Economics of Education Review 36 (2013) 1–11

Contents lists available at SciVerse ScienceDirect

Economics of Education Review

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

The effectiveness of extended day programs: Evidence from a randomized field experiment in the Netherlands☆

Erik Meyer *, Chris Van Klaveren

Maastricht University, Top Institute for Evidence Based Education Research (TIER), P.O. Box 616, 6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 31 August 2012
Received in revised form 3 April 2013
Accepted 15 April 2013

JEL classification:
I21

Keywords:
Extended day
Increased instructional time
Random assignment
Field experiment

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 language achievement. The empirical results show that this three-month program had no significant effect on math or language achievement.

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 language.

While not always explicitly mandated by these policies, instructional time allocated to core subjects is frequently increased in order to improve achievement (Levin & Tsang, 1987). Well known examples of such policies are the *No Child Left Behind* act in the US (Bush, 2001), the *Future for Education and Care* program[1] in Germany (see section 'Development of All-Day School' in Freitag & Schlicht, 2009), and the *Extended School Times* project[2] in the Netherlands (OCW, 2009).

The empirical literature on the effects of extended school days on student achievement can be divided into three main categories. First of all, there are studies that relate instructional time differences to differences in student achievement (e.g. Fitzpatrick, Grissmer, & Hastedt, 2011; Lavy, 2010). Second, there are studies that exploit policy changes to examine how student achievement is affected by changes in instructional time. Bellei (2009), for instance, uses a difference-in-difference strategy to identify the effect of increasing instructional time from half a school day to a full school day on math and language achievement

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

* Corresponding author. Tel.: +31 433884466.
E-mail addresses: e.meyer@maastrichtuniversity.nl (E. Meyer), cp.vanklaveren@maastrichtuniversity.nl (C. Van Klaveren).

[1] Provides funding for all-day schools, 'Ganztagsschulen'.
[2] Provides funds for summer schools, weekend schools and extended day programs.

0272-7757/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.econedurev.2013.04.002

for high school students in Chile. Bellei's (2009) results indicate that the policy had a small positive effect on language achievement. The estimated effect on math achievement, while also positive, was not robust to the specification of different control groups. Third, there are studies that evaluate the effect of specific programs that increase instructional time on student achievement. Programs can be extended day (or year) programs or out-of-school-time programs. Extended day programs are usually organized by the school, using school facilities, and during (extended) school hours. Out-of-school-time programs take place outside of school hours, and are commonly after-school programs or summer school programs. Furthermore, we can distinguish between randomized and non-randomized studies. For example, Zimmer, Hamilton, and Christina (2010) report on the evaluation of two out-of-school tutoring programs in Pittsburgh public schools; a supplemental education services (SES) program and an educational assistance program (EAP). Zimmer et al. (2010) use a fixed effects model to estimate the effect of these programs on math and reading achievement for participants. Their results indicate that participation in both programs or only in SES has a positive effect on math achievement, but not on reading. Participation in EAP results in a small gain for both math and reading.

Zimmer et al. (2010) note that, ideally, a randomized design would be used to examine program effects on achievement. Cook (2002) emphasizes that although randomized experiments provide both a more efficient and unbiased estimate of the causal program effect than quasi-experiments, educational evaluators rarely use them. Indeed, reviews indicate that the literature on extended day programs 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, Charlton, Valentine, & Muhlenbruck, 2000; Lauer et al., 2006; Scott-Little, Hamann, & Jurs, 2002). In the decade since Cook's (2002) examination however, 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 experiments, natural experiments, and regression-discontinuity designs. It is worth discussing the results of James-Burdumy et al. (2005) and Robin, Frede, and Barnett (2006) in more detail because the research question, sample population, research design, and outcome measures of these studies are similar to those of the current study. Both studies conduct a randomized experiment to estimate the effects of increased instructional time on academic outcomes for the US. The first is a final report on the evaluation of the 21st Century Community Learning Centers (21st 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 1748 elementary school students at 26 centers to a treatment and a control group. Treatment students participated in the 21st Century program, while control students could not participate in the 21st Century program but were otherwise free to participate in other after-school programs. During their two year evaluation period, centers were open 3 h a day, four or five days a week, and treatment students spent an average of 81 days at the center within the two year period. Students spent 1 h on homework, one hour on another academic activity, and 1 h 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 21st 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 8 h a day, five days a week, ten months a year, while the half-day programs operated for two and a half to 3 h 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 0.40 standard score points a month in vocabulary score on average, and that program students gained an additional 0.21 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 0.35 standard score points a month, and program students gained an additional 0.35 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 0.12 standard

deviations at the end of preschool to 0.24 standard deviations at the end of kindergarten, and up to 0.27 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 0.08 standard deviations at the end of preschool, and increased to a significant 0.20 standard deviations at the end of kindergarten, and 0.34 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, Cooper, and Allen (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 and language 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 ($mean = 10.6$, $sd = 0.95$), were offered an extended day program consisting, on average, of an additional 2 h of language instruction, 2 h of math instruction, and 1 h of excursions per week. The program received 95% of its funding from the Ministry of Education, Culture, and Science of the Netherlands, and was offered free of charge to students. Program objectives included raising language and math achievement, as well as raising student motivation. The first two objectives, raising language and math achievement, were also mandated as objectives by the subsidy scheme.

The program was housed in one of the participating schools in the neighborhood. While this was an external location for some students, we consider this program to be an extended day program rather than an after-school program because students were taught together with regular-day class mates that were also assigned to the program, and because the program was organized by a group of cooperating schools.

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 (out of 188 total 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 & Willet, 2011). Classes consisted of approximately 10 students from different elementary schools.[3] 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 student 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

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

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 2 h were offered.

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 h a year, an average of 23.5 h a week for the 40 week school year (Eurydice, 2010). Teachers report that they spend around 5 h per week on language development and math each. The effects of an extra 2 h of math or language 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 2500 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 language achievement using standardized tests that are commonly used in Dutch elementary education (Janssen, Verhelst, Engelen, & Scheltens, 2010; Staphorsius, Krom, Kleintjes, & Verhelst, 2004). Language achievement involved tests of reading comprehension, vocabulary, and spelling. 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 language 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–100th percentile, and E is the lowest score, representing the 0–10th percentile. Students with a C, D, or E score below the 50th percentile (i.e., they perform at a level that is below the national average level). To have an idea how participants perform compared to the national average, we present the percentage of students that scored above the 50th percentile in this section.[4] 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.[5] Of the 188 students who were assigned to the treatment and the control group, 153 completed the math pre- and post-tests, 99 completed vocabulary pre- and post-tests, 94 completed reading pre- and post-tests, and 88 completed spelling pre- and post-tests. Of the 188 students, 19 failed to complete pre- and post-tests for any subject, leaving 169 students that completed at least one test. The tables in this section show descriptives for these 169 students.

Given the heterogeneity of programs and effect sizes in the literature we were unsure what effect size to expect from the studied program. We conducted a priori power calculations using Optimal Design Plus software (OD+; Raudenbush et al., 2011). OD+ estimates power at varying levels of effect size, sample size, and R^2. We knew we had many covariates available, including pre-test achievement measures. Pre-test achievement usually correlates highly with post-test achievement (earlier research found $r = 0.75$ for this achievement test; Van Klaveren, 2011), and thus contributes substantially to a high R^2. Using OD+ with an expected effect size of 0.20 *sd*, an expected R^2 of 0.70, and 188 participants yields an estimated power of 0.70.

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. The characteristic *One-parent family* indicates the proportion of students that belong to a one-parent family, and *Class size* indicates the number of students in a regular-day class.

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, and not all schools administer vocabulary and spelling tests, which we indicate with a dash in the table. 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 achievement levels in our sample are often substantially below the national average achievement levels (i.e. the 50th percentile). So the experimental schools are characterized by a high proportion of students that achieve below national levels in one or more subjects.

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,

[4] Additionally, raw test scores are dependent on grade level, and as such they provide little information when averaged over grade levels.

[5] 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)
Parents' 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)
One-parent family	0.07 (0.27)	0.29 (0.46)	0.35 (0.49)	0.29 (0.47)	0.05 (0.22)	0.46 (0.51)	0.00 (0.00)	0.21 (0.41)
Class size	7.50 (0.52)	19.00 (0.00)	12.50 (2.56)	14.00 (0.00)	13.46 (2.44)	12.00 (0.00)	10.90 (2.94)	13.63 (3.76)
Math pre-test >50th pct	29%	58%	25%	7%	36%	21%	30%	34%
Reading pre-test >50th pct	14%	34%	20%	0%	44%	17%	15%	25%
Vocabulary pre-test >50th pct	21%	39%	55%	–	85%	–	65%	67%
Spelling pre-test >50th pct	86%	76%	60%	–	28%	–	45%	66%
Number of obs.	14	38	20	14	39	24	20	169

Note: Standard deviations in parentheses. Numbers of observations for vocabulary and spelling scores can be lower than indicated due to missing values.

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 shows that the treatment and control group have very similar means on the matching variables. The treatment group has a slightly higher proportion of ethnic minorities and a slightly higher percentage of students that score above the 50th percentile on the achievement tests (excluding spelling). The control group has a slightly higher proportion of girls, students with higher educated parents, one-parent families, and a slightly higher percentage of students that score above the 50th percentile in spelling achievement. Our estimation strategy, described later in this section, will correct for these small differences.

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)
One-parent family	0.16 (0.37)	0.27 (0.44)
Class size	13.63 (3.68)	13.64 (3.87)
Math pre-test >50th pct	38%	29%
Reading pre-test >50th pct	27%	24%
Vocabulary pre-test >50th pct	69%	65%
Spelling pre-test >50th pct	64%	67%
Number of obs.	86	83

Note: Standard deviations in parentheses. Numbers of observations for vocabulary and spelling scores can be lower than indicated due to missing values.

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 students 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 ($P = 0$).

Intuitively, 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 & 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 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

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)
Parents' 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)
Class size	13.82 (3.50)	14.24 (3.98)	13.24 (4.05)	13.82 (3.50)
Math pre-test >50th pct	37%	33%	41%	12%
Reading pre-test >50th pct	23%	27%	34%	12%
Vocabulary pre-test >50th pct	67%	61%	72%	–
Spelling pre-test >50th pct	63%	59%	66%	–
Number of obs.	57	66	29	17

Notes: Standard deviations in parentheses. Numbers of observations for vocabulary and spelling scores can be lower than indicated due to missing values, a dash indicates that too few observations were available for that subgroup.

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 & Pischke, 2009).

We estimate the local average treatment effect (LATE; Imbens & Angrist, 1994) using a two-stage least squares regression (2SLS; e.g. see Angrist & 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:

$$D_i = \pi_0 + \pi_1 Z_i + X_i'\pi_2 + \upsilon_i. \quad (1)$$

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:

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

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 υ_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 Eqs. (1) and (2) by using the STATA ivreg2 module, which computes the correct values of these statistics.[6] 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 & 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 language. The first model estimates the effect of receiving a (randomly assigned) voucher on math and language 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 & Willet, 2011). 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.[7] Means are presented only for students whose pre- and post-test scores are available. Test score differences in score between treatment and control group are not significant at the 5% level. It should be noted that, because sample size is limited and there are

[6] Version 03.0.06 for STATA MP 11.2

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

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)
Vocabulary pre-test	88.941 (16.893)	84.979 (21.382)	87.020 (19.205)
Vocabulary post-test	93.961 (18.722)	91.208 (23.012)	92.626 (20.850)
Spelling pre-test	130.783 (7.155)	129.595 (8.249)	130.216 (7.675)
Spelling post-test	134.804 (6.682)	133.191 (8.232)	134.034 (7.462)

Note: Standard deviations in parentheses. Mean math scores are based on 153 observations, mean reading scores are based on 94 observations, mean vocabulary scores are based on 99 observations, mean spelling scores are based on 88 observations.

Table 5
First stage and ITT for math.

	First stage dependent: extended day participation		Intent-to-treat (ITT) dependent: math post-test	
	(1)	(2)	(3)	(4)
Extended day assignment	0.444*** (0.097)	0.456*** (0.095)	0.094 (0.062)	0.087 (0.067)
Math pre-test	−0.001 (0.004)	−0.001 (0.004)	0.063*** (0.004)	0.063*** (0.004)
Girl		−0.060 (0.079)		−0.012 (0.070)
Ethnic minority		0.031 (0.111)		0.105 (0.120)
Parents' education		−0.083 (0.060)		0.050 (0.069)
One-parent family		0.173 (0.155)		−0.074 (0.067)
Class size		−0.017 (0.013)		−0.005 (0.016)
Constant	0.423 (0.343)	0.666 (0.420)	−5.460*** (0.415)	−5.410*** (0.490)
Controls	No	Yes	No	Yes
	$N = 153$	$N = 153$	$N = 153$	$N = 153$
	$F(4,12) = 19.53$	$F(9,12) = 27.04$	$F(4,12) = 120.38$	$F(9,12) = 90.65$
	$R^2 = 0.23$	$R^2 = 0.26$	$R^2 = 0.81$	$R^2 = 0.81$

Notes: Standard errors (SE) in parentheses. All model specifications include dummy variables for grade level, and cluster SE's by class (13 clusters).
*** Statistically significant at the 1% level.

no covariates involved, this analysis is somewhat underpowered. That said, the achievement levels of control and treatment students seem comparable at the start of the program.

Table 5 shows how program assignment affects program participation (i.e. the first stage results) and shows 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.[8] The intent-to-treat estimates show how receiving an extended day voucher affects math achievement. For ease of interpretation, the post-test variable is standardized to mean zero, standard deviation one.

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, $AP\ \chi^2 = 23.36$ and 26.40 for models (1) and (2) respectively, and Stock–Yogo (SY) weak identification tests show that our instruments are not weak (Stock & Yogo, 2005). 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 reports the 2SLS results derived from these first stage and reduced form estimates.[9] The 2SLS estimates of the effect of the extended day program on math achievement range from 0.190 to 0.212, 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 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

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

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

Table 6
OLS and 2SLS estimates for math.

	OLS		IV/2SLS	
	(1)	(2)	(3)	(4)
Extended day participation	−0.093 (0.057)	−0.098 (0.065)	0.212 (0.130)	0.190 (0.136)
Math pre-test	0.065*** (0.005)	0.066*** (0.005)	0.063*** (0.004)	0.063*** (0.004)
Girl		0.002 (0.090)		−0.001 (0.063)
Ethnic minority		0.080 (0.116)		0.099 (0.125)
Parents' education		0.045 (0.086)		0.066 (0.063)
One-parent family		−0.033 (0.056)		−0.106 (0.065)
Class size		−0.047*** (0.015)		−0.002 (0.017)
Constant	−5.486*** (0.487)	−4.997*** (0.480)	−5.550*** (0.461)	−5.537*** (0.547)
Controls	No	Yes	No	Yes
School fixed effects	Yes	Yes	No	No
	$N = 153$	$N = 153$	$N = 153$	$N = 153$
	$F(4,12) = 114.95$	$F(9,12) = 122.99$	$F(4,12) = 118.43$	$F(9,12) = 138.00$
	$R^2 = 0.82$	$R^2 = 0.83$	$R^2 = 0.78$	$R^2 = 0.79$

Notes: Standard errors (SE) in parentheses. All model specifications include dummy variables for grade level, and cluster SE's by class (13 clusters). OLS models, i.e. (1) and (2), include school fixed-effects, the 2SLS models do not because assignment is within classes.
*** Statistically significant at the 1% level.

parents' education positively impacts student achievement (Holmlund, Lindahl, & Plug, 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 makes 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.[10]

The 2SLS estimates of around 0.20 can be converted into an effect size (Cohen's *d*) of approximately 0.12 standard deviations (*sd*). This means that, conditionally on their pre-test score, a program participant's post-test score will increase by 12% 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% of that is approximately 1.54 points. The difference between pre- and post-test means is 5.07 points, and represents 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 0.12 *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.[11] 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 in the lowest quartile of the pre-test score, or students in small classes.

Table 7 presents the first stage and intent-to-treat estimates for language achievement in identical fashion to Table 5. The language models use test scores for comprehensive reading, vocabulary, and spelling, and include the same covariates as the reading models. Post-test scores are standardized to mean zero, standard deviation one.

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. Angrist–Pishke and Stock–Yogo tests show that our models are not underidentified and our instruments are not weak. As was the case with math, the ITT results show that being randomly assigned to the program does not have a significant effect on language achievement. Again, the first stage and ITT estimates are robust to the addition of more covariates to the model.

Table 8 reports the 2SLS and OLS results. The 2SLS estimates show the extended day program did not significantly affect language achievement. Contrary to the math results presented in Table 6, these estimates are close to zero. The, highly similar, estimates produced by ITT and OLS further support the conclusion that the program had no effect on language skills.

Subgroup analyses for language indicate that the 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 in the lowest quartile of the pre-test score, or students in small classes.

As explained in Section 3, not all students completed pre- and post-test for every subject. If missing data are non-randomly distributed over treatment and control students, this could bias our estimates. This is not the case however, as test data are mostly missing for entire classes

[10] 2SLS standard errors of extended day participation were slightly higher when unadjusted for clustering.
[11] 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.

Table 7
First stage and ITT for language.

	First stage dependent: extended day participation		Intent-to-treat (ITT) dependent: language post-test	
	(1)	(2)	(3)	(4)
Extended day assignment	0.515*** (0.103)	0.522*** (0.102)	0.022 (0.086)	0.005 (0.081)
Language pre-test	−0.002 (0.003)	−0.002 (0.003)	0.043*** (0.005)	0.042*** (0.004)
Girl		−0.022 (0.070)		0.152** (0.068)
Ethnic minority		0.047 (0.087)		−0.028 (0.081)
Parents' education		−0.045 (0.070)		0.215** (0.074)
One-parent family		0.101 (0.125)		−0.250** (0.111)
Class size		−0.009 (0.007)		0.012 (0.011)
Constant	0.306** (0.131)	0.419** (0.147)	−1.500*** (0.232)	−1.753*** (0.251)
	$N = 281$	$N = 281$	$N = 281$	$N = 281$
	$F(6,14) = 17.69$	$F(11,14) = 26.66$	$F(6,14) = 54.00$	$F(11,14) = 150.51$
	$R^2 = 0.30$	$R^2 = 0.31$	$R^2 = 0.55$	$R^2 = 0.58$

Notes: Standard errors (SE) in parentheses. All model specifications include dummy variables for language tests and grade level, and cluster SE's by class (15 clusters).
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

Table 8
OLS and 2SLS estimates for language.

	OLS		IV/2SLS	
	(1)	(2)	(3)	(4)
Extended day participation	0.016 (0.056)	0.025 (0.064)	0.043 (0.159)	0.010 (0.146)
Language pre-test	0.040*** (0.005)	0.040*** (0.005)	0.043*** (0.005)	0.042*** (0.004)
Girl		0.154* (0.078)		0.152** (0.064)
Ethnic minority		−0.043 (0.081)		−0.028 (0.075)
Parents' education		0.172** (0.070)		0.216*** (0.068)
One-parent family		−0.199 (0.131)		−0.251** (0.111)
Class size		−0.007 (0.012)		0.012 (0.011)
Constant	−1.216*** (0.188)	−1.227*** (0.295)	−1.513*** (0.209)	−1.757*** (0.233)
School fixed effects	Yes	Yes	No	No
	$N = 281$	$N = 281$	$N = 281$	$N = 281$
	$F(6,14) = 70.86$	$F(11,14) = 268.45$	$F(6,14) = 54.13$	$F(11,14) = 148.79$
	$R^2 = 0.60$	$R^2 = 0.61$	$R^2 = 0.55$	$R^2 = 0.58$

Notes: Standard errors (SE) in parentheses. All model specifications include dummy variables for language tests and grade level, and cluster SE's by class (15 clusters). OLS models, i.e. (1) and (2), include school fixed-effects, the 2SLS models do not because assignment is within classes.
* Statistically significant at the 10% level
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

or entire schools, and randomization took place within classes. Appendix A presents sensitivity analyses that use only matched pairs of whom both students completed pre- and post-tests, and the unmatched students from classes with an uneven number of students (who were individually randomly assigned), thus excluding students that belong to a matched pair of whom one student had complete test data and the other did not.

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 language achievement for the compliers.

The empirical results suggest that receiving an extended day voucher does not influence students' math and language achievement. Also, participation in the extended day program does not significantly influence students' math and language achievement. We can only speculate as to why this program was not effective. While the curriculum has a theoretical foundation, 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 similar programs with longer durations can also be ineffective. Alternatively, it is possible that the educational practices outlined by Marzano (2003) are less effective for the current sample or within the Dutch educational system.

Table A.1
Comparison of IV/2SLS analyses for full samples vs. complete-only samples.

	Math		Language	
	Full sample	Complete-only sample	Full sample	Complete-only sample
Extended day participation	0.190 (0.136)	0.226 (0.153)	0.010 (0.146)	−0.005 (0.150)
Pre-test	0.063*** (0.004)	0.063*** (0.005)	0.042*** (0.004)	0.042*** (0.004)
Girl	−0.001 (0.063)	−0.026 (0.051)	0.152** (0.064)	0.172** (0.067)
Ethnic minority	0.099 (0.125)	0.066 (0.130)	−0.028 (0.075)	−0.026 (0.078)
Parents' education	0.066 (0.063)	0.047 (0.064)	0.216*** (0.068)	0.221*** (0.068)
One-parent family	−0.106 (0.065)	−0.109 (0.071)	−0.251** (0.111)	−0.255** (0.112)
Class size	−0.002 (0.017)	−0.002 (0.018)	0.012 (0.011)	0.013 (0.011)
Constant	−5.537*** (0.547)	−5.480*** (0.578)	−1.757*** (0.233)	−1.802*** (0.256)
	$N = 153$	$N = 145$	$N = 281$	$N = 271$
	$F(9,12) = 138.00$	$F(9,12) = 141.85$	$F(11,14) = 148.79$	$F(11,14) = 109.37$
	$R^2 = 0.79$	$R^2 = 0.79$	$R^2 = 0.58$	$R^2 = 0.56$

Notes: Standard errors (SE) in parentheses. Where relevant, model specifications include dummy variables for language tests and grade level, and cluster SE's by class (13/15 clusters for math/language).
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

Our estimates suggest that the program did not have a significant effect on students' math achievement. 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. The estimate regarding math achievement of five weeks for an 11 week program seems substantive, however, the appeal of such a program to policy makers would also depend on the cost-effectiveness. In addition to encouraging more causal evaluations of extended day programs therefore, we would also encourage cost-effectiveness studies of extended day programs. Such studies could help policy makers compare extended day programs to other approaches in terms of cost-effectiveness.

The current study examines students living in a small city in the Netherlands. While our results are likely to generalize to similar populations in the Netherlands, there are a number of cultural and social differences between our research population and those of studies in other countries that can impede generalization to a broader range of contexts. It is important, therefore, that programs and evaluations are replicated in other contexts before definitive conclusions are reached. That said, our estimates add to the neutral to small positive effects described by Patall et al. (2010) for extended day programs. 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 effective extended day programs to be found, our results, and those of others, suggest that they are the exceptions.

Appendix A

This appendix test the sensitivity of our IV estimates to missing test score data. We repeat the model 4 analyses of Tables 6 and 8 and exclude student pairs if one of the students had incomplete test information. Table A.1 therefore presents the estimation results for student pairs with complete test score data (columns 2 and 4), together with the results from Tables 6 and 8.

A large portion of test scores for certain subjects (e.g. spelling) are missing because schools did not administer that test (for that particular grade). It follows that there are many student pairs who either complete all tests or who did not complete any test on a particular subject. It happens occasionally that one student completed the test while the other student in a student pairing did not. To be more specific, there are eight such pairs for math, zero for reading, seven for vocabulary, and three for spelling. Because of that, the reduction in sample size for the sensitivity analyses is quite small. It is therefore not surprising that the estimates of the sample with complete test score information are very similar to those of the full sample.

References

Angrist, J. (2006). Instrumental variables methods in experimental criminological research: What, why and how. *Journal of Experimental Criminology, 2*(1), 23–44.

Angrist, J., & Pischke, J. (2009, chap. 4). *Instrumental variables in action: Sometimes you get what you need. Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bellei, C. (2009). Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile. *Economics of Education Review, 28*(5), 629–640.

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

Character Connection. (2007). *Character Connection: Parent, child, and school*. http://www.characterconnectionprogram.com.

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

Cook, T. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Cooper, H., Charlton, K., Valentine, J., & Muhlenbruck, L. (2000). 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.

Eurydice. (2010). *National summary sheets on education systems in Europe and ongoing reforms: The Netherlands. Technical report*. European Commission. http://www.eurydice.org.

Fitzpatrick, M., Grissmer, D., & Hastedt, S. (2011). What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review, 30*(2), 269–279.

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

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

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

James-Burdumy, S., Dynarski, M., Moore, M., Deke, J., Mansfield, W., & Pistorino, C. (2005). *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. http://www. ed.gov/ies/ncee.

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *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. http://toetswijzer.kennisnet.nl/html/tg/14.pdf.

Lauer, P., Akiba, M., Wilkerson, S., Apthorp, H., Snow, D., & Martin-Glenn, M. (2006). Out-of-school-time programs: A meta-analysis of effects for at-risk students. *Review of Educational Research, 76*(2), 275–313.

Lavy, V. (2010). *Do differences in school's instruction time explain international achievement gaps in math, science, and reading? Evidence from developed and developing countries. NBER Working Paper 16227* http://www.nber.org/papers/w16227.

Levin, H., & Tsang, M. (1987). The economics of student time. *Economics of Education Review, 6*(4), 357–364.

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

Murnane, R., & Willet, J. (2011, chap. 4). *Investigator-designed randomized experiments. Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

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

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

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

Raudenbush, S., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., et al. (2011). *Optimal design plus empirical evidence (version 3.0)*.

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

Robin, K., Frede, E., & Barnett, W. (2006). *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. http://nieer.org/docs/index.php?DocID=144.

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

Schweinhart, L. (2003). Benefits, costs, and explanation of the High/Scope Perry preschool program. *Paper presented at the Meeting of the Society for Research in Child Development* (pp. 1–10).

Scott-Little, C., Hamann, M., & Jurs, S. (2002). Evaluations of after-school programs: A meta-evaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation, 23*(4), 387–419.

Staphorsius, G., Krom, R., Kleintjes, F., & Verhelst, N. (2004). *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. http://toetswijzer.kennisnet.nl/html/tg/8.pdf.

Stock, J., & Yogo, M. (2005, chap. 5). *Testing for weak instruments in linear IV regression. Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*. Cambridge: Cambridge University Press.

Van Klaveren, C. (2011). Lecturing style teaching and student performance. *Economics of Education Review, 30*(4), 729–739.

Williams, R. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics, 56*(2), 645–646.

Wooldridge, J. (2009, chap. 15). *Instrumental variables and two stage least squares. Introductory econometrics: A modern approach* (4th ed., pp. 506–545). Mason, OH: South-Western Cengage Learning.

Zimmer, R., Hamilton, L., & Christina, R. (2010). After-school tutoring in the context of no Child Left Behind: Effectiveness of two programs in the Pittsburgh Public Schools. *Economics of Education Review, 29*(1), 18–28.