Evaluation of an Extended Day Program in the Netherlands: A Randomized Field Experiment

E. Meyer and C. Van Klaveren

TIER WORKING PAPER SERIES
TIER WP 11/02

Evaluation of an Extended Day Program in the Netherlands: A Randomized Field Experiment*

Erik Meyer†
Chris Van Klaveren‡

Abstract

Policies that aim at improving student achievement frequently increase instructional time, for example by means of an extended day program. There is, however, hardly any evidence that these programs are effective, and the few studies that allow causal inference indicate that we should expect neutral to small effects of such programs. This study conducts a randomized field experiment to estimate the effect of an extended day program in seven Dutch elementary schools on math and reading achievement. The empirical results show that this three-month program had a modest but non-significant effect on math, and no significant effect on reading achievement.

JEL Code: I21

Keywords: Extended Day; Increased Instructional Time; Random Assignment; Field Experiment

*We are grateful to Wim Groot, Henriëtte Maassen van den Brink, Nienke Ruijs, and other colleagues for their comments and suggestions regarding earlier drafts of this paper.

†The corresponding author is affiliated with Maastricht University, Top Institute for Evidence Based Education Research (TIER), P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: e.meyer@maastrichtuniversity.nl.

‡The author is affiliated with Maastricht University, Top Institute for Evidence Based Education Research (TIER), P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: cp.vanklaveren@maastrichtuniversity.nl.

1 Introduction

International comparative studies on student achievement, such as the OECD's *Programme for International Student Assessment* (PISA; OECD, 1999), are frequently designed to give governments insights into the relative performance of their education systems. Since today's students are tomorrow's labor force, such comparisons potentially offer a glimpse into a country's competitive position in tomorrow's knowledge-driven global economy. Under increasing pressure to compete internationally, governments worldwide are enacting policies to improve student achievement, especially in core subjects, such as math and reading.

Central to these policies is frequently that instructional time allocated to core subjects is increased. Well known examples are the *No Child Left Behind* act (Bush, 2001) in the US that stimulates the allocation of extra time to teaching math and reading; the *Future for Education and Care* program in Germany that provides funding for all-day schools (*Ganztagsschulen*) and replaces the traditional half-day schools (see section 'Development of All-Day School' in Freitag and Schlicht, 2009); and the *Extended School Times* project (OCW, 2009) in the Netherlands that provides funds for summer schools, weekend schools and extended day programs aimed at improving math and reading achievement.

It happens more and more that the programs developed to achieve these policy aims are faced with accountability demands, i.e. the effectiveness of these programs must be shown. As a consequence, the demand for studies that evaluate the effectiveness of educational programs has increased. Until recently however, extended day programs have not been rigorously evaluated. Reviews indicate that the field is plagued by a lack of peer-reviewed studies and that many studies do not properly control for selection and composition effects, such that the reported estimates may be biased (Cooper et al., 2000; Lauer et al., 2006; Scott-Little et al., 2002).

In the last decade, policies seem to have encouraged more rigorous evaluations, as an increasing number of programs is evaluated using a research design that focuses on measuring the causal program effect, such as randomized controlled experiments, natural experiments and regression-discontinuity designs. There are two unpublished studies that conduct a randomized experiment to estimate the effects of an extended day program on academic outcomes for the U.S.. The first is a final report on the evaluation of the 21^{st} Century Community Learning Centers (21^{st} CCLC) program (James-Burdumy et al., 2005), where impacts in grades K through 6 are estimated. The second is a working paper that estimates the effect of a full-day compared to half-day preschool program (Robin et al., 2006; also available in Robin, 2005).

James-Burdumy et al. (2005) randomly assigned 1,748 elementary school students at 26 centers to a treatment and a control group. During their two year evaluation period, centers were open three hours a day, four or five days a week, and students spent an average of 81 days at the center within the two year period. Students spent one hour on homework, one hour on another academic activity, and one hour on recreational or cultural activities. James-Burdumy et al. estimated intent-to-treat (ITT) impacts, where participants assigned to the program were compared to those assigned to the control group (regardless of actual participation), as well as the local average treatment effect (LATE) to control for non-participation in the program group (8%) and cross-over from the control to program group (16%). The ITT estimates were similar to the LATE estimates, and both estimates showed that neither the effects on teacher assigned grades in math and English, nor on standardized reading test scores were significant. The direction of effects differed by subject, and the effect sizes seemed to be small, even though they were not reported and could not be calculated from information that was reported. Subgroup estimates of ITT impacts suggested that the program may have improved English grades (but not reading test scores) for students with low initial reading test scores. For reasons that were not specified, subgroup estimates of LATE were not reported such that it remains unknown how these estimates were affected by non-participation and cross-over. Summarizing, the results suggest that the 21^{st} Century Community Learning Centers program did not significantly impact academic outcomes at the participating centers.

Robin et al. (2006) evaluated a preschool program with both an extended day and an extended year. They followed two cohorts of students, starting the program in 1999 and in 2000, during preschool, kindergarten, and first grade (only the 1999 cohort). Admission to the extended day program was based on a lottery: 77 students were randomly assigned to the program group (i.e. full-day preschool), and 217 students to the control group (i.e. half-day preschool). The full-day program operated for eight hours a day, five days a week, ten months a year, while the half-day programs operated for two and a half to three hours a day, five days a week, nine months a year. Both groups used the High/Scope curriculum (described in Schweinhart, 2003), best known from the Perry preschool study. Robin et al. (2006) used a growth curve model to estimate treatment effects on growth in test scores over time, and OLS to estimate treatment-control differences at the end of different grade levels. Using the growth curve model, they found that students gained .40 standard score points a month in vocabulary score on average, and that program students gained an additional .21 standard score points a month compared to control students (i.e. a treatment by time

interaction effect). The average gain in math score was estimated at .35 standard score points a month, and program students gained an additional .35 standard score points a month. In addition to the growth curve model, program effects were estimated cross-sectionally, at the end of each year, by means of OLS. They controlled for pre-program baseline test scores, as well as a number of demographic characteristics. At the end of each year, the program had a significant effect on vocabulary score, and effect sizes increased from .12 standard deviations at the end of preschool to .24 standard deviations at the end of kindergarten, and up to .27 standard deviations at the end of first grade (only the 1999 cohort, $N = 132$). Effects on math score followed a similar pattern, starting at a marginally significant .08 standard deviations at the end of preschool, and increased to a significant .20 standard deviations at the end of kindergarten, and .34 standard deviations at the end of first grade. Interestingly, mother's education was a significant covariate in the preschool analysis, but was no longer significant at the kindergarten or first grade analyses. This may suggest that the influence of parental education diminishes as a student is increasingly exposed to formal education. In contrast to James-Burdumy et al. (2005), Robin et al. (2006) suggested that extended day programs could be effective. An explanation for these contradictory findings could be the timing of the two programs; perhaps intervention in preschool (i.e. early intervention) is more effective than intervention in elementary school.

Recently, Patall et al. (2010) conducted a review of extended day and extended year programs. Like previous reviewers, they noted that rigorous evaluation designs are still very scarce. Based on the results of the few experimental and quasi-experimental studies reviewed in their study, they concluded that we may expect neutral to small positive effects on academic achievement from extended day or year programs. They noted, however, that "the effect of [extended day programs] has yet to be fairly tested using well-controlled experimental or quasiexperimental designs from which strong causal implications could be drawn" (Patall et al., 2010, p. 423).

This paper presents the results of a randomized field experiment and evaluates the impact of an extended day program on math achievement and comprehensive reading (hereafter referred to as reading achievement). During the last three months of the 2009-2010 school year, elementary school students in a small-sized city in the Netherlands participated in an extended day program based on the works of Robert Marzano (e.g. see Marzano, 2003).

The contributions of this study are threefold. First of all, it contributes to the scarce empirical evaluation literature that rigorously estimates the effectiveness of an extended day program. Secondly, it provides, to the best of our knowledge for the first time, empirical

evidence on the effectiveness of an extended day program for a European country. Thirdly, both our sample and estimation strategy are very similar to James-Burdumy et al. (2005), such that the Dutch extended day program can be compared with the US based program.

We proceed as follows. Section 2 outlines the details of the extended day program, and Section 3 describes the data and explains the estimation strategy. In Section 4 the empirical results are presented, and Section 5 concludes.

2 Program Characteristics

The extended day program operated for 11 weeks, from the second week of April 2010 till the end of June 2010. Students, aged 8 through 12 ($mean = 10.6$, $sd = .95$), were offered an extended day program consisting, on average, of an additional two hours of language instruction, two hours of math instruction, and one hour of excursions per week. Parents and students were informed regarding the extended day program by the program staff. Participation in the program was voluntary, and it was offered to 95 randomly selected students in grades five through seven. This design is conceptually identical to a "voucher" system, i.e. students are offered the opportunity to participate in the program, which parents can either use or not (e.g. see Murnane and Willet, 2011a). Classes consisted of approximately 10 students from different elementary schools.[1] Instruction was provided by fully qualified teachers, most of whom were externally contracted for the extended day program, aided by teaching assistants. Teaching assistants supported the teacher in instructional and administrative tasks, supported students in the learning process, kept order in the classroom, and saw to any other needs the students or teacher may have had. Teaching assistants with a relevant vocational education degree and an interest in education were actively recruited.

The program's instruction method was based on the research of education scientist Robert Marzano (e.g. see Marzano, 2003), and was focused on making learning 'meaningful', i.e. relating abstract subject matter to concrete experiences in the outside world. During language classes, for example, students went to a mall to interview shoppers and later wrote small reports based on their interviews – practicing language skills in a realistic context. In advance of the program launch, teachers participated in a training program for the Marzano approach, and during the program received on-the-job coaching and guided feedback. Another focus point of the program was parental involvement. Parents actively participated in their child's learning through take-home assignments – playful learning activities the stu-

[1] Regular class size at these elementary schools is approximately 24 students.

dent and parent do together. The parental involvement component was based on 'Character Connection', a US home-to-school outreach program (Character Connection, 2007).

A typical extended day proceeds as follows. At 3:30 students are welcomed at the program location; they start with an energizer activity, or brain break, to restore energy and attention after the regular school day. Each student, together with the teacher, determines their learning objective(s). The teacher will have prepared a theme, a meaningful context from the outside world, within which he will address the subject matter and the students' learning objectives. Students work interactively in small groups, focused on *doing*, i.e. students present, play with the subject, or physically go outside to apply skills. At the end of the extended day, the class returns to the learning objectives and evaluates. Mondays and Tuesdays one and a half hours of extended day programming were offered, while Wednesdays two hours were offered.

The program was offered free of charge to students, and 95% of costs were funded by the Ministry of Education, Culture and Science of the Netherlands (OCW, 2009). The program budget for the 2009-2010 school year was € 591,045. However, this budget was intended for a full school year, whereas the program only operated during the last three months. Since a large part of the budget consists of labor costs and rent, and only a third of these were realized, we calculate the program's costs per student by dividing the budget by three, and then again by the number of program participants (82). Thus, a realistic approximation of the costs per student for this period comes to € 2,402.62. Compared to other extended day programs, these costs seem high.

The schooling system in the Netherlands is founded on the freedom of education principle, including a freedom of school choice for parents. The government imposes a minimum instruction time norm in elementary education of 940 hours a year, an average of 23.5 hours a week for the 40 week school year (Eurydice, 2010). Teachers report that they spend around 5 hours per week on language development and math each. The effects of an extra two hours of math or reading instruction a week, therefore, represent an increase of approximately 40% over regular instruction time in that subject.

The extended day program was organized by seven elementary schools, located in three neighborhoods in a small city in the Netherlands. The city population of 48,000 has a relatively small proportion ethnic minorities (approximately 8%), and is home to a little over 2,500 students aged 8 through 12. While underachievement is a major concern for education professionals in this area, the extended day program is aimed at improving math and reading achievement of all students at the participating schools, not just underachievers.

Parent informed consent was acquired by the schools before students participated in the program and the evaluation.

3 Data and Identification Strategy

We assessed math and reading achievement using standardized tests that are commonly used in Dutch elementary education (Janssen et al., 2010; Staphorsius et al., 2004). Tests were administered in class by the teacher in February 2010 (pre-test) and again in June 2010 (post-test), which are the standard administration periods for these tests. The math and reading tests each have two outcomes; raw scores, and percentile score categories. The percentile score categories indicate the student's ranking among all Dutch test takers who are in the same grade level. Categories range from A through E, where A is the highest score, representing the 75th to 100th percentile (coded as .875), and E is the lowest score, representing the 0 to 10th percentile (coded as .05). Students who score below 0.5 perform at a level that is below the national average level. To have an idea how participants perform compared to the national average, only the categorical test scores are presented in this section. In the empirical analysis, i.e. Section 4, we use the (more precise) raw scores.

Our data comprises students from seven elementary schools attending grades five through seven.[2] Of the 188 students who were assigned to the treatment and the control group, 153 completed the math pre- and post-tests. For reading only 7^{th} grade scores were available, resulting in 94 completed reading pre- and post-tests. Of the 188 students, 19 failed to complete pre- and post-tests for either subject, leaving 169 students that completed one or the other. The tables in this section show descriptives for these 169 students.

Table 1 describes the means and standard deviations of several demographic variables and test scores for the seven schools, labeled by the Roman numerals I up to VII. The demographic variables were registered data, acquired from the school administration system. *Fifth*, *Sixth* and *Seventh grade* indicate the proportion of students in that grade level, *Girl* indicates the proportion of female students, *Ethnic minority* indicates the proportion of students that belong to an ethnic minority group, and *Parental education* indicates the proportion students of whom at least one parent attained higher vocational credentials and up.

[2]Dutch elementary education has eight grades, and is attended by students that are approximately 4-12 years old.

Table 1: Descriptives: Program schools

	I	II	III	IV	V	VI	VII	Overall
Fifth grade	0.00 (0.00)	0.50 (0.51)	0.00 (0.00)	0.00 (0.00)	0.41 (0.50)	0.00 (0.00)	0.00 (0.00)	0.21 (0.41)
Sixth grade	0.50 (0.52)	0.00 (0.00)	0.50 (0.51)	0.00 (0.00)	0.26 (0.44)	0.00 (0.00)	0.65 (0.49)	0.24 (0.43)
Seventh grade	0.50 (0.52)	0.50 (0.51)	0.50 (0.51)	1.00 (0.00)	0.33 (0.48)	1.00 (0.00)	0.35 (0.49)	0.56 (0.50)
Girl	0.71 (0.47)	0.55 (0.50)	0.45 (0.51)	0.43 (0.51)	0.44 (0.50)	0.54 (0.51)	0.35 (0.49)	0.49 (0.50)
Ethnic minority	0.21 (0.43)	0.39 (0.50)	0.45 (0.51)	0.64 (0.50)	0.18 (0.39)	0.00 (0.00)	0.10 (0.31)	0.27 (0.44)
Parental education	1.00 (0.00)	0.61 (0.50)	0.25 (0.44)	0.29 (0.47)	0.77 (0.43)	0.67 (0.48)	0.85 (0.37)	0.64 (0.48)
Math June 2009	0.39 (0.15)	0.58 (0.11)	0.36 (0.18)	0.42 (0.18)	0.55 (0.20)	0.32 (0.18)	0.48 (0.14)	0.47 (0.19)
Math Feb 2010	0.38 (0.19)	0.51 (0.14)	0.39 (0.16)	0.34 (0.12)	0.47 (0.25)	0.33 (0.18)	0.43 (0.14)	0.43 (0.19)
Reading Feb 2009	0.45 (0.17)	0.44 (0.17)	0.37 (0.16)	0.30 (0.17)	0.46 (0.22)	0.34 (0.16)	0.47 (0.17)	0.41 (0.19)
Reading Feb 2010	0.35 (0.15)	0.39 (0.19)	0.34 (0.17)	0.22 (0.09)	0.44 (0.26)	0.35 (0.16)	0.36 (0.14)	0.37 (0.20)
Number of obs.	14	38	20	14	39	24	20	169

Note: Standard deviations in parentheses.

It should be noted that not all grades participate within each school, indicated in the table by a value of zero for the respective grade level indicator. All variables except *Girl* differ significantly between schools. This shows that the seven schools form a rather heterogeneous group in term of the presented background characteristics; but for analysis this is not problematic because randomization (described in the next paragraph) took place within classes. Table 1 also shows, that the mean achievement levels in our sample are substantially below the national average achievement levels (i.e. the 50th percentile), especially in reading, and that the sample schools have even decreased in rank since the 2008-2009 school year. So the experimental schools are characterized by a high proportion of students that achieve below national levels.

Students were randomized as follows. Matched pairs of students were created within grades and schools using Mahalanobis distances matching (Rubin, 1980), based on the students' two prior math and reading scores and, if possible, their ethnicity, and their parents' highest achieved education level. Although we were aware that the number of students per class was rather small to perform a Mahalanobis matching approach, we deliberately chose to do so. The alternative was to perform matching by hand, which is far less objective. Of the matched pairs, one student was randomly assigned to the treatment, the other to the control group (cf. voucher vs. no voucher). Table 2 shows the means and standard deviations of the matching variables for the treatment and control group. The shows characteristics are the same as in Table 1. We excluded the grade level proportions because pairs were formed within classes, and it follows that the distribution of students over grades is identical for the treatment and control group.

Table 2: Descriptives: Post-randomization

	Treatment	Control
Girl	0.48 (0.50)	0.51 (0.50)
Ethnic minority	0.28 (0.45)	0.25 (0.44)
Parents' education	0.63 (0.49)	0.66 (0.48)
Math June 2009	0.46 (0.20)	0.47 (0.18)
Math Feb 2010	0.44 (0.20)	0.41 (0.18)
Reading Feb 2009	0.43 (0.19)	0.40 (0.18)
Reading Feb 2010	0.37 (0.19)	0.37 (0.20)
Number of obs.	86	83

Note: Standard deviations in parentheses.

Table 2 shows that the randomization was successfully performed as the means of the matching variables are not significantly different at the 5% confidence level. Unfortunately, not all students complied with their assigned treatment. In terms of vouchers, not all students who were offered a voucher made use of it, and some student who were not offered a voucher did participate in the program. This can be problematic, as the non-compliance may impose bias on the estimated average treatment effect such that the true effect may be over- or underestimated. Table 3 shows the means and standard deviations of several descriptive characteristics separately for students who were assigned to the treatment (A=1) or the control (A=0) group, and who participated in the program (P=1) or did not participate in the program (P=0). The characteristic *One-parent family* indicates the proportion of students that belong to a one-parent family, which seemed to play a role in the selective non-compliance we observe.

Naively, one might consider compliers to be those whose participation and assignment match [Columns (1) and (2)]. Unfortunately however, Column (1) additionally represents students who always participate in programs, regardless of their assignment, and Column (2) additionally represents students who never participate in programs (always-takers and never-takers; Angrist and Pischke, 2009). While this complicates comparing the columns in Table 3 somewhat, it, fortunately, poses no problem for our estimation strategy (discussed later).

Table 3: Descriptives: Compliance with assigned treatment

	(1) A=1, P=1	(2) A=0, P=0	(3) A=1, P=0	(4) A=0, P=1
Girl	0.44 (0.50)	0.48 (0.50)	0.55 (0.51)	0.59 (0.51)
Ethnic minority	0.33 (0.48)	0.27 (0.45)	0.17 (0.38)	0.18 (0.39)
Parental education	0.58 (0.50)	0.67 (0.48)	0.72 (0.45)	0.65 (0.49)
One-parent family	0.18 (0.38)	0.20 (0.40)	0.14 (0.35)	0.53 (0.51)
Math June 2009	0.44 (0.20)	0.50 (0.18)	0.52 (0.19)	0.36 (0.15)
Math Feb 2010	0.43 (0.21)	0.43 (0.19)	0.46 (0.18)	0.33 (0.14)
Reading Feb 2009	0.42 (0.20)	0.41 (0.19)	0.44 (0.17)	0.36 (0.15)
Reading Feb 2010	0.35 (0.19)	0.38 (0.21)	0.42 (0.20)	0.33 (0.14)
Number of obs.	57	66	29	17

Notes: Standard deviations in parentheses.

Table 3 shows patterns that may underlie the selection process. Students who were assigned to the treatment group but did not participate [Column (3)], have somewhat higher test scores, as well as slightly higher educated parents. Parents in this group may have decided that program participation was not necessary for their child because they were performing well relative to their classmates (though not that well relative to national levels). In contrast, students who were assigned to the control group but did participate [Column (4)], have lower test scores and come from one-parent families more often than students from other groups. It is possible that parents in this group considered the extended day program as a convenient (and cheaper) alternative to daycare. Finally, it should be noted that the columns that contain compliers [i.e. Columns (1) and (2)] have very similar means and standard deviations despite the non-compliance.

Selective non-compliance may impose a bias on the measured effect of the extended day and to address this problem we make use of the feature that test scores are available for all students, irrespective of their compliance status. To identify the effect of the extended day we use an instrumental variable (IV) method, and instrument the actual program participation by the assigned treatment. The identifying assumption is that the instrument is related to the assignment mechanism, but not directly to the outcome variable of interest, which is true by construction for the instrument 'assigned treatment' in this study. The IV estimate captures the effect of participation of students who participate because they were assigned to the program but who would not otherwise have participated, and excludes always takers and never takers (Angrist and Pischke, 2009).

We estimate the local average treatment effect (LATE; Imbens and Angrist, 1994) using a two-stage least squares regression (2SLS; e.g. see Angrist and Pischke, 2009). In the first stage, the probability of participating in the program is estimated by regressing participation status, D_i, on the instrument assigned treatment, Z_i, and all covariates, X_i, that are also to be included in the second stage regression:

$$D_i = \pi_0 + \pi_1 Z_i + X_i' \pi_2 + \upsilon_i. \tag{1}$$

Subscript i is a student indicator, error term, υ_i, is assumed to be normally distributed with mean zero and variance σ_υ^2, and all explanatory variables are assumed to be independent of the error term. In the second stage regression we plug in the predicted participation probabilities, $\hat{D}_i$, and regress post-test scores, Y_i, on $\hat{D}_i$ and X_i:

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + X_i' \beta_2 + u_i. \tag{2}$$

Again u_i is assumed to be a normally distributed error term with mean zero and variance σ_u^2, and the correlation between u_i and v_i are presumed nonzero.

If we would estimate the two-stage least squares model by performing two separate OLS regressions, this would yield incorrect residuals, as these are computed from the instruments rather than the original variables (Wooldridge, 2009). All statistics computed from those residuals would therefore be incorrect as well (i.e. variances, estimated standard errors of the parameters, etc.). Following Wooldridge, we fit the 2SLS model specified in Equations (1) and (2) by using the STATA ivreg2 module, which computes the correct values of these statistics.[3] Since our sample is clustered at the class level, the observations within classes may not be treated as independent. Therefore, we cluster the standard errors at the class level in all analyses (Williams, 2000). Since we have only a few clusters (13) we tend to underestimate the intra-class correlation (Angrist and Pischke, 2009) and therefore, as a robustness check, we repeated the analyses without clustering the standard errors, but the results remained similar. All tables in Section 4 show the estimation results where we cluster the standard errors.

In this study we estimate two empirical models separately for math and reading. The first model estimates the effect of receiving a (randomly assigned) voucher on math and reading achievement by means of ordinary least squares. This model estimates the so called intent-to-treat (ITT) effect, since there is an intent to treat students who received a voucher (cf. Murnane and Willet, 2011a). However, the student's participation status may be different from the student's assignment status, and, therefore, this model does not estimate the effect of the extended day program. The second model is the 2SLS outlined above and estimates the extended day effect. For completeness we also show the (more precise but biased) OLS estimates that estimate how program participation is associated with achievement.

4 Results

Table 4 shows the means and standard deviations for pre- and post-test scores of participants assigned to treatment and control group.[4] Means are presented only for students whose pre- and post-test scores are available, i.e. 153 out of the initial 169 students for math. As mentioned in Section 3, reading post-test scores were available only for students in 7^{th} grade, leaving only 94 out of 169 students. Test score differences in score between treatment and

[3] Version 03.0.06 for STATA MP 11.2

[4] From Table 5 onward, post-test scores are standardized to mean zero, standard deviation one.

control group are not significant at the 5 percent level. This means that achievement levels of control and treatment students are comparable at the start of the program.

Table 4: Pre- and post-test score means and standard deviations

	Treatment		Control		Overall	
Math pre-test	87.545	(13.685)	84.303	(13.635)	85.935	(13.712)
Math post-test	93.019	(10.855)	89.375	(12.819)	91.209	(11.973)
Reading pre-test	36.300	(9.212)	33.450	(10.355)	34.875	(9.852)
Reading post-test	38.638	(10.910)	37.277	(9.760)	37.957	(10.317)

Note: Standard deviations in parentheses. Mean math scores are based on 153 observations, mean reading scores are based on 94 observations.

Table 5 show how program assignment affects program participation (i.e. the first stage results) and show the intent-to-treat estimates. Columns (1) and (3) show the estimation results when we only include the covariate math pre-test scores. Columns (2) and (4) show the estimation results when we include more covariates to obtain more precise estimators.[5] The intent-to-treat estimates show how receiving an extended day voucher affects math achievement.

Table 5: First stage and ITT for math

	First stage				Intent-to-treat (ITT)			
	dependent: extended day participation				dependent: math post-test			
	(1)		(2)		(3)		(4)	
Extended day assignment	0.426***	(0.102)	0.452***	(0.097)	0.094	(0.066)	0.089	(0.067)
Math pre-test	0.004	(0.003)	-0.001	(0.003)	0.065***	(0.003)	0.064***	(0.005)
Constant	-0.109	(0.243)	0.258	(0.434)	-5.621***	(0.324)	-5.387***	(0.390)
Controls	No		Yes		No		Yes	
	$N = 153$		$N = 153$		$N = 153$		$N = 153$	
	$F(1,12) = 17.58$		$F(1,12) = 21.88$		$F(2,12) = 195.59$		$F(8,12) = 73.31$	
	$R^2 = 0.21$		$R^2 = 0.26$		$R^2 = 0.80$		$R^2 = 0.81$	

Notes: Standard errors (SE) in parentheses. SE's clustered by class in all model specifications (13 clusters).

*/**/*** means statistically significant at the 10/5/1 percent level.

[5]The covariates are dummies for gender, ethnicity, parents' highest achieved education level, coming from a one-parent family, as well as class size, and mean class pre-test score.

The first stage results show that receiving an extended day voucher influences program participation positively and significantly. Angrist-Pishke (AP) first-stage chi-squared tests show that our models are not underidentified, $APChi^2 = 19.30$ and 25.02 for models (1) and (2) respectively, and Stock-Yogo (SY) weak identification tests show that our instruments are not weak (Stock and Yogo, 2005).[6] Columns (3) and (4) show that students who received an extended day voucher do not perform better than students who did not receive an extended day voucher. The first stage and intent-to-treat estimates are robust when more covariates are added to the model. The explanatory power of the model does not increase (much) by the addition of more covariates and therefore estimates are not (much) more precisely estimated, which explains the robustness of the estimation results.

Table 6: OLS and 2SLS estimates for math

	OLS		IV/2SLS	
	(1)	(2)	(3)	(4)
Extended day participation	-0.095 (0.057)	-0.096 (0.063)	0.221 (0.145)	0.197 (0.137)
Math pre-test	0.067*** (0.003)	0.066*** (0.005)	0.064*** (0.003)	0.064*** (0.005)
Constant	-5.694*** (0.286)	-4.922*** (0.388)	-5.597*** (0.304)	-5.438*** (0.402)
Controls	No	Yes	No	Yes
School fixed effects	Yes	Yes	No	No
	$N = 153$	$N = 153$	$N = 153$	$N = 153$
	$F(2,12) = 220.96$	$F(8,12) = 136.76$	$F(2,12) = 168.00$	$F(8,12) = 80.03$
	$R^2 = 0.82$	$R^2 = 0.83$	$R^2 = 0.78$	$R^2 = 0.79$

Notes: Standard errors (SE) in parentheses. SE's clustered by class in all model specifications (13 clusters). OLS models, i.e. (1) and (2), include school fixed-effects, the 2SLS models do not because assignment is within classes. */**/*** means statistically significant at the 10/5/1 percent level.

Table 6 reports the 2SLS results derived from these first stage and reduced form estimates.[7] The 2SLS estimates of the effect of the extended day program on math achievement range from .197 to .221, but do not differ significantly from zero. The estimates, reported in columns (3) and (4) of Table 6, are more positive and much larger than the corresponding OLS estimates, reported in columns (1) and (2) of the same table. The OLS estimates likely

[6] The SY weak ID test compares the F statistic to a critical value. The F statistics are reported in the table.

[7] The 2SLS estimates can be calculated by dividing the intent-to-treat estimates by the first stage estimates.

reflect the selective non-compliance outlined in Table 3. If we compare participants and non-participants in Table 3, we see that parents of non-participants are often higher educated than those of participants. Given that parents' education positively impacts student achievement (Holmlund et al., 2011), this would lead to an under-estimation of the effect using OLS. Due to the non-compliance we also underestimate the intent-to-treat effects (Angrist, 2006). The 2SLS estimates represent the causal effect of extended day participation, and accounts for non-compliance and selection bias. However, the noise that is generated by the non-compliance make the 2SLS less precise (i.e. the standard errors increase). It is possible that the 2SLS estimates are not significantly different from zero due to the increased standard errors and it is therefore useful to consider the magnitude of the effect.[8]

The 2SLS estimates of around .20 can be converted into an effect size (Cohen's d) of approximately .12 standard deviations (sd). This means that, conditionally on their pre-test score, a program participant's post-test score will increase by 12 percent of a standard deviation. The standard deviation of the math post-test score of a student assigned to the control group (see Table 4) is 12.82, and 12 percent of that is approximately 1.54 points. The difference between pre- and post-test means is 5.07 points, and represents a students gain on the test over a period of four months. Therefore, a gain of 1.54 points represents a gain of approximately five weeks. So while an effect size of .12 sd is traditionally considered small (Cohen, 1992), in the context of this particular test it appears meaningful.

The effect of the extended day program on math achievement was also examined for several subgroups.[9] Our results indicate that the extended day program was no more (or less) effective for fifth, sixth, or seventh grade students, nor for girls, ethnic minority students, students from a one-parent family, students with highly educated parents, students with a high pre-test score, or students in small classes.

Table 7 presents the first stage and intent-to-treat estimates for reading achievement in identical fashion to Table 5. It is important to note, however, that we reduced the number of covariates for models (2) and (4) to accommodate the smaller sample size for reading.[10]

[8] 2SLS standard errors of extended day participation were slightly higher when unadjusted for clustering.

[9] For each characteristic considered, we have to show two first-stages and a second-stage. To conserve space, the tables for these results are omitted, but they are available upon request.

[10] The covariates are dummies for ethnicity and parents' highest achieved education level, as well as class size.

Table 7: First stage and ITT for reading

	First stage		Intent-to-treat (ITT)	
	dependent: extended day participation		dependent: reading post-test	
	(1)	(2)	(3)	(4)
Extended day assignment	0.343** (0.140)	0.341** (0.144)	0.002 (0.170)	0.006 (0.163)
Reading pre-test	-0.001 (0.007)	-0.001 (0.007)	0.046*** (0.010)	0.044*** (0.009)
Constant	0.347 (0.212)	0.476* (0.211)	-1.588*** (0.396)	-1.849** (0.613)
Controls	No	No	No	Yes
	$N = 94$	$N = 94$	$N = 94$	$N = 94$
	$F(1,7) = 5.99$	$F(1,7) = 5.60$	$F(2,7) = 9.81$	$F(5,7) = 6.05$
	$R^2 = 0.12$	$R^2 = 0.12$	$R^2 = 0.20$	$R^2 = 0.24$

Notes: Standard errors (SE) in parentheses. SE's clustered by class in all model specifications (8 clusters).

*/**/*** means statistically significant at the 10/5/1 percent level.

As with math, the first stage results show that being randomly assigned to the program has a significantly positive effect on the actual program participation. However, Stock-Yogo weak identification tests suggest that the estimation models for reading are only weakly identified [this is also indicated by the low F statistics in columns (1) and (2)]. When instruments are only weakly correlated with the endogenous explanatory variable, then even a small correlation between the instruments and the error term can bias the estimates (Bound et al., 1995). As was the case with math, the ITT results show that being randomly assigned to the program does not have a significant effect on reading achievement. Again, the first stage and ITT estimates are robust to the addition of more covariates to the model. The ITT results show that the addition of covariates leads to somewhat more precise estimators.

Table 8: OLS and 2SLS estimates for reading

	OLS		IV/2SLS	
	(1)	(2)	(3)	(4)
Extended day participation	0.217** (0.071)	0.232*** (0.056)	0.007 (0.459)	0.018 (0.432)
Reading pre-test	0.046*** (0.009)	0.044*** (0.009)	0.046*** (0.009)	0.044*** (0.009)
Constant	-1.693*** (0.307)	-1.727*** (0.352)	-1.590*** (0.375)	-1.858*** (0.658)
Controls	No	Yes	No	Yes
School fixed effects	Yes	Yes	No	No
	$N = 94$	$N = 94$	$N = 94$	$N = 94$
	$F(2,7) = 14.90$	$F(4,7) = 31.20$	$F(2,7) = 9.83$	$F(5,7) = 6.06$
	$R^2 = 0.35$	$R^2 = 0.36$	$R^2 = 0.20$	$R^2 = 0.24$

Notes: Standard errors (SE) in parentheses. SE's clustered by class in all model specifications (8 clusters). OLS models, i.e. (1) and (2), include school fixed-effects, the 2SLS models do not because assignment is within classes. */**/*** means statistically significant at the 10/5/1 percent level.

Table 8 reports the 2SLS and OLS results.[11] The 2SLS estimates show the extended day program did not significantly affect reading achievement. Contrary to the math results presented in Table 6, these estimates are smaller than the corresponding OLS estimates. As a result of the weak first stage, the 2SLS standard errors for participation are substantially larger than the OLS standard errors. We are confident that our instrument is relevant and exogenous, and so the weak first stage for reading is likely caused by the limited sample size. Small samples may cause large bias, incorrect variances, and different than normal distributions. Considering the first stage results for reading, the 2SLS models are not convincing. Neither are the OLS models, however, as they suffer from selection. The most informative models for reading, perhaps, are the ITT models, which represent the lower-bound estimate (due to non-compliance) of receiving an extended day voucher.

[11] Subgroup effects could not be validly estimated for reading achievement given the modest sample size.

5 Conclusion

This paper reports the results of a randomized field experiment conducted to test the effectiveness of a Dutch extended day program in elementary education. This study examines, first of all, the effect of receiving a voucher that can be used to participate in the extended day program on a voluntary basis. Second, it examines the effect that the extended day program has on math and reading achievement for the compliers.

The empirical results suggest that receiving an extended day voucher does not influence students' math and reading achievement. Also, participation in the extended day program does not influence students' math and reading performance. However, there are two limitations that could potentially have obscured a significant result.

The program was limited by a relatively modest duration. While it's curriculum was evidence-based, it was only offered for 11 weeks, which may have been insufficient to produce the desired improvement in achievement. However, results from a two year program evaluated by James-Burdumy et al. (2005) indicated that programs with longer durations can also be ineffective.

The second limitation was the modest sample size. To measure the program effect on math (reading) achievement we had 153 (94) observations available. Due to non-compliance we could estimate the program effects less precisely (especially for reading), which reduces statistical power of our analysis, which potentially prevents us from finding small program effects. However, much power can be regained by estimating a model with (meaningful) covariates (Murnane and Willet, 2011b). The inclusion of student pre-test scores, and the inclusion of additional control variables, greatly improves the precision of our estimates and, hence, the power of our analyses. Including pre-test scores in the math estimates resulted in an R^2 of around .8, showing that our model was highly explanatory, and suggesting high statistical power.

Our estimates predict a modest but non-significant effect of the program on students' math achievement, i.e. approximately five weeks of extra achievement gain. Even though the estimates appear to be precise (given the model's explanatory power), there is (always) a possibility that we did not reject the null hypothesis due to type II error. The question is, then, if the effect size is of substantive significance to policy makers, and if the intervention is cost-effective. The estimated effect on math achievement of five weeks for an 11 week program seems substantive, but the estimated costs are € 2,402.62 per student, which is also substantive. On average, regular Dutch elementary education costs € 6130 per student per year (Hof et al., 2009), which means that a full-time extended day program costing an

additional € 7,207.86 per student would more than double the costs of educating children. Comparisons of different educational interventions in terms of costs and effects, therefore, can help guide policy makers, and additional causal evaluations of extended day programs are to be encouraged.

Finally, we conclude that our estimates of the extended day effects for the Dutch students in our sample are comparable to the US and South America. Our estimates add to the neutral to small positive effects described by Patall et al. (2010). Our results also mirror those of James-Burdumy et al. (2005), who using a similar sample and estimation strategy, found no significant program effect on math or reading achievement. While there are likely (cost-)effective extended day programs to be found, our results, and those of others, suggest that they are the exceptions (especially when we consider the cost-effectiveness of these programs).

References

J.D. Angrist. Instrumental variables methods in experimental criminological research: What, why and how. *Journal of Experimental Criminology*, 2:23–44, 2006.

J.D. Angrist and J.S. Pischke. *Instrumental variables in action: Sometimes you get what you need*, chapter 4, pages 113–218. Mostly harmless econometrics: An empiricist's companion. Princeton, NJ: Princeton University Press, 2009.

J. Bound, D.A. Jaeger, and R.M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995.

G.W. Bush. *No child left behind*. Washington, DC: Office of the President of the United States, 2001.

Character Connection. Character Connection: Parent, child, and school, October 2007. `http://www.characterconnectionprogram.com`.

J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

H. Cooper, K. Charlton, J.C. Valentine, and L. Muhlenbruck. Making the most of summer school: A meta-analytic and narrative review. *Monographs of the Society for Research in Child Development*, 65(1):1–130, 2000.

Eurydice. National summary sheets on education systems in Europe and ongoing reforms: The Netherlands. Technical report, European Commision, 2010. Retrieved from `http://www.eurydice.org`.

M. Freitag and R. Schlicht. Educational federalism in Germany: Foundations of social inequality in education. *Governance*, 22(1):47–72, 2009.

B. Hof, C. van Klaveren, A. Heyma, and I. van der Veen. Societal benefits to eliminating educational delays [Dutch: Maatschappelijke baten van het opheffen van onderwijsachterstanden]. Technical report, Amsterdam, the Netherlands: SEO, 2009. Retrieved from `http://www.seo.nl/`.

H. Holmlund, M. Lindahl, and E. Plug. The causal effect of parents' schooling on children's schooling: A comparison of estimation methods. *Journal of Economic Literature*, 49(3): 615–651, 2011.

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

S. James-Burdumy, M. Dynarski, M. Moore, J. Deke, W. Mansfield, and C. Pistorino. When schools stay open late: The national evaluation of the 21st Century Community Learning Centers program. Final report, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2005. Retrieved from `http://www.ed.gov/ies/ncee`.

J. Janssen, N. Verhelst, R. Engelen, and F. Scheltens. Scientific accounting of the LOVS mathematics tests for grades 3 through 8 [Dutch: Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8]. Technical report, Arnhem, the Netherlands: CITO, 2010. Retrieved from `http://toetswijzer.kennisnet.nl/html/tg/14.pdf`.

P.A. Lauer, M. Akiba, S.B. Wilkerson, H.S. Apthorp, D. Snow, and M.L. Martin-Glenn. Out-of-school-time programs: A meta-analysis of effects for at-risk students. *Review of Educational Research*, 76:275–313, 2006.

R.J. Marzano. *What works in schools: Translating research into action.* Alexandria, VA: Association for Supervision & Curriculum Development, 2003.

R.J. Murnane and J.B. Willet. *Investigator-designed randomized experiments*, chapter 4, pages 40–60. Methods matter: Improving causal inference in educational and social science research. Oxford University Press, 2011a.

R.J. Murnane and J.B. Willet. *Statistical power and sample size*, chapter 6, pages 82–106. Methods matter: Improving causal inference in educational and social science research. Oxford University Press, 2011b.

OCW. Subsidy scheme extended school times primary education [Dutch: Subsidieregeling onderwijstijdverlenging basisonderwijs]. Document PO-2009/117098, Ministry of Education, Culture and Science of the Netherlands, 2009. Retrieved from `http://www.cfi.nl/public/cfi-online/ocwregelingen/2009/04/po2009117098_onderwijstijdverlenging_bao.aspx?Zoek=JA`.

OECD. Measuring student knowledge and skills: A new framework for assessment. Technical report, Paris, France: OECD Publications Service, 1999. Retrieved from `http://www.pisa.oecd.org/dataoecd/45/32/33693997.pdf`.

E.A. Patall, H. Cooper, and A.B. Allen. Extending the school day or school year: A systematic review of research (1985-2009). *Review of Educational Research*, 80(3):401–436, 2010.

K.B. Robin. *The effects of extended-day, extended-year preschool on learning in literacy and mathematics*. Doctoral dissertation, Rutgers: The State University of New Jersey, GSAPP, 2005. Available from Dissertations and Theses database (UMI No. 3233695).

K.B. Robin, E.C. Frede, and W.S. Barnett. Is more better? The effects of full-day vs. half-day preschool on early school achievement. NIEER Working Paper, National Institute for Early Education Research, 2006. Retrieved from `http://nieer.org/docs/index.php?DocID=144`.

D.B. Rubin. Bias reduction using mahalanobis-metric matching. *Biometrics*, 36(2):293–298, 1980.

L.J. Schweinhart. Benefits, costs, and explanation of the High/Scope Perry preschool program. In *Paper presented at the Meeting of the Society for Research in Child Development*, pages 1–10, Tampa, FL, April 26 2003.

C. Scott-Little, M.S. Hamann, and S.G. Jurs. Evaluations of after-school programs: A meta-evaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation*, 23:387–419, 2002.

G. Staphorsius, R. Krom, F. Kleintjes, and N. Verhelst. Reading comprehension tests: Report of the calibration-, validation-, and standardization-study [Dutch: Toetsen Begrijpend Lezen: Verslag van het kalibratie-, validerings- en normeringsonderzoek]. Technical report, Arnhem, the Netherlands: CITO, 2004. Retrieved from `http://toetswijzer.kennisnet.nl/html/tg/8.pdf`.

J.H. Stock and M. Yogo. *Testing for weak instruments in linear IV regression*, chapter 5, pages 80–108. Identification and inference for econometric models: Essays in honor of Thomas Rothenberg. Cambridge: Cambridge University Press, 2005.

R.L. Williams. A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56:645–646, 2000.

J.M. Wooldridge. *Instrumental variables and two stage least squares*, chapter 15, pages 506–545. Introductory econometrics: A modern approach. Mason, OH: South-Western Cengage Learning, 4 edition, 2009.

TIER WORKING PAPER SERIES
TIER WP 11/02