substitute teacher for the number of hours the headteacher was out of the classroom in training, which was equivalent to a 50 percent position during four months. Including overhead this constituted NOK 89,000 (USD 11,125).

During the preschool year 2016/2017, the trained head teachers implemented the preschool curriculum with the five-year-olds in their preschool center. As noted above, five-year-olds in Norway are typically in mixed child groups with three- to five-year-olds. During curriculum implementation, the five-year-olds were organized in a separate group together with the trained headteacher and the assistant(s). Classrooms were provided with a substitute teacher for the number of hours the headteacher was out of the classroom either preparing or implementing the curriculum. This amounted to NOK 222,000 (USD 27,750), which is equivalent to a 50 percent position for nine months.

Prior to implementation, all centers received the book with the curriculum, in addition to a box with basic material. The box contained materials for implementation of the playful learning activities, such as books, blocks, dices, and scales, with a value equal to NOK 12.000 (USD 1.500). Many preschool centers already had several of the items in the box, but to assure high compliance, we provided the items for all participating centers.

Each trained preschool teacher had one or two assistants when implementing the curriculum, depending on the size of the child group. In groups with more than six five-year-olds, which were most groups, we recommended that the children were divided into two groups, which alternated between the language and mathematics activities. As noted above, we advised the trained teachers to implement the mathematics activities and the assistants to implement the interactive book reading and language games under the guidance of the trained teacher since teachers considered implementation of the mathematics activities more challenging. The trained teachers had the main responsibility to train the assistants. However, assistants also received a one-day training introducing them to the preschool curriculum, and half of this day was devoted to interactive book reading.

The preschool centers in the control group continued as before during treatment implementation, but they received the credit-based university class, funding for substitute teachers during the class, and intervention material in 2017/2018.

We assessed treatment compliance in a brief weekly questionnaire where teachers reported on fidelity of implementation, including spending at least eight hours a week implementing the learning activities. Fidelity was satisfactory, as demonstrated in Appendix A3. Additionally, we conducted surveys among the preschool teachers from which we in Appendix A3 conclude that the teachers perceived the relevance, importance, and benefit of the training and intervention as high.

Attrition is also an important indicator of compliance. Indeed, among the 72 centers that signed up, only one center, randomized to control condition, withdrew from the field experiment. This low attrition is notable given the two-year length of the intervention. Several features in our procedures likely contributed to the low attrition: First, the detailed written and signed agreements with participating centers and teachers; second, that preschool centers received funding for all the expenses in association with the intervention; third, that preschool

centers in the control group received material and training after the field experiment was completed; and fourth, that we involved the teachers in the curriculum design and thereby gave them a sense of ownership, in addition to assuring a careful adaption to the Norwegian context.

VI. Assessments and Data

We conducted assessments at three points in time: Baseline in August 2016 (T1), just before implementation of the intervention; post-intervention in June 2017 (T2), when the intervention was completed, and follow-up assessment in March 2018 (T3), when the children were in 1st grade in school. We assessed the children in language, mathematics, and executive functioning. The T1, T2, and T3 assessments used the same test battery, which took approximately 40 minutes for each child. All assessments were one-to-one with a trained and certified tester, blind to treatment status. The testers used computer tablet instruments with a validated test battery developed for transition between preschool and school. Scales included the Ani Banani Math Test (ten Braak and Størksen 2021) for assessing mathematics skills; the Norwegian Vocabulary Test (Størksen et al. 2013) and The Phonological Awareness Test (The Norwegian Directorate for Education and Training) for assessing language skills; and the Digit Span Test (Wechsler 1991), the Head-Toes-Knees-Shoulders task (McClelland et al. 2014) and the Hearts and Flowers test (Davidson et al. 2006) for executive functioning. From the six tests, we construct three measures for the skill domains Math, Executive functioning, and Language. For each assessment period, the measures are constructed as the mean value across standardized test scores and then re-standardized. Furthermore, we construct a standardized ordinary sum score of the three skill domains. This allows us to evaluate treatment effects on the general skill level, and address concerns of multiple hypothesis testing.[vi]

In assessments T1 and T2 we invited the 71 preschool centers to local science museums. The children engaged in museum activities and, at a scheduled time, each preschool center brought their children to an assessment station. For each assessment day, we invited centers from both the control and the treatment group and testers were blind to treatment status. In T3 the children had finished preschool and were in 1st grade in school. Testers traveled to the schools to conduct the assessment. We collaborated with the school administration who facilitated by guiding the participating children out of the classroom for the assessment. Multiple preschool centers fed into each school and, as in T1 and T2, testers were blind to treatment status.

In total, 665 children participated in the T1 assessment. Missing test scores at T1 are replaced by predicted values (prediction based on gender, birth month, mother and father education and earnings, immigrant status, and an indicator for preschool center).[vii] In the T2 assessment, 650 children participated, and in T3 when the children were in 1st grade in school, we managed to locate and assess 661 children.[viii] Our "gross sample" consists of children assessed in T2 and/or T3, a total of 691 children. Consequently, the analytical sample in analyses on T2 measures is slightly different from T3 measures, but with a major overlap: 620 children were assessed in both T2 and T3. Importantly, attrition was balanced across treatment status, see appendix Table A3.

Table 1 presents pairwise correlations between test scores in T1, T2, and T3 on our gross sample. All correlation coefficients are significant at 1 percent level. We find that T1 test scores are strongly correlated to T2 and T3 test scores on the same measure, ranging from 0.502 to 0.667. Furthermore, we find that all baseline test scores (T1) correlate with the other measures. In particular, the mathematics test score at T1 appears to be strongly predictive of all T2 and T3 measures.

The assessment data was merged to registry data from Statistics Norway on gender (indicator for female), birth month (continuous variable running from 1 (December) to 12 (January)),

mother's and father's education (number of years schooling), earnings, and immigrant status (indicator for whether one or both parents are immigrants from a non-western country).[ix] Furthermore, we added indicators for late consent, missing baseline test scores, and randomization block.

In Table 2 we provide summary statistics and balance tests for the T2 and T3 samples. Even if there are few significant differences, we see that children in treated centers in average score somewhat lower on baseline test scores, and more of them are non-western immigrants, compared to children in control centers. In particular, we find a substantial unbalance in the baseline Language test score in the T3 sample, resulting in jointly significantly different coefficients for this sample. Notably, since the baseline assessment was conducted after the teacher training, lower baseline test scores in the treated group may be related to teacher absence when undergoing training. This leads us to further investigate possible imbalances in the result section, by carefully examining how our estimates of treatment coefficients are robust to the inclusion and exclusion of child characteristics, baseline test scores, and family background.

VII. Results

Table 3 presents our main results. For each outcome (in columns) we estimate four models: In Model 1 we regress the test score on the treatment indicator, controlling for baseline test scores, gender, birth month, indicator for late consent, parental characteristics (mother and father's education level, earnings, and an indicator for non-western country of birth) and randomization block. In Model 2 we exclude baseline test scores in order to investigate robustness with respect to the slight unbalance in baseline test scores. In Model 3 we only include baseline test scores and randomization block. Model 4 controls for randomization

block only. In all models we cluster on block level to adjust for correlated error terms within blocks.

In Model 1 we find evidence of a positive treatment effect on the sum score of the children's skill level (T2) that persists to the follow-up assessment (T3). The effect sizes of the estimates are 12.1 percent of a standard deviation in T2 ($p < 0.10$) and 13.4 percent in T3 ($p < 0.10$). These estimates are robust to excluding controls for baseline test scores in Model 2, and for excluding controls for child and parent characteristics in Model 3, but the estimates lose precision and become smaller in magnitude in Model 4 where all of the above-mentioned controls are excluded.

When investigating effects in specific skill domains, we find that the treatment effect was particularly pronounced in mathematics. Moreover, the treatment effect in mathematics is substantially larger in T3 as compared to T2; 15.5 percent of a standard deviation in T2 (not significant) and 22.8 percent in T3 ($p < 0.01$).[x] There is also an immediate positive treatment effect on executive functioning (EF) of 14.6 percent of a standard deviation ($p < 0.05$), but the effect fades by the follow-up assessment. We find no effects on language in either T2 or T3. The few-cluster correction, reported in the table for all models as "Wild P", leads to only marginally higher P-values.[xi]

Table 4 investigates heterogeneous treatment effects across subsamples by including an interaction term with an indicator for high-quality center (Panel A), high baseline skill level (Panel B), and high parental education (Panel C), all subsamples split at median. Apart from the interaction term, the model specifications are identical to Model 1 in Table 3. Panel A shows that the treatment effect is mainly driven by preschool centers identified as low-quality centers at baseline. In these centers the treatment effect on the sum score was 22 percent of a standard deviation ($p < 0.01$) in the one-year follow-up, which was 19 percent of a standard deviation ($p < 0.10$) larger than in the high-quality centers. For children in these preschool

centers, there are significant treatment impacts on mathematics and language in the T3 follow-up assessment. The treatment effect is particularly strong in mathematics (31.1 percent, $p < 0.01$), but also sizable and significant in language (15.7 percent, $p < 0.05$). There is no significant treatment impact on executive functioning.[xii]

Panels B and C show no significant differences in the treatment effect across baseline skill levels or parental education, suggesting that children benefited equally from the curriculum independent of baseline skill level and family background.[xiii]

VIII. Discussion

This paper seeks to provide evidence that can help us better understand important conditions for preschool effectiveness. Specifically, one concern is the relatively non-specific and unstructured curriculum of many universal preschool programs, which can give rise to large differences in learning across centers. Our field experiment introduced a structured curriculum for five-year-olds and accompanying teacher training into the universal preschool context of Norway, where the business as usual is an unstructured curriculum. We find that our structured curriculum intervention had effects on child development at post-intervention and the effects persist one year following the end of the treatment.

The follow-up results at the end of 1st grade must be interpreted in light of the school experiences the children had in the year following the intervention. As preschool centers are significantly smaller than schools for 1st graders, most schools will only enroll a few treated children. Although teachers are supposed to tailor their instruction to children's skill level, it is unlikely that 1st grade teachers changed their overall curriculum in response to our intervention. The large treatment effects at the end of 1st grade therefore suggest persistent treatment effects from the intervention, and not that children were also treated in 1st grade.

Motivated by the concern of large differences in learning across centers in preschool systems with unstructured curriculum (Bennett and Tayler 2006, Engel et al. 2015), we investigated differential treatment effects across centers identified as low- and high-quality centers at baseline. Interestingly, our heterogeneity analysis demonstrated that the treatment effects are particularly large among preschool centers identified as low quality at baseline. In this sample the treatment effect on the sum score was 16 percent of a standard deviation post-intervention (not significant) and increased to 22 percent in the one-year follow-up ($p < 0.01$).

A survey conducted for control and treated preschool centers the year prior to the implementation of the field experiment, allows us to investigate this heterogeneity further (52 of 71 centers responded). In this survey we asked how many hours per week the center offered planned and structured pedagogical activities exclusively for the five-year-olds prior to the intervention. In regression analyses reported in Appendix Table A4 we can see that a one-hour increase in planned and structured pedagogical activities exclusively for the five-year-olds, predicts a higher value of the fixed effect quality indicator ($b = .112$, $SE = .065$).[xiv] The estimate is imprecise (only significant at the ten percent level), but provides suggestive evidence that there is a larger contrast in practice between treatment and control among centers identified as low-quality. This provides a possible explanation for why our treatment was more effective in low-quality centers. It suggests that centers identified as high-quality had a pre-intervention practice more similar to the treatment, already providing planned and structured pedagogical activities on a regular basis. This illuminates the concern that preschool systems with no structured curriculum, such as the Norwegian preschool system, give centers a large degree of freedom with respect to pedagogical content, which can give rise to large differences in learning across centers (Bennett and Tayler 2006, Engel et al. 2015). Moreover, our field experiment demonstrates that introducing structured curriculum

can reduce inequality in early childhood learning environments by raising center quality at the bottom of the distribution, which is an important new insight.

We assess the cost-effectiveness of our treatment following Kraft (2020), who recently introduced a schema with new empirical benchmarks for assessing cost-effectiveness in experimental investigations of education interventions. In an extensive survey of the literature, Kraft (2020) concludes that standardized effect sizes above .20 should be considered as large, and effects sizes between .05 and .20 as medium. Moreover, interventions more expensive than $4,000 per student in 2016 dollars are benchmarked as high-cost intervention, whereas costs between $500-$4,000 per student are considered as moderate. Appendix A4 calculates the cost of our treatment per child, in addition to the cost-effectiveness ratio; standardized effect size per USD 1000 spent. The cost is $4,298 in 2016 dollars. As such, our intervention provided moderate effects (.13 SD) at high costs, implying a cost-effectiveness ratio of 0.028. However, targeting the intervention to low-quality centers would provide larger effects (.22 SD), implying a cost-effectiveness ratio of 0.051. These cost-effectiveness ratios should be considered as lower bounds, as there are likely spillover effects on the three- and four-year-olds for whom the teacher was also responsible when he/she was not in the group with the five-year-olds only. Moreover, children, and especially five-year-olds, may in years to come benefit from the preschool teacher's training at no additional cost. Accounting for such a benefit of the training for three additional cohorts of five-year-olds, using a discount rate of ten percent, the cost-effectiveness ratio amounts to 0.22 for an intervention targeted towards low-quality centers.

We expected the treatment to be particularly beneficial for children with few learning opportunities at home, and more limited impact for the children with stimulating home learning environments. In contrast, children in our intervention benefited equally from the treatment regardless of their initial skill level or their family background. This could be

because the sample was relatively advantaged, making it more difficult to detect significant differences based on family background or skill level. Moreover, the curriculum provided teachers with suggestions for how to adjust the activities to fit the developmental stage of all children, giving all children – independent of background – equal chances to gain from the intervention.

The effect of our structured curriculum intervention was particularly pronounced in mathematics. The persistent and large impact on mathematics skills is important because previous research has demonstrated that mathematics achievement is a strong predictor of later success in school and high school graduation (Duncan et al. 2007). This suggests that a structured curriculum may be important for children's human capital development in the long run.

There are several possible explanations for why treatment effects were particularly strong for mathematics. Results from U.S. studies show that preschool teachers spend limited time on planned mathematics activities for young children (Engel, Claessens, and Finch 2013), which was also true for this study's teachers prior to the intervention. As such, our intervention may have greater value-added for mathematics skills. This could be further reinforced if the children's home environments are better at stimulating language skills and executive functioning, as compared to mathematics skills, as suggested by research (Cannon and Ginsburg 2008).

In addition, at least four features of the intervention may have shaped these results. First, more of the intervention focused on mathematics content compared to the other skill domains. Second, we advised the trained headteachers to implement the mathematics activities and the assistants to implement the language activities under the guidance of the headteacher. The intervention may have been more effective for language development if the trained teachers also implemented the interactive book reading. Third, the mathematics curriculum was the

most scripted of all the components. Some have argued that scripts support implementation by providing clear models for the adult how to deliver an activity with high fidelity (e.g., Weiland et al. 2018). Finally, researchers argue that executive functioning skills (including working memory and inhibitory control) lay the foundation for children's academic success (Blair and Raver 2015). Thus, it may be that the immediate improvements in executive functioning further helped treated children show greater gains in mathematics skills at the end of 1st grade.

The universal childcare system in Norway provided an excellent field for investigating the effects of a structured curriculum on children's skills, as Norway's current preschool curriculum is very non-specific and unstructured. As in all field experiments, however, it is important to carefully discuss the results considering the specific field context and intervention design, in order to assess external validity and mechanisms, which we do below.

We were limited in this project to design an age-appropriate curriculum for the five-year-olds. However, in Norway classrooms are typically mixed such that the five-year-olds spend the day with three- and four-year-olds. Therefore, in order to deliver the curriculum, they had to be pulled out from their regular classroom. As a consequence, they spent more time exclusively with same-age peers, and typically in a much smaller group than their regular classroom. We were interested in testing a policy-relevant version of the intervention and that meant leaving the control group as business as usual. As a result, we cannot separate the effects of the curriculum from the possible effects from more teacher resources and the structural change in learning environment, and our results need to be interpreted as effects of this bundled treatment.

Furthermore, as we described, teachers were involved in designing the curriculum. This was crucial in order to develop a curriculum that would be relevant, feasible, and attractive for

preschool teachers to implement at scale as discussed in Størksen et al. (2021). Nevertheless, it is possible that this introduced Hawthorne effects and that teachers with less personal investment in the curriculum would implement it differently.

In addition, although our study is the first of a high-quality structured comprehensive curriculum in a universal context, Norway remains a special case. As we have pointed out, Norway has a strong social safety net which may mean that few of the children in the experiment were subject to significant material hardship. The headteachers in Norway all had BA degrees and training in early childhood education. And Norway (along with other Scandinavian countries) has a long-standing tradition of emphasizing play in preschools, rather than didactic learning, which may have affected how they adopted a more structured curriculum. There is also a limitation with respect to external validity within Norway: in order to ensure high implementation quality, participation in the experiment was voluntary. The results may not generalize to preschools who are less motivated or interested in this new type of curriculum and teacher training.

IX. Conclusion

This is the first study to test an intervention that introduces a structured curriculum into a universal preschool context where the existing curriculum is very non-specific and unstructured. We demonstrate that a structured curriculum has effects on child development at post-intervention and the effects persist one year following the end of the treatment. The impact was particularly large for mathematics. This suggests that a structured curriculum is important for children's human capital development in a universal preschool context. However, as discussed above, our results must be interpreted with caution as our treatment was bundled, and we cannot separate the effects of the curriculum from the possible effects of

other mechanisms. These mechanisms include more teacher resources, more time with same-age peers, and a shift in children's focus towards mathematics.

One possible concern about universal preschool system without a structured curriculum is that it gives the preschool centers a large degree of freedom with respect to pedagogical content, which can give rise to large differences in learning across centers (Bennett and Tayler 2006, Engel et al. 2015). Consistent with this concern, we find that treatment impacts were particularly large in preschool centers identified as low quality at baseline. This suggests that a structured curriculum can reduce inequality in early childhood learning environments by substantially raising center quality at the bottom of the distribution.

References

Araujo, M Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher quality and learning outcomes in kindergarten." *The Quarterly Journal of Economics* 131 (3):1415-1453.

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2008. "Universal child care, maternal labor supply, and family well-being." *Journal of Political Economy* 116 (4):709-745.

Bauchmüller, Robert, Mette Gørtz, and Astrid Würtz Rasmussen. 2014. "Long-run benefits from universal high-quality preschooling." *Early Childhood Research Quarterly* 29 (4):457-470.

Bennett, John, and Collette P Tayler. 2006. *Starting strong II: Early childhood education and care*: OECD.

Berlinski, Samuel, Sebastian Galiani, and Paul Gertler. 2009. "The effect of pre-primary education on primary school performance." *Journal of Public Economics* 93 (1-2):219-234.

Berlinski, Samuel, Sebastian Galiani, and Marco Manacorda. 2008. "Giving children a better start: Preschool attendance and school-age profiles." *Journal of public Economics* 92 (5-6):1416-1440.

Best, John R, and Patricia H Miller. 2010. "A developmental perspective on executive function." *Child development* 81 (6):1641-1660.

Blair, Clancy, and C. Cybele Raver. 2015. "School Readiness and Self-Regulation: A Developmental Psychobiological Approach." *Annual Review of Psychology* 66 (1):711-731.

Blau, David, and Janet Currie. 2006. "Pre-school, day care, and after-school care: who's minding the kids?" *Handbook of the Economics of Education* 2:1163-1278.

Blau, David M. 1999. "The effect of child care characteristics on child development." *Journal of Human Resources*:786-822.

Burchinal, Margaret. 2018. "Measuring early care and education quality." Child Development Perspectives 12(1): 3-9.

Bodrova, Elena, and Deborah J Leong. 2007. "Tools of the mind." *Columbus, OH: Pearson*.

California Department of Education. 2016. California Preschool Curriculum Frameworks.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2): 317-372.

Cannon, Joanna, and Herbert P Ginsburg. 2008. ""Doing the math": Maternal beliefs about early mathematics versus language learning." *Early Education and Development* 19 (2):238-260.

Chaisemartin, Clement, and Jaime Ramirez-Cuellar. 2020. "At What Level Should One Cluster Standard Errors in Paired Experimenst, and in Stratified Experiments with Small Strata?" Working paper.

Clements, Douglas H, and Julie Sarama. 2011. "Early childhood mathematics intervention." *Science* 333 (6045):968-970.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2018. "Who benefits from universal child care? Estimating marginal returns to early child care attendance." *Journal of Political Economy* 126 (6):2356-2409.

Currie, Janet, and Matthew Neidell. 2007. "Getting inside the "black box" of Head Start quality: What matters and what doesn't." *Economics of Education review* 26 (1):83-99.

Davidson, Matthew C., Dima Amso, Loren Cruess Anderson, and Adele Diamond. 2006. "Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching." *Neuropsychologia* 44 (11):2037-2078.

Diamond, Adele, W Steven Barnett, Jessica Thomas, and Sarah Munro. 2007. "Preschool program improves cognitive control." *Science* 318 (5855):1387.

Dillon, Moira R, Harini Kannan, Joshua T Dean, Elizabeth S Spelke, and Esther Duflo. 2017. "Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics." *Science* 357 (6346):47-55.

Drange, Nina, and Marte Rønning. 2020. "Child care center quality and early child development." *Journal of Public Economics* 188: 104204.

Duncan, G. J., C. J. Dowsett, A. Claessens, K. Magnuson, A. C. Huston, P. Klebanov, L. S. Pagani, L Feinstein, M. Engel, J. Brooks-Gunn, H. Sexton, K. Duckworth, and C. Japel. 2007. "School Readiness and Later Achievement." *Developmental Psychology* 43 (6):1428-1446.

Engel, Arno, W Steven Barnett, Yvonne Anders, and Miho Taguma. 2015. "Early childhood education and care policy review." OECD.

Engel, Mimi, Amy Claessens, and Maida A Finch. 2013. "Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten." *Educational Evaluation and Policy Analysis* 35 (2):157-178.

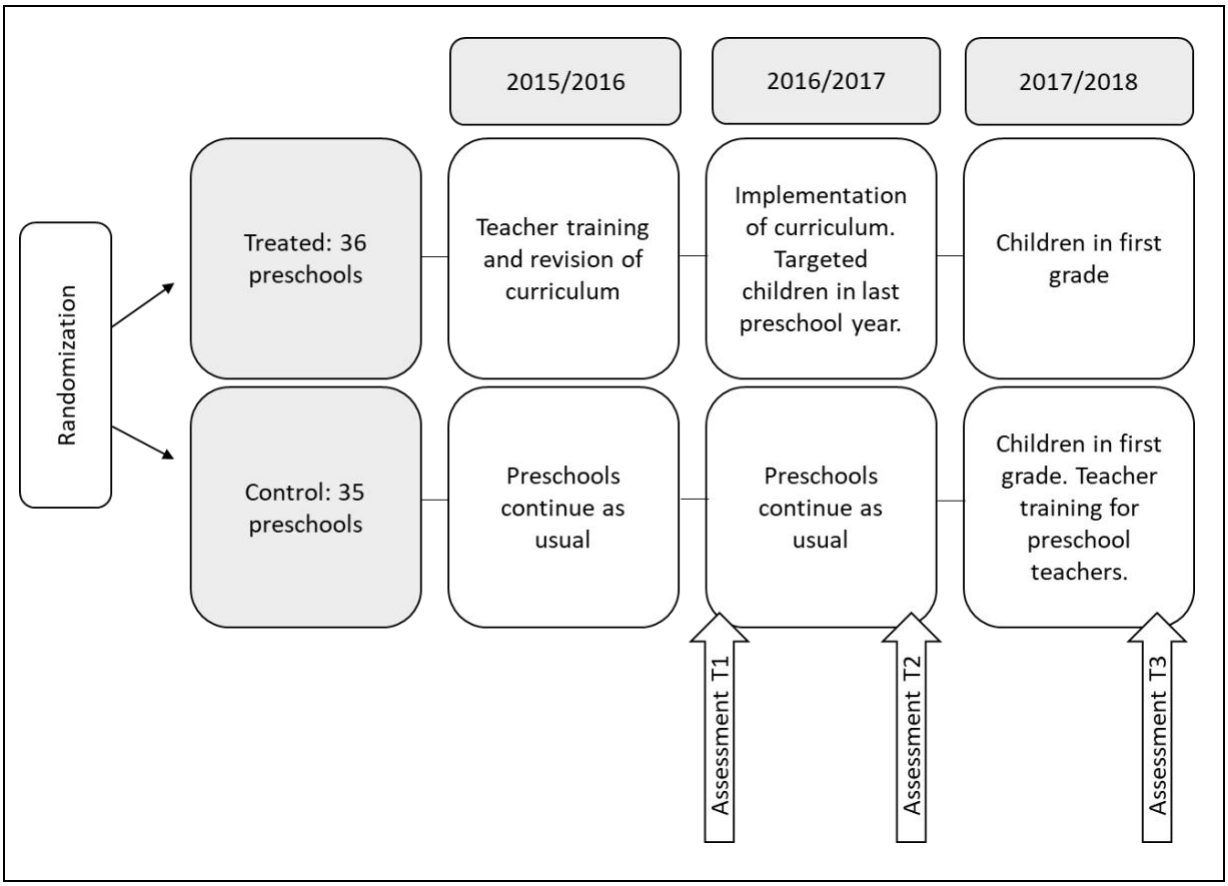
Felfe, Christina, Natalia Nollenberger, and Núria Rodríguez-Planas. 2015. "Can't buy mommy's love? Universal childcare and children's long-term cognitive development." *Journal of Population Economics* 28 (2):393-422.

Gresham, Frank M., and Stephen N. Elliott. 1990. *Social skills rating system: Manual*. American guidance service.
Gupta, Nabanita Datta, and Marianne Simonsen. 2010. "Non-cognitive child outcomes and universal high quality child care." *Journal of Public Economics* 94 (1-2):30-43.
Havnes, Tarjei, and Magne Mogstad. 2011. "No child left behind: Subsidized child care and children's long-run outcomes." *American Economic Journal: Economic Policy* 3 (2):97-129.
Heckman, James J, Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev, and Adam Yavitz. 2010. "The rate of return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94 (1-2):114-128.
Jackson, C Kirabo, Jonah E Rockoff, and Douglas O Staiger. 2014. "Teacher effects and teacher-related policies." *Annu. Rev. Econ.* 6 (1):801-825.
Kalil, Ariel, Kathleen M. Ziol-Guest, Rebecca M. Ryan, and Anna J. Markowitz. 2016. Changes in income-based gaps in parent activities with young children from 1988-2012. *AERA Open 2 (3)*, 1-17.
Kalil, Ariel, Rebecca Ryan, and Michael Corey. 2012. Diverging destinies: Maternal education and investments in children. *Demography*, *49*, 1361-1383
Kraft, Matthew A., 2020. Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), pp.241-253.
Lonigan, Christopher J., et al. 2015. "Impacts of a comprehensive school readiness curriculum for preschool children at risk for educational difficulties." *Child Development* 86(6): 1773-1793.
McClelland, Megan M., Claire E. Cameron, Robert Duncan, Ryan P. Bowles, Alan C. Acock, Alicia Miao, and Megan E. Pratt. 2014. "Predictors of early growth in academic achievement: The Head-Toes-Knees-Shoulders task." *Frontiers in Psychology* 5.
Melhuish, Edward C. 2011. "Preschool matters." *Science* 333 (6040):299-300.
Ministry of Education and Research. 2017. Framework plan for the content and tasks of kindergartens. https://www.udir.no/globalassets/filer/barnehage/rammeplan/ framework-plan-for-kindergartens2-2017.pdf.
Mol, Suzanne E., Adriana G. Bus, and Maria T. de Jong. 2009. "Interactive Book Reading in Early Education: A Tool to Stimulate Print Knowledge as Well as Oral Language." *Review of Educational Research* 79 (2):979-1007.
Montie, Jeanne E., Zongping Xiang, and Lawrence J. Schweinhart. 2006. "Preschool experience in 10 countries: Cognitive and language performance at age 7." *Early Childhood Research Quarterly* 21(3): 313-331.
OECD. 2015. *Starting Strong IV*.
Pianta, Robert C. 1999. *Enhancing relationships between children and teachers*. Washington, DC, US: American Psychological Association.
Rege, Mari, Ingeborg Foldøy Solli, Ingunn Størksen, and Mark Votruba. 2018. "Variation in center quality in a universal publicly subsidized and regulated childcare system." *Labour Economics* 55:230-240.
Sabol, Terri J, SL Soliday Hong, Robert C Pianta, and Margaret Burchinal. 2013. "Can rating pre-K programs predict children's learning?" *Science* 341 (6148):845-846.
Schmitt, Sara A, Megan M McClelland, Shauna L Tominey, and Alan Acock. 2015. "Strengthening school readiness for Head Start children: Evaluation of a self-regulation intervention." *Early Childhood Research Quarterly* 30:20-31.
Shure, Myrna B. 1992. *I can problem solve (kindergarten and primary grades): An interpersonal cognitive problem-solving program for children*: Research Press.
Størksen, Ingunn, Ingunn T. Ellingsen, Maren S. Tvedt, and Ella MC Idsøe. 2013. "Norsk vokabulartest (NVT) for barn i overgangen mellom barnehage og skole: Psykometrisk vurdering av en nettbrettbasert test." *Spesialpedagogikks forskningsdel* 04/13:40 - 54.
Størksen, Ingunn, Dieuwer ten Braak., Svanhild Breive, Ragnhild Lenes, Svanaug Lunde, Martin Carlsen, Ingvald Erfjord, Per Sigurd Hundeland, and Mari Rege. 2018. *Lekbasert læring - et forskningsbasert førskoleopplegg fra Agderprosjektet* Oslo: GAN Aschehoug.
Størksen, Ingunn, Sigrun K. Ertesvåg, and Mari Rege. 2021. "Implementing implementation science in a randomized controlled trial in Norwegian early childhood education and care." *International Journal of Educational Research* 108: 101782.
ten Braak, Dieuwer, and Ingunn Størksen. 2021. "Psychometric properties of the Ani Banani Math Test." *European Journal of Developmental Psychology*: 1-18.
Wechsler, David. 1991. *WISC-III: Wechsler intelligence scale for children: Manual*: Psychological Corporation.
Weiland, Christina, and Hirokazu Yoshikawa. 2013. "Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills." *Child Development* 84 (6):2112-2130.
Weiland, Christina, Meghan McCormick, Shira Mattera, Michelle Maier, and Pamela Morris. 2018."Preschool curricula and professional development features for getting to high-quality implementation at scale: A comparative review across five trials." *AERA Open* 4, no. 1.

Weisberg, Deena Skolnick, Kathy Hirsh-Pasek, and Roberta Michnick Golinkoff. 2013. "Guided Play: Where Curricular Goals Meet a Playful Pedagogy." *Mind, Brain, and Education* 7 (2):104-112.
Yoshikawa, Hirozaku, Christina Weiland, Jeanne Brooks-Gunn, Margaret R. Burchinal, Linda M. Espinosa, William T. Gormley, Jens Ludwig, Katherine A. Magnuson, Deborah Phillips, and Martha J. Zaslow. 2013. *Investing in our future: The evidence base on preschool education*. Report.

Figures

Figure 1: Experimental design



Note: 71 preschool centers randomly split between control and treatment. Preschool year 2015/2016: Teachers in treated centers attended the teacher training and helped revise the curriculum. 2016/2017: Teachers in treated centers implemented the structured curriculum with the five-year-olds in their center. 2017/2018: Teachers in control centers received the teacher training. We assessed children's skills in language, mathematics, and executive functioning in August 2016 (baseline, T1), June 2017 (post-intervention, T2), and March 2018 (follow-up, T3).

Table 1. Test score correlations

T1 test scores:	Sum score	Math	Executive functioning	Language
T1 Test scores:				
Math	0.807			
Executive functioning	0.837	0.545		
Language	0.780	0.411	0.483	
T2 Test scores:				
Sum score	0.764	0.610	0.648	0.595
Math	0.620	0.593	0.507	0.404
Executive functioning	0.667	0.521	0.667	0.428
Language	0.592	0.384	0.418	0.634
T3 test scores:				
Sum score	0.708	0.571	0.604	0.540
Math	0.551	0.502	0.470	0.362
Executive functioning	0.595	0.483	0.586	0.374
Language	0.574	0.402	0.412	0.578

Note: Gross sample, N<=691. Sample varies slightly across test scores and across assessment period. All correlations are significant at 1 percent level.

Table 2. Descriptive statistics and balance test for T2 and T3 sample.

	Post-intervention T2					Follow-up T3				
	Control	Treat	Difference	Wild P	N	Control	Treat	Difference	Wild P	N
T1 Sum score	0.022	-0.021	-0.056	0.501	650	0.028	-0.016	-0.064	0.454	661
	(0.058)	(0.053)	(0.083)			(0.059)	(0.052)	(0.083)		
T1 Executive functioning	-0.015	0.009	0.024	0.799	650	-0.022	0.019	0.040	0.669	661
	(0.059)	(0.053)	(0.090)			(0.058)	(0.052)	(0.093)		
T1 Language	0.058	-0.046	-0.112	0.190	650	0.076	-0.050	-0.147+	0.072	661
	(0.059)	(0.052)	(0.084)			(0.060)	(0.050)	(0.076)		
T1 Math	0.010	-0.013	-0.046	0.573	650	0.014	-0.008	-0.047	0.578	661
	(0.056)	(0.055)	(0.079)			(0.055)	(0.053)	(0.084)		
Female	0.517	0.480	-0.023	0.469	650	0.515	0.475	-0.029	0.435	661
	(0.500)	(0.500)	(0.032)			(0.500)	(0.500)	(0.037)		
Birth month	6.878	6.807	-0.076	0.748	650	6.744	6.826	0.063	0.813	661
	(3.184)	(3.213)	(0.235)			(3.229)	(3.177)	(0.266)		
Mother education	14.377	14.161	-0.169	0.464	626	14.385	14.126	-0.230	0.268	637
	(2.590)	(2.587)	(0.230)			(2.547)	(2.569)	(0.207)		
Father education	13.814	13.696	-0.136	0.658	620	13.782	13.715	-0.085	0.776	628
	(2.563)	(2.504)	(0.265)			(2.536)	(2.469)	(0.252)		
Mother earnings	341,408	320,168	-22,902	0.324	648	339,897	322,267	-21,206	0.329	658
	-221,464	-206,324	-22,751			-214,889	-203,739	-21,584		
Father earnings	544,773	558,596	14,311	0.512	636	547,552	562,885	18,007	0.415	643
	-259,182	-267,057	-21,544			-262,667	-272,631	-21,599		
Non-western immigrant	0.130	0.201	0.068	0.123	650	0.136	0.203	0.064	0.142	661
	(0.337)	(0.401)	(0.040)			(0.343)	(0.403)	(0.039)		
Missing T1 testscores	0.051	0.034	0.015	0.437	650	0.058	0.041	0.016	0.518	661
	(0.013)	(0.010)	(0.019)			(0.014)	(0.010)	(0.024)		
N	292	358	650			293	368	661		
F test			0.295					0.000		

Note: + p<0.1. The columns provide mean (standard deviation) for covariates and T1 test scores for the control group and treatment group in the T2 and T3 analytic samples. The column labeled Difference is the estimated coefficient (standard error) from regressing each covariate against treatment status. Regressions are clustered on and control for randomization block.

Table 3. Main results. Treatment effect on test scores at post-intervention (T2) and in the one-year follow-up (T3).

	Post-intervention (T2)				Follow-up (T3)			
	Sum score	Math	EF	Language	Sum score	Math	EF	Language
Model 1:								
Treat	0.121+	0.155	0.123+	0.011	0.134+	0.228**	0.058	0.043
	(0.063)	(0.093)	(0.058)	(0.068)	(0.072)	(0.063)	(0.055)	(0.093)
Wild P	0.079	0.145	0.057	0.870	0.085	0.004	0.296	0.680
N	652	650	652	648	661	661	660	659
Adj. R2	0.612	0.440	0.492	0.532	0.522	0.364	0.382	0.492
Model 2:								
Treat	0.130	0.154	0.146*	0.012	0.133	0.213*	0.076	0.033
	(0.074)	(0.105)	(0.066)	(0.075)	(0.093)	(0.088)	(0.075)	(0.098)
Wild P	0.122	0.208	0.056	0.867	0.188	0.043	0.348	0.744
N	652	650	652	648	661	661	660	659
Adj. R2	0.147	0.114	0.114	0.101	0.123	0.083	0.084	0.176
Model 3:								
Treat	0.122+	0.187+	0.109	-0.001	0.118	0.236**	0.063	-0.011
	(0.068)	(0.094)	(0.064)	(0.064)	(0.067)	(0.063)	(0.047)	(0.087)
Wild P	0.095	0.079	0.115	0.987	0.108	0.004	0.186	0.912
N	652	650	652	648	661	661	660	659
Adj. R2	0.613	0.433	0.492	0.521	0.517	0.349	0.379	0.456
Model 4:								
Treat	0.091	0.166	0.106	-0.055	0.091	0.213*	0.064	-0.057
	(0.101)	(0.108)	(0.090)	(0.098)	(0.102)	(0.089)	(0.074)	(0.117)
Wild P	0.387	0.165	0.263	0.588	0.377	0.035	0.400	0.632
N	652	650	652	648	661	661	660	659
Adj. R2	0.017	0.004	0.022	0.015	0.015	0.026	0.007	0.031

Note: ** p<0.01, * p<0.05, + p<0.1. Each column in each panel presents regression coefficient of treated (standard error) using ordinary least squares. For both assessment periods: Model 1 regresses outcome on the treatment indicator, controlling for baseline test scores, gender, birth month, parental characteristics (mother and father's education level, earnings, an indicator for non-western country of birth), and indicators for late consent and not having participated in the T1 assessment. In Model 2 we exclude baseline test scores from the model. In Model 3 we restrict controls to baseline test scores and an indicator for not having participated in the T1 assessment. Model 4 has no controls. All regressions are clustered on and control for randomization block. We have utilized the boottest package in Stata to do a few-cluster-correction of the p-value, reported in the table as Wild-P.

Table 4. Heterogeneous treatment effect on test scores at post-intervention (T2) and in the one year follow-up (T3) across high/low preschool quality, baseline skills and parent education.

	Post-intervention (T2)				Follow-up (T3)			
	Sum score	Math	EF	Language	Sum Score	Math	EF	Language
Panel A: Preschool center quality								
Treat	0.156	0.227	0.127	0.026	0.219**	0.311**	0.064	0.157*
	(0.097)	(0.138)	(0.105)	(0.096)	(0.072)	(0.096)	(0.064)	(0.072)
Treat*High	-0.065	-0.113	-0.019	-0.038	-0.186+	-0.153	-0.030	-0.264
	(0.107)	(0.164)	(0.126)	(0.115)	(0.102)	(0.124)	(0.116)	(0.152)
Wild P (treat)	0.142	0.170	0.261	0.791	0.027	0.025	0.330	0.034
Wild P (t*h)	0.545	0.522	0.883	0.736	0.102	0.246	0.796	0.132
N	638	636	638	634	648	648	647	646
Adj. R2	0.605	0.433	0.490	0.526	0.519	0.357	0.385	0.489
Mean Low	-0.171	-0.152	-0.157	-0.105	-0.123	-0.098	-0.126	-0.078
Mean High	0.174	0.152	0.161	0.108	0.124	0.084	0.139	0.081
Panel B: Baseline skills								
Treat	0.066	0.121	0.104	-0.059	0.123	0.205+	0.008	0.010
	(0.093)	(0.125)	(0.089)	(0.084)	(0.111)	(0.109)	(0.093)	(0.089)
Treat*High	0.107	0.049	0.026	0.148	0.023	0.035	0.088	0.068
	(0.123)	(0.143)	(0.100)	(0.137)	(0.108)	(0.133)	(0.133)	(0.135)
Wild P (treat)	0.503	0.401	0.261	0.481	0.282	0.088	0.933	0.908
Wild P (t*h)	0.395	0.741	0.795	0.288	0.821	0.784	0.511	0.608
N	652	650	652	648	661	661	660	659
Adj. R2	0.616	0.444	0.492	0.534	0.527	0.366	0.383	0.491
Mean Low	-0.637	-0.496	-0.553	-0.547	-0.584	-0.442	-0.477	-0.469
Mean High	0.625	0.487	0.543	0.550	0.586	0.441	0.479	0.465
Panel C: Parent education								
Treat	0.086	0.141	0.032	0.040	0.114	0.250*	-0.014	0.038
	(0.102)	(0.115)	(0.098)	(0.105)	(0.127)	(0.105)	(0.108)	(0.137)
Treat*High	0.060	0.010	0.163	-0.044	0.039	-0.042	0.108	0.036
	(0.121)	(0.128)	(0.119)	(0.100)	(0.136)	(0.131)	(0.157)	(0.123)
Wild P (treat)	0.400	0.240	0.759	0.692	0.394	0.038	0.902	0.799
Wild P (t*h)	0.618	0.938	0.206	0.647	0.787	0.756	0.506	0.791
N	641	639	641	637	649	649	648	647
Adj. R2	0.601	0.432	0.486	0.519	0.506	0.358	0.372	0.475
Mean Low	-0.200	-0.192	-0.179	-0.117	-0.192	-0.174	-0.120	-0.176
Mean High	0.222	0.203	0.195	0.147	0.227	0.189	0.151	0.215

Note: ** p<0.01, * p<0.05, + p<0.1. Each column in each panel presents regression coefficients with standard errors in parenthesis, using ordinary least squares. The model specification is in line with Model 1 in Table 3: We add controls for gender, birth month and parental characteristics (education, earnings, and indicator for non-western country of birth), baseline test scores, indicators for late consent and not having participated in the T1 assessment, and randomization block, all regressions clustered on randomization block. Preschool center quality is measured as the preschool center fixed effect (center average covariate-adjusted test score). High/low center quality is split at median value. Parental education is measured as the average of mother's and father's number of years of education. 12 children with no information on parental education (balanced across treatment status) are excluded from the sample. High/low parental education and high/low baseline skills are split at median value. Mean Low and Mean High are means of outcome variable in relevant subsamples. We have utilized the boottest package in Stata to do a few-cluster-correction of the p-value, reported in the table as Wild-P(treat) and Wild-P(t*h).

[i] Our curriculum also included activities to promote social skills. Unfortunately, we were not able to reliably measure social skills in this study due to lack of tests validated in a Norwegian context. See Appendix A1 for more details on curriculum and teacher training.
[ii] See Chaisemartin and Ramirez-Cuellar (2020) for a recent discussion on level of clustering. When clustering on center level, standard errors are marginally smaller, results available on request.
[iii] See Cameron and Miller (2015) for a discussion on challenges when the sample consists of few clusters, and suggested approaches to adjust standard errors correspondingly.
[iv] Notably, when the *High*-indicator refers to high quality center, the correct index should be *c*, not *i*.
[v] Each block consisted of two preschool centers in the same municipality, and of similar center size. After accessing data, we realized that center size was weakly associated with number of children actually participating and consenting. This resulted in large variation in the likelihood of treated (at the child level) across blocks. Consequently, we collapsed blocks in same municipality. Analyses with indicators for non-collapsed blocks provide similar result. See Appendix Table A5 for summary statistics across collapsed blocks.
[vi] See appendix A2 for more details on assessment and measures.
[vii] In all analyses we add a control for not having participated in the T1 assessment. We also investigate if our results are robust to not controlling for any baseline test scores. Furthermore, in a robustness check in Appendix Table A1 Panel B we demonstrate that our findings are robust to excluding from our sample children who did not participate in the T1 assessment.
[viii] The number of children assessed at a given time varies slightly across tests. This is because we ended the assessment session for a few children who were reluctant to continue.
[ix] Data on parental characteristics was not available for 3 percent of the children, likely because these children/families were recent immigrants to Norway at the time, and still not recorded in the Norwegian administrative registers. Missing values were replaced by 0, and indicators for missing were included in the analyses.
[x] As we show in Appendix Table A1 Panel A, the T2 effects on mathematics and executive functioning are somewhat smaller and insignificant when excluding children with late parental consent. However, the T3 effects are robust to excluding these children.
[xi] We have utilized the boottest package in Stata to do a few-cluster-correction of the p-value (wild t-bootstrap).
[xii] Since the quality indicators could not be calculated for four small centers (<5 children), the sample size is somewhat smaller in Panel A.
[xiii] See Appendix Table A2 for treatment effects on each of the six assessments.
[xiv] However, there is an insignificant negative association between total hours of planned and structured pedagogical activities for five-year-olds (in mixed age group or with five-year-olds) and the quality indicator (*b* = -.104, *SE* = .091).