Football to Improve Math and Reading Performance

Chris van Klaveren and Kristof De Witte

TIER WORKING PAPER SERIES
TIER WP 12/07

Football to Improve Math and Reading Performance*

Chris Van Klaveren†
Kristof De Witte‡

Abstract

Schools frequently increase instructional time to improve students' numeric and reading performance, but there is little evidence on the effectiveness of such an increase. This study evaluates 'Playing for Success', an extended day program for underachieving pupils that uses the football environment as a motivating force. Primary school pupils with low motivation and self-esteem are offered practical and sports related teaching content for 30 additional hours. The program is evaluated using a randomized controlled field experiment. The results indicate that Playing for Success does not significantly improve math and reading performance of primary school students.

Keywords: Achievement; Child Development; Evaluation; Motivation; Extended School Day.

*We are grateful to Marieke Heers for her valuable comments and suggestions. We would like to thank Playing for Success Eindhoven and Charlotte Jacobs of CITO for providing data. The usual disclaimer holds.

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

‡The corresponding author is affiliated with Maastricht University, Top Institute for Evidence Based Education Research (TIER), P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: k.dewitte@maastrichtuniversity.nl; and KU Leuven, Faculty of Economics and Business, 3000 Leuven, Belgium. Email: kristof.dewitte@econ.kuleuven.be.

1 Introduction

Motivated students obtain higher educational attainments (e.g., Pintrich and Schunk, 2002). On the contrary, low educational attainments discourage students. The interaction between motivation and learning outcomes creates a vicious circle in which poor results negatively influence motivation, in turn reinforcing the negative outcomes. Eventually, the discouraged students drop out of school (Tinto, 1975). To break the vicious circle, previous literature indicates that already early in the educational career attention should be paid to the motivation of pupils.

This paper focuses on the effects of such an initiative. In particular, it evaluates the effect of the program 'Playing for Success' by means of a field experiment in which children are randomly assigned to a control and an experimental group. Primary school students with low motivation and poor educational attainments are offered an extended school day program which takes place at a football (soccer) stadium. A learning activity taking place in an attractive environment and offering practical and sports related content is expected to motivate students which results in higher educational outcomes. This paper estimates the effect of 'Playing for Success' on the numeracy and literacy outcomes of students.

The 'Playing for Success' (PfS) program originates from the UK where First Division football clubs have been collaborating with schools since 1997. Since its introduction in the UK, it has been developed in the Netherlands (since 2008), Belgium (since 2011), Japan (since 2010) and Spain (since 2010).

This paper is not the first to examine the influence of PfS initiatives, it is, however, the first study that uses a field experimental design which allows for a more causal examination of the effect. Previous literature basically suffers from two major issues. First, previous literature does not focus on measuring the causal effect of PfS as these studies do not rely on a proper control group. For example, Sharp et al. (2001; 2003) found that participating primary school pupils improved their math performance with an effect that is equal to 17 months of regular primary school education. This is a substantial and significant progress in numeracy. The authors state that "gains in numeracy are particularly impressive, given the relatively short periods of time for which pupils attend (most pupils attend for less than 20 hours)." (Sharp, for Educational Research in England and Wales, 2003, p. 7). This favorable finding may, however, reflect that motivated pupils were selected to participate in the program which may drive the results. In Sharp's (2003) evaluation, the control group students were "selected to be as similar as possible to pupils attending PfS" (p.13). This corresponds to a matching design in which treated students are matched to control students

based on their observed characteristics. However, students' ethnicity and living environment is not a one-to-one relationship with motivation. Moreover, and worse, unobserved pupil characteristics, such as parental education and math test scores at the start of the program probably play a more important role than (the few) mentioned observed pupil characteristics. These issues are avoided if a field experiment is conducted in which students who are willing to participate in the program are randomly assigned to a treatment and a control group. The program effect can then be determined by comparing the math and reading performance between the treatment group and the control group.

A second drawback of previous studies is that they estimate a constant program effect, while there are many different locations (the so-called Centers) where the program is executed. Although all Centers aim to raise attainments in literacy, numeracy and ICT, they have their own peculiarities in, for example, the nature of the learning program, the course length, the capacity and resources (e.g., Sharp, Kendall and Schagen, 2003). This study avoids this pitfall by focusing on one implementation in the Netherlands (PSV Eindhoven, a successful and well-known football club from Eindhoven). We note that Schagen et al. (2002) apply multilevel models in their study to take into account that children are clustered in classes and locations. Nevertheless, the appliance of a multilevel model does not guarantee that proper control group is used and, as a consequence, the results of this study cannot be interpreted causally.

As mentioned above, this paper contributes to the literature by evaluating the PfS program by means of a field experiment. In a first step, schools were asked to list eligible students (i.e., with low motivation and self-esteem). Second, eligible students who wanted to enroll in the program were randomly assigned to a control and a Playing for Success (PfS) group conditional on past math and reading test scores. The PfS group participated in the program in the first 20 weeks of the school year, while the control group participated in the program in the last 20 weeks of the school year (and after a post test). The experimental setting with random assignment of children to the program ensures that both observed (e.g., past math performance) and unobserved (e.g., motivation) characteristics are accounted for.

The conclusions of this paper go beyond the PfS-setting. It can be compared to other initiatives which encourage learning outcomes and motivation through sports and role models (e.g., Bricheno and Thornton, 2007, Skelton, 2000 and Perkins, 2000). It can also be related to alternative extended school day programs (e.g., additional teaching hours or cultural classes).

The paper unfolds as follows. Section 2 describes the Playing for Success program.

Section 3 outlines the experimental design, while section 4 discusses the data and results. Finally, section 5 concludes.

2 Playing for Success

Playing for Success (PfS) targets underachieving students in primary/lower secondary education (aged – 9 till 14 years old).[1] Students who are offered the opportunity to participate in the program are selected by the school based on their (lack of) motivation, difficulties at home, socio-economic problems or low self-esteem. It is clear that this group of students is 'at risk' (for early references of programs targeting 'at risk' students: Manning and Baruth, 1995; Slavin et al., 1989; Finn and Rock, 1997). Given the voluntary basis of the PfS program, it is part of an extended school day. It does not replace the time at school, but takes place on a voluntary and additional basis. The program duration is 10 weeks with each week 2 till 3 hours at the soccer stadium.[2]

The underlying mechanism of the program originates from the social learning theory (Bandura, 1969), which argues that an improvement in self-efficacy is necessary for motivation and scholarly success. Therefore, the PfS program works around successful role models. The teaching is practical and related to the sport. For example, students are asked to write training programs, communicate with journalists, compute the size of the field, etc. In a real-life environment related to sports, both languages and numeracy are explicitly practiced, as well as ICT skills. The stadium replaces the traditional school environment which is for most participating pupils connected with fears and low esteem. It is expected that this increased self-esteem improves the educational attainments. During the program, participating pupils usually meet the first-team players, which works as a motivating force (see also Sharp et al., 2003b).

The program runs in close cooperation with the schools, sport club and the parents. Schools are asked to select the eligible students for the PfS program. This increases the probability for underachieving students with a low motivation to participate the program. Selection by the schools avoids selection bias (e.g., from highly motivated students, students from higher educated parents, or students loving football). PSV Eindhoven foresees in the facilities. The latter is regarded as a sign of social commitment by the club.

The program consists of an intensive guidance. For each 4 students there is a teacher,

[1] The UK website is http://www.playingforsuccessonline.org.uk.

[2] As discussed in Section 3, the program is implemented in two treatment groups such that the control group can only start after 20 weeks.

a social worker or a trainee (i.e., a student in vocational education or higher education - in particular students from teacher training programs, social & cultural work, youth helpers and students from sport studies). The different background of the employees guarantees a multidisciplinary team and a multidisciplinary experience for the participants. The project costs are funded for 25% by the sports club, while the remainder consists of public subsidies. The program costs that are not covered by PSV Eindhoven or firms that sponsor the initiative are about 350 euro per pupil.

The Dutch PfS program differs from the UK program by a less intensive use of ICT. While all students have their own computer, in the Dutch program the numeric and language skills are not offered by the use of ICT. As an additional difference, the Dutch PfS is more targeted at the individual participant. Participants determine their own learning goals, on which the employees work extensively. Despite these differences, we think that the outcomes of the paper in terms of literacy and math can still be compared to the implementation in other countries.

3 Experimental Design

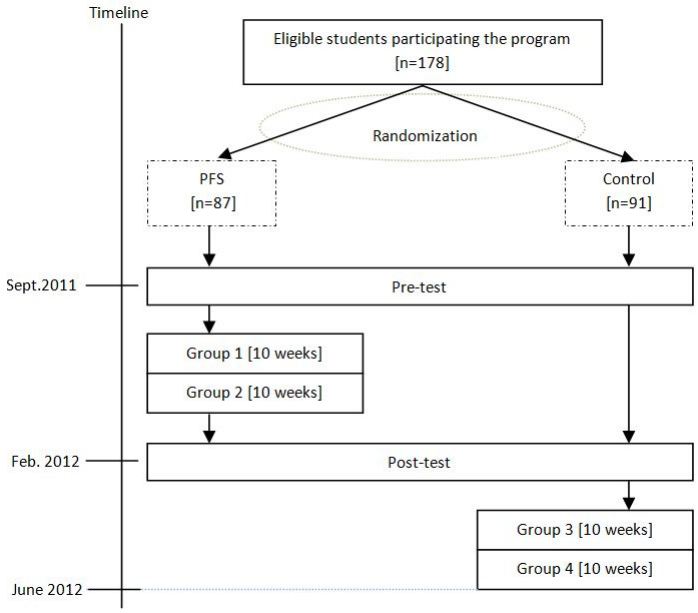
24 primary schools participate in the evaluated PfS program. Within each school a team of experts was formed to examine the cognitive and social-emotional development of children. Children who were labeled as underachievers were invited to participate in the experiment. Eventually there were 178 eligible children who wanted to participate in the program and whose parents gave their consent.

The experimental design is graphically illustrated in Figure 1. To avoid ethical issues, all eligible students who wanted to participate eventually receive the PfS treatment. However, students in the experimental group (later also referred to as PfS group) receive the treatment in the first 20 weeks of the school year (before the post-test), while students in the control group receive the treatment in the last 20 weeks of the school year (after the post-test). A national numeracy and literacy test taken in February 2012 serves as the post-test. Frequently, randomized field experiments are considered as unethical because it withholds children from program participation (see, for example, Rodrik, 2008). However, the conducted experiment was not deemed as unethical by schools and parents because all eligible children could participate in the program.

The 178 eligible children who participate in the experiment were paired within schools and classes conditionally on their math and comprehensive reading test scores of a national

and standardized test taken in February 2011 (i.e., the year before the post-test). For each child we drew a random number from a uniform distribution on the interval [0.1]. The child with the highest number within each pair was assigned to the PfS program in the first 20 weeks of the school year (i.e., the experimental group), while the child with the lowest number was assigned to the PfS program after the post-test was taken (i.e., the control group). The figure shows that children were unevenly divided over the experimental and the control group. This is due to the uneven number of potential participants in some classes. The adopted rule for these children was that they are assigned to the experimental group if the drawn number was higher than 0.5, and to the control group otherwise.

Figure 1. *Graphical Illustration of the Experimental Design*



To foster the attention of each student, both the experimental group and control group students are randomly divided in two groups. Working in small groups guarantees sufficient attention to the individual targets of the students (see Section 2). Group 1 and 2 completed the PfS program in the first 20 school weeks, while group 3 and 4 completed the program in after the post-test was taken. The learning activities and content was equal for all groups.

While an experimental design does not necessarily require a pre-test (given the randomization), we took a pre-test to improve the internal validity and efficiency of the results. The pre-test is taken in September 2011. The test consists of a standardized math and reading

test specifically designed for this study by the CITO (www.cito.com), a testing and assessment company that designs the national standardized tests that students normally make in primary education.

4 Experimental Data

Randomization within schools and classes conditionally on past math and reading performance does not ensure comparability of PfS group and the control group in other background characteristics. Table 1, therefore, presents background characteristics for children in the PfS and the control group. The table shows the mean and standard deviation of observed child characteristics. The last column presents the p-values and indicates if mean differences between the PfS and control group are significant.

In the Netherlands, girls and boys are usually evenly distributed over classes (see also Table 2). In the PfS experiment, Table 1 indicates a relatively large proportion of boys participating the PfS program. Because girls aged around 9 typically perform better on cognitive tests then boys, they might be less eligible to the program. It may also be that girls are less attracted to football and therefore, despite being eligible, do not enroll for the PfS program. The over-representation of boys does not affect the internal validity of the experiment (as pupils are randomly assigned to PfS and control group), but limits the external validity of the conducted experiment. The over-representation is in line with previous literature. Sharp and Schagen (2010), who examine for a similar PfS program in the UK if learning gains for boys are different than for girls within the PfS program, conclude that there is no evidence that the football emphasis in PfS prevents girls from participating in the program. They, however, mention that the program attracted a proportion of *pupils* (and not girls) from ethnic minority backgrounds that reflected the composition of the local population and comment that girls and boys were equally enthusiastic in their responses to the PfS experience. The latter comment refers to the group of children who already were interested to participate in the PfS program. Finally, it should be noted that Schagen et al. (2002) observed higher scores for reading and writing enjoyment.

Participating eligible students (i.e. the experimental group and the control group) are, on average, 10.5 years old. This is consistent with the observed grade proportions: eligible children mostly came from grades 4 and 5. Observed differences in grade proportions between the PfS and control group occur because the randomization happened within classes while sometimes there are an uneven number of children within classes. The proportion

of native Dutch children (i.e. children whose both parents are born in the Netherlands) is slightly higher in the control group than the proportion of native Dutch children in the PfS group, although the difference is not statistically significant. Note that the proportion of native Dutch children among eligible students participating in the experiment is somewhat lower than the proportion of Dutch children on the 24 participating schools (0.76; see Table 2). Apparently, Dutch children are less likely to participate in PfS experiment, either because they are less often labeled as underachievers or because they are less motivated to participate in the program. Nevertheless, based on the high proportion of Dutch children, we can conclude that the PfS program clearly does not primarily focus on non-Dutch students, which differs from many existing extended school day programs that tend to focus on disadvantageous and non-native students (see, e.g., Patall et al., 2010; Meyer and Van Klaveren, 2011).

Table 1. *Comparing children in the experimental and control group*

	Eligible students participating in the experiment				
	Experimental group		Control group		Differences
	Mean	Std. Dev	Mean	Std. Dev	Prob > |z|
Girl (= 1)	0.35	0.48	0.37	0.49	0.81
Age	10.43	0.98	10.57	0.92	0.56
Dutch (=1)	0.53	0.53	0.59	0.49	0.39
Grade 3 (US-terminology)	0.03	0.18	0.02	0.15	0.61
Grade 4	0.39	0.49	0.43	0.50	0.61
Grade 5	0.33	0.47	0.34	0.48	0.92
Grade 6	0.24	0.43	0.21	0.41	0.61
Father's education low	0.21	0.41	0.17	0.38	0.60
Father's education middle	0.38	0.49	0.42	0.50	0.60
Father's education high	0.33	0.47	0.30	0.46	0.60
Mother's education low	0.31	0.47	0.31	0.46	0.97
Mother's education middle	0.36	0.48	0.42	0.50	0.40
Mother's education high	0.23	0.42	0.22	0.42	0.87
Reading Feb. 2011 (assignment test)	26.54	15.95	28.74	14.72	0.34
Reading Sept. 2011 (pre-test)	27.81	18.52	27.58	17.93	0.98
Math Feb. 2011 (assignment test)	75.32	20.39	77.97	18.71	0.53
Math Sept. 2011 (pre-test)	75.44	20.56	77.44	17.54	0.83
Number of Observations	87		91		

The highest attained parental education level is categorized by three education levels: low (i.e., without formal education, primary or lower secondary education), middle (i.e., higher secondary or vocational education) and high (i.e., bachelor or academic education). The descriptive statistics show that the observed education levels of the parents are approximately evenly distributed over the three categories. The latter shows again that the PfS program does not focus on disadvantageous children who often have lower educated parents (and come from economically disadvantaged neighborhoods). Given that there is a positive causal effect of parent's schooling on children's schooling (Holmlund et al., 2011) and that the program welcomes but is not aimed specifically at children with a diagnosed disorder, we would expect that children of higher educated parents are more likely to be underachievers. When we compare the proportions of mothers and fathers in the three education categories between the PfS and the control group we find that the means (and medians) are not statistically

different between the two groups.

The lower panel of Table 1 shows the descriptive statistics on achieved math and reading test scores in February 2011 and September 2011. The February test scores were used to assign children to the experimental group and the control group. The September test scores are standardized tests developed specifically for this study (see also Section 3), and measure the math and reading ability of children just before the PfS program started for the experimental group. The September test scores serve therefore as pre-test scores.

In Appendix A we describe in detail how observed test scores of children relate to the achievement levels that these children are supposed to obtain. In general and in line with the program, we find that the achievement levels of the eligible children are below the appropriate achievement levels. While this does not intricate the (internal and external) validity of the analysis, it might influence the estimated effect size as low performing students typically learn slower. This would result in a lower bound estimate.

The descriptive statistics in Table 1 show that the mean test scores in February 2011 and September 2011 are not statistically different between the PfS (experimental) and the control group. Morover, the September 2011 test scores appear to be very similar to the February 2011 test scores, which gives the impression that children have not (or only marginally) improved their math and reading skills. There are three explanations for this observation. First of all, children made the math and reading test in September during a test day that was organized at the football stadium of PSV Eindhoven.[3] The test day was organized as a festive day, such that the first experience with the Playing for Success program was a positive one. It is therefore likely that the children's math and reading tests scores were negatively influenced by the relatively more 'noisy' test day environment. That the test day was indeed a more 'noisy' environment is also reflected by the standard deviation of the reading test in February that is lower than that of the reading test in September. The standard deviations for the math tests appear to be very similar, which might be due to the moment of testing: the math test was taken in the morning, while the reading test was taken in the afternoon. However, children were randomly assigned to the program, such that the produced noise by the test day environment was randomly distributed as well. So even though the September test scores are downward biased, they still indicate properly that the math and reading skills of PfS and control group children are not statistically different at the beginning of the PfS program.

The second explanation for the low test scores in September is that the acquired math

[3]To be precise: the September test was taken on August 30th, 2011.

and reading skills during the school year are partly lost during the summer holidays. This is generally referred to as the summer dip and has led to many summer school programs that aim at reducing the effects of such a summer dip (see, for example, Cooper et al., 2000; Jacob and Lefgren, 2004; Borman and Dowling, 2006).

As a final, but unlikely, explanation, the test might have a different difficulty level. This is unlikely because both tests are made by CITO, which also designs the regular national math and reading tests that are taken in primary and secondary education. Moreover, children are assigned to secondary education tracks based on the CITO test that children take at the end of grade 6. Therefore, the tests can be considered as valid and reliable.

Table 2. *Comparing children in the experimental and control group*

	Mean	Std. Dev
Proportion child weight .3	0.11	0.08
Proportion child weight 1.2	0.15	0.17
Proportion children in disadvantageous areas	0.33	0.47
Proportion Dutch	0.76	0.08
Proportion boys	0.51	0.51
Number of students at the school	194.75	143.91
Number of Observations	178	

Table 4 shows the average school characteristics for all children participating the experiment. Because the randomization happened within schools and classes there is no need to present these statistics separately for children in the PfS and the control group. The first and second row denote the proportion of children with a disadvantageous background. Schools receive more subsidies if there are more children with higher needs. The Ministry of Education distinguishes between two types (and thus two weights) based on the education level of the parents. A weight of .3 is assigned to children who have parents without at least a higher secondary diploma, while a weight of 1.2 is assigned to children who have one parent without at least lower secondary diploma, and another parent with at most a lower secondary diploma. Table 2 shows that 11% of the children receive a .3 weight and that 15% of the children receive a 1.2 weight. Official school-level registration data on all Dutch primary schools indicate that at the average primary school 9% of the children receive a .3 weight and 6% of the children receive a 1.2 weight. It is therefore clear that the schools under study have a rather disadvantageous student population.

In a similar fashion we can compare other observed characteristics relatively to the Dutch average. Disadvantageous areas denote neighborhoods with an above average combination of unemployment, early school leaving, criminality and low income. They receive additional funding by the central government. The proportion of children living in a disadvantageous area is with 0.33 relatively high compared to that of the average primary school (0.15). The proportion of native Dutch children is with 0.76 slightly higher than that of the average primary school (0.71). This means that even though children are more often living in disadvantageous neighborhoods the proportion of non-Dutch students is relatively small. This might be due to selection in schools along ideology (see also De Witte and Van Klaveren, 2012). Finally, the proportion of boys is similar to that of the average primary school.

For 13 out of the 178 participating children we do not observe post-test scores. Consequently, these children cannot be considered in the empirical analysis. Table 3 characterizes these children and examines the selective nature of their experimental dropout. The 13 children are evenly distributed over the PfS and the control group which seems to indicate that the randomization process was successful in the sense that the incidence of not observing post-test scores was randomly distributed over the two groups. 7 of the 87 children in the PfS group dropped out of the program. Two of these children dropped out because of practical reasons and will enroll in the program again in the second semester, one child dropped out of the program because she moved to another school and 4 children did not want to continue with the program for unknown reasons. For 6 of the 91 children in the control group we do not observe the post-test scores. One child moved to another school, but for the other children in the control group it is unclear why they did not take the post-test. Because of the small number of children dropping out of the program, in Table 3, we cannot determine (parametrically nor non-parametrically) if the means are significantly different between the two groups. Consequently, we do not show the standard deviations. Table 3 reveals that parental education of dropouts in the experimental group is, on average, higher than parental education of dropouts in the control group. On the other hand, children who drop out of the control group are more often native Dutch and higher achieving. Nevertheless, considering that the pre-test scores of children assigned to the control group are somewhat (but not significantly) higher than those of children assigned to the PfS group, and given the small percentage of experimental dropout, it is unlikely that the observed differences of children who dropped out of the program will bias the estimated program effect.

Table 3. *Comparing dropout children in the PfS and control group*

	Dropouts	
	PfS	**Control**
	Mean	Mean
Girl (boy = 0)	0.29	0.33
Age	10.33	10.70
Dutch (=1)	0.29	0.83
Grade 3 (US-terminology)	.	0.17
Grade 4	0.43	0.33
Grade 5	0.43	0.33
Grade 6	0.14	0.17
Father's education average*	2.50	1.5
Mother's education average	2.00	1.67
Reading Feb. 2011 (assignment test)	28.00	36.17
Reading Sept. 2011 (pre-test)	32.77	25.17
Math Feb. 2011 (assignment test)	73.57	80.00
Math Sept. 2011 (pre-test)	76.27	79.33
Number dropouts	7	6
Number of assigned children	87	91

* Father's education has 2 missing observation.

5 Estimation Results

The evaluation of the results proceeds in three steps. First, we examine by simple t-tests the difference between the pre- and post-test scores in the control and PfS (experimental) group. The upper panel of Table 4 presents the reading estimates, while the lower panel presents the math estimates. The first two rows of each panel show the mean pre- and post-test scores, while the last column shows if the post-test scores differ significantly from the pre-test scores. The third row of each panel shows if the mean pre- and post-test scores and the difference between the two significantly differs between the PfS group and the control group.[4]

[4]Note that we ran the analysis also separately for the two groups within the experimental group as it might be argued that the learning outcomes of group 1 have already been fade out by the time of the post-test. The results for those axillary regressions were robust.

Table 4. *Mean pre- and post-test scores for PfS and control group*

	Pre-test scores		Post-test scores		Pre-post differences	
Reading:						
PfS Group	26.23	(2.10)	36.04	(1.92)	9.81 ***	(1.80)
Control Group	28.14	(2.10)	39.01	(1.69)	10.88***	(1.49)
PfS-control difference	-1.90	(2.98)	-2.97	(2.55)	-1.07	(2.32)
Math:						
PfS Group	75.42	(2.43)	88.64	(2.14)	13.21***	(1.84)
Control Group	77.80	(1.95)	91.52	(1.63)	13.72***	(1.31)
PfS-control difference	-2.37	(3.11)	-2.88	(2.68)	-0.51	(2.25)

Note 1: Standard errors are shown in parentheses.

Note 2: */**/*** means statistically significant at the 10/5/1 percent level.

The results indicate that the reading performance of children in the PfS group increased with 9.8 points, while the reading performance of children in the control group increased with 10.9 points. To put those figures in perspective, we relate them to the expected performance increase as indicated by the testing institute CITO. Table 5 shows for each grade how much children should improve their math and reading skills per year, according to the standards that are set by testing institute CITO. As younger children learn faster, Table 5 indicates that children in lower grades should improve their achievement levels more than children in higher grades. The average improvement over all grades amounts to 12.75 points for math and 10.10 points for reading. Unfortunately, these improvement points cannot directly be related to the increase in math and reading performance of children in PfS and control group, as these children are unevenly distributed over grades. We avoid this drawback by constructing weighted averages indicating how much children in the PfS and control group should have improved their math and reading performance conditional on the numbers of children that enrolled in each grade. This is presented on the right hand side of Table 5. The weighted average performance increases in Table 5 reveal that children in the control group should obtain a higher performance increase on numeracy and reading than students in the PfS group.

Relating the theoretical average performance increase of Table 5 to the observed performance increase of Table 4, one could conclude that the observed average improvement of the children in the experimental group is more than sufficient, considering the standards. It is, however, important to recognize that the pre-test scores of these students are lower than the pre-test scores that children should achieve according to the standard. Therefore it is only

natural that the performance increase is more than sufficient.

Test score improvements of children in the control group are somewhat higher than those of children in the PfS group, but this is mainly due to a small, and statistically insignificant, difference in pre-test scores.

Table 5. *Average expected performance increase according to testing institute CITO*

	Assumed increase			Proportion	
	Math	Reading		PfS (87)	Controls (91)
Grade 3	19.75	16.34		0.03	0.02
Grade 4	13.5	8.71		0.39	0.43
Grade 5	12.25	8.84		0.33	0.34
Grade 6	5.50	6.50		0.24	0.21
Average	12.75	10.10	Weighted average Reading	8.36	8.44
			Weighted average Math	11.22	11.52

As a second evaluation procedure, we estimate the treatment effect more precisely by an ordinary least squares (OLS) regression. Light et al. (1990) showed that the statistical power is maintained at half the sample size if a regression analysis is performed with a set of covariates that predicts about half of the variation in the outcome jointly. The following model is estimated:

$$Y^s_{post,i} = \alpha_0 + \alpha_1 X_i + \alpha_2 PFS_i + \alpha_3 Y^s_{Sept,i} + \alpha_4 Y^s_{Feb,i} + \varepsilon_i \tag{1}$$

where $Y^s_{post,i}$ denotes the standardized post-test score of student i for subject s (s=$m$ath, $r$eading), X_i represents a vector of student and school characteristics, and PFS indicates if the student participated in the playing for success program. $Y^s_{Sept,i}$ denotes the September pre-test score and $Y^s_{Feb,i}$ represents the February test score which is included to account for the difference between the September test score and the February test score. As usual, the error term, ϵ_i, is assumed to be normally distributed with mean zero and variance σ^2_ϵ and all explanatory variables are assumed independent of the error term. The estimation results for math and reading are presented in Table 6.

Table 6. *Regression estimates of the Playing for Success effect for the treated*

	Math				Reading			
	(1)		(2)		(3)		(4)	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
PfS participation (1=yes)	-0.054	(0.098)	-0.042	(0.098)	-0.058	(0.101)	-0.065	(0.103)
$Y^s_{Sept,i}$	0.016 ***	(0.004)	0.016 ***	(0.004)	0.019***	(0.003)	0.020***	(0.004)
$Y^s_{Feb,i}$	0.026 ***	(0.004)	0.027 ***	(0.004)	-0.026***	(0.005)	-0.027***	(0.005)
Constant	-3.940***	(0.967)	-3.013*	(1.595)	-1.585*	(1.778)	-1.980	(1.788)
Child Controls	Yes		Yes		Yes		Yes	
School Controls	No		Yes		No		Yes	
N	165		165		165		165	
Adj.R^2	0.66		0.66		0.61		0.60	

Note: Standard errors are printed in parentheses and */**/*** means statistically significant at the 10/5/1 percent level.

The left panel of Table 6 shows the results for math, while the right panel presents the results for comprehensive reading. For each subject two model specifications are estimated that control for achieved September and February test scores as well as child control variables. The latter include gender, age, ethnicity, a set of grade dummies and a set of dummy variables that indicate the education level of the father and mother (i.e. those mentioned in Table 1). Models (1) and (3) show the estimation results when we do not include school-level controls, while school-level control variables are included in models (2) and (4). The included school-level variables are similar to those in Table 2. Including school-level control variables does not improve the explained variation in the outcome variables nor does it influence the estimated effect of the treatment.

To be more precise: in estimating the model with only child-level background characteristics, an adjusted-R^2 of 0.30 for math and 0.26 for reading is observed. By adding the pre-test score variables, the adjusted-R^2 increases to 0.66 and 0.61 for respectively math and reading. Adding additional school-level variables does not increase or even decreases (for reading) the adjusted-R^2. The estimation results become thus less efficient by including school-level variables. In the discussion below, we therefore focus on the estimation results of models (1) and (3). Note that the estimation results are similar when the standard errors are clustered at the class or the school level. Moreover, we performed quantile regressions to take into account possible non-linear effects of the PfS program. However, the estimation results were similar to those reported in Table 5.

The results further indicate that children have higher post-test scores if they have higher test scores in September and February 2011. The estimation outcomes reveal that the February 2011 test scores correlate better with the post-test scores than the September test scores. In Section 4 we explained that this is likely caused by the noise that was produced by the festive test day environment in September. The pre- and post-test correlation tables are shown in Appendix B.

Finally, and most importantly, the estimation results show that children in the experimental group do not perform significantly better than children in the control group. The estimate is even negative, which is likely due to the higher pre-test scores of the control group children (note that these were not significantly different from the experimental group children). The effect size of about -0.06 for both math and reading is close to zero. Both the difference-in-differences in the simple t-test in Table 4 and the linear regression estimates therefore indicate the PfS program did not effectively increased the math and reading performance of the program participants.

6 Conclusion and Discussion

This paper reports the effects an extended school day program. The program targets underachieving students (i.e. students with low motivation, difficulties at home, socio-economic problems or low self-esteem) and aims to improve their learning outcomes on math and reading by working in a football stadium around role models and practical exercises. The program called 'Playing for Success' (PfS) is popular in the UK, Belgium, Japan, Spain and the Netherlands. This paper evaluates the effectiveness of the Dutch program. For the first time, a randomized field experiment is conducted to determine the causal impact of the intervention. The randomization of eligible students avoids any selection biases on observed and unobserved characteristics.

The empirical results suggest that the PfS program did not improved the math or reading performance of the participating children. We see three possible explanations for the absence of an effect. First of all, the program duration of 10 weeks might be too short. Second, we did not test for the effects on motivation but only for educational attainments. Besides reading and numeracy outcomes, the program explicitly targets motivation and self-esteem of students. While the former might not have been changed, the latter might have changed. Third, playing for success may have increased the math and reading performance of all eligible children if control group children were contaminated by children in the experimental

group. Such a contamination effect means that the estimation results indicate that the PfS program is not effective, while in reality the program does have positive effect.

As a note for a broader discussion, this paper illustrates that studies that focus on the internal validity of estimators tend to find more modest outcomes than studies that do not focus in particular on the internal validity of estimators. While previous papers on Playing for Success observed (huge) positive outcomes of the program, no effect on learning outcomes was observed in the first field experimental study in which students were randomly assigned to a treatment and control group. Given the absence of any selection effects, the modest outcomes might not be a surprise. Nevertheless, there is often a pressure by policy makers, stakeholders and even academic journals to 'find' and report significant outcomes (which results in a publication bias). Therefore, this paper would like to trigger the debate on the comparison of effect sizes in both randomized and non-randomized studies.

7 Appendix A

In Dutch primary schools, children take a national standardized math and reading test (usually twice a year). This Appendix, and in particular Table A.1, explains how children's test scores are linked to achievement levels within primary school grades. The second column of Table A.1 shows to which level a certain test score belongs. The numbers in this column refer to the primary school grade, *Mid* refers to the achievement levels that children should have in the middle of the school year and *End* refers to the achievement levels that children should have at the end of the school year.

Table A.1. *Relation between achieved test-scores and achievement level*

		Math		Reading	
Nr.	Level	Lower Bound	Upper Bound	Lower Bound	Upper Bound
1	< Mid1	-99.00	38.99	-99.00	-5.24
2	Mid1	39.00	44.00	-5.23	0.43
3	Mid1 - End1	44.01	48.99	0.44	6.07
4	End1	49.00	54.25	6.08	9.98
5	End1 - Mid2	54.26	59.74	9.99	12.10
6	Mid2	59.75	64.50	12.11	15.46
7	Mid2 - End2	64.51	68.49	15.47	20.02
8	End2	68.50	73.50	20.03	24.37
9	End2-Mid3	73.51	79.49	24.38	28.44
10	Mid3	79.50	84.00	28.45	31.16
11	Mid3-End3	84.01	86.99	31.17	32.46
12	End3	87.00	90.00	32.47	34.47
13	End3 - Mid4	90.01	92.99	34.48	37.15
14	Mid4	93.00	96.25	37.16	40.00
15	Mid4 - End4	96.26	99.74	40.01	42.99
16	End4	99.75	102.75	43.00	45.00
17	End4 - Mid5	102.76	105.24	45.01	45.99
18	Mid5	105.25	107.50	46.00	47.25
19	Mid5 - End5	107.51	109.49	47.26	48.74
20	End5	109.50	112.00	48.75	50.50
21	End5 - Mid6	112.01	114.99	50.51	52.49
22	Mid6	115.00	118.00	52.50	54.50
23	> Mid6	118.01	999.00	54.51	999.00

We are grateful to Charlotte Jacobs (CITO) for providing us with these statistics.

Let us consider the average math test score in February 2011 of PfS children who were enrolled in grade 4 in school year 2011-2012. Table 4 only shows an average math test score of 75.32, which represents the test scores of all PfS children who were enrolled in grades 3, 4, 5 and 6 in school year 2011-2012. The average math test scores achieved in February 2011 by PfS children who were enrolled in grade 4 is 69.58. It is important to recognize that these children where still in grade 3 in February 2011, and so the achievement levels that these children are supposed to have are shown in Table A.1 in row number 10, where column 2

indicates $Mid3$. The lower and upper bound math score interval for $Mid3$ is [79.5;84] and therefore we can conclude that the math level that PfS children were supposed to have was below level. The average test score of 69.58 indicates that these children had an achievement level that resembles the math level that children should have at the end of grade 2 (see Table A.1, rows 7 and 8).

In a similar fashion we can also examine if test scores of children who were enrolled in other grades were at the appropriate level. On average, we find for the eligible students that the achievement levels are below the appropriate achievement levels, which is in line with the program design.

In the main text we have not elaborately discussed if children in the experimental group are at the appropriate achievement levels, because this is not of importance for the evaluation of the playing for success program, given that children are randomly assigned to the PfS group and the control group. We do, however, emphasize in Section 4 that the achievement levels of children in the experimental group are below the appropriate achievement levels and that the focus of the program on children who are labeled as underachievers can influence the size of program effect that we will find.

8 Appendix B

In this appendix we show the correlations between the reading and math post-test scores and the pre-test scores in February and September in Table B.1. The table shows for each subject that both pre-tests correlate highly with the post-test. Moreover, it shows that the pre-test in February 2011, that is used to assign children to the PfS program, correlates better with the post-test scores than the September pre-tests.

Table B.1. *Pre- and post-test correlations*

	Math			**Reading**		
	Post-test	Pre-test Feb.	Pre-test Sept.	Post-test	Pre-test Feb.	Pre-test Sept.
Post-test	1.00	.	.	1.00	.	.
Pre-test Feb.	.77	1.00	.	.72	1.00	.
Pre-test Sept.	.70	.70	1.00	0.65	.59	1.00

References

Bandura, A. (1969), 'Social-learning theory of identificatory processes', *Handbook of socialization theory and research* pp. 213–262.

Borman, G.D. and N.M. Dowling (2006), 'Longitudinal achievement effects of multiyear summer school: Evidence from the teach baltimore randomized field trial', *Educational Evaluation and Policy Analysis* **28**, 25–48.

Bricheno, P. and M. Thornton (2007), 'Role model, hero or champion? children's views concerning role models', *Educational research* **49**(4), 383–396.

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

De Witte, K. and C. Van Klaveren (2012), 'Comparing students by a matching analysis–on early school leaving in dutch cities', *Applied Economics* **44**(28), 3679–3690.

Finn, J.D. and D.A. Rock (1997), 'Academic success among students at risk for school failure.', *Journal of Applied Psychology* **82**(2), 221.

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

Jacob, B.A. and L. Lefgren (2004), 'Remedial education and student achievement: A regression-discontinuity analysis', *Review of Economics and Statistics* **86**, 226–244.

Light, R.J., J.D. Singer and J.B. Willett (1990), *By design: Planning research on higher education*, Harvard Univ Pr.

Manning, M.L. and L.G. Baruth (1995), *Students at Risk.*, Allyn and Bacon, 160 Gould Street, Needham Heights, MA 02194-231.

Meyer, E. and C. Van Klaveren (2011), Evaluation of an extended day program in the netherlands: A randomized field experiment, TIER Working Paper Series 11-02.

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

Perkins, S. (2000), 'Exploring future relationships between football clubs and local government', *Soccer and Society* **1**(1), 102–113.

Pintrich, P.R. and D.H. Schunk (2002), *Motivation in education: Theory, research, and applications*, Merrill Upper Saddle River^ eN. J NJ.

Rodrik, D. (2008), The new development economics: We shall experiment, but how shall we learn?, Working Paper Series 08-055, Harvard University, John F. Kennedy School of Government.

Schagen, I., L. Kendall and C. Sharp (2002), 'Measuring the success of'playing for success", *Educational Research* **44**(3), 255–267.

Sharp, C., Kendall L and I. Schagen (2010), 'Different for girls? An exporation of the impact of Playing for Succes', *Educational Research* **45**(3), 309–324.

Sharp, C., L. Kendall and I. Schagen (2003), 'Different for girls? an exploration of the impact of playing for success', *Educational Research* **45**(3), 309–324.

Sharp, C., L. Kendall, S. Bhabra, I. Schagen and J. Duff (2001), 'Playing for success: an evaluation of the second year', *DfES Research Brief No RB291, September* .

Sharp, C., National Foundation for Educational Research in England and Wales (2003), *Playing for success: An evaluation of the fourth year*, DfES Publications.

Skelton, C. (2000), ''a passion for football': Dominant masculinities and primary schooling', *Sport, Education and Society* **5**(1), 5–18.

Slavin, R.E., N.L. Karweit and N.A. Madden (1989), *Effective programs for students at risk.*, Allyn & Bacon.

Tinto, V. (1975), 'Dropout from higher education: A theoretical synthesis of recent research', *Review of educational research* **45**(1), 89–125.

TIER WORKING PAPER SERIES
TIER WP 12/07

ISBN 978-94-003-0043-9