Pre-analysis Plan: Finland's Preschool Expansion Experiment

Ramin Izadi[*] VATT

Matti Sarvimäki[†] Aalto University

Oskari Harjunen[‡] Aalto University

November 16, 2022

Abstract

Together with Finland's Ministry of Education and Culture, we designed a randomized experiment to test the impact of extending preschool education on children's social-emotional, numeric and language skills. in the current Finnish system, children attend preschool at age six and continue to elementary school at age seven. Our intervention advances the starting age of mandatory preschool by one year, from age six to five. The treatment group (N ≈ 15,000) attends preschool for two years instead of one, following a curriculum specifically created by the Finnish National Agency for Education for the two-year preschool. The control group (N ≈ 20,000) either continues with the business-as-usual early childhood education at daycare centers or stays at home. We evaluate each child three times between the ages of five and seven using standardized tests and teacher evaluations. Our primary outcomes are indices of social-emotional, numeric and language skills. These data are augmented using register-based data on the children and their parents. The broader research project also includes online surveys for teachers, parents and civil servants, in-depth interviews with children and parents, and text analysis of administrative records.

Keywords: education, early childhood, preschool, kindergarten
JEL Codes: I20, I24, J01, J24

[*] VATT Institute for Economic Research and Helsinki GSE, ramin.izadi@vatt.fi.

[†] Aalto University School of Business, VATT Institute for Economic Research and Helsinki GSE, matti.sarvimaki@aalto.fi.

[‡] Aalto University School of Engineering and VATT Institute for Economic Research, oskari.harjunen@aalto.fi

Revisions

Nov. 16th, 2022 Child's sex was added as a primary dimension for heterogeneity analysis. Child's sex was also added in the list of control variables. These revisions were made during the first follow-up survey *before* seeing or receiving outcome data from the electronic platform in which the survey is conducted.

Oct. 4th, 2022 Original submission.

1 Intervention

In the current Finnish education system, children attend preschool education at age six for one year (preceding the start of elementary school starting at age seven). This experiment was designed to examine the effects of advancing the starting age of preschool by one year, i.e., from age six to five, so that the treatment group attends preschool for two years instead of one. The experiment is regulated by a temporary law, *Act on a Two-Year Pre-primary Education Trial* (1046/2020), which the Finnish Parliament passed on December 17th, 2020. The law was prepared by the Ministry of Education and Culture drawing inputs from an initial research plan prepared by the research team (Izadi et al., 2020). The law defines the target population and randomization protocol, and provides authority for necessary data collection. Importantly, the law also mandates children randomized into the treatment group to start preschool during the year they turn 5-years-old. Participation is not strictly enforced or compulsory, but children who do not attend preschool are required to achieve the goals of preschool by other means. This requirement is identical to the one currently in place for one-year preschool. The control group remains in the business-as-usual system, i.e., they are not mandated to participate in any activities outside of home but most of them nevertheless attend early childhood education at daycare centers during the year they turn 5-years-old. The control group starts in one-year-preschool during the year they turn 6-years-old.

The two-year preschool is implemented based on a curriculum created by the Finnish National Agency for Education (2021). The curriculum is centrally directed by the Ministry of Education, but locally adapted by the municipal boards of education. Thus, there is likely variation across municipalities in the exact content of the curriculum. Preschool is provided for four hours a day and the curricula are more structured compared to daycare. Typically, children stay in daycare after preschool is finished, i.e., their day is divided between four hours of preschool in the morning and daycare in the afternoon. Preschool is free of charge, while parent's pay income dependent fees for daycare. Children living more than five kilometers away from their preschool are entitled to free transportation.

2 Research Questions

The experiment seeks to answer the following questions:

1. What is the impact of two-year preschool (in comparison to one-year preschool) at the start of mandatory education on average, and on the distribution of children's social-emotional, numeric and language skills?

2. Does the impact vary across different background characteristics?

3 Research Design

3.1 Population

The experiment targets children born in 2016 and 2017 living in Finland, who are enrolled in or live close to eligible daycare centers. A daycare center is eligible if it provides both preschool and daycare (in order to ensure that children can stay in daycare in the same premises after preschool). Eligible daycare centers were identified through a survey to municipality administrators. Data on the number of daycare centers were in some cases missing or contradicting, but we estimated that there were a total of 3,379 daycare centers in 288 municipalities[1], of which 2,296 were deemed eligible.

3.2 Randomization

Our randomization procedure divided daycare centers and children into treatment and control groups using a combination of stratified sampling and re-randomization. We used a slightly different procedure for small and large municipalities. We defined a municipality to be small if it had less than five eligible daycare centers, while the rest were defined as large municipalities.

We included all small municipalities into the randomization, which we executed at the municipality-level so that all eligible daycare centers within a municipality were allocated either to the treatment or the control group. To guarantee geographical representation, we used the six regional state administrative agencies as strata. We performed a re-randomization algorithm with 1,000,000 replications, allocating half of the municipalities within each stratum to the treatment and the other half to the control group in each replication. The final assignment vector was chosen randomly out of 2,000 assignment vectors that produced the best balance between pre-treatment municipality characteristics.[2] The characteristics were: (i) the share of foreign-born population, (ii) the share of people with secondary education, (iii) the share of people with higher education, and (iv) average income per capita. The resulting sample consists of 83 municipalities, out of which 41 were assigned into the treatment group.

Large municipalities were first randomized into participating and non-participating municipalities. We conducted the randomization using 22 strata based on the regional geography and number of eligible daycare centers in the municipality. The exception

[1] This is less than the total number of municipalities in Finland, because some municipalities organize early childhood and preschool education jointly with adjacent municipalities.

[2] We made sure that our balance criterion was loose enough so that the distribution of pairwise correlations of the 2000 potential assignment vectors was similar to the unrestricted design

to this rule were the majority-Swedish-speaking municipalities, which formed their own stratum. The share of municipalities from each stratum that were allocated to participate in the experiment was set to 65%. The share was calibrated so that the expected total costs would pass below budget. To guarantee this we re-randomized the sample draw until the budget goal was achieved. The resulting sample consists of 63 large municipalities participating in the experiment.

Once we had selected the participating large municipalities, we conducted the final randomization into treatment and control groups at the daycare center level. Specifically, 40% of the eligible daycare centers within each municipality were assigned to the treatment group. Similar to the small municipalities, we first performed a re-randomization algorithm 1,000,000 times and then chose from them the 2,000 assignment vectors that produced the best balance between the covariates at the level of the daycare center. At the end we picked the final assignment vector randomly from these 2,000 vectors. The daycare center level covariates used were: (i) the Herfindahl index for language concentration, (ii) the share of parents with primary education, (iii) the share of parents with secondary education, (iv) average parental income, and (v) the share of single parents. The resulting sample consists of 752 daycare centers.

4 Data

We will evaluate the effect of the preschool expansion using a combination of children assessments (including teacher's evaluation), register-based data and survey data collected from daycare centers. The assessment and survey data are collected by the research team.

4.1 Assessments

Children's social-emotional, language and numeric skills are measured through assessments conducted annually during fall, between 2021–2024. Each child is assessed three times: at daycare center/preschool at ages 5 and 6, and at primary school at age 7. The assessments are carried out in a digital platform using tablet computers at the online learning environment *ViLLE*, developed and maintained by the Centre of Learning Analytics at the University of Turku. Each assessment is conducted during a span of one month (with additional time for children who have been absent). Before the assessments, teachers participate in online training.

The assessed skills will remain the same throughout the experiment, but the items of language and numeric assessments will become more challenging as the children grow older. The assessment combines items from various existing instruments, some of which were piloted with Finnish children before the start of the data collection. Specifically, we measure the following skills:

- social-emotional skills (teacher ratings)
    - Behavioral Strategy Rating Scale (BSRS)
    - Multisource Assessment of Children's Social Competence (MASCS)
    - Child Behavior Rating Scale (CBRS)
    - Strengths and Difficulties Questionnaire (SDQ)
    - Survey on Social and Emotional Skills (OECD)
- language skills
    - vocabulary
    - phonology
    - knowledge of letters
    - reading
- numeric skills
    - counting numbers
    - comparison of quantities and numbers
    - producing quantities
    - arithmetic operations
- self-rated well being (Kiddy-KINDL)
- spatial recognition

Baseline The baseline assessments were conducted in fall 2021 and 2022, at the start of the two-year preschool (treatment group) or as part of standard daycare (control group). At the time, most of the participating children were five years old. The assessment was conducted in two parts. In the first part, children answered a set of questions assessing their numeric skills, language skills, cognitive ability and subjective well being. The assessments were completed at their own daycare center with one-on-one interviews. This part took approximately 30 minutes to complete. In the second part, teachers independently answered the social-emotional skill questionnaires about the children. The approximated length of this part was 15 minutes. The baseline assessments cannot reach children who did not participate in daycare or preschool at the time of the assessments.

Follow-up 1 The first follow-up assessments will be conducted in October 2022 and 2023, roughly a year and two months after the start of two-year preschool (treatment group) and roughly two months after the start of one-year preschool (control group). The format of these assessments is similar to the baseline, but part of the questions are adjusted to take into account that the children are now older than during the baseline.

Follow-up 2 The second follow-up assessment will take place in fall 2023 and 2024 at the start of primary education. The assessments will be conducted in classrooms during standard 45-minute class / classes. The instruments will be similar to those used in baseline and follow-up 1, but adjusted to take into account children's age and classroom setting.

4.2 Register data

Statistics Finland will augment the assessment data with the following register-based information from their database:

- child's native language, place and time of birth
- parents' sex, native language, place and time of birth, year of arrival to Finland (if born abroad), municipality of residence, annual income, occupation, education, number of children, children's sex, children's time of birth, job characteristics, neighborhood characteristics

The precise definition of job and neighborhood characteristics will have to be decided later, but they will likely include information such as the share of immigrants, average income and education at the workplace and immediate residential neighborhood.

4.3 Data Processing and Security

We process data according to the privacy guidelines agreed upon prior to the data processing with the Ministry of Education and Culture and Statistics Finland. Ministry of Education and Culture of Finland is the owner of the data gathered in this study, while Statistics Finland is the owner of the register data. Data are only available to members of the research group, and stored in a secure location. The research team analyzes the final (pseudonymized) data in Statistics Finland's secure remote access system. After the study is completed, all data will be archived for later research use by the National Archives of Finland and will be made available to all researchers subject to standard approval processes of the National Archives and Statistics Finland.

4.4 Attrition

There are two possible sources for attrition. First, children may not respond to assessments. We mitigate this challenge by conducting the assessments in daycare centers and schools (see Section 4). Furthermore, the legislation governing the experiment mandates daycare centers and schools to conduct data collection. Thus, we expect attrition due to an entire daycare center or school refusing to collect data to be rare. However, we will not reach children who are at home during the time of the assessments, which is likely to create differential attrition between treatment and control groups at baseline because a larger share of the control group is at home at age 5. This will limit the analysis controlling for baseline measures (see section 5.3.1), but the this difference will disappear in later assessments where all children attend preschool/school. Second, we will not conduct assessments in municipalities excluded from the experiment (see above). Thus, the assessment data will not follow children who move from participating to non-participating municipalities during the experiment. We expect also this form of attrition to be limited due to the large number of municipalities participating in the experiment and relatively small geographical mobility among families with 5–7 year-old children.

5 Empirical Analysis

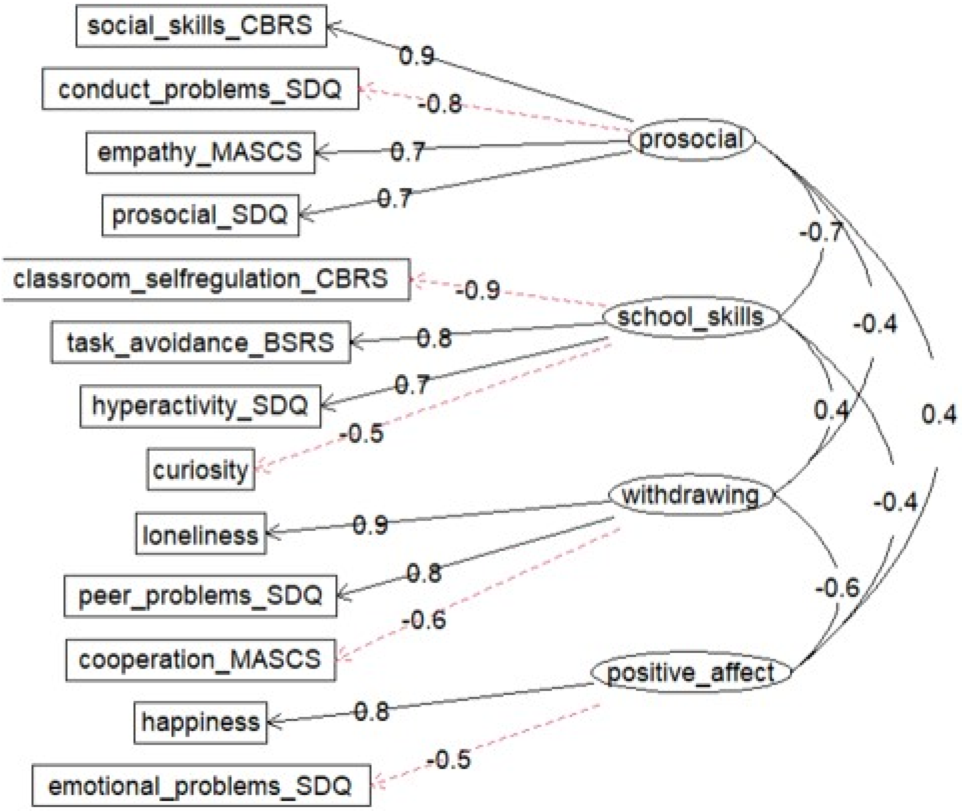
5.1 Primary and secondary outcomes

For the first follow-up (beginning of last year of preschool), our primary outcomes are summary indices which are constructed by combining different subscales from the assessment presented in section 4.1. For social-emotional skills, the formation of primary outcomes is guided by exploratory factor analysis presented in Figure 1:

1. **Prosociability** includes the following subscales: social skills (CBRS), conduct problems (SDQ), empathy (MASCS), prosocial behavior (SDQ).
2. **Schools skills** includes the following subscales: classroom selfregulation (CBRS), task avoidance (BSRS), hyperactivity (SDQ), curiosity (OECD).
3. **Peer problems** includes the following subscales: peer problems (SDQ), cooperation (MASCS), loneliness (OECD).
4. **Positive affect** includes the following subscales: emotional problems (SDQ), happiness (OECD).

Each subscale consists of several items and is scored by summing over the appropriate items according their respective manuals. The primary outcomes are constructed by

Figure 1: Exploratory factor analysis



averaging over the normalized subscales (using the opposite number when appropriate for the scale).

For academic skills, primary outcomes are constructed by using item scores estimated with a graded response model. Two separate models are estimated with the following structure:

5. **Numeric skills** includes the following subscales: counting numbers, comparison of quantities, comparison of numbers, producing quantities, arithmetic operations.
6. **Language skills** includes the following subscales: vocabulary, phonology, knowledge of letters, reading.

The items in each subscale are first aggregated as intended by the task design: Counting up is coded by the highest number reached (30 is max). Counting down is coded as a binary indicator for successful effort. Rest of the subscales are sums over all correct answers to the subscale items. For verbal skills, all subscales are sums over all correct answers to the subscale items. The aggregated subscales are then entered into a graded response model that determines the weighting scheme used in the final index.

Our secondary outcomes are:

- self-reported self-esteem, happiness, wellbeing
- teacher-reported child behavior subscales
- visual-spatial reasoning
- subscales of the numeric and verbal skill indices

At the time of writing, the questions for the second follow-up have not been finalized and thus we will define these outcome variables in fall 2023 (before the start of the second follow-up).

5.2 Balancing Checks

Since we re-randomized based on demographic characteristics of the randomization unit (daycare center or municipality), we already have balance by construction at that level along the variables used in re-randomization. In order to check balance along other covariates, we report averages for the treatment and control groups, and use the main estimating equation to estimate "treatment effects" for predetermined background characteristics (see section 5.3.1). Using identical methodology, we also test for differences in attrition rates between treatment and control groups.

5.3 Treatment Effects

5.3.1 Intent to Treat

We estimate the intent-to-treat effects on average outcomes using a standard regression framework:

$$Y_{ist} = \beta_0 + \beta_1 \mathbb{1}[treatment_i] + \mu_{st} + \beta_2 X_i + \epsilon_{ist} \tag{1}$$

where Y_{ist} is the outcome of interest for child i and $\mathbb{1}[treatment_i]$ is an indicator variable equal to one if the child was either mapped to a day care center which was assigned to the treatment group (in "large" municipalities) or the child was a resident of a "small" municipality randomized to treatment group at the time of randomization. We include stratum×birth cohort fixed effects μ_{st} (stratum is municipality or regional administrative area and birth cohort is 2016 or 2017). We start by reporting unconditional estimates, where the control vector X_i is omitted. To improve precision, we also report results from a specification where X_i includes the following control variables:

- parental income
- mother's highest degree
- father's highest degree
- mother's native language
- father's native language
- child's language
- single parent
- immigrant parents
- number of siblings
- child's age at arrival to Finland
- child's sex

Our third specification additionally controls for the following information gathered during the baseline

- baseline measures of the outcome variables
- group composition and child/adult ratio at the daycare center / preschool

- teachers' formal qualifications

A limitation of the last specification is that we do not observe baseline measures for all children. Thus interpretation of the results from this analysis will require more caution compared to the first two specifications. However, controlling for baseline measures may substantially improve the precision of the estimates.

We also examine the effect of the treatment on the distributions of our outcome variables. Here, our primary interest is on the 90/10 inequality, i.e., the difference between 90th and 10th percentiles of the outcome variable distributions. Our first estimate for the effect on 90/10 inequality is the difference in this measure between treatment and control groups. In order to improve precision, we also use the methods of Firpo et al. (2009) to estimate unconditional quantile treatment effects at the 90th and 10th percentiles of the outcome distributions conditional on the background characteristics listed above. Using these estimates, we then construct treatment effects on 90/10 inequality. In secondary analysis, we report unconditional quantile treatment effects (with and without controlling for background characteristics) over the entire distributions of the outcome variables.

5.3.2 Treatment on the Treated

In addition to the intent-to-treat estimates, we will estimate the effect of actually participating in the treatment curriculum in a treatment day care center.[3] Here, we use the random assignment into the treatment group as an instrumental variable for program participation. We measure treatment participation by asking municipalities to report once a year on whether each child is participating in the treatment curriculum or not. This typically corresponds to whether a child attends a day care center from the treatment group. In our main specification, we define participation status in September 2021 (birth cohort 2016) or September 2022 (birth cohort 2017).

In our interpretation of the results, the complier population consists of two distinct groups:

1. "Daycare compliers": children who participate in two-year preschool because they were randomized to the treatment group and would have participated in business-as-usual daycare otherwise.

[3] Non-compliance can take place for at least two reasons. First, families may move between randomization and the start of the treatment. The importance of these moves likely differ across municipalities, but in most cases families who move in are either offered a day care seat from a nearby day care center with vacancies or they apply to a daycare center of their choice. The intervention does not impose limits on post-randomization seat allocation. Second, some families randomized into the treatment group may decide not to participate in the treatment, while some families randomized into the control group may participate in the treatment.

2. "Home compliers": children who participate in two-year preschool because they were randomized to the treatment group and would have stayed at home otherwise.

While the combined effect of these two groups is the relevant parameter for Finnish policy-makers, we aim to follow Kline & Walters (2016) and estimate the local average treatment effects separately for both groups. However, the details of this analysis are hard to anticipate, and we are not able to include a detailed pre-analysis plan for this part of the project.

5.4 Heterogeneous Effects

Our primary heterogeneity analysis focuses on the following background characteristics:

1. Whether the child was in daycare at the time of randomization (spring preceding the start of two-year preschool)
2. Parents' immigrant status (1 if both parents were born abroad, 0 otherwise)
3. Parents' income percentile (continuous variable)
4. Child's sex

The motivation for the first heterogeneity analysis is that preschool curriculum may not be dramatically different to day care curriculum for children age five. Thus, the treatment effect is likely to be substantially lower for children who would have participated in regular day care than for those who would have stayed at home had they not been assigned to the treatment. We recognize that this heterogeneity analysis does not directly measure the effects for "daycare compliers" and "home compliers" (see the previous subsection) because some of the children who were at home at the time of randomization would have moved to out-of-home care even if they had not been assigned to the treatment group. Nevertheless, we expect the share of "home compliers" to be substantially larger among the children who were at home at the time of randomization. Thus this heterogeneity analysis complements the approaches discussed in the previous subsection. The motivation for examining treatment effect heterogeneity by parents' immigrant status is the expectation that children of immigrants may benefit more from more structured curriculum, particularly in improving their language skills. The motivation for the third primary heterogeneity analysis is the hypothesis that children coming from more disadvantaged households may benefit more from the treatment than those from more affluent households. The motivation for examining treatment effects by child's sex are earlier results showing heterogeneity by sex for comparable interventions (e.g. Gray-Lobe et al., 2022).

We also report secondary heterogeneity analysis along the following dimensions:

1. Parents' country of birth
2. Daycare center or preschool characteristics
   (group composition, teachers formal qualifications, child/adult ratio etc.)
3. Parent's education
4. Mother's labor market attachment
5. Number and age of siblings
6. Small vs large municipalities (as defined in section 3.2)
7. Child's birth cohort (2016 vs 2017)

5.4.1 Intent to Treat

We estimate heterogeneous treatment effects using the following regression analysis:

$$Y_{ist} = \beta_0 + \beta_1 \mathbb{1}[treatment_i] + \beta_2 \mathbb{1}[treatment_i] \times W_{ist} + \mu_{st} + \beta_3 X_{ist} + \epsilon_{ist} \qquad (2)$$

where W_{ist} is one of the background characteristics listed above and other notation is as before. We report results from the same specifications for the heterogeneity analysis as for our main analysis.

5.4.2 Treatment on the Treated

We estimate local average treatment effects for the heterogeneity analysis using the same approach as for the main analysis, i.e., use assignment status (and its interaction with W_i) as an instrument for the treatment participation (and its interaction with W_i).

5.5 Adjustments for Multiple Testing

We divide our estimates into families based on their ex-ante importance for the experimental results. We report the coefficient estimates, individual standard errors and p-values for each family in one table or in a single panel which may be part of a larger table. Family wise error rate and false discovery rate for each family of tests is controlled for and reported in each table/panel. The guiding principle is that primary and secondary outcomes belong to different families and that pooled analysis and heterogeneity analysis belong to different families. Also, one family never includes the same estimand twice (for example with and without control variables).

Specifically, our main results are reported in a table that includes only the pooled intention to treat estimates for the six primary outcomes presented in section 5.1. The table will contain three panels corresponding to the specifications estimated with and without control variables as outlined in section 5.3.1. Heterogeneity analysis for the primary outcomes are reported in a table consisting of three 3×6 panels with otherwise identical structure. The three heterogeneity dimensions are presented in section 5.4.

For each family (panel) we report an exact p-value for the null hypothesis of no treatment effect on any outcomes (Young, 2018). For all individual estimates, we report standard errors clustered at the level of randomization (daycare center or municipality). Additionally, we report four p-values for each estimate:

1. standard p-values implied by the standard errors
2. exact two-sided p-values calculated as the proportion of t-statistics in the 2000 potential assignment vectors below the actual t-statistics (in absolute values)
3. sharpened False Discovery Rate (FDR) q-values (Anderson, 2008)
4. family wise adjusted p-values calculated by the Romano-Wolf step down algorithm

We note that best practices for multiple test adjustments are still evolving. Furthermore, we cannot perform a fully reliable power calculations at the time of writing because we cannot base them on comparable earlier data. Most importantly, we face substantial uncertainty on how much precision we gain from conditioning on register-based background characteristics and baseline measures of the outcome variables. Furthermore, we note that in our context type II errors may lead to policy mistakes that are of similar order of magnitude as type I errors. That is, we are concerned of using overly "conservative" approaches guarding against type I error and failing to reject the null of no effect when an economically significant effect exists. In short, it may become appropriate to adjust our inference approach as the literature on multiple hypothesis adjustments and our understanding of our data evolves.

At the time of writing, we also considered using only two primary outcomes: (i) an index of socio-emotional skills (pooling together the subscales from outcomes 1–4, see section 5.1), and (ii) an index of academic skills (pooling together outcomes 5–6). While we opted for the six primary outcomes—which we believe better map into well-founded psychological concepts—we remain uncertain whether we correctly judged this trade-off between statistical power and interpretability of the outcomes. Thus, we will also report complementary analysis using only these two indices as primary outcomes. In this reporting, we will be clear that these results were not chosen as the primary outcomes at the September 2022 pre-analysis stage and should hence be interpreted as exploratory analysis.

Analysis of secondary outcomes and secondary heterogeneity is exploratory in nature. Here, we consider each group of outcome variables listed in section 5.1 as a single family (e.g. self-reported self-esteem, happiness, wellbeing) and report the same standard errors and p-values as for our main analysis within each family. For the secondary heterogeneity analysis, we consider the seven secondary background variables specified in section 5.4 as a single family.

References

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484), 1481–1495.

Finnish National Agency for Education (2021). Kaksivuotisen esiopetuksen kokeilun opetussuunnitelman perusteet 2021. ISBN: 978-952-13-6734-2.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3), 953–973.

Gray-Lobe, G., Pathak, P. A., & Walters, C. R. (2022). The Long-term Effects of Universal Preschool in Boston. *The Quarterly Journal of Economics, forthcoming.*

Izadi, R., Luukkonen, E., Nokso-Koivisto, O., & Sarvimäki, M. (2020). Kaksivuotinen esiopetus -kokeilu tutkimussuunnitelma. Ministry of Education and Culture.

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

Young, A. (2018). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*. *The Quarterly Journal of Economics*, 134(2), 557–598.