Labour Economics 55 (2018) 230–240

Contents lists available at ScienceDirect

Labour Economics

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

Variation in center quality in a universal publicly subsidized and regulated childcare system

Mari Rege[a], Ingeborg Foldøy Solli[b,*], Ingunn Størksen[c], Mark Votruba[d]

[a] University of Stavanger, Norway
[b] University of Stavanger, Norway
[c] University of Stavanger, Norway
[d] Case Western Reserve University, United States

ARTICLE INFO

Keywords:
Child development
Childcare quality
Early childhood education and care
JEL codes: I20
I24
J13

ABSTRACT

A large literature suggests that high quality childcare programs can produce positive and lasting effects by promoting math, language and social-emotional skills, referred to as school readiness skills, especially for children of parents with low education. Hence, a universal childcare system with easy access has the potential to make a substantial difference in children's lives and reduce socio-economic disparities in educational outcomes. However, if childcare quality varies across centers, universal childcare systems can also potentially increase disparities in school readiness if the children of more highly-educated parents select into centers of higher quality. Using a unique dataset with one-to-one assessments of school readiness skills among 627 five-year-olds attending 67 different childcare centers, we investigate differences in childcare quality by testing whether covariate adjusted assessments scores are clustered by center. Through fixed effect and random effect analyses, we demonstrate significant variation in school readiness across centers. However, selection into centers of different quality appears to be limited.

1. Introduction

There is a large literature linking early childhood language, math and socio-emotional skills as foundational for future learning and development (Duncan et al., 2007; Hall et al., 2016; Rabiner et al., 2016). Moreover, studies show that high quality childcare programs stimulating these school readiness skills can produce positive and lasting effects by promoting school success and fostering workforce productivity (Gupta and Simonsen, 2016; Havnes and Mogstad, 2009; Heckman and Kautz, 2013; Melhuish, 2011; Reynolds et al., 2011; Weiland and Yoshikawa, 2013). This is especially true for children of parents with low education.

Based on the evidence from the early-childhood-education-and-care (ECEC) literature, publicly subsidized childcare with easy access for low-income children is often accentuated as a key policy to provide children more equal opportunities. Norway is frequently considered a frontrunner in this respect, with publicly subsidized universal childcare for all children ages one-to-five and free access for children of families with low income (e.g. Bennett and Tayler, 2006). Indeed, Norway is among the OECD countries with the highest public spending on early childhood education and care (Engel et al., 2015). However, the Norwegian childcare system has also been criticized because the regulatory standards for structural and process quality are lenient and imprecise (Bennett and Tayler, 2006; Engel et al., 2015), which could contribute to large quality differences across childcare centers. Such variation is concerning since it suggests missed opportunities to improve school readiness among children attending lower quality centers. Moreover, if childcare quality varies across centers, universal childcare systems could increase disparities in school readiness if the children of parents with high education select into centers of higher quality.

This paper investigates differences in childcare quality across centers in the universal childcare system in Norway. We utilize a unique dataset collected in the Agder-project[1] with one-to-one assessments of literacy, math and self-regulation of 627 five-year-olds in 67 different childcare centers in Norway. The assessment data is matched with indicators for childcare center and registry data on parental education,

[1] The Agder-project is a randomized field experiment investigating effects of a preschool intervention for five-year-olds at Norwegian daycare centers. Treated childcare centers receive teacher education, a focused curriculum with concrete examples of playful learning activities cultivating school readiness skills, and resources providing teachers with time for engaging the children in the playful learning activities. We use the pre-intervention data wave of the Agder-project in this paper.

* Corresponding author.
E-mail addresses: mari.rege@uis.no (M. Rege), ingeborg.f.solli@uis.no (I.F. Solli), ingunn.storksen@uis.no (I. Størksen), mxv27@case.edu (M. Votruba).

https://doi.org/10.1016/j.labeco.2018.10.003
Received 1 November 2017; Received in revised form 28 September 2018; Accepted 12 October 2018
Available online 12 October 2018

earnings and immigrant status from Statistics Norway. Our empirical analyses test whether covariate-adjusted assessment scores vary across childcare centers by more than would be predicted by random variation in children's skills. Our fixed effect and random effect analyses demonstrate significant variation in school readiness skills at age five. The difference in school readiness skills in centers at the 90th and 10th percentile in the center effect distribution is estimated to be over one half (0.55) of a standard deviation. Thus, the differences in school readiness skills (not accounted for by covariates) across childcare centers in Norway appears to be quite substantial. Differential selection of children into higher/lower quality centers on the basis of socio-economic background appears quite limited. However, we cannot rule out that the differences in center effects capture unobserved similarities in children's prior skills and parental background.

Notably, the variance in school readiness scores attributable to centers captures the combined effect of teacher quality, peers, childcare pedagogy, in addition to structural quality such as child-teacher ratios. Moreover, children in the same center may also be co-located residentially, and may experience common shocks affecting child development. A limitation of our study is that the mechanism(s) behind the variation across centers can be explored only partially. We extend our analyses by investigating how measures of structural quality – the child-teacher ratio, center size and the tenure of the director – predict the development of school readiness skills. This evidence should be interpreted with caution as the structural indicators are not random and may be associated with other unobserved indicators for quality. Nevertheless, we find that the teacher-child ratio is associated with a large and significant increase in school readiness skills, and this single characteristic explains a meaningful portion (about 30 percent) of the variance in school readiness across centers. The other structural characteristics of centers fail to significantly predict school readiness scores.

Our study relates to the large literature in economics investigating quality differences across classrooms and teachers in schools (e.g. Chetty et al., 2011; Rivkin et al., 2005; Rockoff, 2004). This literature has demonstrated large variation across classrooms and teachers in children's learning, which cannot be explained by observables such as class size, child-teacher ratios, and teacher experience and education. As we do not observe pre-measures of children's skills, our empirical approach is less rigorous than the value-added approaches in the above-cited studies. Nevertheless, we are not familiar with any other paper investigating quality differences in learning across centers in a universal childcare system. Moreover, our rich dataset allows us to do a rigorous selection analysis, which suggests the center effects are not driven by selection.

Additionally, our analyses contribute to the emerging economic literature investigating how childcare structural quality indicators affect child development (e.g. Bauchmüller et al., 2014; Blau, 1999; Currie and Neidell, 2007; Drange and Rønning, 2017). The literature provides mixed evidence of the effect of structural parameters, and more research is needed. In general, the evidence from these studies seems to mimic evidence from similar studies in schools (see review in Jackson et al., 2014), which suggest limited potential for improving child development by merely investing in structural characteristics. Nevertheless, consistent with our findings, Bauchmüller et al. (2014) and Currie and Neidell (2007) demonstrate positive associations between teacher-child ratios and child outcomes. In contrast, Blau (1999) find no significant associations between these measures.

Our paper proceeds as follows. In Section 2, we provide some background information on key school readiness skills, and existing evidence on how childcare centers can promote these skills. In Section 3, we describe the Norwegian childcare system and context. In Section 4, we present our measures of key school readiness skills, procedures for data collection and sample. In Section 5, we present our empirical analyses. Finally, in Section 6 we conclude by discussing the results.

2. Childcare quality and key school readiness skills

Childcare quality is often defined and measured according to two basic aspects —*structural quality* and *process quality* (OECD, 2015). Structural quality includes easily observable aspects, such as teacher education, child-staff ratios, teacher and management experience, and class size. Process quality represents the direct experiences of children, and include factors such as the sensitivity and responsiveness of caregivers to individual needs, the pedagogical approaches and materials available for learning, and interactions with teachers and peers. Evidence suggest that increasing structural quality in isolation has a modest or no effect on children's development (Bauchmüller et al., 2014; Blau, 1999). However, increased structural quality is often needed to improve process quality, which several studies document as crucial for early childhood learning, development, and well-being (Hamre 2014; Lee and Bierman, 2015; Sabol et al., 2013).

An important aspect of process quality is compassionate and systematic cultivation of language, math and socio-emotional skills. The ECEC literature demonstrates that these skills are critical for school success and malleable in early childhood (Jensen et al., 2017; Melhuish, 2016; Tominey and McClelland, 2011; Yoshikawa et al., 2013). Below we summarize the evidence on these key school readiness skills:

2.1. Language

Young children with advanced language skills are more likely to be prepared for formal schooling compared to children with weak language skills (Duncan et al., 2007; McGinty and Justice, 2010). In particular, vocabulary at school entry explains a substantial portion of the variance in children's early literacy, including growth in literacy (Lervåg and Aukrust, 2010). Early vocabulary also significantly predicts future reading comprehension. Research indicates that interactive book reading, when combined with professional development activities designed to improve the quality of teacher's language, can be effective in promoting richer conversational exchanges in the classroom and gains in child vocabulary and oral comprehension skills (Bierman et al., 2008). Furthermore, exploring rhymes, letters and sounds at an early age is generally seen as an important pre-reading activity (Muter et al., 2004).

2.2. Math

The importance of early math skills is increasingly evident. A widely cited U.S. study of children's grade-school achievement used data from six different large-scale surveys and showed that the strongest predictors of later achievement were school-entry mathematical, attention (an aspect of self-regulation), and reading skills (Duncan et al., 2007). Among the possible predictors, mathematical skills seemed to be the most prominent. These findings have since been replicated using Canadian data (Romano et al., 2010). The quantitative, spatial, and logical reasoning competencies of early mathematics may form a cognitive foundation for thinking and learning across subjects (Clements and Sarama, 2011). Randomized controlled trials have demonstrated that interventions designed to facilitate children's mathematical learning during ages 3 to 5 years can have a strong effect on children's mathematical achievement (Clements and Sarama, 2011).

2.3. Socio-emotional skills

Research consistently confirms the importance of children's early abilities to control their behavior and use their cognitive abilities to successfully navigate school settings (Heckman and Kautz, 2013). Self-regulation can be defined as the ability to control thoughts, feelings and behaviors in order to adapt effectively with demands and social standards and expectations in the environment, and to be able to reach future goals (Berger, 2011). Preschool self-regulation has been linked to emergent literacy, vocabulary, and math skills (McClelland et al.,

2007) and its importance has been demonstrated in various cultures (von Suchodoletz et al., 2013; Wanless et al., 2011a). Several early childhood interventions have demonstrated significant impacts on self-regulation (Diamond and Lee, 2011; Merritt et al., 2012; Raver et al., 2011; Schmitt et al., 2018; Tominey and McClelland, 2011, 2013), and it seems like teacher sensitivity (high process quality) is an essential aspect for children's development of self-regulation (Blair and Raver, 2015).

In addition to self-regulation, children's emotion comprehension or understanding about their own emotions and the emotions of others helps lay the foundation for strong socio-emotional skills (Denham & Brown, 2010). Children's early relationships with others are significantly related to future school success. For example, children who are rejected and isolated are at risk of future school avoidance (Buhs et al., 2006). Importantly, emotion comprehension and relationship skills can be enhanced in early childhood and promote lasting effects (Nix et al., 2016). Unfortunately, we do not measure emotion comprehension and relationship skills in this study due to lack of tests validated in a Norwegian context. Still, even if these skills are hard to measure, they belong in a list of key school readiness skills.

3. The Norwegian context, childcare system and hypotheses

Norway has a strong welfare state with many family policies facilitating both child well-being and a strong labor market attachment for parents of young children. In association with childbirth or adoption, parents have the right to 11 months parental leave with full wage compensation[2] and job security. All children ages one to five years old have the right to publicly regulated and subsidized childcare. The utilization of the childcare system is very high, with an uptake rate of 97 percent among five-year-olds. Children in Norway start primary school in August the year they turn six. All children are obliged to attend primary school in Norway, and most children go to public school; only 3.6 percent go to private schools.

The Norwegian childcare system has been criticized because it gives childcare centers a large degree of freedom with respect to pedagogical content, which can give rise to large differences in process quality across centers (Bennett and Tayler, 2006). The system was originally established as a response to a need for high quality childcare as mothers entered the labor market. Even if the educational and developmental purpose is now prevailing, the program is still dominated by the social pedagogical tradition, which has a limited curricular focus. The program's pedagogy builds on a belief that preschool age children have their best learning experiences through free play and activities that build on the preschoolers own initiatives. The centers' pedagogical content is regulated by the National Framework Plan for Content and Tasks of Kindergartens (Ministry of Education and Research, 2011)[3] which defines seven learning areas: 1. Communication, language and text; 2. Body, movement and health; 3. Art, culture and creativity; 4. Nature, environment and technology; 5. Ethics, religion and philosophy, 6. Local community and society; 7. Numbers, spaces and shapes. However, these learning areas are only loosely described and are the same for all children ages 1–5. Moreover, there are no specific guidelines for how the childcare centers should implement the learning areas and teachers have no benchmarks for children.

Additionally, the Norwegian childcare system has been criticized because the standards for structural quality are lenient (Engel et al., 2015), allowing variation which might also facilitate quality differences across childcare centers. In Norway an ECEC teacher has a bachelor degree in early childhood education. The adult-child ratio is regulated so that the youngest children have at least one ECEC teacher per 7 – 9 children, whilst the older children have at least one ECEC teacher per 14 – 18 children. However, centers often apply for exemptions because of a shortage of qualified personnel. Moreover, there are no mandatory continuing education programs for childcare teachers. In addition to the childcare teacher, each child group has two assistants. However, there are no formal qualification requirements for theses assistants – it is not even required that they have completed high school.

The imprecise and lenient standards for process and structural quality in the Norwegian childcare system, discussed above, lead us to hypothesize:

Hypothesis 1. There is significant variation in center quality measured by children's school readiness skills.

If there are significant differences in learning opportunities across centers, we might expect children of parents with low education to be especially vulnerable to these differences. Several studies demonstrate that parents' socio-economic background is a strong predictor of early childhood development. This is also true in Norway despite a generous welfare system and limited child poverty (Bøe et al., 2016; Schjølberg et al., 2008; Størksen et al., 2013; Størksen et al., 2015; Størksen and Mosvold, 2013). Several studies suggest that the gaps in skill development across family background can partially be explained by parents' ability to create a home environment that stimulates learning and development (Guryan et al., 2008; Harris et al., 1999; Kalil et al., 2012). Moreover, research suggests that the positive impact of quality childcare is most pronounced for low-income children, who generally have few alternative learning opportunities (Dearing and McCartney). As such, we hypothesize:

Hypothesis 2. The relative advantage in school readiness skills for children from higher SES families will decrease for children in higher quality centers.

4. Assessments, data collection and sample

In August 2016, the Agder-project assessed school readiness skills by inviting all the five-year-olds attending 71 childcare centers in southern Norway to local science museums. The children had access to all the activities at the museum. At a scheduled time, each childcare center brought their children to an assessment station, where each child completed playful tasks on computer tablets in a room with a trained and certified tester. Assessments took approximately 40 minutes for each child.

The testers used computer tablet instruments developed for the Skoleklar-project (Størksen et al., 2013; Størksen and Mosvold, 2013) and further refined for the Agder-project. The tablets were loaded with a specially-designed application containing a battery of six tests designed to assess math, literacy and self-regulation skills, which are considered critical for successful school adjustment (see Section 2).

Math skills were assessed via the Ani Banani Math Test (ABMT) (Størksen and Mosvold, 2013). The ABMT is an 18-item digital math assessment on a tablet application, which includes items covering three areas of mathematics – numeracy, geometry and problem solving. Children help the monkey on the screen with different tasks, such as counting bananas and setting the table with enough plates for the guests in a birthday party. The measure correlates strongly with an existing school based math assessment $r = 0.69$ (Utdanningsdirektoratet, 2017) and with another validated early numeracy task $r = 0.74$ (*Number Sense Task*) (Van Luit and Van de Rijt, 2009). Cronbach's alpha in this study is $\alpha = 0.72$.

Two assessments were conducted to measure literacy, one pertaining to vocabulary and the other to phonological awareness. Vocabulary was tested with the Norwegian Vocabulary Test (NVT) (Størksen et al.,

[2] The compensation from the government has a ceiling. However, this ceiling is set above average earnings and for most employees with earnings above the ceiling, the employer will compensate for the differential.

[3] The Framework Plan was revised in 2017. However, we describe the Framework Plan of 2011, as this is the relevant plan for the children in our sample. It is worth noting that the key concerns with the Plan of 2011 remains in the plan of 2017: The learning areas are only vaguely described and are the same for all children ages one-to-five. There is no detailed age appropriate curricular focus.

2013). The NVT is a "naming test" (total 20 words) where an illustration appears on the computer tablet screen and the child is subsequently asked to name it. Cronbach's alpha in this study is $\alpha = 0.81$. Assessments of children's phonological awareness were constructed from the official literacy screening battery from The Norwegian Directorate for Education and Training. The measure consists of a 12-item blending task. For each task, a target word is 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 and summed up, resulting in scores ranging from 0–12. Notably, the test is difficult for the children at this age, and we see considerable floor effects. This was expected because most childcare centers do not work to stimulate this skill, since it is not emphasized in the Framework plan. Nevertheless, it was included among the pre-intervention measures of school readiness captured by the Agder-project.

Three assessments were conducted to measure children's self-regulation skills. The Head-Toes-Knees-Shoulders task (HTKS) integrates *attention, inhibitory control, body control and working memory* demands into a short task of behavioral self-regulation appropriate for children aged 4–8 years (McClelland et al., 2014). It has strong reliability and validity, is significantly related to other measures of self-regulation, and to children's academic outcomes in diverse samples (Cameron Ponitz et al., 2009; Fuhs et al., 2014; McClelland et al., 2014; Wanless et al., 2011b) including Norwegian children (Storksen et al., 2015). Cronbach's alpha in this study $\alpha = 0.76$. Second, in the Hearts and Flowers task (Davidson et al., 2006), children have to respond by pressing a key on the same side of the stimulus when they see a heart and by pressing a key on the opposite side when the stimulus is a flower. The measure has strong reliability and validity (Davidson et al., 2006). This test provides a more narrow measure of cognitive control, as opposed to the behavioral self-regulation and working memory skills also at work in the HTKS. The third self-regulation assessment uses the Forward/Backward Digit Span subtests from the Wechsler Intelligence Scales for children-III (Wechsler, 1991). Digits were read aloud, one digit per second, and the children were asked to repeat the sequence of digits. First they had to repeat sequences in the same order as they were read aloud, then in reversed order. The test was automatically discontinued after two subsequent errors. It measures working memory and the ability to focus, considered aspects of self-regulation. The measure has strong reliability and validity (Davidson et al., 2006).

71 child centers from 17 municipalities participated in the project. There were around 855 five-year olds (children in their last year of ECEC) in these centers, ranging from 4 to 29 five-year olds in each center. Among these, 701 children had parental consent (82 percent), of which 669 showed up for testing. Through the collection of personal identifiers, we were able to match children's assessment scores with registry data from Statistics Norway, used to construct measures of the families' socio-economic status.

Due to our interest in center-level variation, we excluded from our sample 14 children from four centers which had fewer than 5 children represented. We additionally excluded three children who could not be matched to Norwegian registry files in 2015.[4] These criteria yielded a sample with 648 children from 67 childcare centers, with registry data on child and family background variables (birth month, gender, parental education and earnings, parent's country of birth). All children completed the math test, but only 601 children completed all six assessments because we ended the assessment early for children who became uncomfortable with the test situation. Twenty-six children were missing one assessment, and 21 were missing two or more.

For our main analysis, we utilize an aggregate "school readiness" score as a weighted average of the six individual assessments. We employ confirmatory factor analysis to determine the weight applied to each assessment score (see Brown, 2014 for details).[5] In doing so, we effectively assume the six assessments share a common component ("school readiness") which is best approximated employing the empirically-derived weights. Unfortunately, producing this index score is only possible for children that have scores on all six assessments. To maximize our sample, we run this analysis *including* the 26 children with a single missing score ($N = 627$ total) by generating a predicted value (via OLS) for the missing score as a function of the child's non-missing scores, estimated over those with all six assessments. This reduced our estimates of center-level variation slightly, but modestly improved the power of the relevant statistical tests.[6] While our main analysis focuses on this aggregate score of school readiness, results are also provided for the individual assessments.

In Table 1, we see that the six assessment scores are strongly correlated to one another, indicating children with high competency within one developmental area are also likely to have higher competency in other areas. The six assessment scores also, by construction, correlate strongly with our index score for school readiness. The difference in those correlations reveal that the index score placed somewhat higher weight on the math and working memory assessments, with relatively less weight on phonological awareness.

The 67 childcare centers represent a selected sample, which matters in interpreting our results. The sample includes childcare centers in the Agder counties of southern Norway that self-selected for participation in the Agder-project. The Agder region is not representative of Norway as a whole. In this region, work force participation is lower and welfare participation higher than the rest of Norway. Moreover, it is more common for mothers to stay home with young children than in the rest of the country. In addition, the childcare centers selecting to participate in the Agder-project are likely not representative of centers in the Agder region. As the Agder-project was an intervention study involving continuing education and more systematic curriculum use, the centers selecting to participate may have been those already working more systematically to improve school readiness skills than the average childcare center. Many of the centers who declined our invitation to participate stated that they believed the Agder-project had too much structured activities. Our focus on centers in a particular region that self-selected into the Adger-project is therefore expected to produce a more homogeneous sample of centers than a random sampling of Norwegian centers would produce, causing estimates of center-level variation to be smaller than what exists in the broader universe of Norwegian centers. Additionally, we had selection at the individual level, because not all parents consented for their child to participate in the study. Among the mothers of the children in our sample 52 percent has university or college education, which is slightly higher than the full population of mothers of five year olds in Agder (49 percent), and slightly lower than comparable mothers in all of Norway (56 percent).[7]

Table 2 presents summary statistics of our covariates. Slightly more than half of our main analytic sample ($N = 627$) is female.[8] Birth month is a continuous variable, taking the variable 1 for the youngest (born in December 2011), and 12 for the oldest (born in January 2011). An average birth month of 6.8 implies that average age at assessment was 5.15 years. Mothers average 14.3 years of education, while fathers average 13.7 years. Mothers are more likely to have completed a college than fathers in our sample. Fathers' mean earnings are about 70 percent higher than mothers' mean earnings. A sizable fraction of children had a non-western immigrant mother (13.6%) or father (10.5%). On average, centers had 11.7 children included in our sample, ranging from 5 to 22.

[4] The 3 children with no available registry data are likely to be families recently immigrated to Norway.

[5] Implemented using Stata's "confa" command.

[6] If we also predict the missing scores of those missing two assessments, results are unchanged.

[7] Own calculation from registry data provided by Statistics Norway.

[8] Summary statistics pertaining to the full sample (N=648) are presented in Appendix Table A1.

Table 1
Pairwise correlation between test scores.

	Index	Math	Self-regulation HTKS	Vocabulary	Working memory	Phonological awareness
Math	0.847					
Self-regulation HTKS	0.675	0.426				
Vocabulary	0.662	0.466	0.364			
Working memory	0.756	0.490	0.418	0.401		
Phonological awareness	0.432	0.273	0.234	0.249	0.277	
Self-regulation H&F	0.592	0.475	0.334	0.210	0.368	0.174

Note: All correlation coefficients are significant at 1 percent level. "Index" is a weighted mean of the six assessment scores, with weights determined by confirmatory factor analysis (see text for details).

Table 2
Summary statistics.

	Mean	Std. dev.	Obs
Female	0.507	0.500	627
Birth month	6.83	3.194	627
Mother education (# years)	14.25	2.560	607
Father education (# years)	13.68	2.477	599
Mother drop out	0.188	0.391	607
Mother high school	0.283	0.451	607
Mother college	0.529	0.500	607
Father drop out	0.207	0.406	599
Father high school	0.409	0.492	599
Father college	0.384	0.487	599
Mother earning (NOK)	331,260	214,649	627
Father earning (NOK)	563,132	266,978	613
Mother immigrant	0.136	0.343	627
Father immigrant	0.105	0.307	627
# 5 yrs old in center	11.66	5.05	67

5. Analyses and results

5.1. Empirical strategy

We investigate the magnitude of center-level variation in school readiness by estimating a series of fixed and random effects models, represented as follows:

$$Y_{ic} = \alpha_c + \beta X_i + \epsilon_{ic} \tag{1}$$

$$Y_{ic} = \alpha + \beta X_i + \varphi_c + \mu_{ic} \tag{2}$$

where Y_{ic} is the score for child i in child center c, and X_i is a vector of child and parent characteristics. In the fixed effects (FE) model (1), tests of significant center-level variation in scores are based on an *F test* of the joint significance of the estimated fixed effect terms ($\hat{\alpha}_c$). In the random effects (RE) specification (2), center levels effects are assumed to be normally distributed and independent of the included covariates, but are not explicitly estimated.[9] Instead, only the variance in center effects is estimated, and the test of significant center-level variation is formed from the likelihood ratio test comparing the RE model to one where $Var(\varphi_c)$ is assumed to be zero.

In most of our analyses, we focus on a parsimonious set of covariates found to be the strongest individual predictors of child outcomes – child sex, child birth month, mother's education[10] and indicators of each parent's immigrant status. "Full covariate" models additionally include covariates for father's education, mother earnings and father earnings.[11] For some children, parental earnings or education was missing in the registry data. In the regressions, indicators for "missingness" were included as appropriate.

Beyond testing for significant center level effects, we also have an interest in quantifying the *magnitude* of center-level differences in school readiness. Due to the modest numbers of children drawn from each center, the estimated center fixed effects suffer from "over-fitting." The variation in center fixed effects therefore predictably overstates the true level of center-level variation. Our estimates for the magnitude of center-level variation are therefore based on estimates of $Var(\varphi_c)$ in the RE models. Since estimates of $Var(\varphi_c)$ are not particularly intuitive, we use this estimate to infer the predicted difference in school readiness scores across centers at the 90th and 10th percentile of the φ_c distribution. In our results tables, this result is presented as the "good/bad difference."

The RE model is also extended to explore whether center quality has differential effects on children based on their family's socio-economic status, for which we use mother's education as a proxy. Specifically, Eq. (2) is extended to allow the influence of mother's education to vary across centers in different parts of the φ_c distribution.[12] The model we estimate is frequently called a mixed-effects linear regression model and takes the form

$$Y_{ic} = \alpha + \beta X_i + \gamma_c Med_i + \varphi_c + \mu_{ic} \tag{3}$$

with variance terms estimated for $Var(\varphi_c)$, $Var(\gamma_c)$, and $Cov(\varphi_c, \gamma_c)$.[13] A negative value on the covariance term would indicate that the positive gradient between maternal education (Med_i) and child readiness scores is smaller for children in centers at higher points on the φ_c distribution.

5.2. Center-level variation in assessment scores

As documented in Table 3, the fixed effects models support the notion of substantial heterogeneity in the school readiness of children in different centers. Column 1 reports results with no covariates, and the estimated fixed effects are highly significant ($p < .01$). The estimated center variation decreases when the parsimonious set of covariates are included (column 2), likely reflecting selection into neighborhoods and that most children attend childcare at a center in their neighborhood, but the estimated fixed effects remain statistically significant ($p = .04$). Notably, the test of including the richer set of covariates has virtually no effect on the variance in estimated center effects, though it does weaken the finding of significant center effects ($p = .07$). In light of our small sample size, including extraneous covariates inflicts a particularly high cost in terms of the power of statistical tests, so we focus on models containing the parsimonious covariates in the remaining columns of Table 3. [14]

[9] Estimation of the RE model is performed via MLE using Stata's "mixed" or "xtreg/mle" command. GLS estimation of the RE models yield very similar estimates of the variance in center effects, but doesn't allow us to test the significance of that variance.

[10] Education controls were included as identifiers for high school and college completion.

[11] Covariates for earnings are entered as indicators for each quartile in the earnings distribution in our sample.

[12] This analysis required us to control for mother's education linearly, instead of in categories. This difference had no effect on our main results.

[13] Estimation is performed using Stata's "mixed" command with the "covariance(unstructured)" option.

[14] Coefficient estimates for the included covariates are presented in Appendix Table A2.

Table 3
Center-level variation in school readiness scores.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Model	FE	FE	FE	RE	RE	RE	RE
levels:	center	center	center	center	municipality, center	testing time, center	testing day, center
covariates	none	parsimonious	full	parsimonious	parsimonious	parsimonious	parsimonious
Adj. R2	0.059	0.217	0.220				
$Sd(\hat{\alpha}_c)$	0.431	0.365	0.362				
F test	0.003**	0.037*	0.068+				
$\widehat{Sd(\varphi_c)}$				0.168	0.106	0.168	0.115
$\widehat{Sd(\varphi_m)}$					0.154		
$\widehat{Sd(\varphi_{tt})}$						0.000	
$\widehat{Sd(\varphi_{td})}$							0.128
Hausman test				0.376			
LR test				0.045*	0.299	–	0.112
Good/bad difference				0.553			0.546
Obs	627	627	627	627	627	627	627

Notes: **, * and + denote significance at 1, 5 and 10 percent level, respectively. Parsimonious covariates include child gender and birth month, mother education level and parent immigrant status. The full set of covariates additionally includes father education level and parent earnings. $Sd(\hat{\alpha}_c)$ is the standard deviation of the estimated center fixed effects. *Ftest* reports the significance level (p-value) of the fixed effects. $\widehat{Sd(\varphi_c)}$ is the estimate standard deviation in center random effects. $\widehat{Sd(\varphi_m)}$, $\widehat{Sd(\varphi_{tt})}$ and $\widehat{Sd(\varphi_{td})}$ are the corresponding standard deviation in municipality, testing time (hour) and testing date random effects. In columns 4 and 8, the *LR test* reports the significance level for variation in center random effects. In columns 5 and 7, the *LR test* reports the significance of the improved fit offered by inclusion of the additional level of random effect (i.e relative to column 4). The *Good/Bad* value calculates the difference in expected assessment scores between the 90th and 10th percentiles on the distribution of child center effects, calculated as (+/-)1.645• $\widehat{Sd(\varphi_c)}$.

In column 4 of Table 3, we estimate the random effects analogue to the fixed effects model in column 2. The Hausman test fails to reject the RE model in favor of the FE specification, a sign that the estimated coefficients are not much affected by the inclusion of fixed (versus random) effects. The absence of center-level variation can again be rejected at conventional levels ($p = .045$). The estimated variation in center effects was then used to predict the difference in scores expected across children in centers at the 90th and 10th percentile in the center effect distribution. The difference in scores amounts to over one half (0.55) of a standard deviation. While this suggests sizable center variation, the specific estimate should be interpreted with caution, since we have only a modest numbers of centers and few children per center. The 95% confidence interval around the estimated standard deviation is [0.083, 0.338].

In columns 5–7 we explore a number of concerns pertaining to the interpretation of these findings as evidence of variation in center quality.

First, in column 5, we augment the RE specification to allow random effects at the level of municipality as well, of which 17 are represented. The likelihood ratio test fails to reject the RE specification with center effects in favor of the specification with random effects at both levels. Nonetheless, most of variation originally attributed to centers appears to reflect cross-municipality variation. While this is an interesting finding, we do not believe it undermines the plausibility of center quality driving the variation in school readiness we observe. To the extent local policies and differential funding influence center quality, these would emerge as variation at the municipality, not center, level.

A larger concern relates to the fact that classmates were almost always tested on the same day and at the same time.[15] If children test better at certain times of day, or if the atmosphere for testing was better on some days than others, these forces could artificially inflate the impression of center-level variation. In column 6, we evaluate this concern in relation to the time of day the children from a particular center were tested, augmenting the RE specification to allow random effects related to the time students were tested (in hourly groupings). No evidence of testing time effects is supported.

Column 7 includes random effects at the level of "testing day" (and center). Here we find sizable estimates for variation at the level of testing day, which suggests our original estimate of center variation may be substantially inflated by testing day effects. Unfortunately, the assessment data were not collected with an eye towards differentiating center and testing day effects. The fact that children from the same center mostly tested on the same day, and only a handful of centers (6–9) were tested on any given day (see Appendix Table A3), prevents us from generating a meaningful decomposition of the center and testing day effects. By chance, mean center quality is expected to vary across different testing days, and, to the extent it does, would cause the appearance of testing day effects that are actually a reflection of variation in center quality.

We are able shed more light on the existence of potential testing day effects by utilizing data from a second round of assessments collected on the same children 10 months later (T2). Importantly, the centers testing on the same day at baseline (T1) did not always test on the same day for the follow up assessments, though they often did (see Appendix Table A3).

In Model 1 of Table 4, we show by way of simple OLS regression how assessment scores at baseline (T1) vary by testing day. If these estimates reflect "true" testing day effects, we would expect those effects to be transitory; the day on which the baseline assessment was conducted would not be expected to affect subsequent (T2) scores. If, instead, these estimates reflect differences in the (true) school readiness of children who happened to test on different days, we would expect T1 testing day effects to be similar across the two periods of outcomes. In Model 2, we find strong evidence for the latter. Testing day at T1 significantly predicts T2 test scores, and the pattern of estimates are very similar across the two outcomes. An F-test for the joint significance of the difference in testing day effects strongly fails to reject their equivalence (p-value of 0.98). In contrast, testing day at T2 has a more modest correlation with T2 outcomes (see Model 3). When both sets of testing days are controlled for, it is the testing day at T1 which more strongly predicts T2 outcomes (see Model 4).

Together, these results suggest our estimate of center variation in Table 3, Model 4 is not substantially contaminated (biased upwards) by legitimate testing day effects. Instead, the variation in assessment scores across testing days appears to be an artifact of the small count of centers testing on each day and on the idiosyncratic groupings of centers who happened to test on the same days. While we cannot rule out some upwards bias in our estimate of center variation arising through different

[15] In four cases the children were tested on a different day than the rest of their peers in their center. Dropping these had no effect on our estimates.

Table 4
Testing day effects by testing period.

Test score period	(1) T1	(2) T2	(3) T2	(4) T2
T1 testing day:				
Monday	—	—		—
Tuesday	0.083	0.267		−0.013
	(0.159)	(0.180)		(0.230)
Wednesday	0.245	0.341*		0.046
	(0.154)	(0.169)		(0.228)
Thursday	−0.104	−0.052		−0.298
	(0.157)	(0.170)		(0.236)
Friday	0.334*	0.440*		0.265
	(0.160)	(0.171)		(0.194)
Monday	0.372	0.348		0.083
	(0.319)	(0.333)		(0.362)
Tuesday	0.200	0.204		−0.212
	(0.151)	(0.165)		(0.958)
Wednesday	0.061	0.191		−0.134
	(0.149)	(0.161)		(0.970)
Thursday	−0.235	−0.190		−0.916
	(0.168)	(0.185)		(1.015)
Friday	0.327*	0.335+		−0.356
	(0.164)	(0.179)		(1.019)
T2 testing day:				
Tuesday			—	—
Wednesday			−0.246	−0.255
			(0.158)	(0.159)
Thursday			−0.184	0.289
			(0.158)	(0.317)
Friday			−0.131	0.219
			(0.166)	(0.371)
Monday			−0.333*	−0.545
			(0.157)	(0.964)
Tuesday			0.083	−0.049
			(0.160)	(0.953)
Wednesday			−0.183	−0.257
			(0.159)	(0.953)
Thursday			−0.114	−0.176
			(0.149)	(0.937)
Friday			0.160	−0.176
			(0.161)	(0.965)
F tests				
T1 testing days	0.006	0.009		0.037
T2 testing days			0.072	0.239
T1/T2 difference	0.982			
Adj R2	0.226	0.171	0.160	0.186
Obs	627	578	578	578

Note: Results from OLS regressions on aggregate school readiness scores in baseline period (T1) and at follow up (T2). Parsimonious covariates included in all specifications (results not shown), as well as a student-specific "treatment" indicator in T2 models. *F tests* provide p-value for joint significance of included testing day indicators and, across columns 1 and 2, the p-value testing whether (T1) testing day has different effects on T1 outcomes than on T2 outcomes.

Table 5
Selection into childcare centers.

	(1)	(2)	(3)
Predicted test score	0.048		
	(0.042)		
Mother immigrant		0.017	0.014
		(0.051)	(0.050)
Father immigrant		0.090+	0.108+
		(0.052)	(0.061)
Mother high school degree		0.050	0.030
		(0.043)	(0.046)
Mother college degree		0.030	−0.017
		(0.051)	(0.050)
Father high school degree			0.076+
			(0.041)
Father college degree			0.089*
			(0.038)
Mother 2. quartile earning			−0.013
			(0.048)
Mother 3. quartile earning			−0.003
			(0.038)
Mother 4. quartile earning			0.039
			(0.054)
Father 2. quartile earning			−0.064
			(0.041)
Father 3. quartile earning			−0.039
			(0.047)
Father 4. quartile earning			0.010
			(0.046)
Observations	627	627	627
Adj R2	0.004	0.010	0.030
F test		0.241	0.0849

Note: * and + denote significance at 5 and 10 percent level respectively. Estimates from OLS regressions on the leave-out-mean residual score calculated over each child's classmates. All models are clustered on center level. *F test* reports the joint significance (p-value) of each set of included covariates. Predicted test scores are constructed from the full set of covariates.

assessment conditions, our evidence suggests testing day effects are not a serious source of bias.

A final robustness test of this sort was estimated to verify that the matching of children to individual testers is not biasing our results. Including "tester" identifiers had virtually no effect on our estimates of center-level variation, and consequently appear to be of no concern. In the RE specification, the estimated standard deviation of the center effects increased from 0.168 to 0.170 when tester effects are included.

5.3. Evidence of selection

The robustness of the FE results to richer covariates and the failure of the Hausman test to reject both suggest that selection effects are unlikely to be important drivers biasing our estimates of center variation. To investigate selection more formally, for each child we constructed a proxy for center quality as the leave-out residual score (i.e. conditional on the full covariate set) over each child's classmates. We then tested via OLS regression whether children with certain characteristics were more likely to attend a center with "over-performing" children. The results of this exercise are presented in Table 5.

In column 1, we find a weakly positive and statistically insignificant relationship between a child's own predicted score and the LOM residual score of her classmates. Focusing on our parsimonious covariates in column 2, higher maternal education is only weakly predictive of selection into a center with higher performing classmates. Somewhat contrary to expectations, we find that father's immigration status is a positive predictor of selection into a center with higher performing classmates, with an estimate that reaches marginal significance. This finding is robust to the inclusion of the full set of covariates in column 3. Moreover, the estimated effects of father's education indicate children of higher educated fathers tend to select into higher quality centers. That said, as we documented earlier, the inclusion of the richer covariates had little effect on our estimates of center variation. While we find some evidence of selection, there is little support for the notion that children with the strongest SES backgrounds are differentially selecting into better centers to any great degree. This may suggest that there is limited differences in center quality across different types of neighborhoods, and that location is the primary concern for parents when choosing childcare center. Moreover, it may be hard for parents to observe center quality.

5.4. Mechanism investigation

Table 6 explores potential mechanisms behind center-level variation in scores focusing on a limited number of structural characteristics we were able to obtain for a subset of the centers in our study. This evidence

Table 6
Impact of center characteristics on school readiness.

	(1)	(2)	(3)	(4)	(5)
Director #yrs tenure		0.007			0.007
		(0.007)			(0.007)
Teacher/Child ratio			2.211**		2.289**
			(0.750)		(0.650)
Center size				0.007	0.007
				(0.008)	(0.007)
$\widehat{Sd(\varphi_c)}$	0.134	0.113	0.113	0.125	0.067
LR test	0.171	0.255	0.242	0.204	0.405
Observations	482	482	482	482	482
# centers	52	52	52	52	52

Notes: ** denotes significance at 1 percent level. Estimates from random effects regressions analogous to Table 3, column 4. All regressions include covariates for child gender and birth month, mother education level and parent immigrant status (not reported in table). Reported standards errors are corrected for clustering at the level of centers. $\widehat{Sd(\varphi_c)}$ is the estimated standard deviation of the random center effects. *LR test* reports the significance level (p-value) of center random effects

Table 7
Mixed effects by center and mother education.

	(1)	(2)
random effect:	center	Center
mixed effect:		mother education
covariates:	parsimonious	parsimonious
$\widehat{Var(\varphi_c)}$	0.028	0.094
	(0.020)	(0.193)
$\widehat{Var(\gamma_c)}$		0.001
		(0.002)
$\widehat{Cov(\varphi_c, \gamma_c)}$		−0.011
		(0.018)
LR test	0.048	0.333
Obs	627	627

Note: Estimates from mixed effects regressions on standardized assessment score. Parsimonious set of covariates include child gender and birth month, mother number of years education and parent immigrant status. $\widehat{Var(\varphi_c)}$ is the estimated variance in center effects (and standard error), $\widehat{Var(\gamma_c)}$ is the estimated variance in the effects of mother education, and $\widehat{Cov(\varphi_c, \gamma_c)}$ is the estimated covariance between the two. *LR test* in column 1 reports the significance level for variation in center random effects. *LR test* in column 2 reports the significance of the improved fit offered by modeling the additional variance terms (i.e. relative to column 1).

should not be interpreted as casual as the structural indicators are not random and may be associated with other unobserved indicators for quality. For 52 centers in our study, we obtained data on the following characteristics: the tenure of the center director, the teacher-child ratio and center size. Due to the much reduced sample size, we replicate the results of our preferred RE specification in column 1, finding the variation to be somewhat smaller and (unsurprisingly) no longer significant. In columns 2–4, we augment this model with controls for each of the structural characteristics, and in column 5 we include all three simultaneously. The teacher-child ratio proves to be a strong positive predictor of school readiness scores. The result in column 3 suggests that 30% of the originally estimated variance in center effects is potentially attributable to variation in child-teacher ratios. Column 5 suggest almost 75% of the originally estimated variance in center effects is potentially explained by the three structural characteristics, though this should be regarded cautiously considering two of the three coefficients are statistically insignificant. Nonetheless, child-teacher ratios (at least) appears to be a strong contributor to the center variation we estimate.

5.5. Heterogeneous center effects by maternal education

Mixed effects models were estimated to investigate whether the influence of centers on school readiness is mediated by family SES, for which mother's education serves as proxy. If higher quality centers are effective at reducing the school readiness gap between high and low SES children, we would expect the γ_c and φ_c terms in Eq. (3) to covary negatively.[16]

Estimation results are presented in Table 7. Unfortunately, the mixed effects models proved too weakly-powered to provide substantive evidence towards this hypothesis. The bottom of column 2 reports the likelihood ratio test for whether the allowance for the additional variance terms, Var(γ_c) and Cov(φ_c, γ_c), significantly improves model fit over the original random effects specification, reported to column 1.[17] We fail to find evidence for significantly improved fit ($p = .33$). So while the directional evidence suggests children of lower SES might benefit more from placement in a higher quality center, we are unable to draw strong conclusion in that regard. The same is true when if we perform this exercise over the individual assessments; while estimates of the covariance term are generally negative, they never approach levels of statistical significance.

5.6. Estimation by subject

Table 8 reports the results of FE and RE specifications with parsimonious controls for each of the six individual assessment tests. Several important distinctions are revealed as far as the influence of family characteristics. Maternal education appears to be an especially strong predictor of self-regulation HTKS scores. The children of college educated mothers also score strongly on math, vocabulary and working memory, but the effects of maternal education are more muted for phonological awareness and the self-regulation hearts and flowers test. Not surprisingly, the children of immigrant parents perform especially poorly on the vocabulary test. Female effects are generally positive and significant, but insignificant (and negative) for vocabulary and the self-regulation hearts and flowers test. Age effects are consistently positive, but substantially smaller for phonological awareness.

Turning to the main findings of interest, for each assessment the Hausman test fails to reject the RE specification in favor of the FE model. For math, self-regulation HTKS and vocabulary we find significant evidence of center-level variation. While the results for the working memory test fail to achieve statistical significance, the estimated variance terms are comparable in size to the math test. In contrast, center level variation appears weaker for phonological awareness and is especially weak for the self-regulation hearts and flowers test.

6. Discussion

Our paper investigates differences in childcare quality across centers in the universal childcare system of Norway. This is important because the ECEC literature suggests that if childcare quality varies across centers, universal childcare systems can potentially increase disparities in school readiness, particularly if the children of parents with high education select into centers of high quality. Our analysis demonstrates large and significant variation in school readiness across centers. Indeed, the difference in school readiness skills in centers at the 90th and 10th percentile in the center effect distribution was estimated to be over one half (0.55) of a standard deviation. These results are robust to numerous robustness tests and do not seem to be driven by parental selection into centers of different quality.

[16] If children of higher educated mothers had their kids in childcare longer, we would expect any effect of center quality to be higher among those, and that would bias the estimated magnitude of a negative covariance term downwards.

[17] As noted earlier, for this specification maternal education is controlled for linearly to facilitate estimation of the mixed model.

Table 8
Center-level variation in individual assessments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Self-regulation HTKS		Math		Vocabulary		Working memory		Phonological awareness		Self regulation H&F	
Model	FE	RE	FE	RE	FE	RE	FE	RE	FE	RE	FE	RE
Female	0.138*	0.155**	0.184*	0.160*	−0.079	−0.067	0.199*	0.201*	0.233*	0.235**	−0.074	−0.082
	(0.061)	(0.058)	(0.084)	(0.077)	(0.075)	(0.071)	(0.081)	(0.080)	(0.089)	(0.087)	(0.075)	(0.073)
Birth month	0.052**	0.055**	0.075**	0.082**	0.074**	0.081**	0.060**	0.063**	0.028*	0.033**	0.080**	0.078**
	(0.012)	(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.013)	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
Mother immigrant	−0.385*	−0.404**	−0.125	−0.102	−0.524**	−0.537**	−0.113	−0.143	−0.330*	−0.256+	0.016	0.116
	(0.162)	(0.148)	(0.177)	(0.156)	(0.151)	(0.143)	(0.132)	(0.129)	(0.162)	(0.148)	(0.167)	(0.140)
Father immigrant	−0.075	0.054	−0.181	−0.152	−0.459**	−0.444**	0.052	0.115	0.168	0.164	0.012	0.098
	(0.187)	(0.186)	(0.140)	(0.130)	(0.142)	(0.134)	(0.170)	(0.160)	(0.130)	(0.136)	(0.171)	(0.165)
Mother high sch.	0.324*	0.325**	0.118	0.126	0.171	0.175	0.070	0.076	0.112	0.147	−0.001	0.051
	(0.131)	(0.121)	(0.115)	(0.107)	(0.112)	(0.110)	(0.103)	(0.093)	(0.110)	(0.104)	(0.116)	(0.109)
Mother college	0.560**	0.558**	0.404**	0.426**	0.348**	0.360**	0.415**	0.409**	0.200+	0.231*	0.112	0.163+
	(0.128)	(0.114)	(0.127)	(0.119)	(0.131)	(0.128)	(0.107)	(0.100)	(0.103)	(0.100)	(0.097)	(0.094)
$Sd(\hat{\alpha}_c)/\widehat{Sd(\varphi_c)}$	0.388	0.186	0.366	0.187	0.386	0.235	0.379	0.174	0.361	0.128	0.332	0.000
F test/LR test	0.033*	0.056+	0.032*	0.020*	0.009**	0.021*	0.263	0.458	0.225	0.256	0.402	1.000
Hausman test		0.162		0.678		0.714		0.771		0.701		0.149
Adj R2	0.101		0.112		0.167		0.070		0.034		0.065	
Observations	624	624	648	648	628	628	629	629	629	629	623	623

Note: **, * and + denote significance at 1, 5 and 10 percent level respectively. Estimates result from fixed (FE)/random (RE) effects regressions on standardized assessment scores. Models replicate the FE Model (2) and the RE Model (4) in Table 3 on each assessment. See Table 3 for additional details. Standard errors corrected for clustering at level of childcare center. Clustering omitted (by necessity) in performing the *F test, LR test,* and *Hausman Test*.

We extend our analyses by investigating how measures of structural quality predict the development of school readiness skills. The analysis demonstrates that the teacher-child ratio is associated with a large and significant increase in school readiness skills and that the ratio can explain a meaningful portion (about 30 percent) of the variance in center effects. Notably, this evidence is not causal as the structural indicators are not random and may be associated with other unobserved indicators for quality.

Our analysis of center quality must be interpreted with caution. First, even if differential selection of children into higher/lower quality centers on the basis of socio-economic background appears quite limited, we cannot rule out that the differences in center effects capture unobserved similarities in children's prior skills and parental background. Another possible concern is that children in the same center are co-located residentially, and may experience common shocks affecting child development. Moreover, the estimated variance in learning across center captures the combined effect of teacher quality, peers, childcare pedagogy, in addition to structural quality such as child-staff ratios, and class size. While our study demonstrates that which childcare center a child attends is a strong predictor of school readiness skills, it provides limited causal evidence for the mechanisms of the substantial variation across centers we observe. The substantial variation in childcare center effects documented in this paper highlight the importance of research identifying factors that contribute to a good learning environment in universal childcare systems.

Acknowledgement

Financial support from Aust-Agder utviklings- og kompetansefond, the Sørlandet Knowledge Foundation, Aust-Agder and Vest-Agder county administrations, and the Norwegian Research Council [237973] is gratefully acknowledged.

Data statement

The data we use in the analysis are subject to strict regulation by the Norwegian Data Inspectorate. This is because the data contain detailed information about each child and indirect identification is possible. We will help as best we can scholars interesting in replicating the results to get the necessary permission from the Norwegian Data Inspectorate. If permission is granted, we will be happy to share the data.

Appendix A

Tables A1, A2, A3.

Table A1
Summary statistics, full sample.

	Mean	Std. dev.	Obs
Female	0.504	0.500	648
Birth month	6.81	3.206	648
Mother education (# years)	14.25	2.534	628
Father education (# years)	13.70	2.468	619
Mother drop out	0.182	0.386	628
Mother high school	0.291	0.455	628
Mother college	0.527	0.500	628
Father drop out	0.204	0.402	619
Father high school	0.409	0.492	619
Father college	0.388	0.488	619
Mother earning (NOK)	331,096	212,844	648
Father earning (NOK)	562,875	264,163	633
Mother immigrant	0.133	0.343	648
Father immigrant	0.105	0.307	648
Math	0	1	648
Self- regulation HTKS	0	1	624
Vocabulary	0	1	628
Working memory	0	1	629
Phonological awareness	0	1	629
Self- regulation Hearts and Flowers	0	1	623
# 5-years-old in center	11.64	5.01	67

Table A2
Coefficient estimates.

	(1)	(2)
Female	0.160*	0.166*
	(0.077)	(0.076)
Birth month	0.091**	0.093**
	(0.011)	(0.012)
Mother immigrant	−0.318*	−0.361*
	(0.156)	(0.158)
Father immigrant	−0.139	0.006
	(0.160)	(0.190)
Mother high sch.	0.199+	0.119
	(0.108)	(0.108)
Mother college	0.526**	0.406**
	(0.120)	(0.121)
Father high sch.		0.148
		(0.096)
Father college		0.171
		(0.119)
Mother earning q2		0.208+
		(0.116)
Mother earning q3		0.185
		(0.118)
Mother earning q4		0.154
		(0.151)
Father earning q2		0.091
		(0.113)
Father earning q3		−0.007
		(0.114)
Father earning q4		0.064
		(0.114)
$Sd(\hat{\alpha}_c)$	0.365	0.362
F test	0.037*	0.068+
Adj. R2	0.217	0.220
Obs	627	627

Note: **, * and + denote significance at 1, 5 and 10 percent level respectively. Reports coefficient estimates from Models (2) and (3) in Table 3. See Table 3 for details. Indicators for missing values on education and earnings are also included as covariates, but not reported in table. Standard errors corrected for clustering at the level of centers.

Table A3
Center counts on each testing day in baseline period (T1) and follow up (T2).

		Follow up (T2) testing day									Total
		Tue	Wed	Thu	Fri	Mon	Tue	Wed	Thu	Fri	
Baseline (T1) testing day	Mon					6		1		2	9
	Tue					2	3		2		7
	Wed						2	2	2	1	7
	Thu						1	4	4		9
	Fri					1		1		4	6
	Mon								1	1	2
	Tue	4	2	1							7
	Wed	3	5								8
	Thu			3	3						6
	Fri			2	4						6
	Total	7	7	6	7	9	6	8	9	8	67

References

Bauchmüller, R., Gørtz, M., Rasmussen, A.W., 2014. Long-run benefits from universal high-quality preschooling. *Early Childhood Res. Quart.* 29 (4), 457–470.

Bennett, J., Tayler, C.P., 2006. Starting Strong II: Early childhood Education and Care. OECD.

Berger, A., 2011. Self-regulation: Brain, Cognition, and Development. American Psychological Association, Washington, DC.

Bierman, K.L., Domitrovich, C.E., Nix, R.L., Gest, S.D., Welsh, J.A., Greenberg, M.T., Gill, S., 2008. Promoting academic and social-emotional school readiness: the head start REDI program. *Child Dev.* 79 (6), 1802–1817. doi:10.1111/j.1467-8624.2008.01227.x.

Blair, C., Raver, C.C., 2015. School readiness and self-regulation: a developmental psychobiological approach. *Annu. Rev. Psychol.* 66 (1), 711–731. doi:10.1146/annurev-psych-010814-015221.

Blau, D.M., 1999. The effect of child care characteristics on child development. *J. Hum. Resour.* 786–822.

Brown, T.A., 2014. Confirmatory Factor Analysis For Applied Research. Guilford Publications.

Buhs, E.S., Ladd, G.W., Herald, S.L., 2006. Peer exclusion and victimization: processes that mediate the relation between peer group rejection and children's classroom engagement and achievement. *J. Educ. Psychol.* 98 (1), 1–13.

Bøe, T., Hysing, M., Zachrisson, H.D., 2016. Low family income and behavior problems in norwegian preschoolers: is child emotionality a marker for sensitivity of influence. *J. Develop. Behav. Ped.* 37 (3), 213–222. doi:10.1097/dbp.0000000000000282.

Ponitz, Cameron, E., C., McClelland, M.M., Matthews, J.M., Morrison, F.J., 2009. A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Dev. Psychol.* 45 (3), 605–619. doi:10.1037/a0015365.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quart. J. Econ.* 126 (4), 1593–1660.

Clements, D.H., Sarama, J., 2011. Early childhood mathematics intervention. *Science* 333 (6045), 968–970. doi:10.1126/science.1204537.

Currie, J., Neidell, M., 2007. Getting inside the "black box" of head start quality: What matters and what doesn't. *Econ. Edu. Rev.* 26 (1), 83–99.

Davidson, M.C., Amso, D., Anderson, L.C., Diamond, A., 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. http://dx.doi.org/10.1016/j.neuropsychologia.2006.02.006.

Diamond, A., Lee, K., 2011. Interventions shown to aid executive function development in children 4 to 12 years old. *Science* 333 (6045), 959–964.

Drange, N., & Rønning, M. (2017). Child Care Center Staff Composition and Early Child Development. *Working Paper, Statistics Norway*.

Duncan, G.J., Dowsett, C.J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., Japel, C., 2007. School readiness and later achievement. *Dev. Psychol.* 43 (6), 1428–1446.

Engel, A., Barnett, W.S., Anders, Y., Taguma, M., 2015. Early Childhood Education and Care Policy Review. OECD, Norway https://www.regjeringen.no/contentassets/6372d4f3c219436e990a5b980447192e/oecd_norway_ecec_review_final_web.pdf.

Fuhs, M.W., Nesbitt, K.T., Farran, D.C., Dong, N., 2014. Longitudinal associations between executive functioning and academic skills across content areas. *Dev. Psychol.* 50 (6), 1698–1709. doi:10.1037/a0036633.

Gupta, N.D., Simonsen, M., 2016. Academic performance and type of early childhood care. *Econ. Edu. Rev.* 53, 217–229.

Guryan, J., Hurst, E., Kearney, M., 2008. Parental education and parental time with children. *J. Econ. Perspect.* 22 (3), 23–46. doi:10.1257/jep.22.3.23.

Hall, C.M., Welsh, J.A., Bierman, K.L., Nix, R., 2016. Kindergarten social withdrawal and reading achievement: a cross-lagged path model for at-risk learners. *Psychol. School.* 53 (7), 751–759. doi:10.1002/pits.21939.

Hamre, B.K., 2014. Teachers' daily interactions with children: an essential ingredient in effective early childhood programs. *Child Develop. Perspect.* 8 (4), 223–230. doi:10.1111/cdep.12090.

Harris, Y.R., Terrel, D., Allen, G., 1999. The influence of education context and beliefs on the teaching behavior of African American mothers. *J. Black Psychol.* 25 (4), 490–503. doi:10.1177/0095798499025004002.

Havnes, T., & Mogstad, M. (2009). No child left behind: universal child care and children's long-run outcomes.

Heckman, J., Kautz, T., 2013. Fostering and measuring skills: interventions that improve character and cognition. *Nation. Bureau Econ. Res.. Working Paper 19656*. Retrieved from http://www.nber.org/papers/w19656 .

Jackson, C.K., Rockoff, J.E., Staiger, D.O., 2014. Teacher effects and teacher-related policies. Annu. Rev. Econ. 6 (1), 801–825.

Jensen, B., Jensen, P., Rasmussen, A.W., 2017. Does professional development of preschool teachers improve children's socio-emotional outcomes. *Lab. Econ.* 45, 26–39.

Kalil, A., Ryan, R., Corey, M., 2012. Diverging destinies: maternal education and the developmental gradient in time with children. *Demography* 49 (4), 1361–1383. doi:10.1007/s13524-012-0129-5.

Lee, P., Bierman, K.L., 2015. Classroom and teacher support in kindergarten: associations with the behavioral and academic adjustment of low-income students. Merrill-Palmer Quarterly (Wayne State University. Press) 61 (3), 383–411. doi:10.13110/merrpalmquar1982.61.3.0383.

Lervåg, A., Aukrust, V.G., 2010. Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *J. Child Psychol. Psychiat.* 51 (5), 612–620.

McClelland, M.M., Cameron, C.E., Duncan, R., Bowles, R.P., Acock, A.C., Miao, A., Pratt, M.E., 2014. Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Front. Psychol.* 5. doi:10.3389/fpsyg.2014.00599.

McClelland, M.M., Cameron, C.E., McDonald Connor, C., Farris, C.L., Jewkes, A.M., Morrison, F.J., 2007. Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Dev. Psychol.* 43 (4), 947–959.

McGinty, A.S., Justice, L.M., 2010. Language facilitation in the preschool classroom: Rationale, goals, and strategies. In: Mckenna, M.C., Walpole, S., Conradi, K. (Eds.), Promoting Early reading: Research, resources, and Best Practice. The Guilford Press, New York.

Melhuish, E., 2016. Longitudinal research and early years policy development in the UK. *Intern. J. Child Care Edu. Policy* 10 (1), 3.
Melhuish, E.C., 2011. Preschool matters. *Science* 333 (6040), 299–300.
Merritt, E.G., Wanless, S.B., Rimm-Kaufman, S., Cameron, C.E., Peugh, J.L., 2012. The contributions of teachers' emotional support to children's social behaviors and self-regulatory skills in first grade. *School Psychol. Rev.* 41 (2), 141–159.
Ministry of Education and Research (2011). National Framework Plan for the Content and Tasks of Kindergartens.
Muter, V., Hulme, C., Snowling, M.J., Stevenson, J., 2004. Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Dev. Psychol.* 40 (5), 665–681. doi:10.1037/0012-1649.40.5.665.
Nix, R.L., Bierman, K.L., Heinrichs, B.S., Gest, S.D., Welsh, J.A., Domitrovich, C.E., 2016. The randomized controlled trial of Head Start REDI: Sustained effects on developmental trajectories of social–emotional functioning. *J. Consult. Clin. Psychol.* 84 (4), 310–322. doi:10.1037/a0039937.
OECD, 2015. Starting Strong IV.
Rabiner, D.L., Godwin, J., Dodge, K.A., 2016. Predicting academic achievement and attainment: the contribution of early academic skills, attention difficulties, and social competence. *School Psychol. Rev.* 45 (2), 250–267. doi:10.17105/spr45-2.250-267.
Raver, C.C., Jones, S.M., Li-Grining, C., Zhai, F., Bub, K., Pressler, E., 2011. CSRP's impact on low-income preschoolers' preacademic skills: self-regulation as a mediating mechanism. *Child Dev.* 82 (1), 362–378. doi:10.1111/j.1467-8624.2010.01561.x.
Reynolds, A.J., Temple, J.A., Ou, S.-R., Arteaga, I.A., White, B.A.B., 2011. School-based early childhood education and age-28 well-being: effects by timing, dosage, and subgroups. *Science* 333 (6040), 360–364. doi:10.1126/science.1203618.
Rivkin, S.G., Hanushek, E.A., Kain, J.F., 2005. Teachers, schools, and academic achievement. *Econometrica* 73 (2), 417–458.
Rockoff, J.E., 2004. The impact of individual teachers on student achievement: Evidence from panel data. *Am. Econ. Rev.* 94 (2), 247–252.
Romano, E., Babchishin, L., Pagani, L.S., Kohen, D., 2010. School readiness and later achievement: Replication and extension using a nationwide Canadian survey. *Dev. Psychol.* 46 (5), 995–1007. doi:10.1037/a0018880.
Sabol, T.J., Soliday Hong, S.L., Pianta, R.C., Burchinal, M.R., 2013. Can rating pre-k programs predict children's learning. *Science* 341 (6148), 845–846. doi:10.1126/science.1233517.
Schjølberg, S., Lekhal, R., Wang, M.V., Zambrana, I.M., Mathiesen, K.S., Magnus, P., & Roth, C. (2008). *Forsinket språkutvikling: En foreløpig oversikt basert på data fra Den norske mor og barn undersøkelsen.* Retrieved from Nasjonalt folkehelseinstitutt:
Schmitt, S.A., McClelland, M.M., Tominey, S.L., & Acock, A. 2018. (Submitted). Strengthening school readiness for children at-risk: Evaluation of a self-regulation intervention.
Storksen, I., Ellingsen, I.T., Wanless, S.B., McClelland, M.M., 2015. The influence of parental socioeconomic background and gender on self-regulation among 5-year-old children in Norway. *Early Edu. Develop.* 26 (5-6), 663–684. doi:10.1080/10409289.2014.932238.
Størksen, I., Ellingsen, I.T., Tvedt, M.S., Idsøe, E.M.C., 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, I., Ellingsen, I.T., Wanless, S.B., McClelland, M.M., 2015. The influence of parental socioeconomic background and gender on self-regulation among 5-year-old children in Norway. *Early Edu. Develop.* 26 (5-6), 663–684. doi:10.1080/10409289.2014.932238.
Størksen, I., Mosvold, R., 2013. Assessing early math skills with tablet computers: Development of the Ani Banani Math Test (ABMT) for young children. Paper presented at the Utdanning2020.
Tominey, S.L., McClelland, M.M., 2011. Red light, purple light: findings from a randomized trial using circle time games to improve behavioral self-regulation in preschool. *Early Edu. Develop.* 22 (3), 489–519. doi:10.1080/10409289.2011.574258.
Tominey, S.L., McClelland, M.M., 2013. Improving behavioral regulation in preschool. Dialog, Research-to-practice-summary 16 (3), 149–154.
Utdanningsdirektoratet Eksempelsider for kartleggingsprøver i regning på 1. trinn. Retrieved from https://www.udir.no/globalassets/filer/vurdering/kartlegging/ eksempeloppgaver-kpregning1-trinn.pdf. 2017.
Van Luit, J.E.H., Van de Rijt, B.A.M., 2009. Utrechtse Getalbegrip toets [Utrecht Early Numeracy Test]. Graviant, Doetinchem.
von Suchodoletz, A., Gestsdottir, S., Wanless, S.B., McClelland, M.M., Birgisdottir, F., Gunzenhauser, C., Ragnarsdottir, H., 2013. Behavioral self-regulation and relations to emergent academic skills among children in Germany and Iceland. *Early Childhood Res. Quart.* 28 (1), 62–73. http://dx.doi.org/10.1016/j.ecresq.2012.05.003.
Wanless, S.B., McClelland, M.M., Acock, A., Ponitz, C.C., Son, S.-H., Lan, X., Li, S., 2011a. Measuring behavioral regulation in four societies. *Psychol. Assess.* 23 (2), 364–378.
Wanless, S.B., McClelland, M.M., Tominey, S.L., Acock, A.C., 2011b. The influence of demographic risk factors on children's behavioral regulation in prekindergarten and kindergarten. *Early Edu. Develop.* 22 (3), 461–488. Retrieved from http://www.informaworld.com/10.1080/10409289.2011.536132 .
Weiland, C., Yoshikawa, H., 2013. Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Dev.* 84 (6), 2112–2130.
Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M.R., Espinosa, L.M., Gormley, W.T., Zaslow, M.J. (2013). Investing in our future: The evidence base on preschool education: Society for Research in Child Development.