CESifo WORKING PAPERS

7775
2019

August 2019

Promoting Child Development in a Universal Preschool System: A Field Experiment

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

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7775*
Category 5: Economics of Education

Promoting Child Development in a Universal Preschool System: A Field Experiment

Abstract

This study tests an intervention that introduces a structured curriculum for five-year olds into the universal preschool context of Norway. 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. Compared to business as usual, the nine-month curriculum intervention has effects on child development at post-intervention and the effects persist one year following the end of the treatment. The effects are entirely driven by the preschool centers identified as low-quality at baseline, suggesting that a structured curriculum can reduce inequality in early childhood learning environments.

JEL-Codes: I200, H420.

Keywords: universal preschool, intervention, randomized controlled trial, field experiment, child development.

*Mari Rege**
University of Stavanger / Norway
mari.rege@uis.no

*Ingunn Størksen**
University of Stavanger / Norway
ingunn.storksen@uis.no

Ingeborg F. Solli
University of Stavanger / Norway
ingeborg.f.solli@uis.no

Ariel Kalil
University of Chicago / IL / USA
akalil@uchicago.edu

Megan McClelland
Oregon State University / Corvallis / USA
megan.mcclelland@oregonstate.edu

Dieuwer ten Braak
University of Stavanger / Norway
dieuwer.t.braak@uis.no

Ragnhild Lenes
University of Stavanger / Norway
ragnhild.lenes@uis.no

Svanaug Lunde
University of Stavanger / Norway
svanaug.lunde@uis.no

Svanhild Breive
University of Agder / Kristiansand / Norway
svanhild.breive@uia.no

Martin Carlsen
University of Agder / Kristiansand / Norway
martin.carlsen@uia.no

Ingvald Erfjord
University of Agder / Kristiansand / Norway
ingvald.erfjord@uia.no

Per S. Hundeland
University of Agder / Kristiansand / Norway
per.s.hundeland@uia.no

*corresponding authors

Acknowledgments: We are grateful to project coordinator Åse Lea who has juggled all the different logistics in this intervention study, 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, The Sørlandet Knowledge Foundations and The Agder County administrations. This study is registered in the registry of the American Economic Association (0002241).

1. 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 universal programs 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 universal preschool is effective. Specifically, one quality concern is the relatively non-specific and unstructured curriculum of many universal preschool programs (Bennett and Tayler 2006, Engel et al. 2015). This gives preschool centers a large degree of freedom with respect to pedagogical content, which can give rise to large differences in learning across centers.

The present study tests an intervention that introduced a comprehensive 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 and executive functioning. A playful learning approach permeated all the activities (Weisberg, Hirsh-Pasek, and Golinkoff 2013), and the curriculum emphasized a warm and responsive child-teacher relationship (Pianta 1999). This new curriculum was accompanied by teacher training in how to implement it.

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

received the comprehensive structured curriculum for the five-year olds in addition to the teacher training. The teachers were committed to spending at least eight hours a week for nine months engaging the five-year-olds in this curriculum. The centers in the control group continued with business-as-usual, but teachers received the teacher training and intervention material two years later. We assessed the children's skills in language, mathematics and executive functioning at baseline, post-intervention, and in a one-year follow up. At all assessments, the testers were trained, certified and blind to treatment status. Tests on baseline skills and background characteristics demonstrated that our sample was well-balanced across treatment status.

The structured curriculum intervention had a significant positive effect on a summary score of children's skills in math, language and executive functioning at post-intervention. Importantly, a significant treatment impact, with a magnitude of about 13 percent of a standard deviation, persisted one year following the end of the treatment. Investigating effects in specific skill domains, the treatment effect was particularly pronounced in mathematics. In the other skill domains, the treatment had an immediate significant impact on executive functioning, of about 11 percent of a standard deviation, but no significant impact on language. Impacts on executive functioning and language were no longer apparent at the one year follow up.

The effect on mathematics skills in the one-year follow up was quite large – 23 percent of a standard deviation. 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, and the difference in average mathematics skills among children of mothers with and without a college degree is 36 percent of a standard deviation.

We also investigated differential effects of treatment across centers identified as low- and high-quality centers at baseline. We utilize center fixed effects at baseline as a proxy for quality, i.e. the center mean difference between observed and predicted assessment scores, given their observables. As 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. Our structured curriculum intervention should be particularly effective in these centers as there is much learning on which to catch up. Consistent with this conjecture, our analyses demonstrate that the treatment effect was entirely driven by low-quality centers at baseline (median split). In these centers the treatment effect on the sum score was 15 percent of a standard deviation post-intervention, and increased to 27 percent in the one-year follow up. Moreover, we found significant and sizable treatment impacts in all skill domains at the one-year follow up; in language and executive functioning it was about 14 percent, whereas in math it was 37 percent of a standard deviation. 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, which is an important new insight as variation in center quality has given rise to widespread scientific and policy concern (Bennett and Tayler 2006).

Our field experiment makes several important contributions. Despite an enormous policy interest in universal preschool, evidence of the effectiveness of universal preschool programs is scarce and far from unified (Cornelissen et al. 2018). Some papers 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 papers showing that preschool participation can be detrimental for child development (Baker, Gruber, and

Milligan 2008, 2015), 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, the children's age, and the populations differ across the studies. As such, this literature does not help us understand the conditions under which universal preschool is effective. In contrast, in our field experiment all children participated in preschool. This allows us to investigate the causal impact 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 2017). In general, the evidence from these studies seems to mimic evidence from similar studies in schools, which suggests limited potential for improving 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).

The paper most closely related to our study 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 utilizing random classroom assignment for identification, Araujo et al. (2016) provide convincing evidence that these teacher practices are important predictors of child development. Still, as noted by the authors, the measures may be correlated with other unmeasured teacher attributes, which could themselves affect student 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 is 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 targets school readiness skills through play-based activities are a key determinant of child development in preschool. These studies, however, are conducted in the context of low-income populations in the U.S. or in developing countries. Moreover, the curriculum often targets one specific skill domain. 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.

2. 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 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).

Given the lack of a specific and intentional curriculum, systematic attention directed to children's potential learning opportunities may be low in many preschool centers (Bennett

and Tayler 2006, Engel et al. 2015), and Norway may be missing a critical opportunity to improve the human capital development of its children. Research in psychology and education suggests that detailed age appropriate curricular foci, that intentionally and systematically targets school readiness skills through play-based activities, is a key determinant of effective preschools and kindergartens (Dillon et al. 2017, Clements and Sarama 2011, Weiland and Yoshikawa 2013, Schmitt et al. 2015, Diamond et al. 2007).

3. Intervention: Structured Preschool Curriculum and Accompanying Teacher Training

Our intervention consisted of a comprehensive curriculum with age-appropriate intentional skill-building activities in mathematics, language and executive functioning, and an accompanying teacher training.[1] 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.

The curriculum has 130 learning activities, which we developed in collaboration with Norwegian preschool teachers. The learning activities are inspired by existing U.S. curricula with evidence of positive effects, 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).

[1] Our curriculum also stimulated social skills. Unfortunately, we did not 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.

Importantly, the curriculum is not a scripted program intended to dictate teacher practice on a daily basis. Instead, teachers are encouraged to develop their own approach to the curriculum 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 level. 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.

The accompanying 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 to 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 (Domitrovich et al. 2008). The coaching during 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.

4. 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. 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 first 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 and analysis plan in the registry of the American Economic Association (0002241).

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

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

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. $T_{i,c}$ is an indicator for the child's treatment status, 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 preschool center level to adjust for correlated error terms within centers.

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

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

where $High_i$ is an indicator for high-quality center, high baseline skills, or high parental education. 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. Moreover, we measure parental education as the median education level of mother and father, and high/low is split by the median of parental education. As an indicator for center quality we use the center mean difference between observed and predicted assessment 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 order to define high- and low-quality centers, we split the sample at median value of the center fixed effects.

5. 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. At the time of the randomization, the only available information about preschool centers was location and size (number of children). Each block consisted of two preschool centers in the same municipality, and of similar size. In municipalities with uneven number of preschool centers, including small municipalities with only one preschool center, we blocked centers in neighboring municipalities with similar center size.

We collected parental consents in spring 2015 when the children were three to four years old, prior to the randomization of 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. In order to maintain a large sample size, we include children with late consent in our main analyses, while adding a control for this. In a robustness check (Table A2) we demonstrate that our findings are robust 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 teachers to participate in the training during work hours, the centers received funding that compensated for their time spent on class work. Including overhead this constituted NOK 89,000 (USD 11,125), which was supposed to cover a substitute teacher in a 50 percent position during four months.

During the preschool year 2016/2017, the trained 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 teacher and the assistant(s). To enable the teachers to focus their time on the five-year-olds during curriculum implementation, the centers received NOK 222,000 (USD 27,750). This was supposed to cover a substitute teacher, including overhead, in a 50 percent position during nine months to compensate for the preschool teacher's time spent on preparation and curricular implementation. The substitute teacher would typically take charge of the younger children, so that the participating teachers could spend sufficient time with the five-year-olds during implementation.

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 teacher 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.

6. 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 (Best and Miller 2010). 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 (Størksen and Mosvold 2013) 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.

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). 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. 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.

The assessment data was merged to registry data on gender, birth month, mother's and father's education, earnings and immigrant status. Furthermore, we added indicators for late consent and randomization block. For our regression analyses, we standardize measures within each period to mean 0 and standard deviation 1 for all three skill domains. 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. [2]

7. Results

In Table 1 we provide summary statistics and balance test for the T2 and T3 samples. We find that child and parent characteristics are well balanced across treatment status. There is a higher number of non-western immigrants in the treatment group. In addition, we find, as expected, that the sample of children with late parental consent is significantly higher in the treated group. Because the preschool teachers were in charge of collecting the parental consents, a plausible explanation for this imbalance is that teachers in the treated group were more engaged in the project and worked harder to get the remaining parental consents. In the analyses, we add a control for late consent and we conduct robustness analyses excluding children with late consent (Appendix Table A2).

Table 2 presents our main results. For each outcome (in columns) we estimate three models: In Model 1 we regress the test score on the treatment indicator, controlling for baseline scores, indicators for randomization block, gender, birth month, parental characteristics

[2] See appendix A2 for more details on assessment and measures.

(mother and father's education level, earnings, and an indicator for non-western country of birth). In Model 2 we only include baseline test scores and indicators for randomization block as controls. Model 3 has no controls. In all models we cluster on preschool center level to adjust for correlated error terms within centers.

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 9.5 percent of a standard deviation in T2 and 12.6 percent in T3. These estimates are robust to excluding controls for child and parent characteristics in Model 2 and 3, but the estimates lose precision and are no longer significant.

Investigating effects in specific skill domains, the treatment effect was particularly pronounced in mathematics. Moreover, the treatment effect in mathematics is nearly twice as large in T3 as compared to T2; 12.6 percent of a standard deviation in T2 and 23.0 percent in T3. There is also an immediate positive treatment effect on executive functioning (EF), but the effect fades by the follow-up assessment. We find no effects on language in either T2 or T3.

Table 3 reports 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 2. Panel A shows that our main results are entirely driven by the preschool centers identified as low-quality centers at baseline. For these preschool centers, there is a significant treatment effect on all three skill domains in the T3 follow-up assessment. The treatment effect is particularly strong in math (37.4 percent), but also sizable and significant in executive functioning (EF) (13.3 percent) and language (14.5 percent).

Panels B and C show no significant differences in the treatment effect across baseline skill levels or across parental education.

8. Discussion

Our structured curriculum intervention was particularly pronounced in mathematics. There are several possible explanations for this. Results from U.S. studies show that preschool teachers spend limited time on structured 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 two features of the implementation 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 teachers to implement the mathematics activities and the assistants to implement the language activities under the guidance of the head teacher. The intervention may have been more effective for language development if the trained teachers also implemented the interactive book reading. Third, 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 improvements in executive functioning further helped treated children show greater gains in mathematics skills at the end of first grade.

The lasting treatment effects is entirely driven by the preschool centers identified as low quality at baseline, suggesting that a structured curriculum can reduce inequality in early

childhood learning environments. Figure 2 illustrates this point by showing the raw difference in sum achievement score between centers of high and low quality by treatment status in T1 (baseline), T2 (post-intervention) and T3 (follow-up). As expected, there is a large gap in achievement score between children in low and high quality centers in T1. For centers in the control group, it remains large in T3 demonstrating that the quality of the early childhood learning environment has lasting effects. For centers in the treatment group, however, the gap is nearly eliminated in T3. This is consistent with a concern that non-specific and unstructured curriculum gives rise to large differences in learning across centers (Bennett and Tayler 2006, Engel et al. 2015). In centers with limited stimulation there is much learning on which to catch up, and this may explain why the intervention was more effective in low quality centers. Consistent with our conjecture, a structured curriculum seems to be an effective tool to reduce inequality in early childhood learning environments.

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.

9. Conclusion

This is the first study to test an intervention that introduces a comprehensive structured curriculum for five-year olds into the universal preschool context of Norway. The treatment impacts on children's skills persisted one year following the end of the treatment. The impact was particularly large for math – 23 percent of a standard deviation in the one-year follow up. 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 preschool curriculum is important for children's human capital development in a universal preschool context.

Investigating differential treatment effects suggested that the treatment impacts were entirely driven by the preschool centers identified as low quality at baseline. In these centers, the treatment impact was significant and sizable in all skill domains at the one-year follow up; in language and executive functioning it was about 14 percent, whereas in math it was 37 percent of a standard deviation. 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. This is important new insight as variation in center quality has given rise to widespread scientific and policy concern focused on increasing quality in early childhood learning environments (Bennett and Tayler 2006).

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.

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2015. Non-cognitive deficits and young adult outcomes: The long-run impacts of a universal child care program. National Bureau of Economic Research.

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.

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

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

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.

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.

Domitrovich, Celene E, Catherine P Bradshaw, Jeanne M Poduska, Kimberly Hoagwood, Jacquelyn A Buckley, Serene Olin, Lisa Hunter Romanelli, Philip J Leaf, Mark T Greenberg, and Nicholas S Ialongo. 2008. "Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework." *Advances in School Mental Health Promotion* 1 (3):6-28.

Drange, Nina, and Marte Rønning. 2017. "Child Care Center Staff Composition and Early Child Development." *Working Paper, Statistics Norway*.

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."

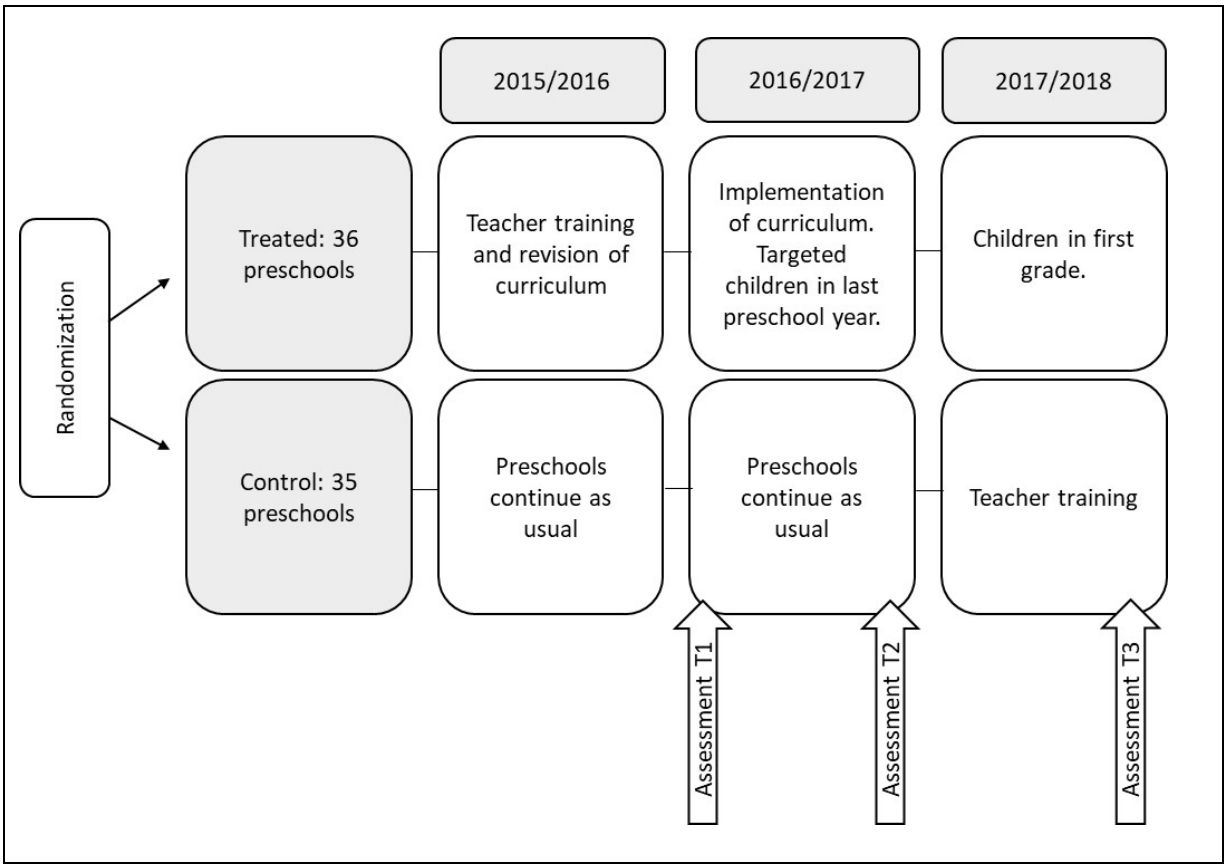
Engel, Mimi, Amy Claessens, and Maida A Finch. 2013. "Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten." *Educational Evaluation and Policy Analysis* 35 (2):157-178.

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.

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.
McClelland, M. M., C. E. Cameron, R. Duncan, R. P. Bowles, A. C. Acock, A. Miao, and M. 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.
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.
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, I. T. Ellingsen, M. S. Tvedt, and E. M. C. 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, and R. Mosvold. 2013. "Assessing early math skills with tablet computers: Development of the Ani Banani Math Test (ABMT) for young children." Utdanning2020, The Norwegian Research Council, Oslo, March 18th 2013.
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.
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.
Wells, Gordon. 1999. *Dialogic inquiry: Towards a socio-cultural practice and theory of education*: Cambridge University Press.

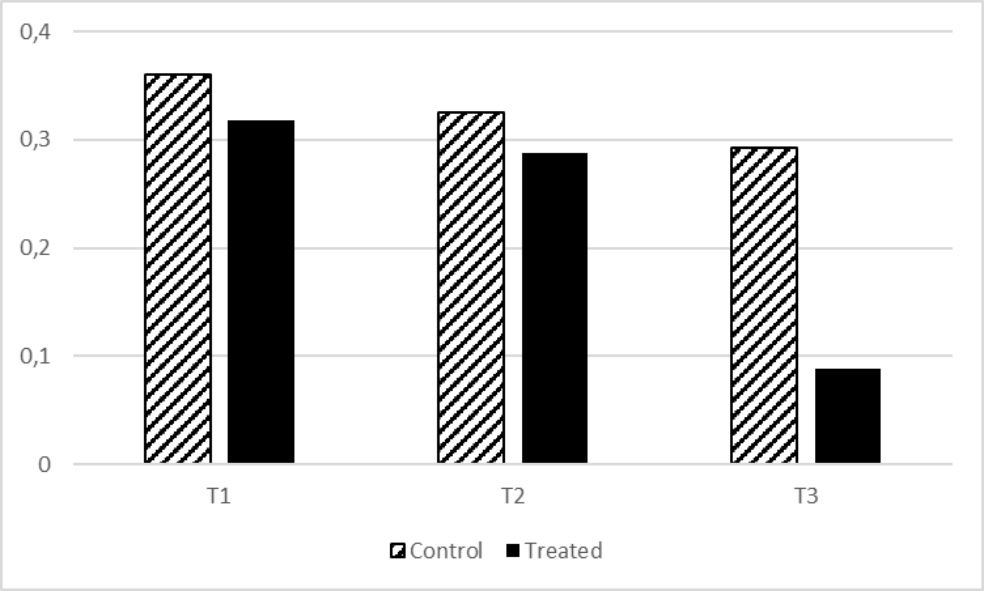
Figures

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 implemented the structured curriculum with the five-year-olds in their center. 2017/2018: Teachers in control received attended 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).

Figure 2: Closing the Learning Gap



Note: Gap in mean achievement score between centers of high and low quality by treatment status in T1 (baseline), T2 (post-intervention) and T3 (follow-up).

Table 1. Descriptive statistics and balance test for T2 and T3 sample.

	Post-intervention T2				Follow-up T3			
	Control	Treat	Difference	N	Control	Treat	Difference	N
T1 Sum score	0.022	-0.021	-0.043	650	0.028	-0.016	-0.045	661
	(0.058)	(0.053)	(0.095)		(0.059)	(0.052)	(0.097)	
T1 Executive functioning	-0.015	0.009	0.024	650	-0.022	0.019	0.042	661
	(0.059)	(0.053)	(0.096)		(0.058)	(0.052)	(0.096)	
T1 Language	0.058	-0.046	-0.104	650	0.076	-0.050	-0.128	661
	(0.059)	(0.052)	(0.096)		(0.060)	(0.050)	(0.097)	
T1 Math	0.010	-0.013	-0.024	650	0.014	-0.008	-0.024	661
	(0.056)	(0.055)	(0.094)		(0.055)	(0.053)	(0.093)	
Female	0.517	0.480	-0.037	650	0.515	0.475	-0.040	661
	(0.500)	(0.500)	(0.036)		(0.500)	(0.500)	(0.037)	
Birth month	6.878	6.807	-0.066	650	6.744	6.826	0.082	661
	(3.184)	(3.213)	(0.277)		(3.229)	(3.177)	(0.278)	
Mother education	14.377	14.161	-0.216	626	14.385	14.126	-0.259	637
	(2.590)	(2.587)	(0.279)		(2.547)	(2.569)	(0.273)	
Father education	13.814	13.696	-0.119	620	13.782	13.715	-0.067	628
	(2.563)	(2.504)	(0.281)		(2.536)	(2.469)	(0.282)	
Mother earnings	341,408	320,168	-21,240	648	339,897	322,267	-17,629	658
	(221,464)	(206,324)	(23,109)		(214,889)	(203,739)	(22,792)	
Father earnings	544,773	558,596	13,822	636	547,552	562,885	15,332	643
	(259,182)	(267,057)	(26,281)		(262,667)	(272,631)	(27,479)	
Non-western immigrant	0.130	0.201	0.071+	650	0.136	0.203	0.067+	661
	(0.337)	(0.401)	(0.036)		(0.343)	(0.403)	(0.037)	
Late consent	0.113	0.243	0.130**	650	0.116	0.252	0.137**	661
	(0.317)	(0.429)	(0.045)		(0.320)	(0.435)	(0.048)	
N	292	358	650		293	368	661	

Note: 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 child center level. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 2. 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	0.095*	0.126*	0.114*	-0.013	0.126*	0.230**	0.039	0.039
	(0.047)	(0.061)	(0.050)	(0.052)	(0.048)	(0.044)	(0.047)	(0.061)
N	652	650	652	648	661	661	660	659
Adj. R2	0.611	0.443	0.483	0.537	0.520	0.364	0.396	0.489
Model 2:								
Treat	0.100*	0.157*	0.100*	-0.017	0.111*	0.238**	0.047	-0.012
	(0.046)	(0.062)	(0.050)	(0.048)	(0.043)	(0.043)	(0.042)	(0.057)
N	652	650	652	648	661	661	660	659
Adj. R2	0.611	0.434	0.483	0.527	0.512	0.347	0.388	0.451
Model 3:								
Treat	0.107	0.183+	0.117	-0.043	0.096	0.223*	0.052	-0.044
	(0.101)	(0.101)	(0.095)	(0.095)	(0.096)	(0.091)	(0.092)	(0.096)
N	652	650	652	648	661	661	660	659
Adj. R2	0.001	0.007	0.002	-0.001	0.001	0.011	-0.001	-0.001

Note: 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, indicators for randomization block, gender, birth month, parental characteristics (mother and father's education level, earnings, and an indicator for non-western country of birth). In Model 2 we restrict controls to baseline test scores and indicators for randomization block. Model 3 has no controls. All regressions are clustered on child center level. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 3. 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	0.154+	0.252*	0.065	0.055	0.267**	0.374**	0.133*	0.145*
	(0.079)	(0.113)	(0.081)	(0.073)	(0.061)	(0.069)	(0.062)	(0.066)
Treat*High	-0.131	-0.228	0.070	-0.175+	-0.322**	-0.289**	-0.201	-0.292*
	(0.112)	(0.153)	(0.119)	(0.104)	(0.097)	(0.103)	(0.124)	(0.121)
N	638	636	638	634	648	648	647	646
Adj. R2	0.604	0.437	0.482	0.535	0.519	0.358	0.398	0.488
Panel B: Baseline skills								
Treat	0.064	0.137+	0.080	-0.094	0.139+	0.225**	-0.017	-0.005
	(0.072)	(0.080)	(0.081)	(0.070)	(0.081)	(0.071)	(0.080)	(0.078)
Treat*High	0.066	-0.034	0.057	0.193+	-0.028	0.002	0.107	0.093
	(0.113)	(0.108)	(0.120)	(0.111)	(0.114)	(0.118)	(0.128)	(0.136)
N	652	650	652	648	661	661	660	659
Adj. R2	0.614	0.442	0.484	0.540	0.522	0.363	0.397	0.488
Panel C: Parent education								
Treat	0.058	0.121	0.014	0.009	0.103	0.252**	-0.035	0.030
	(0.081)	(0.093)	(0.092)	(0.081)	(0.078)	(0.082)	(0.086)	(0.085)
Treat*High	0.061	-0.009	0.172	-0.031	0.042	-0.041	0.109	0.040
	(0.122)	(0.132)	(0.131)	(0.119)	(0.105)	(0.118)	(0.131)	(0.110)
N	641	639	641	637	649	649	648	647
Adj. R2	0.601	0.436	0.478	0.522	0.505	0.359	0.379	0.474

Note: 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 2: We add controls for gender, birth month and parental characteristics (education, earnings and indicator for non-western country of birth), baseline test scores and randomization block, all regressions clustered at the preschool level. Preschool center quality is measured as the child center fixed effect (center average covariate adjusted assessment 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 sample. High/low parental education and high/low baseline skills are split at median value. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Promoting Child Development in a Universal Preschool System: A Field Experiment

Online Appendix Materials

A1. Intervention

Preschool curriculum

Scientifically based: The preschool curriculum consists of age-appropriate skill-building activities in mathematics, language, executive functioning and social skills. The pedagogical principles for the curriculum builds on research related to teacher-child relationships (Pianta 1999) and to playful learning (Weisberg, Hirsh-Pasek, and Golinkoff 2013). For young children, learning takes place during play, both during free play and guided play. During guided play, teachers intentionally prepare and introduce books, games, activities or toys, and engage children in exploring specific content themes. Additionally, the activities were designed with inquiry as a pedagogical principle which emphasizes children's own explorations (Wells 1999).

The activities were inspired by existing curricula with promise for positive effects. A multi-disciplinary team of researchers (including researchers with many years of experience working as preschool teachers in Norway) studied existing curricula in detail. We selected curricula based on two criteria: a) Documented gain in executive functioning, language and mathematics development, and b) Promise for implementation in the Norwegian culture and preschool context, i.e. a playful learning approach (Weisberg, Hirsh-Pasek, and Golinkoff 2013). This narrowed down the list of curricula and pedagogical approaches we used as inspiration to the following: 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).

Based on the selected curricula, we drafted outlines for more than 130 learning activities, stimulating the targeted skill domains. The main objective was to provide material and inspiration for teachers to create a more intentional and structured practice. For each activity, we organized the text according to the following headings: Intention, Preparation, Implementation and Materials needed. Our goal was that the teachers would use the activities thoughtfully and intentionally, and not just as fun games. Still, all activities were designed to be engaging and meaningful for the children, stimulating them to be active and collaborative, in accordance with theory on playful learning.

Embedded in Norwegian culture and context: As part of the teacher training, the teachers were asked to try out the drafted activities with the current five-year-olds in their own preschool centers, and to give the researchers feedback and suggestions. Notably, teachers were instructed that this piloting should not be done with our treatment children. Throughout the year, each teacher had to provide us with oral and written feedback on multiple activities. At the end of the school year, we had critical and constructive feedback from many teachers on each of the activities. They also suggested alternative activities that we integrated into the curriculum. This resulted in a substantial revision embedding the playful learning activities to Norwegian culture and context.

Published as a Book: The curriculum was published as a book called "Lekbasert læring" in Norwegian (playful learning). The book is written in Norwegian to fit with the Norwegian preschool context. In addition to 130 learning activities, the book contains an introduction to

the theoretical and empirical evidence on which the curriculum builds. In particular, it emphasizes curricula that intentionally and systematically target skills in language, mathematics, executive functioning and social skills, as key ingredients in quality preschool. Moreover, it gives an introduction to the importance of positive and stimulating relationships between teachers and children, and the playful learning approach.

The book includes the following:

- Brief introduction to evidence base (importance of child-teacher relationship, the playful learning approach and important skills to stimulate in early childhood)
- Practical implementation guidelines, including templates to plan each day, week, and month.
- 130 playful learning activities stimulating skills in mathematics, language, executive functioning and social skills.
- Fall and spring curricula referring to relevant pages for all 130 activities in the book
- Templates for activities

Importantly, the curriculum was not a detailed program intended to dictate teacher practice on an everyday basis. We provided a suggestion for activities suited for each month (fall and spring plan). However, teachers were encouraged to develop their own unique approach to the curriculum and to augment it with their own ideas. The activities were flexible in terms of challenge and complexity, allowing teachers to match their practice to their different children's skill level. Teachers where encouraged to enhance children's learning opportunities by continuously giving them new challenges. The treated preschool centers signed a contract to spend at least eight hours a week engaging the five-year-olds in the curriculum, and split the time between activities stimulating skills in mathematics, language, executive functioning and social skills.

In the following, we will give detailed examples of specific learning activities in the curriculum.

Mathematics activities: The curriculum covers activities stimulating number and quantitative thinking, in addition to measurement, geometry and statistics. For example, within numeracy, children are engaged in the game *Marve Larve* (*Marve the Caterpillar*). In this activity, children make their own Caterpillar with beads and pearls on a string. They also make a "house" out of a matchbox labelled with the same numeral as well as the same number of dots as the number of beads. Furthermore, children compare the lengths of all the children's caterpillars (smaller, longer, smallest etc.). Then, the teacher mixes all the children's caterpillars and houses, and the children are asked to match them into the right houses. The teacher plays a game and lets the Caterpillar partly pop out of the house, and children and adults count the number of beads outside the house and try to figure out how many beads are still inside the house. This is an example of how a fun game in the book corresponds with quite advanced mathematics such as equations (e.g. $3 + x = 6$) introduced later in school. Another example is the *Geometric photo safari.* In this game*,* the group discusses various geometrical shapes and identify such shapes in the environment outside the preschool center. The children take photos of geometrical shapes and then discuss the result during circle time at the end of the session.

Language activities: Important for the language activities, is the theory section of the book that describes how language can be divided into three separate but overlapping components: content, form and use. In this way, teachers are conscious of different developmental areas within language. Furthermore, this section reviews the principles of interactive book reading (Mol, Bus, and de Jong 2009), including teacher preparation, pre-reading with children to

create engagement for the book, reading sessions, focus words, retelling and dialogue, and finally, book related activities such as drawing and drama. In these drama activities, children are given props related to the story that has been read so that they can dramatize the content and continue to practice new words and concepts. Language activities for children are related to either interactive book reading or other kinds of language games. For example, in a language game called *I am a letter!,* children are given one letter each on a piece of paper, and then the teacher challenges the them to form words with these letters by moving around so that the letters come in correct position. In another language game, each child is given a short word, and then the group is challenged to move around so that they form a sentence.

Activities related to executive functioning: The theoretical part of the book explains how self-regulation relies on underlying executive functioning processes (including attentional or cognitive flexibility, working memory, and inhibitory control) (Blair and Raver 2015). In accordance with this, the activities involve children's ability to use attentional flexibility, working memory, and inhibitory control. For example, in the activity *The bear is sleeping, the snail is sleeping* teachers rewrite a traditional Norwegian nursery rhyme so that not only does the bear fall asleep and wake up, other animals fall asleep and wake up again too. The children pay attention and listen carefully to the song and to what kind of animals appear as the teacher sings. During the song, they dramatize the sleeping and awakening of these different animals with different speed and movements. In this game, children use attention, working memory and inhibition to be able to follow the song and mime the animals and their movements. In another activity called *The ready, steady, go game* children are to run on "ready, steady, go!" and to inhibit their impulse to run on other instructions such as "ready, steady, gorilla!".

Activities related to social competence: The book explains how social competence relies on skills such as self-control, assertiveness, responsibility, co-operation and empathy (emotional competence), and games are included that stimulate these skills. For example, in the activity called *Mailing a hug*, children think of a relative or a friend that might enjoy encouragement, and write or draw a message that can encourage the receiver. Through this activity, the children imagine the experiences and emotions of another person. Likewise, children identify and express their own emotions through drawing and drama activities. In an activity called *The gingerbread man* children express their emotions through colors within the outlines of a gingerbread man and talk about these emotions and how they feel inside.

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. The class provided the preschool teachers with key insights from the theoretical and empirical research literature on which the curriculum builds. Importantly, the class was practice-oriented. The class consisted of four two-day lecture gatherings over a period of eight months. Between class gatherings teachers practiced playful learning activities with the current five-year-olds in their preschool center (not the children in our study), and reported on feasibility and reflected on how their experiences aligned with the theoretical and empirical literature covered in class.

Since our baseline assessments with teachers told us that they spent much less time on playful learning within mathematics compared to other skill domains, we chose to give more attention to mathematics in the teacher training. More precisely almost 40 percent of teaching hours were spent on mathematics.

Teachers were spread across a large region in the southern part of Norway called Agder. In order to make the class feasible for all teachers, we arranged all lecture gatherings twice, once

in Eastern Agder and once in Western Agder. Fulfillment of class participation, practice and assignments gave 15 credit points in the Norwegian university system, in which full time students are supposed to complete 60 credit points a year. All the preschool teachers passed the class. Absence from sessions or classes was low, with less than 10 full day absences due to health issues across all teachers.

In addition to the credit base university class, teachers were coached during the intervention year in two gatherings and phone meetings. In the two gatherings (September 2016 and March 2017), teachers reviewed insights from research on systematic curricular focus, and were challenged to reflect in groups on how their curriculum implementation matched with the intentions in the project. They were also asked to reflect on challenges and successes. Thereafter, the teachers shared ideas and experiences with the entire class, and the instructors participated in the discussion by listening to the preschool teachers concerns and guiding them on how to address these concerns.

Additionally, the teachers had scheduled phone meetings with their coach two times each semester, and could schedule additional phone meetings to discuss any immediate questions or concerns. One of the scheduled phone meetings per semester was dedicated to mathematics and the other to language, executive functioning, social skills and pedagogical approaches. Again, mathematics was emphasized, since we knew from our baseline assessments with teachers that mathematics was less emphasized in preschool centers. The scheduled phone meetings were conducted as semi-structured interviews. For example, in the first meeting we asked: 1. a very broad opening question allowing them to come up with whatever they felt relevant; 2. whether their center administration gave practical support and facilitated for the group; 3. their experiences with guiding the assistants; 4. whether they believed they succeeded in building trusting relationship with children; 5. whether activities met criteria for playful learning, and finally; 6. experiences from daily activities. Throughout the conversations, we listened to their concerns and guided them on how to address these concerns. All treated teachers participated in the scheduled phone meetings.

Typically, each preschool teacher had one or two assistants when implementing the curriculum, depending on the size of the child group. The trained preschool teacher had the main responsibility to train the assistants. However, assistants also received a one-day training in the intervention material and research project in general (1/2 day), and more specifically on interactive book reading (1/2 day). This is because 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, with the assistant in charge of the language activities.

A2: Assessment, Measures and Control Variables

Assessment:

We assessed the children in August 2016 (baseline, T1), June 2017 (post-intervention, T2), and March 2018 (follow-up, T3). 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. All testers had to hand in a police certificate stating that they had no record of offences that would make them unsuitable for working with children. The testers used computer tablet instruments with a validated test battery developed for transition between preschool and school.

Tester training consisted of one full day of theory and practice related to our computer tablet test battery. Testers where then instructed to visit pilot preschool centers (T1 and T2) or

schools (T3) to practice the test on the computer tablets with children in relevant age group. A week later, the testers came back to discuss their experiences and to take the certification. The certification involved conducting all tests in the battery while one of the researchers made systematic notes according to a certification form. For minor mistakes, testers got reminders and feedback and subsequently received their certification.

The T1 and T2 assessments were conducted at three central locations, including a large Science museum. For the other two locations, we hired personnel from the science museum to come and arrange an activity day for the children. All children in the participating preschool centers in the Agder project were invited to these activity days with assessments, and all preschool centers participated. Children were allowed to use the facilities and activities provided by the science museum for the full day. At a scheduled time the preschool centers met for assessment. Children's names were replaced with personal codes on stickers attached to their clothes, and the children were then guided to individual test stations for assessment. This way each preschool center was exposed to many different testers, which made it possible to detangle center effects from tester effects in analyses. In T3 (spring of 1st grade), 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. All children received a small gift for their participation (e.g. a ruler, a gym bag or a pencil case with illustrations from the computer tablet assessment printed on them).

In all assessments, we ended the assessment if children expressed that they could not manage more. If the assessment was ended in the middle of a test, the remaining items in this test were coded as incorrect. If a test was never started (for example if the child was upset), we coded it as missing. Consequently, the number of children with registered test scores on the different measures varies slightly within T2 and T3.

Measures:

We assessed skills in mathematics, language, working memory and inhibitory control. The latter two are important components of executive functioning. Unfortunately, we did not measure social competence in this study due to lack of tests validated in a Norwegian context.

Mathematics skills were assessed with the *Ani Banani Math Test* (ABMT; Størksen and Mosvold 2013). The ABMT is a playful mathematics assessment on a tablet application, which includes items covering three areas of mathematics – numeracy, geometry and problem solving. Children help a monkey with different tasks, such as counting bananas and setting the table with enough plates for birthday party guests. All correct answers were given one point. Due to technical problems with the tablet application at T3, data for 5 out of the 18 items of the ABMT was not recorded correctly and therefore omitted in the analyses for all assessment periods. We calculate the total score as the mean across the 13 items. This short version correlates strongly (r = .58) with the Preschool Early Numeracy Skills test (PENS) in kindergarten and significantly predicts mathematic achievement in 1st (r = .529) and 5th grade (r = .553). Internal consistency was considered adequate (Cronbach's alpha = .60).

Two assessments were conducted to measure language; one pertaining to vocabulary and the other to phonological awareness. Vocabulary was assessed with the *Norwegian Vocabulary Test* (NVT; Størksen et al. 2013). The NVT is a typical expressive vocabulary task including 20 words. Illustrations appeared on a tablet screen and the child was subsequently asked to name it. Cronbach's alpha was high; α = .81. Children's phonological awareness was assessed with a 12-item *blending task* that is part of the official literacy screening battery from The Norwegian Directorate for Education and Training. For each task, a target word was presented

in its individual phonemes by the experimenter and children had to indicate the corresponding alternative from four presented images on a tablet screen. All correct answers were given one point. For both tests, the total score was calculated as the sum across all items.

Three assessments were conducted to measure executive functioning. The *Head-Toes-Knees-Shoulders task* (HTKS; McClelland et al. 2014) integrates attention, inhibitory control, body control and working memory demands into a short task of behavioral self-regulation appropriate for children aged 4 to 8 years. It has strong reliability and validity, and is significantly related to other measures of self-regulation and to children's academic outcomes (26). These results have been replicated in many recent studies across the world. The task includes three blocks with 10 items each. Responses were scored with two points when correct, one point when the child made an incorrect movement but ended up with the correct response, and zero points for incorrect responses. Cronbach's alpha was sufficient; $\alpha = .76$. In the *Hearts and Flowers task* (Davidson et al. 2006), children had to press a key on the same side of the stimulus when they saw a heart and on the opposite side when the stimulus was a flower. The task has 57 items and number of correct responses were counted. The measure is designed to assess inhibitory control and cognitive flexibility skills and has been widely used with young children. Cronbach's alpha was $\alpha = .89$. For both tests, the total score was calculated as the sum across all items. The third assessment of executive functioning was the *Forward/Backward Digit Span* subtest from the Wechsler Intelligence Scales for children-III (Wechsler, 1991), which measures working memory. Digits were read aloud, one digit per second, and the children were asked to repeat the sequence of digits. First, they had to repeat digit sequences in the same order as they heard them, and then in reversed order. The number of digits in each sequence increased as the test continued. The test was automatically discontinued after two subsequent errors. The total score reflects the highest number of repeated digits forward plus the highest number of repeated digits backwards.

From these six tests conducted at each assessment, we created three outcome measures:

- Math: Percent correct answers at the ABMT test.
- Executive Functioning: Mean score of the standardized "HTKS", "Hearts and Flowers" and "Digit Span" tests. If missing on one of the tests, the other tests constitute Executive Functioning.
- Language: Mean score of the standardized "Phonological Awareness" and NVT tests. If missing on one of the tests, the other test alone constitutes Language.

All three outcome measures are standardized within each period to mean 0 and standard deviation 1.

Table A1 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 (coefficients in italic). 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 particularly predictive of all T2 and T3 measures.

Table A1. 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.

Control variables:

From registry data we constructed the following control variables that entered into our analyses:

- Gender: Indicator for female
- Birth month: Continuous variable running from 1 (December born) to 12 (January born)
- Mother's and father's education: Continuous variable for number of years education, running from 10 (compulsory schooling) to 18 (Master's degree). We standardized the mean of mother's and father's (standardized) education.
- Mother's and father's earnings: Income from work (employment and self-employment). In categories of 50.000 NOK, but recoded as a continuous variable for the analyses.
- Immigrant status: Indicator for whether one or both parents are immigrants from a non-western country.

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. Finally, we also constructed indicators for late consent (parental consent received after the preschool center's treatment status was known); for randomization block; and for tester ID.

A3. Compliance

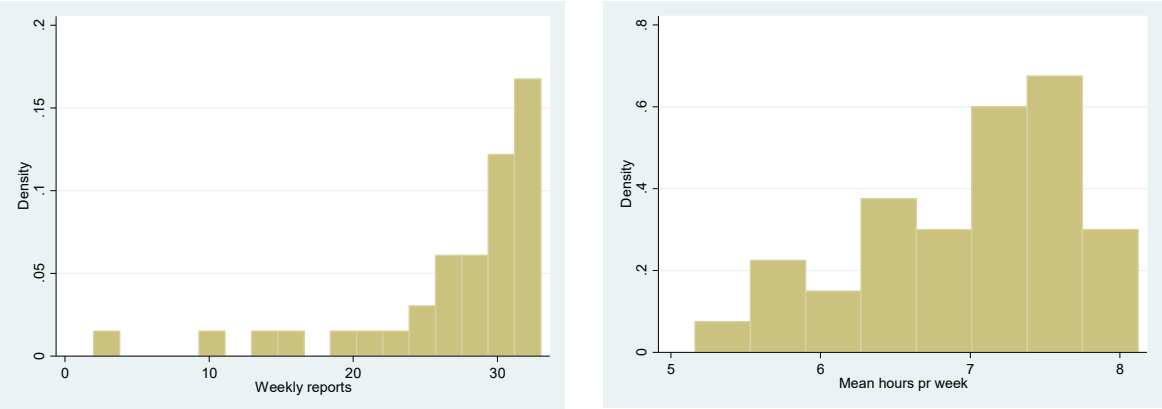
We assessed treatment compliance in a brief weekly electronic questionnaire where teachers reported on fidelity of implementation, including how many hours spent implementing the learning activities. The teachers were told to spend eight hours per week on the activities in mathematics, language, executive functioning and social skills. We requested response to the questionnaire for a total of 34 weeks, which excluded five vacation weeks during fall, Christmas, winter and Easter. The first questionnaire was in early September 2016, and the

last in June 2017. Figure A1 (left) reports number of responses submitted by each preschool center. As can be seen, the majority of centers submitted reports in most of the 34 weeks.

In the weekly reports, teachers were asked how many hours were spent on the learning activities during the previous week. Among the 974 weekly reports submitted, 67 percent reported spending eight or more hours on the learning activities the previous week, and only in 16 percent of the reports, the teacher reported they spent less than six hours on learning activities. In the open comment field that was included in the questionnaire teachers gave reasons for not complying with the eight hours they had committed to. Reasons typically included teacher absence due to health issues or other practical issues that prevented them from following their plans.

Finally, Figure A1 (right) shows the distribution of average number of hours per week across the year spent on learnings activities for each preschool center. We find that 60 percent of all centers spent at least 7 hours per week on the learning activities.

Figure A1: Distributions of number of weekly reports (left) and average weekly hours spent on learnings activities (right) for each preschool center. N = 36 centers.



Additionally, we assessed the teachers' perceived relevance, importance and benefit of the intervention in anonymous evaluations of the credit base university class (spring 2016, response rate 85 percent) and the curriculum (spring 2017, response rate 80 percent). All teachers agreed that they found the material and activities in the class relevant for their work as practitioners. Their overall rating of the course was on average 4.9 on a five-point scale. When assessing the curriculum, all teachers agreed to the statements "The children have enjoyed working with the learning activities" and "The children have learned a lot from working with the learning activities." All but one teacher agreed that they would continue to use the curriculum with the five-year-olds in the next preschool year.

A4. Robustness

In Table A2 we report our main results when children with late consent is excluded and in Table A3 we report our main results for each of the six assessment tests.

Table A2. Treatment effect on test scores at post-intervention (T2) and in the one year follow-up (T3). Sample excluding children with late consent.

	Post-intervention (T2)				Follow-up (T3)			
	Sum score	Math	EF	Language	Sum score	Math	EF	Language
Treat	0.053	0.077	0.063	-0.017	0.111*	0.212**	0.037	0.022
	(0.049)	(0.060)	(0.054)	(0.057)	(0.052)	(0.053)	(0.060)	(0.056)
N	532	530	532	528	534	534	533	532
Adj. R2	0.596	0.412	0.501	0.488	0.511	0.361	0.382	0.467

Note: Each column presents regression coefficient of treated (standard error) using ordinary least squares. For both assessment periods: We control 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 randomization block. All regressions are clustered on child center level. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table A3: Treatment effect on all six assessment scores at post-intervention (T2) and in the one year follow-up (T3).

	Post-intervention (T2)						Follow-up (T3)					
	Math	HTKS	Digit Span	Hearts & Flowers	Vocabulary	Phon. awareness	Math	HTKS	Digit Span	Hearts & Flowers	Vocabulary	Phon. awareness
Panel A: Full sample												
Treat	0.126*	0.041	0.188**	0.072	-0.050	0.038	0.230**	0.018	0.021	0.058	0.037	0.028
	(0.061)	(0.060)	(0.063)	(0.057)	(0.039)	(0.072)	(0.044)	(0.066)	(0.056)	(0.054)	(0.041)	(0.078)
N	650	645	641	635	648	645	661	659	653	660	659	658
Adj. R2	0.443	0.282	0.394	0.387	0.643	0.203	0.364	0.230	0.291	0.260	0.608	0.182
Panel B: Preschool center quality												
Treat	0.252*	-0.026	0.185+	0.048	0.017	0.084	0.374**	-0.046	0.179*	0.179*	0.120+	0.118
	(0.113)	(0.089)	(0.096)	(0.099)	(0.051)	(0.103)	(0.069)	(0.081)	(0.082)	(0.081)	(0.063)	(0.086)
Treat*	-0.228	0.109	-0.038	0.041	-0.156+	-0.128	-0.289**	0.114	-0.269+	-0.294*	-0.210*	-0.269
high_quality	(0.153)	(0.124)	(0.174)	(0.153)	(0.081)	(0.161)	(0.103)	(0.138)	(0.147)	(0.138)	(0.092)	(0.165)
N	636	631	627	621	634	631	648	646	640	647	646	645
Adj. R2	0.437	0.285	0.389	0.384	0.637	0.206	0.358	0.227	0.300	0.259	0.606	0.181
Panel C: Parent education												
Treat	0.121	-0.055	0.171+	-0.017	-0.093	0.117	0.252**	0.018	-0.026	-0.076	0.054	-0.002
	(0.093)	(0.098)	(0.096)	(0.091)	(0.062)	(0.111)	(0.082)	(0.104)	(0.097)	(0.097)	(0.062)	(0.108)
Treat*	-0.009	0.151	0.044	0.139	0.078	-0.127	-0.041	-0.029	0.049	0.244	-0.024	0.090
high_education	(0.132)	(0.146)	(0.116)	(0.136)	(0.093)	(0.170)	(0.118)	(0.151)	(0.150)	(0.155)	(0.091)	(0.133)
N	639	634	630	624	637	634	649	647	642	648	647	646
Adj. R2	0.436	0.275	0.392	0.378	0.628	0.205	0.359	0.211	0.270	0.250	0.587	0.180
Panel D: Baseline skills												
Treat	0.133	0.166+	0.171+	0.013	-0.039	-0.144*	0.221**	-0.010	0.101	-0.057	0.017	-0.037
	(0.081)	(0.098)	(0.093)	(0.087)	(0.080)	(0.064)	(0.073)	(0.101)	(0.073)	(0.103)	(0.098)	(0.069)
Treat*	-0.022	-0.256+	0.042	0.117	0.173	0.197*	0.015	0.057	-0.146	0.194	0.012	0.159
high_pre skills	(0.110)	(0.149)	(0.122)	(0.133)	(0.125)	(0.095)	(0.121)	(0.134)	(0.142)	(0.145)	(0.147)	(0.106)
N	650	645	641	635	645	648	661	659	653	660	658	659
Adj. R2	0.441	0.284	0.405	0.386	0.204	0.644	0.363	0.228	0.302	0.262	0.180	0.609

Note: Each column in each panel presents regression coefficients with standard errors in parenthesis, using ordinary least squares. The model controls for gender, birth month and parental characteristics (education, earnings and indicator for non-western country of birth), baseline test scores and randomization block, all regressions clustered at the preschool level. Preschool center quality is measured as the child center fixed effect (center average covariate adjusted assessment 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. High/low parental education and high/low baseline skills are split at median value. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.