CPB Netherlands Bureau for Economic Policy Analysis

CPB Discussion Paper | 264

The impact of a comprehensive school reform policy for failing schools on educational achievement;

Results of the first four years

Roel van Elk
Suzanne Kok

The impact of a comprehensive school reform policy for failing schools on educational achievement;

Results of the first four years[1]

Roel van Elk[2]
CPB, Netherlands Bureau for Economic Policy Analysis
r.a.van.elk@cpb.nl

Suzanne Kok
CPB, Netherlands Bureau for Economic Policy Analysis
s.j.kok@cpb.nl

Abstract

This paper estimates the effects of a comprehensive school reform program on high-stakes test scores in Amsterdam. The program implements a systematic and performance-based way of working within weakly performing primary schools and integrates measures such as staff coaching, teacher evaluations and teacher schooling, and the use of new instruction methods. Difference-in-differences estimates show substantial negative effects on test scores for pupils in their final year of primary school. The program decreased test scores with 0.17 standard deviations in the first four years after its introduction. A potential explanation for this finding is the intensive and rigorous approach that caused an unstable work climate with increased teacher replacement.

JEL Codes: I2
Keywords: comprehensive school reform, educational achievement, difference-in-differences

[1] The authors would like to thank Chris van Klaveren, Dinand Webbink, Bas ter Weel, Debby Lanser, Ted Reininga, Jacqueline Visser and seminar participants at the Ministry of Education and CPB The Hague for their valuable comments. The authors also thank the Dutch Inspectorate of Education, the CITO organisation, and the municipality of Amsterdam for supplying the data used in this paper. The data and programs that the authors use for the analysis in this paper are available upon request at r.a.van.elk@cpb.nl.

[2] Corresponding author.

1. Introduction

The improvement of weakly performing schools is an important issue in many countries. Comprehensive school reform (CSR) methods have been widely used to turn around failing schools.[3] These programs involve integrated changes at all levels within schools rather than incremental changes targeted at single aspects. The school is considered as the level of improvement and a program's content is tailored to its specific needs.[4] CSR models typically include various elements such as professional development of educators, an increased attention to instruction methods and the individual needs of pupils, improvement in classroom or school management, parental involvement, curriculum improvements and setting high achievement goals. Proponents of CSR methods argue that comprehensive changes are needed since the impact of isolated interventions can be distorted by dysfunction in other areas (Borman et al., 2004). However, the effectiveness of CSR programs in increasing student outcomes is by and large unclear.

The purpose of this paper is to estimate the causal effects of a CSR policy on educational performance. We investigate the impact of the Amsterdam School Improvement Program (ASIP) in the Netherlands, which was introduced in 2008. The goal of the program is to improve the educational quality of failing primary schools in Amsterdam. The ASIP is an intensive two-year program that aims to implement a systematic and performance-based way of working. This includes data-driven teaching with educators that systematically measure pupil performance. This information is used to respond to the individual needs of the pupils. The ASIP applies school-specific improvement plans that typically integrate measures such as staff coaching, teacher quality evaluations, teacher schooling, and the use of new instruction methods. The implementation of each improvement plan is guided by an expert team with strong educational experience.

As of April 2008, all primary schools in Amsterdam that were judged to perform below national quality standards were invited to voluntarily participate in the program. To examine the impact of the program we compare the development of pupil achievement in failing schools in Amsterdam to that in failing schools outside Amsterdam. Our assessment of educational achievement is based on the CITO test. This is a nationwide, high-stakes test that pupils take in the highest grade (eight) of primary education. The test includes questions on language, math, and information processing. We make use of administrative data on CITO test scores from 2005 to 2012, which enables us to compare the change in performance in Amsterdam before and after the introduction of the CSR policy with the change in performance in other cities that did not introduce the CSR policy.

[3] The U.S. government has devoted over 2 billion dollars to the implementation of CSR programs in the 1990s and early 2000s (U.S. Department of Education, 2004; 2006).

[4] Over 800 variations of CSR models have been implemented in more than 5,000 schools in the US in the past decades (Rowan et al., 2004).

This study relates to the literature on the effects of CSR models. Borman et al. (2003) provide an extensive overview of existing studies with respect to 29 of the most widely implemented CSR models in the U.S. Although they find overall promising results, the authors conclude that both quality of research designs and quantity of studies are insufficient to draw strong conclusions on the effectiveness of CSR models. More recent studies show ambiguous effects of CSR models. Whereas some studies find positive effects on student performance (e.g. May and Supovitz, 2006), others find no significant effects (Gross et al., 2009; Bifulco et al., 2005) or mixed results across grades and subjects (Schwartz et al., 2004). While the overall record of CSR models appears to be encouraging, most results come from studies that do not use credible research designs to handle potential selection problems. Schools that adopt a comprehensive school model are likely to differ from other schools in a number of aspects, such as student composition, educational quality, or desire for innovation. Even if one controls for student characteristics or school fixed effects, estimated effects may still be biased because of unobserved heterogeneity.

Our main contribution is that we use a quasi-experimental research design to address potential endogeneity. We estimate difference-in-differences models to identify the effects of the introduction of the CSR policy. Difference-in-differences models have been frequently used in previous economic evaluation studies (see e.g. Ashenfelter and Card, 1985; Card and Krueger, 1994; Blundell et al., 1998; Jacob, 2005). We present intention-to-treat estimates and find substantially negative effects on test scores in the first four years after the introduction of the program. This result is robust to a variety of sensitivity analyses. In our preferred specification the introduction of the ASIP decreases test scores by 0.17 standard deviations. The detrimental impact on test scores is largest for language and is generally larger at the lower part of the test score distribution. Interviews with school-leaders of participating schools provide a candidate explanation for our findings. The rigorous and demanding approach appears to have caused an increase in teacher replacement. The resulting loss of school specific knowledge, increase in recruitment and hiring costs, and uncertain work atmosphere felt by teachers may have negatively affected pupil achievement. The results concern the first four years after the start of the policy. We cannot exclude that our findings reflect adjustment costs during the transition from a failing to a successful school and that it takes longer before beneficial effects become manifest. In any case, we conclude that the introduction of the ASIP induced major costs in terms of substantial test score losses for at least four cohorts of pupils.

The rest of this paper is organised as follows. Section 2 provides a brief overview of previous studies. Section 3 describes the ASIP. Sections 4 and 5 discuss the empirical strategy and the data. Section 6 presents the results and Section 7 discusses potential mechanisms that could explain our findings. Section 8 concludes.

2. Previous studies

The literature on CSR models largely consists of practitioner-oriented studies (see e.g. Herman et al., 1999; Traub, 1999; Slavin and Fashiola, 1998). Although some of these studies provide an assessment of CSR models, none of them makes use of control groups to identify causal effects of CSR models on student achievement. Borman et al. (2003) provide an overview of existing studies with respect to 29 of the most widely implemented CSR models in the U.S. Considering only studies that are performed by an independent third party and that make use of some form of control groups, the strongest evidence for positive effects is provided by three programs: the Direct Instruction Program (DIP), the School Development Program (SDP) and Success for All (SFA). Remarkably, two of these, DIP and SFA, are relatively narrow-targeted interventions mainly focusing on better instruction methods and curriculum improvements. For one program, Edison, statistically significant negative effects have been found. This program intended to create innovative schools with a challenging curriculum, instruction methods tailored to the needs of the pupils and an emphasis on computer technology. In addition, the authors report large heterogeneity in the estimated effects between CSR models that cannot be explained by the differences in the specific measures included in the program. This suggests that school-specific requirements and/or the level and quality of implementation are more important for determining success. The average school across all studies reviewed had implemented its CSR model for around three years. The authors point out that, if cumulative effects exist, the analyses may underestimate the impact of CSR models. Although they find overall effects that appear promising, Borman et al. (2003) conclude that both quantity and quality of studies are insufficient to draw reliable conclusions on the effectiveness of CSR models yet. They advocate new programs to be evaluated making use of (quasi-) experimental research designs, to obtain more evidence-based knowledge on the effects of CSR.

Some more recent, also non-experimental, studies show ambiguous results. In an 11-year longitudinal study May and Supovitz (2006) evaluate the impact of a CSR design, called 'America's Choice', on student test performance in Rochester, New York. The authors find significant positive effects on student performance which accumulate over time. The impact of the reform seemed to be larger in later grades than in the early grades. The authors argue that the positive impact of the program is likely to be caused by instruction methods targeted towards the needs of individual students, ambitious expectations for student performance and a supporting organisational structure within schools that facilitate this tailored way of working. Gross et al. (2009) investigate the effects of federal CSR funds on student achievement in Texas making use of student-level panel data. Since schools have to apply for these funds, selection bias is a concern for the identification of the effects. The authors deal with this issue by controlling for school fixed effects and find that CSR funds did not significantly affect student's reading performance. The effects on math performance varied across

different student types. Bifulco et al. (2005) evaluate three reform programs in New York City, including the School Development Program (SDP), Success for All (SFA) and the More Effective Schools (MES) program. The authors selected control groups by a random sample of troubled schools and then adjusted the sample so that the treatment and control groups have a similar propensity to start a CSR program. In contrast to some previous studies (see Borman et al., 2003) they do not find evidence that the SDP or SFA program significantly contributed to student performance. These findings, however, do support results from other earlier evaluations of SDP in Maryland, Chicago and Detroit (Cook et al., 1998; 1999; Millsap et al., 2001). Positive effects on reading scores are found for the MES program in the short run, but these do not persist when the external program trainers leave the school. This finding suggests that schools face difficulties to maintain progress on their own, after the end of the program. Schwartz et al. (2004) assess the impact of a CSR model on student performance in New York, called the New York Networks for School Renewal Project. Student test scores in these schools are compared to a control group of students attending a set of randomly selected New York public schools. The authors make use of three cohorts of students, who are in grades 4, 5 or 6 at the start of the program in 1995-1996. The authors find mixed effects across grades after two to three years: in grade 4 CSR significantly increased both math and reading test scores, while the effects in grade 5 are insignificant for reading and negative for math. In grade 6, the program did not significantly affect performance.

In sum, the number of studies that use a credible research design to examine the effect of CSR models on student achievement is limited and existing studies show mixed findings. Our study adds to the literature by investigating the impact of a CSR policy on high-stakes test scores in the Netherlands. The ASIP shares some of the elements of other promising programs, including the increased attention to the individual needs of pupils and the tailor-made improvement plans that are guided by external experts. Our main contribution lies in the use of a quasi-experimental difference-in-differences approach to identify the effects of the program on pupil achievement.

3. The Amsterdam School Improvement Program

3.1 Background

The Dutch Inspectorate of Education judges the quality of primary schools in the Netherlands. This governmental agency periodically investigates whether schools provide an acceptable standard of education. Each primary school is judged on a yearly basis by means of a risk analysis. The risk analysis includes several aspects such as student test results, the level of exams, personnel management, the financial position of the school, and compliance with Dutch educational laws. If the outcomes of the risk analysis provide evidence of weak performance, a more extensive quality analysis follows to determine whether schools are failing to meet the required educational quality

standards. Based on these analyses the Inspectorate of Education classifies schools as 'basic', 'weak' or 'very weak'. Schools that satisfy national quality standards are classified as basic, while schools classified as (very) weak perform below national standards. Inspection reports in 2006 and 2008 showed that the educational quality of a relatively large proportion of schools in Amsterdam was below the national standards. In Amsterdam 13 percent of all primary schools was classified as weak and 2.4 percent as very weak in 2008. The nationwide fraction of weak and very weak schools was 9.2 and 1.4, respectively (Inspectorate of Education, 2009). The primary schools in Amsterdam also performed worse compared to the ones in other large cities in the Netherlands (Inspectorate of Education, 2008). Concerns on those weakly performing schools have led to an intensive policy effort by the municipality of Amsterdam to improve educational quality (see e.g. Municipality of Amsterdam, 2009). It invested in a comprehensive school reform program that was introduced in 2008. All schools in Amsterdam that were classified as weak or very weak in the beginning of 2008 were invited to voluntarily participate in the program. After the school year 2008-2009, all primary schools in Amsterdam became eligible for the program. Hence, the program was initially targeted at failing schools, but became also accessible for sound performing schools later on.

3.2 Content

The Amsterdam School Improvement Program (ASIP) is an intensive two-year program designed to improve educational quality of participating schools. It aims to implement a systematic and performance-based way of working within the whole school. This includes 'data-driven teaching', meaning that teachers systematically measure pupil performance and use this information to adjust lessons to the individual needs of the pupils. Schooling of teachers is used to improve their classroom practices, school-leaders and other school personnel take courses to improve their skills in performance-based working, and instruction methods are often replaced. A consistent way of working throughout the school should create an efficient organisation in which teachers are optimally facilitated in their primary teaching tasks. The program is guided by an expert team with strong educational experience.[5]

The ASIP consists of three steps. First, the educational experts conduct a profound quality analysis in cooperation with the school. This analysis includes instruction methods, student performance, student care, didactical routines and management performance, including leadership, communication skills and the existence of a coherent vision of education. An important aspect of the analysis is the evaluation of teacher quality through observations of lessons. The experts use a specific teacher evaluation system to judge teacher quality on a variety of classroom practices, including pedagogical,

[5] The expert team largely consists of former inspectors of the Dutch Inspectorate of Education.

didactical, and organisational competences.[6] Second, the school sets up an improvement plan in collaboration with the experts. The improvement plan states the specific measures that have to be implemented to improve educational quality within two years. These measures are suited to the specific needs of the schools and typically involve schooling of teachers, coaching of school-leaders and the use of new instruction methods. Third, the improvement plan is implemented after it has been tested and approved by the expert team. The expert team supports the schools during the implementation period. The experts serve as critical advisors and visit the school at least once every three months to regularly assess the progress. Every six months the improvement in educational quality is measured in classes based on the teacher evaluation system. The expert team evaluates the progress and educational quality of the school more broadly and extensively during the first formal audit one year after the start of the program. If needed, the plan can be adjusted. After two years, a second audit takes place, which can be considered as the end of the program.[7]

As part of the program the municipality of Amsterdam developed instruction courses for professional development of school personnel. These accredited courses focus on the development of performance-based working skills and became available in the school year 2009-2010. [8] The courses are not restricted to educators of schools that participate in the ASIP, but are accessible for all primary schools in Amsterdam.

At the start of the ASIP, the municipality of Amsterdam also introduced new achievement goals for primary schools. The standards aimed for are above the nationwide standards of the Inspectorate of Education. [9] The announcement of the achievement goals reflects the ambitious plans of the municipality. The goals serve as a signal, but no explicit sanctions follow if a school does not satisfy the standards.

3.3 Costs and participation

The costs of the ASIP depend on the specific measures in the improvement plan. The average costs of the two-year ASIP amount to around 300,000 Euros per school, of which 250,000 Euros for the implementation of the interventions and 50,000 Euros for counseling by the expert team. Costs are shared by the municipality and the schools. The amount of resources to be paid by the school is

[6] The teacher evaluation system (TES) is called '*Kijkwijzer*'. Van der Steeg and Gerritsen (2013) find that high teacher quality scores on this TES are associated with better pupil test scores. This suggests that the TES measures teacher practices that are important for the educational performance of pupils.

[7] In exceptional cases where educational goals are not achieved yet, the program can be extended by an additional third year.

[8] These courses are mainly targeted at school-leaders and supportive school personnel.

[9] The achievement goals include (i) an average CITO test score of at least 534, (ii) at least 25% percent of the pupils assigned to higher secondary education, and (iii) at most 20% of the disadvantaged pupils assigned to specific secondary education levels that provide special care because of learning difficulties (called '*Praktijkonderwijs'* and '*Leerwegondersteunend Onderwijs'*) .

dependent on its financial position. On average, the matching percentage of the schools is around 25%. The total costs involved are substantial: on average the yearly ASIP investment is more than 10% of the total government funding for an average primary school. [10]

There are 209 primary schools in Amsterdam, of which 50 participated in the ASIP by the end of the 2010-2011school year: 16 schools started in the program during or before the 2008-2009 school year, 14 during the 2009-2010 school year, and 20 during the 2010-2011 school year. It should be noted that these concern both schools that are classified as (very) weak and schools that are classified as basic. In our main analyses we focus on the sample of weak-performing schools in the beginning of 2008, which was the initial target population of the program (see Sections 4 and 5).

4. Empirical Strategy

To assess the impact of the ASIP on educational performance, we adopt a difference-in-differences (DID) estimation approach. This approach essentially compares the change in educational performance after and before the start of the program in Amsterdam to the same change in other Dutch cities that did not implement a CSR program. We implement this strategy by estimating the following model:

$$Y_{ist} = \beta_0 + \beta_1 A_{ist} + \beta_2 T_t + \beta_3 A_{ist} * T_t + X_{ist} + \alpha_s + \tau_t + \varepsilon_{ist}, \quad (1)$$

where Y_{ist} is the test score of pupil i in school s in year t, A_{ist} is a dummy variable that takes value 1 if pupil i is at school in Amsterdam and 0 otherwise, T_t is a dummy variable that takes value 1 in case of a post-treatment year and value 0 in case of a pre-treatment year, X_{ist} is a vector with individual background characteristics, α_s are school fixed effects, τ_t are year dummies and ε_{ist} is the error term.

The estimated coefficient β_3 is the parameter of interest. For our main analysis we use data on the CITO test scores from 2005 to 2012. This is a nationwide, high-stakes test that pupils take in their final year of primary school (see Section 5). We define the years 2005 till 2008, before the start of the ASIP, as pre-treatment years. The years 2009 till 2012 are the post-treatment years. Regarding the two-year program duration, one might argue whether 2009 is an appropriate post-treatment year. Therefore, we will also present results in case of only including later post-treatment years.

The main analysis focuses on the sample of all schools in the Netherlands that were classified as 'weak' or 'very weak' by the Inspectorate of Education on 1 January 2008. These failing schools were the initial target group of the ASIP and constitute a homogeneous sample with respect to school

[10] With an average primary school size of 220 pupils, the average yearly ASIP investment per pupil is around 680 Euros. This is more than 10% of the per-pupil government funding of around 5,000 Euros.

quality according to the nationwide standards of the Inspectorate of Education. All of these schools in Amsterdam were eligible for participation in the ASIP, while similar weak-performing schools outside Amsterdam were not allowed to participate. We use all (very) weak schools outside Amsterdam as our main control group. In addition, we construct two alternative control groups that consist of only larger cities in the Netherlands. Schools in other large cities may be more similar to the schools in Amsterdam. The crucial assumption for identification of the treatment effect is the common trend assumption, which implies that the development of test scores in Amsterdam would have been the development in test scores in the control group in the absence of the ASIP. This assumption rules out city-specific trends and composition effects. To investigate the validity of the common trend assumption we compare the pre-treatment test score development in Amsterdam to the pre-treatment test score development in the control groups. We address the potential effect of the policy on the composition of schools by presenting a sensitivity analysis that includes only those schools that participated in the CITO test both before and after the introduction of the policy. Furthermore, we argue and provide supportive evidence that our estimation results are not likely to be affected by changes in the composition of pupils within schools.

The estimated treatment effect should be interpreted as the effect of the introduction of the ASIP policy. The introduction of the policy offered all failing schools in Amsterdam the opportunity to participate in the program. Since not all of the eligible schools participated in the program, we estimate an intention-to-treat (ITT) effect. The ITT effect differs from the effect of actual participation in the program. The standard approach to estimate the effect of participation for those who participate would be to use eligibility for the program as an instrumental variable for participation in the program (Imbens and Angrist, 1994; Angrist et al., 1996). Estimation by two-stage-least-squares then yields the treatment-on-the-treated effect, which is essentially equal to the ITT effect divided by the compliance rate (Bloom, 1984). This analysis assumes that eligibility for the program, i.e. being at school in Amsterdam in a post-treatment year, does not affect the outcomes of non-participating schools. This assumption is not likely to hold here since schools that do not participate in the ASIP have the opportunity to take the professional development courses from the 2009-2010 school year onwards, which may affect outcomes. It turns out that three schools in our estimation sample participated in the courses without following the complete ASIP program (see Section 5). This makes it impossible to strictly disentangle the effect of participation in the complete ASIP from participation in only the professional development courses. We therefore only present the estimated ITT effects, which pick up both effects.

The potential influence of other implemented policies in Amsterdam does not seem to be a main concern in our analysis, because of the relatively large size of the ASIP. One project with the goal of

raising pupil's math performance was implemented in the south-east district of Amsterdam in 2008.[11] Outside Amsterdam, the municipality of Rotterdam started an action program to improve the educational quality of primary schools in 2011.[12] To address the potential impact of these other programs we present sensitivity analyses in which we leave out the schools in the corresponding areas.

When calculating standard errors we take into account the presence of common group errors (Moulton, 1986). In all estimation results we present robust standard errors corrected for clustering at the school-year level. Still, standard errors might be too small in our case where we use multiple years of data. Bertrand et al. (2004) show that ignoring serial correlation in outcomes can lead to over-rejection of the null hypothesis of no effect. To address this potential issue of serial correlation, we also present estimation results of a model collapsing the data before and after the introduction of the policy (Bertrand et al., 2004).

5. Data

We received information on schools that were classified as 'weak' or 'very weak' on 1 January 2008 from the Inspectorate of Education. For this sample of schools we obtained data on CITO test scores from the CITO organisation. The CITO test is a nationwide, high-stakes test that pupils take in the highest grade of primary education (grade eight). It contains questions on language, math, and information processing. The test takes place during three days in the beginning of February and forms an important input for the assignment of pupils to different levels of secondary education. Teachers use the test results to advice pupils on the most appropriate secondary education level and secondary schools often use threshold values for enrolment in more advanced types of secondary education. Test results at the school level are used by the Inspectorate of Education to judge the quality of primary schools. Each year more than 80 percent of all primary schools participate in the CITO test. When a school chooses to participate in the CITO test, in principle all pupils in grade eight take the test. [13]

Our dataset contains information on CITO test scores at the pupil level from 2005 to 2012 for all failing schools on the reference date 1 January 2008. The total sample includes 614 schools that are comparable with respect to educational quality according to the nationwide standards of the Inspectorate of Education. It should be noted that our sample contains only schools that have

[11] This program was called '*Omdat elk kind telt in Zuidoost*'.
[12] This action program is called '*Beter Presteren'* and focuses on additional school time, professionalisation of schools and parental involvement.
[13] An exception is made for pupils in special categories such as foreign students that have been in the Netherlands for a short time and students that are expected to be assigned to secondary education types with special care (see also Table 3).

participated in the CITO test at least once during the period 2005-2012.[14] The total number of observations equals 78,545. Our estimation sample contains 35 schools in Amsterdam that were eligible for the ASIP. A total of 24 of these 35 schools have participated in the ASIP. Table 1 provides a more detailed overview of the timing of relevant events. The CITO tests of 2005, 2006, 2007, and 2008 have taken place before the introduction of the policy. As of April 2008, all failing schools in Amsterdam were invited to participate in the ASIP. It turns out that 12 schools started in the ASIP before or during the 2008-2009 school year, of which 7 before the CITO test of 2009; 6 schools started in the ASIP during the 2009-2010 school year, of which 3 before the CITO test of 2010; and 6 schools started in the ASIP during the 2010-2011 school year, of which 5 before the CITO test in 2011. From the 2009-2010 school year onwards, schools in Amsterdam could also participate in the professional development courses. In total 16 of the 35 schools in Amsterdam participated in these courses; 13 schools followed the courses as part of the complete program, and 3 schools only followed the courses without participation in the ASIP.

Table 1. Timing of Events.

Time	Event
February 2005	CITO test 2005
February 2006	CITO test 2006
February 2007	CITO test 2007
February 2008	CITO test 2008
April 2008	Municipality of Amsterdam introduces the ASIP
April 2008 - February 2009	7 schools start in the ASIP
February 2009	CITO test 2009
February 2009 - September 2009	5 schools start in the ASIP
September 2009	Municipality of Amsterdam introduces the professional development courses
September 2009 - February 2010	3 schools start in the ASIP
February 2010	CITO test 2010
February 2010 - September 2010	3 schools start in the ASIP
September 2010 - February 2011	5 schools start in the ASIP
February 2011	CITO test 2011
February 2011 - September 2011	1 school starts in the ASIP
February 2012	CITO test 2012

Notes. The presented number of schools that have started in the ASIP concern only those in our estimation sample.

Table 2 presents summary statistics of the CITO test scores in our sample. The test consists of 200 questions: 100 questions on language, 60 questions on math and 40 questions on information

[14] The total number of primary schools that were classified as 'weak' or 'very weak' on 1 January 2008 equals 751. We do not observe the schools that never participated in the CITO test.

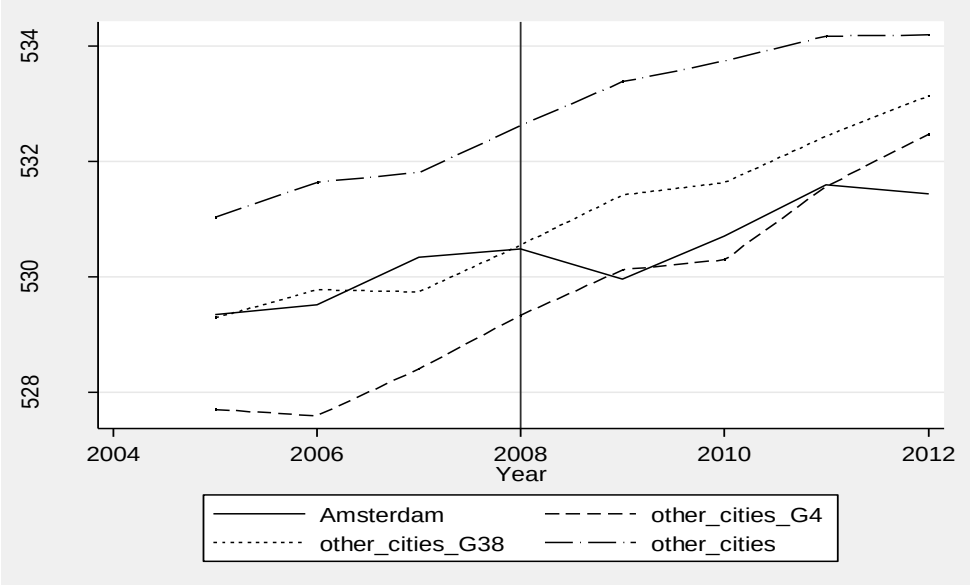
processing. The test scores on language, math and information processing equal the number of correctly answered questions. The total score is a linear transformation of the total number of correctly answered questions and ranges from 501 (lowest score) to 550 (highest score) each year. [15] The linear transformation is such that the total scores are comparable across years.

Table 2. Summary statistics of CITO test scores (total sample 2005-2012).

	Average	Standard deviation	Min	Max	Observations
Total score	532.60	10.30	501	550	78,545
Language	70.27	14.08	10	100	78,545
Math	40.16	11.60	1	60	78,545
Information processing	28.42	6.57	0	40	78,545

Figure 1 presents the total test score developments from 2005 to 2012 for schools in Amsterdam and the three control groups. The control groups contain failing schools outside Amsterdam. The main control group consists of all schools outside Amsterdam that were classified as (very) weak by the Inspectorate of Education on reference date 1 January 2008. In addition, we use two alternative control groups that include subsamples of large cities in the Netherlands: the so-called 'G38' and 'G4' cities. The G38 consists of 38 medium- and large cities and the G4 consists of the four largest cities in the Netherlands (Amsterdam, Rotterdam, The Hague and Utrecht). Schools in other large cities may be more similar to the schools in Amsterdam.

Figure 1. CITO test scores for weakly performing schools in the Netherlands



[15] A five point increase in CITO test score corresponds roughly to a one level higher secondary education type.

The availability of CITO test scores back to 2005, 3 years prior to the introduction of the policy, allows us to compare the pre-treatment test score trend in Amsterdam with that in the control groups. The performance of pupils in Amsterdam increased from 2005 till 2008. The increase in test scores is well comparable to that in other cities.[16] The test score trend indicates that the schools in Amsterdam did not experience a pre-treatment performance dip that would invalidate the common trend assumption (Ashenfelter, 1978).[17] The level of test scores in Amsterdam is larger compared to that in the other three large cities in the Netherlands, almost identical to that in the group of 37 other medium and large cities, and smaller compared to the nationwide test score level. To test for differential trends between Amsterdam and the control group, we regressed the first-difference of the test score on a time trend and an interaction term between the time trend and a dummy variable for Amsterdam. This yields an insignificant coefficient for the interaction term (with a *t*-value of -0.46), indicating that there is no evidence for differential trends. This supports the credibility of our identifying common trend assumption. A comparison of the post-treatment development of test scores in Amsterdam to the post-treatment development of test scores outside Amsterdam provides a first impression of the impact of the program. After the introduction of the ASIP in 2008, we observe that the increase in test scores in Amsterdam becomes smaller compared to that in the other cities.

The CITO dataset also contains information on individual background characteristics, such as gender, birth date and subsidy factor. In our empirical analyses we use these as covariates to control for observable changes in the pupil population over time. Table 3 presents the sample means of both the covariates and the outcome variables in 2008 (the most recent pre-treatment year) and 2012 (the latest available post-treatment year) for Amsterdam and the three control groups. The variable gender takes the value of one in case of a male and the value of zero in case of a female. We dispose of the year and month of birth of each pupil, from which we construct the age in years at the time of the test. We lack data on gender for 372 observations and on age for 359 observations. For this small fraction of our estimation sample, we impute missing values by the average value in the estimation sample. Furthermore, we dispose of categorical variables for the language spoken at home (seven categories), subsidy factor (six categories) and pupil category (six categories). The subsidy factor is an indicator of socioeconomic background. The Dutch funding scheme for primary schools distinguishes several groups of disadvantaged pupils, for whom primary schools receive additional funding.

[16] All primary schools that are classified as weak or very weak are subject to the nationwide interventions of the Dutch Inspectorate of Education. This may explain the improvement in test scores of failing schools over time.

[17] In case of a so-called 'Ashenfelter dip' one would expect schools in Amsterdam to improve after the introduction of the program, because of mean reversion. One might then incorrectly conclude that this improvement would be caused by the program.

Table 3. Sample means in 2008 and 2012.

	Amsterdam		Rest Netherlands		G38		G4	
	2008	2012	2008	2012	2008	2012	2008	2012
Covariates								
Gender (male = 1)	0.48	0.47	0.50	0.49	0.50	0.49	0.48	0.50
Age	12.12	12.00	12.06***	12.00	12.14	12.07***	12.21***	12.09***
Home language								
Dutch	0.51	0.00	0.81***	0.00	0.66***	0.00	0.59***	0.00
Other Western-Europe	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00
Arabic	0.13	0.00	0.02	0.00	0.05	0.00	0.06	0.00
Surinam	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.00
Turkish	0.09	0.00	0.03	0.00	0.09	0.00	0.10	0.00
Other	0.06	0.00	0.03	0.00	0.04	0.00	0.05	0.00
Unknown	0.18	1.00	0.10	1.00	0.14	1.00	0.18	1.00
Subsidy factor								
0	0.30	0.61	0.62***	0.71***	0.47***	0.63***	0.33***	0.57***
0.3	0.00	0.10	0.00	0.09	0.00	0.11	0.00	0.12
0.4	0.05	0.00	0.12	0.00	0.12	0.00	0.12	0.00
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.2	0.00	0.25	0.00	0.08	0.00	0.17	0.00	0.23
Unknown	0.66	0.04	0.26	0.11	0.41	0.10	0.55	0.08
Pupil category								
No special category	0.80	0.79	0.92***	0.90***	0.88***	0.89***	0.85	0.86***
I	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.03
IJ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
J	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
K	0.17	0.18	0.07	0.07	0.10	0.08	0.14	0.09
Outcome variables								
Total score	-0.21	-0.11	0.00***	0.15***	-0.20	0.05***	-0.32**	-0.01**
Language	-0.18	-0.36	0.02***	-0.08***	-0.18	-0.19***	-0.29**	-0.26**
Math	-0.05	0.03	0.08***	0.19***	-0.06	0.14***	-0.18**	0.11*
Information processing	-0.25	-0.14	-0.02***	0.18***	-0.19	0.04***	-0.26	-0.04**
Schools	34	34	500	487	112	132	32	38
Observations	901	991	8,739	9,275	2,411	2,996	733	1,023

Notes. Asterisks indicate that the sample mean in the control group differ significantly from that in Amsterdam in the corresponding year at a *10% level, **5% level, and ***1% level. Tests of significant differences for gender, age, and the outcome variables are based on a two-tailed *t*-test. Tests of significant differences for the categorical variables are based on a chi-squared test.

The subsidy factor depends on parental education level and can take values 0.3 and 1.2.[18] A subsidy factor of 0.3 implies that a school receives 30% of additional funding and a subsidy factor of 1.2

[18] The subsidy factor equals 1.2 in case the highest completed education level is primary education for at least one of the parents and lower secondary education for the other. The subsidy factor equals 0.3 in case lower

implies 120% of additional funding. The subsidy factor takes value 0 in case of a non-disadvantaged pupil. In earlier years, before 2010, other rules for the subsidy factor were used that depended not only on parental education level, but also on profession and ethnic background. This explains that the factors can also take values 0.4 or 0.9 in the years 2005-2009. The pupil category refers to specific groups of pupils for whom participation in the CITO test is not compulsory. Category 'I' stands for foreign pupils that have been in the Netherlands for less than four years; category 'J' for pupils that are expected to be assigned to special education; and category 'K' for pupils that are expected to be assigned to vocational secondary education with additional care.[19]All of these categorical variables contain one category to refer to an unknown value. The language spoken at home is unknown for around 12% of the observations in the years 2005-2011. In 2012 this information was not collected, implying that it is unknown for all observations. The subsidy factor is unknown for around 30% of the observations in the years 2005-2009 and for around 12% of the observations in the years 2010-2012.

Asterisks indicate that the sample mean in the control group differs significantly from that in Amsterdam in the corresponding year. Pupils in Amsterdam are well comparable to those in the other three groups with respect to gender. The age of the pupils when taking the CITO test in Amsterdam is comparable to that in the G38, but somewhat below (above) that in the G4 (rest of the Netherlands) in the pre-treatment year. Amsterdam is less comparable to the control groups with respect to the other covariates. Amsterdam, the largest city in the Netherlands, typically has a large fraction of disadvantaged pupils. This explains the smaller fraction of pupils with Dutch as their home language, the larger fraction of pupils with a high subsidy factor and the larger fraction of pupils belonging to a special category. We observe that in the G38 and the G4, differences on these variables are smaller, though still statistically significant in most cases. Only the pre-treatment difference in pupil category between Amsterdam and the G4 is insignificant. The empirical analyses include these variables as covariates to control for observable changes in the student population over time in Amsterdam and the control groups.

The bottom panel presents the CITO test scores that have been standardised to have a mean of zero and standard deviation of one in the full sample. Pre-treatment test scores in Amsterdam are very similar to those in the G38, and below (above) those in the rest of the Netherlands (G4). We observe that the total test scores have improved over time, both in Amsterdam and in the three control groups. The difference in test scores between 2012 and 2008 is smaller in Amsterdam compared to the three control groups for each of the test subjects.

---

secondary education is the highest completed education level for both parents (or the parent which is responsible for daily care).

[19] This type of education is called '*Leerwegondersteunend Onderwijs' (LWOO)*. Pupils in this category generally suffer from learning arrears, low IQ and/or social or emotional problems.

6. Results

6.1 Main findings

Table 4 presents the estimates of the effect of the introduction of the ASIP on CITO test scores for three model specifications. The first model (column 1) regresses the standardised CITO test score on a dummy for Amsterdam, a dummy for a post-treatment year, an interaction term that indicates a post-treatment year in Amsterdam, and year dummies.[20] The second model adds individual pupil background characteristics such as gender, age, age squared, home language, subsidy factor and pupil category. The third model additionally includes school fixed effects.

The top panel reports the estimation results for the complete sample. The middle panel shows similar results for the subsample of 38 middle and large Dutch cities and the bottom panel presents them for the subsample containing the four large cities in the Netherlands. For each of these samples we present the estimated effects on the total CITO score as well as on the specific subjects language, math, and information processing. Since all test scores are standardised, the estimated effects can be interpreted in terms of standard deviations.

In the complete sample we find negative effects of the introduction of the ASIP on CITO test scores in all model specifications. The addition of controls increases the size of the point estimates. This suggests a more favourable development of covariates (related to higher test scores) in Amsterdam compared to the control groups.[21] When estimated with all individual background characteristics and school fixed effects, the estimated effect implies that the introduction of the ASIP decreases the total CITO test score by 0.17 standard deviations. The estimated coefficient is statistically significant at the 1% level. Statistically significant negative effects are also found for each of the subjects of the test. The negative impact is largest for language and smallest for math, with estimated effects of -0.19 and -0.09, respectively. The G38 and G4 samples yield similar results. We find negative and statistically significant effects in all models, within a range from -0.11 to -0.20. These results indicate that the ASIP has negatively affected educational performance in the highest grade of primary education in the first four years after its introduction.

[20] Estimated treatment effects in models without the year dummies are very similar.

[21] For example, the increase in the share of non-disadvantaged pupils with a subsidy factor of 0 between 2008 and 2012 is larger in Amsterdam than that in other cities (see Table 3). Since non-disadvantaged pupils are more likely to perform well on the CITO test, inclusion of this control variable decreases the estimated treatment effect.

Table 4. Difference-in-differences estimates of the introduction of the ASIP.

	(1)	(2)	(3)
A. Complete Sample			
Total score	-0.105*(0.062)	-0.144***(0.049)	-0.170***(0.036)
Language	-0.109*(0.060)	-0.152***(0.046)	-0.188***(0.034)
Math	-0.052 (0.056)	-0.078*(0.047)	-0.086**(0.036)
Information processing	-0.127**(0.061)	-0.160***(0.049)	-0.179***(0.035)
Observations	78,545	78,545	78,545
Schools	614	614	614
B. G38			
Total score	-0.127*(0.068)	-0.154***(0.053)	-0.190***(0.040)
Language	-0.117*(0.067)	-0.155***(0.050)	-0.200***(0.037)
Math	-0.107*(0.061)	-0.114**(0.052)	-0.136***(0.040)
Information processing	-0.114*(0.067)	-0.139***(0.053)	-0.164***(0.039)
Observations	27,882	27,882	27,882
Schools	173	173	173
C. G4			
Total score	-0.181**(0.084)	-0.145**(0.066)	-0.149***(0.049)
Language	-0.162*(0.085)	-0.140**(0.063)	-0.158***(0.046)
Math	-0.175**(0.074)	-0.131**(0.064)	-0.120**(0.049)
Information processing	-0.143*(0.086)	-0.113*(0.068)	-0.107**(0.050)
Observations	13,597	13,597	13,597
Schools	74	74	74
Individual characteristics	no	yes	yes
School fixed effects	no	no	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

6.2 Heterogeneity

We proceed by investigating the impact of the introduction of the ASIP across different groups of pupils. We distinguish between male and female, higher and lower socio-economic status, and Dutch speaking and non-Dutch speaking pupils. We define all pupils with a subsidy factor larger than 0 to have low socio-economic status and all other pupils to have high socio-economic status. Pupils with home language other than Dutch are defined as 'non-Dutch speaking' and all other pupils as 'Dutch speaking'. We leave out pupils with missing values on these variables. Table 5 reports the estimated full model effects for each of the subgroups, taking into account the complete sample containing all weakly performing schools in the Netherlands. The estimated effects on the test scores are reasonably in line with the total sample estimates and are similar across subgroups.[22] With respect to the specific

[22] In models (3) and (4) the estimated effects are both larger (in absolute value) than the total sample estimates, while models (5) and (6) both yield smaller estimates. This can be explained by the fact that we leave out pupils with missing values on socioeconomic status or home language.

subjects, the estimated effects on math turn insignificant in models (3), (5) and (6), but the differences in effect sizes across the subsamples are not large. Table A.1 in the appendix presents the results for the G38 and G4 samples. The results in the G38 sample are in line with our findings of no differential effects across subgroups in the complete sample. The results in the G4 suggest that the introduction of the policy has been more detrimental for non-disadvantaged pupils than it has been for disadvantaged pupils. This finding holds for each of the subjects and differs from the findings in the other samples. In sum, we conclude that we find no strong evidence that specific groups of pupils are particularly affected by the policy. When taking into account only the four largest cities, the policy seems to have had a more detrimental impact for non-disadvantaged pupils.

Table 5. Heterogeneous treatment effects: Estimated effects of the introduction of the ASIP.

	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Low socio-economic status	High socio-economic status	Non-Dutch speaking	Dutch speaking
Total score	-0.164***	-0.164***	-0.205***	-0.196***	-0.106*	-0.130***
	(0.042)	(0.040)	(0.069)	(0.046)	(0.057)	(0.048)
Language	-0.186***	-0.178***	-0.231***	-0.230***	-0.107*	-0.151***
	(0.041)	(0.037)	(0.066)	(0.045)	(0.058)	(0.046)
Math	-0.074*	-0.088**	-0.075	-0.091**	-0.068	-0.038
	(0.040)	(0.043)	(0.078)	(0.046)	(0.058)	(0.046)
Information processing	-0.178***	-0.174***	-0.277***	-0.211***	-0.123**	-0.171***
	(0.045)	(0.039)	(0.073)	(0.048)	(0.061)	(0.050)
Observations	38,607	39,566	12,006	48,527	7,150	53,385
Schools	614	614	556	608	419	601
Individual characteristics	yes	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

In addition to the effects for different groups, we investigate the impact of the ASIP on different parts of the test score distribution by estimating quantile regressions. Table 6 presents the estimated effects for various quantiles of the test score distributions. With respect to the total test score, the estimated coefficients differ across quantiles. The impact is most detrimental at the lower tale of the distribution. The introduction of the ASIP decreases the lower quartile of the total test score distribution by 0.19 standard deviations, the median by 0.18 standard deviations (which is close to the OLS coefficient of -0.17) and the upper quartile by 0.11 standard deviations. The impact is smallest at the upper decile, -0.06, and largest at the lower decile, -0.25. This pattern of decreasing estimated effect sizes with

quantile also shows up for the specific subjects language and information processing. The pattern is less clear for math, though again the estimated effects are smaller (in absolute value) in the upper tail of the test score distribution. Table A.2 in the appendix presents quantile regression results for the G38 and G4 samples. The findings in the G38 are consistent with the observed pattern in the complete sample. The picture in the G4 is less clear. The largest effect sizes are not always found to be at the lowest decile, but in most cases the smallest effect sizes are found at the upper decile. Hence, our finding of a less detrimental impact at the upper part of the test score distribution is confirmed in both other samples as well.

Table 6. Quantile regressions results: Estimated effects of the introduction of the ASIP for quantiles of the test score distributions.

	(1) quantile regression	(2) quantile regression	(3) quantile regression	(4) quantile regression	(5) quantile regression
	0.1	0.25	0.50	0.75	0.90
Total score	-0.247***	-0.193***	-0.183***	-0.108***	-0.062**
	(0.038)	(0.032)	(0.029)	(0.03)	(0.029)
Language	-0.246***	-0.192***	-0.212***	-0.161***	-0.093***
	(0.041)	(0.032)	(0.028)	(0.027)	(0.029)
Math	-0.097**	-0.101***	-0.111***	-0.063**	-0.052**
	(0.041)	(0.036)	(0.031)	(0.026)	(0.026)
Information	-0.309***	-0.240***	-0.186***	-0.138***	-0.077***
processing	(0.043)	(0.035)	(0.029)	(0.025)	(0.026)
Observations	78,545	785,45	78,545	78,545	78,545
Schools	614	614	614	614	614
Individual characteristics	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes

Notes. Each cell represents a separate quantile regression. Standard errors are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

6.3 Sensitivity

Table 7 presents several sensitivity analyses to probe the robustness of our main findings. First, we restrict our sample to schools that participated in the CITO test in each of the years 2005 to 2012. A potential concern is that the introduction of the policy affected participation in the CITO test. For instance, it might be that non-treated schools outside Amsterdam were closed due to persistently bad performance and do not show up in the data in all post-treatment years. This might bias our estimates downwards if the weakest performing schools outside Amsterdam drop out of the estimation sample.

Column (1) presents estimation results for the sample consisting of only schools that participate each year. This includes 399 schools containing 63,165 pupils. The estimated effects are slightly larger (in absolute value) than the main estimates. This indicates that our results are unlikely to be biased by selective attrition of schools.

Table 7. Sensitivity: Estimated effects of the introduction of the ASIP on CITO test score (full model).

	(1) Sample of schools that participate in the CITO test in all years 2005-2012	(2) Sample excluding the south-east district in Amsterdam	(3) Sample excluding schools in Rotterdam	(4) Sample excluding the year 2009	(5) Sample excluding the years 2009, and 2010	(6) Sample excluding the years 2009, 2010 and 2011	(7) Sample collapsed to before/after observations at school level
Total score	-0.190***	-0.128***	-0.169***	-0.169***	-0.208***	-0.299***	-0.144**
	(0.037)	(0.039)	(0.035)	(0.039)	(0.043)	(0.057)	(0.063)
Language	-0.200***	-0.155***	-0.189***	-0.187***	-0.217***	-0.321***	-0.177**
	(0.035)	(0.037)	(0.034)	(0.037)	(0.041)	(0.053)	(0.059)
Math	-0.109***	-0.042	-0.081**	-0.100**	-0.153***	-0.204**	-0.075
	(0.037)	(0.040)	(0.036)	(0.039)	(0.043)	(0.059)	(0.064)
Information processing	-0.201***	-0.140***	-0.184***	-0.165***	-0.181***	-0.274***	-0.144**
	(0.037)	(0.036)	(0.035)	(0.038)	(0.044)	(0.057)	(0.065)
Observations	63,165	76,956	75,955	69,030	59,438	49,462	1,145
Schools	399	607	601	612	612	609	614
Individual characteristics	yes	yes	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

Second, we exclude schools whose results might have been affected by other programs from the sample. We leave out schools in the south-east of Amsterdam to address the potential influence of the program that was raised to improve pupil's math performance in this district. The district contains 7 primary schools in our sample including 1,589 pupils. If the program has improved results, we would expect the estimated effects to become more detrimental. Instead, we find negative point estimates that are smaller (in absolute value) than our main estimates and the estimated effect for math turns insignificant (see column 2). This suggests that our main results are not biased upwards because of the impact of the program in the south-east district. Furthermore, we exclude schools in Rotterdam and

find estimated effects that are very similar to our main estimates (see column 3). This suggests that the action program launched by the municipality of Rotterdam in 2011 does not affect our findings.[23]

Third, we perform similar analyses on a sample in which we leave out the year 2009. Since the first schools that participated in the ASIP started in 2008, one might argue whether 2009 is an appropriate post-treatment year. Excluding 2009 from the sample, using only 2010 to 2012 as the post-treatment years, yields similar result (see column 4). Regarding the two-year program duration, it might take even longer before the impact of the program was felt in the CITO test scores. To provide insight into the timing of effects, we further reduce the number of post-treatment years taken into account. In model (5) we leave out the years 2009 and 2010 whereas model (6) additionally excludes the year 2011 from our sample. In case of an improvement in test scores over time, we expect to find better results in models that include only the most recent post-treatment years. Instead, the negative point estimates of the effect of the ASIP become larger in models (5) and (6). Hence, we find no evidence for cumulative effects over time in the first four years after the introduction of the program.

A complicating factor for analysing cumulative effects is that not all schools started in the program at the same time. If participation in the ASIP causes an initial decrease in test scores that is followed by an upward development, a concern might be that an improvement in test scores of participating schools that started in 2008-2009 is negated by a decrease in test scores in schools that started in later years. This, however, seems a less plausible interpretation of our findings because of the relatively large share of schools in the sample that started in 2008-2009 (see Table 1). We further address this issue by excluding all schools in Amsterdam that did not start in the ASIP before or during the 2008-2009 school year. Including only the 12 schools in Amsterdam that participated in 2008-2009 (and all other schools outside Amsterdam), we would expect to find increasing effects when restricting the sample to more recent post-treatment years in case of cumulative effects. Table A.3 in the appendix presents the estimation results for the four sets of post-treatment years. We find that the negative point estimates of the impact of the ASIP become larger when we restrict the sample to more recent post-treatment years. This is consistent with our conclusion of no improvement in test scores over time.

In addition, we have estimated four separate models that each include one post-treatment year. The results are shown in Table A.4 in the appendix. We find mostly statistically significant negative estimates for each of the years. The estimated effects on test scores are smallest (in absolute value) in the years 2010 and 2011, and largest in 2012. Table A.5 presents similar results when estimated on the sample that excludes schools in Amsterdam that did not start in the ASIP before or during the 2008-2009 school year. The estimated effects on test scores are smaller (in absolute value) and

[23] Excluding the schools in Rotterdam in the G4 sample yields an estimated effect of -0.167***(0.061) on the total CITO test score.

statistically insignificant for the year 2010, while larger and statistically significant for the other post-treatment years. We conclude that also these separate estimates for each post-treatment year do not provide evidence of increasing effects over time.

Fourth, we address concerns on over-significance of our estimates because of potential serial correlation problems, by collapsing the data before and after the introduction of the ASIP at the school level (Bertrand et al., 2004). This leaves us with 1145 observations.[24] We find a statistically significant effect on the total test score of -0.14 at the 5%-level.

A final concern might be that the policy has affected the testing pool within schools that participate in the CITO test. More specifically, if the policy caused a relative increase in the CITO test participation of weak students (with low unobserved ability) in Amsterdam compared to other cities, this might have biased our estimated effects downwards. A change in the tested population induced by the program, however, does not seem very likely since a larger CITO test participation was no clear element or goal of the ASIP. Hence, the program did not explicitly stimulate additional participation of (low ability) students in the test. Still, one might be concerned that the ASIP implicitly affected the testing pool if increased scrutiny limited opportunities to exclude weak students from the test. [25] In our data we do not observe a particularly large increase in the number of students that take the test in Amsterdam relative to other cities after the introduction of the policy. In addition, the share of pupils placed in special categories stays reasonably constant over time, both in Amsterdam and other cities (see also Table 3). Ideally, we would have disposed of the total number of students in grade eight for each school and year. This would have enabled us to compare the development of participation shares in the CITO test between Amsterdam and other cities, and to estimate the effect of the policy on CITO test participation. However, our CITO data only contain information on the pupils that participated in the test. Therefore, we have performed an additional analysis making use of another dataset, called COOL, that includes a representative sample of around 10% of all Dutch primary schools in the pre-treatment year 2007-2008 and post-treatment year 2010-2011 (Driessen at al., 2009; 2012). These data include information on CITO test participation for the pupils in grade eight. Regressing a dummy variable for CITO test participation on a dummy variable for Amsterdam, a dummy variable for the post-treatment year, an interaction term between Amsterdam and the post-treatment year, and a set of pupil background characteristics yields an insignificant estimated effect for the interaction term that is

[24] The sample contains 614 schools and 2 time periods. Not all 614 schools are present in both the before and the after period: there are 561 schools in the before period and 584 schools in the after period.

[25] Since the CITO test scores are important for judging educational quality, schools may have an incentive for shaping the testing pool. Previous studies have shown that schools can respond strategically to the implementation of accountability policies by excluding weak students from the test (e.g. Jacob, 2005).

close to zero.[26] Although this analysis is performed on a different sample that concerns not only failing schools but also sound performing schools, we interpret our finding of no effect as supportive evidence that the introduction of the policy did not affect participation in the CITO test.[27]

A related issue is the impact of grade retention. If the policy has affected retention, this could have changed the pupil composition in grade eight. Since we do not dispose of formal information on grade retention, we investigate this issue further by comparing the age of the pupils after and before the introduction of the policy in Amsterdam and in the other cities. We perform two analyses. First, we regress age on a dummy for Amsterdam, a dummy for a post-treatment year, an interaction term that indicates a post-treatment year in Amsterdam, year dummies and all other covariates. Second, we use a similar specification to estimate the effect of the introduction of the policy on a dummy variable indicating whether a pupil is older than 12.5. We use this dummy variable as an indicator for grade retention since pupils aged above 12.5 are most likely to have retained in grade. The first column of Table A.6 in the appendix presents the estimation results. We find statistically significant effects in both models: the age at which pupils take the test decreases with around 0.08 years (row 1) and the probability of grade retention decreases with around 3 percentage points (row 2). These results suggest that the retention probability has decreased following the introduction of the policy. This may have biased our estimates downwards if an additional year in education increases test scores for weak pupils. To address this issue, we proceed with two robustness analyses that are presented in columns 2 - 5 of Table A.6. First, we estimate the effect of the introduction of the policy on the CITO test score for the subsample of pupils who have not been retained. Excluding pupils who are aged above 12.5 yields a statistically significant negative effect, -0.14, which is somewhat below (in absolute value) the estimated effect in the complete sample, -0.17 (see column 2).[28] Second, we estimate the impact of the policy for three subsamples that exclude schools in Amsterdam with the largest decrease in age (see columns 3-5). We define the decrease in age as the average age of test taking in the years after the introduction of the policy minus the average age of test taking in the years before the introduction of the policy. In the first subsample we exclude all schools in Amsterdam for which the average age of test taking decreases with 0.2 years or more (eight schools). The second subsample leaves out all

[26] The estimated effect for the interaction term is 0.001 (0.029). The included pupil background characteristics are gender, age, age squared, and a categorical socioeconomic status variable that distinguishes six categories based on parental education level and ethnic origin. The total sample includes 18,887 pupils in grade eight divided over 676 different schools.

[27] Restricting the sample to only those that were classified as weak or very weak on 1 January 2008 leaves us with only 6 schools in Amsterdam, of which 3 participated in the ASIP. A similar analysis on this subsample yields an estimated effect for the interaction term of 0.005 (0.006).

[28] Please note that this analysis is not fully informative on the magnitude of potential bias caused by the impact of grade retention. After all, it does not exclude those pupils in Amsterdam that have not retained after the introduction of the policy, but that would have retained in the absence of the policy. In addition, the lower estimated effect may well be explained by the exclusion of weak performing pupils for whom the impact of the policy on test scores is likely to be more detrimental (see Table 6).

schools in Amsterdam with a decrease in age of at least 0.15 years (twelve schools) and the third subsample leaves out all schools in Amsterdam with a decrease in age of at least 0.10 years (sixteen schools). In this way schools that are most likely to have faced a reduction in retention following the introduction of the policy are excluded from the analyses. Leaving out such schools obviously reduces the estimated impact of the policy on age and on the indicator for grade retention (see rows 1 and 2 in columns 3, 4, and 5). The estimated effect on the indicator for grade retention is statistically insignificant and close to zero in all models. In each of the three subsamples we find statistically significant negative effects of the introduction of the policy on the CITO test score, ranging from -0.14 to -0.17 (see row 3 in columns 3, 4, and 5). These results are close to our main estimates and suggest that our findings are not importantly affected by the potential impact of grade retention.

We conclude that our finding that the introduction of the ASIP negatively affected test scores in grade eight of primary education is robust to a variety of sensitivity tests.

7. Interviews and potential mechanisms

To gain further insight into the effects of the policy, we have taken interviews with school-leaders of participating schools. We focused on the schools that started in the ASIP before or during the 2008-2009 school year and found seven school-leaders who were willing to provide information on their experiences with the program. The outcomes of these interviews reveal a potential explanation for our empirical findings. The overall opinions of the school-leaders were very similar and yield a consistent picture that is two-fold. Most of the school-leaders expected that the program would result in better educational quality in the longer term. They especially appreciated the use of teacher quality evaluations which provided insight into teacher behaviour that revealed current weaknesses in classroom practices and provided a clear view on potential improvements. In addition, the courses for school-leaders on performance-based working and the use of new instruction methods were mentioned as valuable elements of the program. At the same time, they experienced the ASIP as an intensive and radical program with a rigorous approach. Most of them reported severe resistance among teachers, for whom the program was especially demanding. The teachers were confronted with direct feedback on their classroom behaviour, changes in their instruction materials, and were expected to put in effort to improve their competences in addition to their regular teaching tasks. Almost all school-leaders report the exit of school personnel after the introduction of the program. Some of them left the school voluntarily because they did not want to go along with the changes induced by the program, but others were forced to leave because they appeared not capable to satisfy the required standards. The proportion of replaced teachers seems to be substantial. Three school-leaders explicitly mention the number of replaced teachers. Two of them report that around 25% of the initial teacher population was replaced; the other one reports that even around 90% of the initial

teacher population was replaced. According to most school-leaders, the replacement of teachers also led to uncertainty among school personnel.

The outcomes of the interviews provide suggestive evidence that the ASIP has increased teacher mobility. This finding is consistent with existing literature on the impact of school reforms on school personnel. Figlio and Loeb (2011) discuss the relationship between school accountability and teacher labour markets. They refer to interview and survey research providing evidence that teachers value a cohesive and supportive work environment that acknowledges their efforts and competences, while they interpret increased scrutiny and/or high-stakes testing as a reduction in their classroom autonomy and a message of being viewed as incompetent (e.g. Luna and Turner, 2001). The authors state that school reforms that influence these aspects of the work place are likely to affect teacher mobility. In addition, they discuss empirical studies providing evidence that accountability systems especially increase teacher attrition in schools that are labeled as low performing (see e.g. Feng et al., 2010). The ASIP, initially targeted at weak performing schools, increased scrutiny by lesson observations and may have contributed to the feeling of an unsupportive work atmosphere. In addition, the teacher evaluations may have helped the school-leaders to identify and replace ineffective teachers.

The increased mobility can negatively affect pupil's test scores in the short term via two mechanisms (Figlio and Loeb, 2011). First, recruitment and hiring of new teachers can take time and resources away from the regular instruction tasks. Second, the leave of (experienced) teachers implies a loss of specific knowledge on the school's way of working, instruction program and pupils. It takes time before new teachers have developed this knowledge. In addition to these mechanisms, the changing and uncertain work environment may have disturbed an optimal focus on primary instruction tasks among teachers.

In sum, the program appears to have created an unstable work atmosphere with increased teacher mobility. This may explain the negative impact on educational achievement in the first four years after its introduction. In the longer term, however, teacher mobility need not be detrimental if the least effective teachers are replaced by more effective teachers. In that case one might expect better performances once the more effective teachers are hired and the new work environment within schools has been stabilised for a while. Our main analyses concern the impact on a high-stakes test during the first four years after the introduction of the program. We do not find evidence for an improvement in the CITO test scores over these years. Nevertheless, we cannot exclude that these findings reflect adjustment costs of the reform policy and that it takes longer before more beneficial effects become manifest in the CITO test scores.

8. Conclusion

CSR methods have been widely used as an instrument to improve failing schools, but the evidence on its effectiveness remains limited. We estimate the effects of the ASIP, a CSR policy introduced in the Netherlands in 2008 with the goal of improving the educational quality of weak-performing primary schools in Amsterdam. The program implements performance-based working at all levels within the school and typically integrates measures such as staff coaching, teacher observations and teacher schooling, and the use of new instruction methods. Each program is tailored towards the specific needs of the school and is guided by educational experts.

Difference-in-differences estimates show substantial and significant detrimental effects on the educational achievement of pupils in the highest grade of primary education. This finding is robust to a broad range of sensitivity tests. In our preferred specification, test scores decrease by 0.17 standard deviations in the first four years after the introduction of the policy. The overall negative effects are larger for language scores than for math scores. The size of the estimated effects varies across different parts of the test score distribution. The largest negative effects are generally found at the left part of the test score distribution, and the least detrimental effects at the upper tail of the test score distribution.

Interviews with school-leaders of participating primary schools reveal a candidate explanation for our findings. Although most of the school-leaders expected that the program would result in better educational quality in the longer term, they experienced the ASIP as an intensive program with a rigorous approach. It was especially confronting and demanding for teachers, who were judged based on lesson observations and expected to improve their competences. All required efforts had to be made in addition to their primary teaching tasks. Almost all school-leaders report the replacement of teachers after the introduction of the program. Some of them left the school voluntarily because of disagreement with the program, but others were forced to leave because they appeared not capable to satisfy the required standards. The outflow of teachers implies a loss of school specific knowledge and an increased focus on hiring new personnel that may have gone at the cost of instruction tasks. In addition, it seems to have created uncertainty among school personnel which can have disturbed an optimal focus on instruction tasks. Altogether, an increased teacher mobility induced by the program is a potential explanation for our negative findings on educational achievement. In that case, one might expect more beneficial effects in the longer term if less effective teachers are replaced by more effective ones and once the work environment within schools has been stabilised for a while. We do not find evidence for increasing effects over time in the first four years after the introduction of the policy. Nevertheless, we still cannot exclude that our findings reflect adjustment costs of the reform, and that it takes longer before beneficial effects become manifest in the CITO test scores. Even in

such a case, one may question whether the future gains will outweigh the initial losses. In any case, we conclude that the introduction of the comprehensive school reform induced large costs in terms of a substantial decrease in educational performance for at least four cohorts of pupils.

References

Angrist, J.D., G. Imbens, D. Rubin, 1996, Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association*, 91(434): 444–55.

Ashenfelter, O., 1978, Estimating the effect of training programs on earnings, *Review of Economics and Statistics*, 60: 47-50.

Ashenfelter, O., D. Card, 1985, Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs, *Review of Economics and Statistics*, 67(4): 648-660.

Bertrand, M., E. Duflo, S. Mullainathan, 2004, How Much Should We Trust Differences-in-Differences Estimates, *Quarterly Journal of Economics*, 119, 249-275.

Bifulco, R., W. Duncombe, J. Yinger, 2005, Does whole-school reform boost student performance? The case of New York City, *Journal of Policy Analysis and Management*, 24, 47-72.

Bloom, H., 1984, Accounting for No-shows in Experimental Evaluation Designs, *Evaluation Review*, 8: 225-246.

Blundell, R., A. Duncan, C. Meghir, 1998, Estimating Labor Supply Responses Using Tax Reforms, *Econometrica*, 66(4): 827-861.

Borman, G. D., Hewes, G. M., Overman, L. T., Brown, S., 2003, Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125–230.

Borman, G. D., Hewes, G. M., Overman, L. T., Brown, S., 2004, Comprehensive school reform and achievement: A meta-analysis. In C. T. Cross (Ed.), *Putting the pieces together: Lessons from comprehensive school reform research* (pp. 53–108). Washington, DC: The National Clearinghouse for Comprehensive School Reform.

Card, D., A. Krueger, 1994, Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, *American Economic Review*, 84(4): 772-793.

Cook, T.D., Habib, E, Phillips, M., Settersten, R., Shagle, S.C., & Degirmencioglu, S.M., 1999, Comers school development program in Prince George's County, Maryland: A theory-based evaluation. American Education Research Journal, 36(3), 543-597.

Cook, T.D., Hunt, H.D., Murphy, R.E, 1998, Comer's School Development Program in Chicago: A theory-based evaluation. WP-98-24. Evanston, IL: Institute for Policy Research, Northwestern University.

Driessen, G., L. Mulder, G. Ledoux, J. Roeleveld, I. van der Veen, 2009, *Cohortonderzoek COOL5-18. Technisch rapport basisonderwijs, eerste meting 2007/08.* Nijmegen: ITS / Amsterdam: SCO-Kohnstamm Instituut.

Driessen, G., L. Mulder, J. Roeleveld, 2012, *Cohortonderzoek COOL5-18. Technisch rapport basisonderwijs, tweede meting 2010/11.* Nijmegen: ITS / Amsterdam: SCO-Kohnstamm Instituut.

Feng, L., D. Figlio, T. Sass, 2010, School Accountability and Teacher Mobility, NBER Working Paper No. 16070.

Figlio, D., S. Loeb, 2011, School Accountability. In E. Hanushek, S. Machin, L. Woessmann, editor: *Handbook of the Economics of Education,*Vol. 3: 383-421.

Gross, B., T. K. Brooker, D. Goldhaber, 2009, Boosting Student Achievement: The Effect of Comprehensive School Reform on Student Achievement, *Educational Evaluation and Policy Analysis*, 31 (2): 111-126.

Herman, R., Aladjem, D., McMahon, P., Masem, E., Mulligan, I., O'Malley, A. S., Quinones., S., Reeve, A., Woodruff, D., 1999, *An educator's guide to schoolwide reform*. Arlington, VA: American Institutes for Research.

Imbens, G.W., J.D. Angrist, 1994, Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62(2): 467–75.

Jacob, B.A., 2005, Accountability, incentives and behavior: the impact of high-stakess testing in the Chicago Public Schools, *Journal of Public Economics*, vol. 89: 761-796.

Inspectorate of Education, 2008, *Regionale analyse; Een analyse van het Amsterdamse basisonderwijs*. Utrecht.

Inspectorate of Education, 2009, *De Staat van het Onderwijs, Onderwijsverslag 2007/2008.* Utrecht.

Luna, C., Turner, C.L., 2001. The impact of the MCAS: Teachers talk about high-stakes testing, *English Journal*, 91 (1): 79–87.

May, H., J.A. Supovitz, 2006, Capturing the Cumulative Effects of School Reform: An 11-Year Study of the Impacts of America's Choice on Student Achievement, *Educational Evaluation and Policy Analyis*, 2006, 28 (3): 231-257.

Millsap, M.A., Chase, A., Obiedallah, D., Perez-Smith, A., 2001, Evaluation of the Comer School Development Program in Detroit, 1994-1999: Methods and results. Washington, DC.

Moulton, B., 1986, Random Group Effects and the Precision of Regression Estimates, *Journal of Econometrics*, 32(3), 385–97.

Municipality of Amsterdam, 2009, *Kwaliteitsaanpak Basisonderwijs Amsterdam; programmaplan 2009-2014.* Amsterdam.

Rowan, B., Barnes, C., Camburn, E., 2004, Benefiting from comprehensive school reform: A review of research on CSR implementation. In C. T. Cross (Ed.), *Putting the pieces together: Lessons from comprehensive school reform research* (pp. 1–52). Washington, DC: National Clearinghouse for Comprehensive School Reform.

Schwartz, A.E., Stiefel, L.S., Kim, D.Y., 2004, The impact of school reform on student performance: Evidence from the New York Network for School Renewal Project, *Journal of Human Resources*, 39 (2), 500-522.

Slavin, R. E., & Fashola, O. S., 1998, *Show me the evidence!* Thousand Oaks, CA: Corwin Press.

Traub, J., 1999, *Better by design? A consumer's guide to schoolwide reform*. Washington, DC: Thomas B. Fordham Foundation.

U.S. Department of Education, 2004, *Implementation and early outcomes of the Comprehensive School Reform Demonstration (CSRD) Program* (No. 2004-15). Jessup, MD: Policy and Program Studies Service, U.S. Department of Education.

U.S. Department of Education, 2006, *Comprehensive school reform program: Funding status*. Washington, DC: Author. Retrieved July 6, 2006, from http://www.ed.gov/programs/compreform/funding.html.

Van der Steeg, M., S. Gerritsen, 2013, Teacher evaluations and pupil achievement; Evidence from classroom observations, CPB Discussion Paper 230.

Appendix

Table A.1. Heterogeneous treatment effects: Estimated effects of the introduction of the ASIP in the G38 and G4 samples.

	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Low socio-economic status	High socio-economic status	Foreign	Dutch
A. G38						
Total score	-0.190***	-0.184***	-0.200***	-0.208***	-0.103*	-0.130**
	(0.046)	(0.044)	(0.076)	(0.052)	(0.060)	(0.052)
Language	-0.200***	-0.192***	-0.211***	-0.226***	-0.091	-0.147***
	(0.045)	(0.040)	(0.072)	(0.050)	(0.060)	(0.049)
Math	-0.135***	-0.134***	-0.098	-0.144**	-0.077	-0.063
	(0.044)	(0.048)	(0.085)	(0.053)	(0.061)	(0.051)
Information processing	-0.166***	-0.158***	-0.263***	-0.182***	-0.130**	-0.143***
	(0.049)	(0.043)	(0.081)	(0.053)	(0.064)	(0.054)
Observations	13,534	14,198	5,374	13,892	5,108	14,553
Schools	173	173	169	171	156	169
B. G4						
Total score	-0.169***	-0.120**	-0.051	-0.200***	-0.124	-0.095
	(0.058)	(0.055)	(0.095)	(0.064)	(0.076)	(0.062)
Language	-0.171***	-0.137***	-0.092	-0.202***	-0.081	-0.124**
	(0.058)	(0.049)	(0.086)	(0.062)	(0.073)	(0.059)
Math	-0.143***	-0.088	0.038	-0.152**	-0.146*	-0.027
	(0.054)	(0.060)	(0.102)	(0.063)	(0.079)	(0.060)
Information processing	-0.131**	-0.080***	-0.100	-0.193***	-0.126	-0.119*
	(0.061)	(0.056)	(0.105)	(0.065)	(0.079)	(0.064)
Observations	6,692	6,817	2,877	5,630	3,034	6,047
Schools	74	74	72	73	72	73
Individual characteristics	yes	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

Table A.2. Quantile regressions results: Estimated effects of the introduction of the ASIP for quantiles of the test score distributions in the G38 and G4 samples.

	(1) quantile regression	(2) quantile regression	(3) quantile regression	(4) quantile regression	(5) quantile regression
	0.1	0.25	0.50	0.75	0.90
A. G38					
Total score	-0.246***	-0.208***	-0.220***	-0.130***	-0.093***
	(0.042)	(0.033)	(0.031)	(0.029)	(0.034)
Language	-0.236***	-0.211***	-0.227***	-0.174***	-0.099***
	(0.047)	(0.035)	(0.030)	(0.030)	(0.032)
Math	-0.142***	-0.158***	-0.166***	-0.127***	-0.067**
	(0.042)	(0.040)	(0.034)	(0.030)	(0.030)
Information	-0.266***	-0.201***	-0.164***	-0.140***	-0.066**
processing	(0.047)	(0.036)	(0.033)	(0.029)	(0.029)
Observations	27,822	27,822	27,822	27,822	27,822
Schools	173	173	173	173	173
B. G4					
Total score	-0.156***	-0.159***	-0.172***	-0.130***	-0.086**
	(0.055)	(0.041)	(0.039)	(0.039)	(0.040)
Language	-0.135***	-0.166***	-0.219***	-0.144***	-0.065
	(0.050)	(0.043)	(0.039)	(0.039)	(0.040)
Math	-0.058	-0.155***	-0.166***	-0.116***	-0.086**
	(0.052)	(0.050)	(0.041)	(0.038)	(0.039)
Information	-0.217***	-0.117***	-0.090**	-0.132***	-0.039
processing	(0.058)	(0.046)	(0.041)	(0.039)	(0.038)
Observations	13,597	13,597	13,597	13,597	13,597
Schools	74	74	74	74	74
Individual characteristics	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes

Notes. Each cell represents a separate quantile regression. Standard errors are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

Table A.3. Estimated effects of the introduction of the ASIP: Different post-treatment years included. Sample without schools in Amsterdam that did not start in the ASIP in 2008-2009.

	(1)	(2)	(3)	(4)
	Post-treatment years: 2009, 2010, 2011, 2012	Post-treatment years: 2010, 2011, 2012	Post-treatment years: 2011, 2012	Post-treatment years: 2012
Total score	-0.237***	-0.208***	-0.287***	-0.353***
	(0.067)	(0.071)	(0.081)	(0.122)
Language	-0.236***	-0.221***	-0.283***	-0.380***
	(0.065)	(0.069)	(0.081)	(0.116)
Math	-0.179***	-0.171**	-0.262***	-0.279**
	(0.066)	(0.069)	(0.076)	(0.117)
Information processing	-0.198***	-0.143**	-0.200***	-0.270**
	(0.063)	(0.069)	(0.078)	(0.114)
Observations	73,182	64,352	55,469	46,214
Schools	591	589	589	587
Individual characteristics	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

Table A.4. Estimated effects of the introduction of the ASIP for separate post-treatment years.

	(1)	(2)	(3)	(4)
	Post-treatment year: 2009	Post-treatment year: 2010	Post-treatment year: 2011	Post-treatment year: 2012
Total score	-0.189***	-0.094	-0.128***	-0.299***
	(0.053)	(0.059)	(0.050)	(0.057)
Language	-0.206***	-0.130**	-0.125***	-0.321***
	(0.049)	(0.056)	(0.045)	(0.053)
Math	-0.071	0.002	-0.109**	-0.204**
	(0.051)	(0.058)	(0.051)	(0.059)
Information processing	-0.225***	-0.143**	-0.100*	-0.274***
	(0.051)	(0.059)	(0.053)	(0.057)
Observations	48,711	48,788	49,172	49,462
Schools	575	586	599	609
Individual characteristics	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

Table A.5. Estimated effects of the introduction of the ASIP for separate post-treatment years. Sample without schools in Amsterdam that did not start in the ASIP in 2008-2009.

	(1)	(2)	(3)	(4)
	Post-treatment year: 2009	Post-treatment year: 2010	Post-treatment year: 2011	Post-treatment year: 2012
Total score	-0.339***	-0.053	-0.226***	-0.353***
	(0.092)	(0.081)	(0.085)	(0.122)
Language	-0.299***	-0.088	-0.187**	-0.380***
	(0.093)	(0.074)	(0.079)	(0.116)
Math	-0.237**	-0.001	-0.256***	-0.279**
	(0.099)	(0.091)	(0.091)	(0.117)
Information processing	-0.361***	-0.042	-0.137*	-0.270**
	(0.070)	(0.088)	(0.078)	(0.114)
Observations	45,475	45,528	45,900	46,214
Schools	553	564	576	587
Individual characteristics	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

Table A.6. Additional analyses: Grade retention.

	(1)	(2)	(3)	(4)	(5)
	Total sample	Sample excluding pupils with age > 12.5	Sample excluding schools in Amsterdam with decrease in age > 0.2	Sample excluding schools in Amsterdam with decrease in age > 0.15	Sample excluding schools in Amsterdam with decrease in age > 0.10
	Dependent variable: age				
Introduction of the ASIP	-0.079***		-0.039**	-0.026	-0.017
	(0.015)		(0.015)	(0.017)	(0.018)
	Dependent variable: dummy variable indicating age > 12.5				
Introduction of the ASIP	-0.032***		-0.011	-0.001	0.002
	(0.011)		(0.012)	(0.013)	(0.014)
	Dependent variable: total CITO test score				
Introduction of the ASIP	-0.170***	-0.139***	-0.174***	-0.140***	-0.159***
	(0.036)	(0.037)	(0.038)	(0.039)	(0.044)
Observations	78,545	66,823	76,820	76,027	75,195
Schools	614	614	606	602	598
Individual characteristics	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level. The individual characteristics include gender, age, age squared, home language, subsidy factor and pupil category.

Publisher:

CPB Netherlands Bureau for Economic Policy Analysis
P.O. Box 80510 | 2508 GM The Hague
T (070) 3383 380

January 2014 | ISBN 978-90-5833-627-9