Heterogeneity in Short- and Long-term Impacts of School-Wide Positive Behavior Support (SWPBS) on Academic outcomes, Behavioral outcomes, and Criminal Activity

Nicolai Topstad Borgen\textsuperscript{1,2}, Oddbjørn Raaum\textsuperscript{2}, Lars Johannessen Kirkebøen\textsuperscript{3}, Mari-Anne Sørlie\textsuperscript{4}, Terje Ogden\textsuperscript{4}, and Ivar Frønes\textsuperscript{1,4}

\textsuperscript{1} Department of Sociology and Human Geography, University of Oslo, Norway
\textsuperscript{2} Ragnar Frisch Centre for Economic Research, Oslo, Norway
\textsuperscript{3} Statistics Norway, Oslo, Norway
\textsuperscript{4} Norwegian Center for Child Behavioral Development, Oslo, Norway

Corresponding author
Nicolai T. Borgen, University of Oslo, P.O. Box 1140 Blindern, 0318 Oslo, Norway.
E-mail: n.t.borgen@isp.uio.no

Acknowledgments
We thank Catherine Bradshaw, Dan Olweus, Kyrre Breivik, Mona Elin Solberg, and Bo Vinnerljung for comments on an earlier version of this manuscript, presented at the PALSOBPP workshop at the Norwegian Center for Child Behavioral Development in 2019. We thank two anonymous reviewers for useful comments and suggestions. We thank Asgeir Røyrhus Olseth for help on preparing the implementation data. The register data were made available by Statistics Norway.

Declaration of interest
This study was financed by a grant from the Research Council of Norway (Grant #238050). Borgen, Kirkebøen, and Raaum have no interests to declare. They are responsible for choosing the research design and analyses. Most data management and estimations have been performed by Borgen, who also wrote the first complete draft in cooperation with Kirkebøen and Raaum. Ogden and Sørlie have provided the school-level data on the SWPBS intervention, the presentation of the model, and contributed to the choice of design as well as editing of the paper. Frønes have contributed to the choice of design and with the editing of the paper. The development department at the Norwegian Center for Child Behavioral Development is responsible for the adaptation, training, and implementation of SWPBS in Norway. Frønes, Ogden, and Sørlie work at the research department at the Norwegian Center for Child Behavioral Development, which is not involved in any parts of the implementation.
Abstract
To address social and behavioral problems in schools, more than 26,000 schools around the world have implemented School-Wide Positive Behavior Support (SWPBS). Previous studies have focused on the effects of SWPBS on short-term teacher-rated behavior such as office discipline referrals or academic outcomes, but no study has yet investigated effects on long-term student outcomes. We use population-wide longitudinal register data, including all Norwegian students that are exposed to SWPBS, and examine effects on short- and long-term academic outcomes, as well as long-term school behavior and youth crime. Both when we evaluate average program effects for all students and when looking at at-risk students only, we find no indications that the Norwegian SWPBS affected any of these outcomes.

Keywords: SWPBS, behavioral problems, high-risk students, register data, long-term, scale-up

Introduction
Students with behavioral problems are a vulnerable group, with a major risk of academic failure, criminal careers, health problems, and poor social relationships (Evensen, Lyngstad, Melkevik, & Mykletun, 2016; Hinshaw, 1992; McLeod & Kaiser, 2004; Wertz et al., 2018). Additionally, these students’ disruptive classroom behavior adversely affects their classmates’ learning outcomes (Aizer, 2008; Carrell & Hoekstra, 2010; Figlio, 2007; Fletcher, 2010; Lazear, 2001). Persistent adverse effects of exposure to students with behavioral problems have been found in adult earnings (Carrell, Hoekstra, & Kuka, 2016). There are also some suggestions that peers influence individuals’ likelihood of engaging in antisocial and criminal behavior (Billings & Hoekstra, 2019; Kim & Fletcher, 2018), although the findings are mixed (Ludwig & Kling, 2007).

The consequences of behavioral problems in schools and neighborhoods have led to an increasing interest among policymakers in interventions to reduce antisocial behavior (Caspi et al., 2017; Duncan & Magnuson, 2013). The School-Wide Positive Behavior Supports (acronymized as SWPBS or SWPBIS) model is an evidence- and theory-based, data-driven, and systematic whole-school prevention framework implemented in more than 26,000 schools in
the United States and internationally, and which in the United States is funded by the Department of Education (Gage, Whitford, & Katsiyannis, 2018; Pas, Ryoo, Musci, & Bradshaw, 2019). The primary aims of SWPBS (referred to as PALS in Norway) are to reduce and prevent externalizing problem behaviors in school, including frequent but minor problems such as unruly, noisy behavior, and “mental absence” during lessons, as well as less frequent but more severe problem behaviors such as aggression, truancy, bullying, harassments, and vandalism. By promoting a positive, predictable, and supportive learning environment, SWPBS is commonly hypothesized to improve academic achievements, although this has not been the main focus of the intervention model (Gage, Sugai, Lewis, & Brzozowy, 2015).

An extensive literature has found that SWPBS successfully addresses social and behavioral problems in the short-term (e.g., Bradshaw, Waasdorp, & Leaf, 2012). These short-term improvements are the main objectives of the model and are essential for the wellbeing and safety of students and teachers. Our contribution is a study of whether the model has effects that extend beyond these primary goals. In particular, we examine the intervention’s scope for preventing students’ long-term academic failure and marginalization. The ultimate goal of interventions such as SWPBS is to produce durable outcomes (Horner, Sugai, & Anderson, 2010), and thereby reduce the likelihood of chronic and more intense problems (Mitchell, Hatton, & Lewis, 2018). However, most U.S. studies only investigate effects on SWPBS constructs such as office discipline referrals (ODRs) (Chitiyo, May, & Chitiyo, 2012), and they have a short follow-up period (Madigan, Cross, Smolkowski, & Strycker, 2016). No study has yet investigated whether the short-term effects of SWPBS have lasting effects on student behavior after students leave an SWPBS school. From a policy perspective, persistent individual student effects that extend beyond the main goals would strengthen the case for implementing the SWPBS model.
In this paper, we compare the outcomes of students from SWPBS schools with other students after they have graduated. We examine short- and long-term effects of SWPBS on individual-level variables using population-wide Norwegian register data. Using a difference-in-difference (DiD) design, we investigate whether SWPBS affects short-term test scores and long-term academic grades, high school dropout, school behavior, and youth crime. While previous research from Norway has not found any short-term effect of SWPBS on bullying and school wellbeing (Borgen, Kirkebøen, Ogden, Raaum, & Sørlie, 2019), research has found effects on other behavioral outcomes (Sørlie, Idsoe, Ogden, Olseth, & Torsheim, 2018; Sørlie & Ogden, 2015) and classroom order (Borgen et al., 2019). Consequently, this paper investigates whether the short-term effects of SWPBS on behavioral outcomes in Norway produce a durable impact on academic failure and marginalization.

Furthermore, we examine whether program effects are more substantial for students at risk of academic failure and behavioral problems. SWPBS is a multi-tiered model with universal interventions targeting all students (tier I) and individually tailored interventions targeting at-risk students (tiers II and III). Students at high risk of academic failure and behavioral problems may benefit more from the universal intervention than other students, and they also receive targeted interventions. However, most studies have investigated the effects of solely the universal tier, and only two studies have looked at differential effects for students with varying needs and levels of risk (Bradshaw, Waasdorp, & Leaf, 2015; Sørlie et al., 2018). The present study examines the intervention effects of the combined three-tiered model. The width of the register data allows us to forecast academic failure and behavioral problems from childhood risk (Caspi et al., 2017). We build upon the main principles of machine learning to identify students at various risk levels and use a DiD model to estimate program effects.
A major concern in the intervention literature is the extent to which the evidence from controlled environments is reproduced in actual educational settings (Hulleman & Cordray, 2009). Results in controlled trials, where the program is evaluated under optimal conditions of delivery, are likely to be stronger than in interventions that are scaled-up under standard institutions (Bradshaw et al., 2015; Flay et al., 2005; Hulleman & Cordray, 2009). For instance, units who volunteer to participate in RCTs may differ from the target population in important aspects (Stuart, Bradshaw, & Leaf, 2015), such as being more likely to implement interventions in accordance with the model (Pas & Bradshaw, 2012). An intervention’s effects may decrease by 25-50% when going to scale due to deviations or dilutions from the original model (i.e., low fidelity) (Welsh, Sullivan, & Olds, 2010). Thus, it is crucial to examine the effectiveness of interventions in a scale-up (Flay et al., 2005). The population-wide register data allows us to explore the effects of SWPBS having been operated under normal conditions and scaled up to include about 10% of the student population.

**Literature review**

**Behavioral problems**

A large body of research has examined the effects of SWPBS on behavioral problems in U.S. schools, with results reviewed in Horner et al. (2010), Chitiyo et al. (2012), and Mitchell et al. (2018). Based on a review of 46 studies between 2000 and 2009, Horner et al. (2010) found strong evidence for positive effects of SWPBS and argued for a large scale implementation. However, Chitiyo et al. (2012), reviewing studies between 1990 and 2011, argued that less than a third of effect studies are grounded in rigorous designs, with most being case studies and cross-sectional studies. Based on ten rigorous studies, they concluded that substantial evidence in favor of SWPBS is still lacking and more research is needed. The most recent
review study, carried out by Mitchell et al. (2018), examined the evidence of SWPBS based on studies that used group experimental designs (quasi-experimental or RCT). Based on 12 published papers from five separate studies, they found strong support for positive short-term effects of SWPBS on externalizing problems in schools such as office discipline referrals, suspension rates, and attendance.¹

Previous research has shown that school interventions may have effects on behavioral problems that emerge late, such as criminal behavior (Flay et al., 2005; Griffin, Botvin, & Nichols, 2004). Based on various developmental models (Muthén et al., 2002; Nagin & Tremblay, 1999), reducing the likelihood of minor behavioral problems and difficulties occurring and, if they occur, prevent them from escalating to more severe problem behavior, may alter the direction of youth at risk of adolescent and adult behavioral problems. By promoting the social skills of students in addition to effective prevention of problem behaviors, the SWPBS model has been hypothesized to reduce the risk that students enter a dysfunctional developmental trajectory that in the long run may result in criminal activity, mental health problems, and labor market exclusion (Ogden, Sørlie, & Hagen, 2007).

Contrary to this optimistic perspective of developmental trajectories, other perspectives suggest that short-term early interventions cannot produce sizable long-term gains and that short-term program effects may fade-out over time. For example, according to biological-rooted perspectives, student outcomes are a result of a complex dynamic interplay between individuals’ genetic predispositions and their environmental exposure. Changes to

¹ The manuscripts that were reviewed was: Sprague et al. (2001), Bradshaw, Koth, Bevans, Ialongo, and Leaf (2008), Bradshaw, Reinke, Brown, Bevans, and Leaf (2008), Bradshaw, Koth, Thornton, and Leaf (2009), Bradshaw, Mitchell, and Leaf (2010), Bradshaw, Waasdorp, et al. (2012), Waasdorp, Bradshaw, and Leaf (2012), Pas, Waasdorp, and Bradshaw (2015), Bradshaw et al. (2015), Horner et al. (2009), Bradshaw, Pas, Goldweber, Rosenberg, and Leaf (2012), and Ward and Gersten (2013).
the school environment by interventions such as SWPBS may essentially serve as a dimmer switch for genetic predispositions (Moore, 2017; Sokolowski & Ansari, 2018). As this gene-environment interplay is an ongoing process, however, students may revert to the pre-intervention level once they leave an SWPBS school.

**Academic outcomes**

Students’ academic outcomes are an important indicator of school effectiveness (Freeman et al., 2016). Apart from the detrimental effects of disengagement for students with behavioral problems, several studies have demonstrated that disruptive peers affect the learning of their classmates (Aizer, 2008; Carrell & Hoekstra, 2010; Figlio, 2007; Fletcher, 2010; Kristoffersen, Krægpøth, Nielsen, & Simonsen, 2015). Disruptive peers may impede a learning-oriented peer culture (Jencks & Mayer, 1990) and force teachers to discipline the students at the expense of classroom instruction (Coleman et al., 1966; Lavy & Schlosser, 2011). Since problem behavior in schools interferes with teaching, it has been hypothesized that reducing disruption may increase students’ exposure to classroom instruction and lead to increased academic engagement and learning (Muscott, Mann, & LeBrun, 2008; Scott & Barrett, 2004). By reducing social and behavioral problems, teachers can spend less time on discipline problems and more time delivering effective instruction that will benefit all pupils (Gage et al., 2015).

A recent meta-analysis suggests that overall, SWPBS has a small positive impact on academic achievement (Lee & Gage, 2019). Looking at individual studies, however, the results are mixed. Typically, experimental studies have not found any effect on academic achievement in the United States (Benner, Nelson, Sanders, & Ralston, 2012; Bradshaw et al., 2010; Horner et al., 2009), while quasi-experiments have reported mixed effects (Caldarella, Shatzer, Gray, Young, & Young, 2011; Freeman et al., 2016; Gage, Leite, Childs, & Kincaid,
Likewise, no effects of SWPBS were found on academic school level one year after primary school (8th grade) in Norway (Borgen et al., 2019).

There may be several explanations for the failure to find effects of SWPBS on academic achievements in Norway, such as the lack of interventions targeting academic improvements, partly different antecedents for behavioral and academic problems, and that any indirect effect through exposure to classroom instructions is small. In the U.S. studies, the outcomes' validity is limited by small sample sizes and short study periods (Madigan et al., 2016). Additionally, no study has looked at long-term effects on academic tests or school dropout. Thus, whether the model impacts long-term academic failure is still an open question.

**Lasting effects on school environment versus long term effects on student outcomes**

Whether SWPBS has lasting effects on the school environment is linked to the question of durable implementation (Flay et al., 2005; George Sugai & Horner, 2006). In contrast, long-term effects on student outcomes relate to whether the school environment at a certain age has a persistent impact on the individual (after he/she has left). Even if the impact on the school environment dilutes over time – for instance, due to lack of booster sessions and refreshment activities – the cohorts of students who experienced the program during its peak performance may nevertheless benefit in terms of favorable outcomes after having left the SWPBS school. In general, most studies of sustainability of school prevention programs have focused on persistent program impact on the school environment, rather than on the sustained behavioral change of students (e.g., Han & Weiss, 2005). Concerning SWPBS studies, sustainability studies have dealt with the significant enablers and barriers to the sustainability
of the model in schools, rather than the sustainability of student outcomes (McIntosh, Kim, Mercer, Strickland-Cohen, & Horner, 2015; Pinkelman, McIntosh, Rasplica, Berg, & Strickland-Cohen, 2015).

To our knowledge, no studies of SWPBS have followed students after they left the program school, and we do not know whether the positive intervention effects diminish, persist, or even increase as students age. In our study, we examine whether altering the direction of students at risk of (severe) behavioral problems turns up in better school behavior, as well as a reduction in the likelihood of dropout and criminal charges by late adolescences. The economic and social burden of dropout and criminal activity is substantial, which means that from a policy perspective, even small improvements may be cost-beneficial (Levin & Belfield, 2007). Establishing whether interventions have effects that are of practical importance on long-term policy-relevant outcomes is important for the overall effectiveness of interventions.

**Heterogeneous treatment effects**

The effects of SWPBS are unlikely to be the same for all students. Since SWPBS primarily aims to improve students’ behavior, students who are unlikely to exhibit problematic behavior in primary school will be less affected by the intervention. While even these students may benefit from the intervention if SWPBS improves the school environment (e.g., their peers' behavior), such spillover effects are expected to be minor, compared to any direct impact on high-risk students frequently exhibiting externalizing problem behaviors in primary school.

While differential treatment effects have long been recognized as relevant for any intervention (Farrell, Henry, & Bettencourt, 2013), it is particularly important in the SWPBS model as differential treatment effects are built into the model (Bradshaw et al., 2015). While
Tier I interventions are universal measures affecting all students, Tier II interventions are for students who continue exhibiting problem behavior, and Tier III interventions are highly intensive and individualized interventions for students with chronic and severe problem behavior (Mitchell et al., 2018). Thus, students targeted by Tier II and Tier III interventions are more likely to experience adverse outcomes such as academic failure or having a criminal career and thereby potentially benefit more from the intervention, and they receive more support and interventions.

In the Standards of Evidence for identifying effective prevention programs by the Society for Prevention Research, Flay et al. (2005) include identifying for whom and under what conditions interventions work as a criterion. However, only two studies have examined variation in treatment effects of SWPBS. Bradshaw et al. (2015) find that high-risk (6.6%) and at-risk (23.3%) students benefit most from exposure to SWPBS in the United States, despite that the schools they studied only received formal training in Tier I interventions and support. In Norway, Sørlie et al. (2018) find an indication of intervention effects for students with persistent high behavioral problems (2.5%), but no effects for students on a persistent low (84.4%), decreasing (7.9%), or increasing (5.3%) trajectory of externalizing behavior problems.

The Norwegian context and the SWPBS model

Compulsory education in Norway starts at the age of six and lasts for ten years, with primary education in grades 1 to 7 and lower secondary education in grades 8 to 10. After compulsory education, all students have by law the right to three years in upper secondary school. Schools are publicly funded, and very few are private. Compared to other European countries, Norway has an inclusive school setting and few special schools (Ogden, 2014). Studies typically find small differences in educational achievement between schools in Norway. For instance,
studies of TIMSS, PIRLS, and PISA suggest that primary schools explain about 10 percent of the variation in test scores in Norway, compared to about 25-30 percent for all included countries (Caponera & Losito, 2016; Marks, 2006; Martin, Foy, Mullis, & O’Dwyer, 2011).

The prevalence of behavioral problems such as conduct disorders and oppositional disorder in Norway is in the lower end of the spectrum along with Germany, China, and the other Nordic countries (Rescorla et al., 2012). A limitation of cross-national studies is, however, that one cannot rule out differences in the interpretation or the standards of describing behavior as problematic. Nevertheless, behavioral problems in schools is a major challenge also in Norway. Compared to other western countries, Norway has similar levels of behavioral problems such as bullying (Due et al., 2009) and similar levels of dropout from upper secondary education (OECD, 2017). Reports by Norwegian students of disruption and learning conditions in class have been among the worst of the OECD countries, with considerable differences between schools (OECD, 2009). The prevalence of problem behavior in Norwegian schools has resulted in a widespread use of school programs that aims at improving the school environment via addressing behavioral problems, with about 75 percent of the primary schools in 2013 reported having implemented a school program in the last five years (Eriksen, Hegna, Bakken, & Lyng, 2014).

Since 2002, SWPBS (called N-PALS in Norway) has been implemented in 244 primary schools in Norway (10 percent of the schools) (Borgen et al., 2019). SWPBS is a structured yet flexible whole-school approach with the primary goals to prevent and reduce school problem behavior and promote an inclusive learning environment that can facilitate safety and the psychosocial functioning and learning of all students (Sørlie & Ogden, 2015). The Norwegian model is an adapted version of SWPBS, as developed by Sprague and Walker (2005). The core model components, basic training, and implementation features of N-PALS are, however,
identical to the U.S. version. Except for minor adaptations of the training materials, no changes were made to the original model when SWPBS was transported to Norway.

The Norwegian three-tiered SWPBS model has a team-structured organization of the implementation leadership. A team of five to six representatives is appointed at each school, and this team is trained and supervised by a local coach for two years (2 hours/10 training sessions per year). The school teams are trained to plan, inform, conduct, monitor, and report on their schools' implementation process and outcomes. The school teams inform and train the rest of the staff in the key model and implementation features. The prevention model involves all staff and students and takes approximately three to five years to implement fully. Weaker program effects are thus expected the first couple of years compared to three to five years after the intervention is initiated. The Norwegian Center for Child Behavioral Development is responsible for drafting handbooks and training local coaches.

The focus of SWPBS is on positive, systematic, data-driven, educative, and reinforcement-based practices conducted within a framework of research-based, collective (school-wide), proactive, and predictable approaches. Direct behavioral teaching and interventions are combined with school-wide modifications of the social learning environment. As in the original version of the SWPBS model (Sprague & Walker, 2005), no evidence-based interventions to promote academic performance are included. The core model components are 1) school-wide positive behavior support strategies, in which 3-5 positively formulated school rules are taught and followed up by systematic praise and supervision from staff; 2) monitoring of student behavior across all school areas with a web-based assessment and evaluation system (School-Wide-Information System, SWIS); 3) immediate corrections of problem behavior by all staff using mild consequences; 4) classroom management skills training for teachers; 5) parent information and collaboration strategies;
6) small-group instruction or training in academic or social skills for students at risk, and 7) individually tailored interventions and support for high-risk students (further described below). The intervention schools use web-based assessment and feedback to ensure that relevant data support decisions about actions and that the interventions are implemented and sustained with high fidelity.

The SWPBS model’s multi-theoretical basis is based on social learning theory (Bandura, 1986), social interaction learning theory and coercion theory (Patterson, 1982), and social-ecological theory (Bronfenbrenner, 1979). Overall, they make up the model’s theory of change, which conveys that student school behavior is influenced by changes in adult behavior. The school staff is systematically trained in promoting positive behavior support by the generous use of common praise, communicating positive learning expectations, and teaching school-specific behavioral skills. Together, these systematic interventions are assumed to increase the impact on the students who observe, listen, and imitate teachers and other students in formal and informal teaching and learning activities. The school intervention activities are further hypothesized to prevent the escalation of conflicts (Patterson, 1982) and the fragmentation of contextual influences (Bronfenbrenner, 1979).

The SWPBS is organized according to the principle of matching interventions to students’ risk level (Sørlie & Ogden, 2015). More specifically, the intervention model relies on a three-tiered system of evidence-based preventions and supports. Tier I interventions (universal, primary prevention) apply to everyone and all settings in the school with the goal to “prevent problems by defining and teaching consistent behavioral expectations across the school setting and recognizing students for expected and appropriate behaviors” (Lohrmann, Forman, Martin, & Palmieri, 2008). Tier II interventions (selected, secondary prevention) are designed for students at moderate risk for severe behavior problems and who might not
respond sufficiently to the universal interventions. Students at moderate risk often receive time-limited small group instruction or training in academic or social skills or the Check-In/Check-Out (CICO) program (Todd, Campbell, Meyer, & Horner, 2008). Tier III (indicated, tertiary prevention) targets the few students with or at high risk of conduct disorder. Interventions for high-risk students include a functional behavior support plan with individual special education and/or intensive individual social skills training programs, such as the Stop-Now-and-Plan (SNAP) program (Augimeri, Farrington, Koegl, & Day, 2007).

Importantly, the SWPBS model does not include any follow-up activities such as booster sessions and refreshment activities for students after they complete 7th grade and leave an SWPBS primary school; however, the model is designed to apply to all grades through high school (Sugai and Horner, 2009).

**Data and methods**

**Data and variables**

In this study, we use population-wide Norwegian register data covering full school cohorts of students born between 1986 and 2002 (N=964924). All Norwegian primary schools (grades 1-7) were included in the study (N=2366), whereof 216 of the schools in our sample (9%) had implemented SWPBS. Each school and student has a unique identifier that allows matching of SWPBS schools to student-level population-wide register data from residential address (further described below). We define students in grades 4-7 from the time of SWPBS implementation at their school (and onwards) as exposed to the model. The share of SWPBS

---

2 The total number of intervention schools are 244. We exclude 11 lower secondary schools that do not have grades 7 or lower, 8 schools that we did not find in the school registers, and 9 schools that were not the likely attended school for any student (see Appendix 1).
exposed students has increased from close to zero until the 1994 birth cohort to about 11 percent of those born in 2000 and later (Borgen et al., 2019).

Note that the national register data does not include information about practices in the control schools. In 2013, about 75 percent of the primary and lower secondary schools in Norway reported having implemented a school program the last five years, which means that a large proportion of Norwegian schools are implementing other interventions and programs such as the Olweus Bullying Prevention Program, Second Step, Connect, and Aggression Replacement Training (Eriksen et al., 2014; Sørlie & Ogden, 2015). Additionally, we cannot rule out that one or more components of SWPBS have been adopted in nearby control schools (i.e., program contamination).

We distinguish between two types of outcomes: academic and behavioral outcomes. Descriptive statistics are presented in panel A of Table 1. Available birth cohorts and control variables for different outcome variables are shown in Appendix Table A5.1 and Appendix Table A5.2.

**Academic outcomes.** Standardized national tests was measured as the average score on standardized tests in literacy, English (foreign language), and numeracy. All tests were performed early in 8th grade, the first semester after leaving an SWPBS school. Examination grades was measured as the average of all examination grades, taken at the end of compulsory education (10th grade). Both standardized national tests and examination grades were for the analyses standardized to mean=0 and standard deviation=1. Dropout was measured as non-completion of the first year of upper secondary education (11th grade).

**Behavioral outcomes.** Marks in order and conduct in 10th grade are included at the school-leaving certificate from the compulsory school (10th grade) and distinguish between good, fair, and poor. The order and conduct are graded by the students' teachers to reflect
behavior such as being late to class, not doing homework, being violent, and cheating on tests. We measure whether students have fair or poor school behavior (=1) as opposed to good (=0).\(^3\) Criminal charges measures whether students are charged for a criminal act by the age of 17 (11\(^{th}\) grade), and includes all types of charges. Appendix 9 shows effect estimates with different types of charges (Figure A8.1) and by the severity of the charge (misdemeanor and felony offenses) (Figure A8.2).

**Treatment (intervention).** The treatment indicator of each student is defined by his/her school grade cohort at the year of program intervention, from four years before the intervention to four years after the intervention. The treatment variable contains eight unique values (-4, -3, -2, -1, 1, 2, 3, 4), and we include seven dummy variables in the regression analyses (with -1 as the reference category). Thus, students finishing primary school (7\(^{th}\) grade) the school-year before SWPBS was implemented is labeled -1. The next cohort, exposed to SWPBS for one year (7\(^{th}\) grade), is labeled 1. Treatment 4 is the first cohort exposed through all four grades (grades 4-7). Since exposure equals time since implementation, any differential effects across cohorts will capture the joint impact of length of exposure for the individual and length of implementation at the school. Outcomes for students were analyzed in grades 8, 10, or 11.

**Control variables.** We include several control variables not affected by the program that account for confounding and improve the efficiency of the estimated program effects by reducing residual variance in the outcome variable. The available control variables vary by outcome variable, shown in tabular form in Appendix Table A5.1. Descriptive statistics are

---

\(^3\) The SWPBS intervention may influence teachers’ perception of poor school behavior and raise their expectations to students behavior, which may negatively bias the program effects (e.g., Gage et al., 2018). Different from poor school behavior, the other outcome variables are objective measures and not influenced by this limitation.
presented in panel B of Table 1. The following variables were included as covariates in the
effect estimation in all analyses: gender, mother’s and father’s level of education (9 dummies
for each parent), mother’s and father’s earnings and earnings squared between age 11 and
15, father receiving social welfare between age 7 and 9 (yes/no), mother receiving social
welfare between age 7 and 9 (yes/no), father being charged for a criminal offense between
age 7 and 9 (yes/no), mother being charged for a criminal offense between age 7 and 9
(yes/no), father’s marital status (unmarried, married, widow(er), divorced, separated,
missing), mother’s marital status, immigrant background (6 dummies), age of immigration
(dummies), father’s and mother’s year of birth and year of birth squared, number of siblings
and number of siblings squared, number of half-siblings and number of half-siblings squared,
birth order and birth order squared, year of birth (dummies).

Additionally, all models include dummies for whether an older sibling has been
charged with a criminal offense by age 17, whether an older sibling has completed upper
secondary education by the age of 21, and whether an older sibling has been in contact with
the child welfare services at age 18. These three variables have three values each; no older
sibling has reached the age limit, the older sibling has experienced the event, and the older
sibling has not experienced the event.

For academic outcomes in 8th grade (standardized test scores), we also observe prior
standardized test scores (5th grade) and visits to the general practitioner (GP). Since we control
for test scores early in 5th-grade, we restrict the analysis for short-term academic outcomes
to cohorts exposed to SWPBS in 5th, 6th, and 7th grade (1-3 years of exposure). Standardized
test scores in 5th and 8th grade are highly correlated ($\rho = 0.819$), and controlling for prior test
scores allow us to improve the precision of the effect estimates, account for potential bias,
and more precisely identify at-risk students (see below). When looking at academic outcomes
in 8th grade, we control for low test scores in 5th grade, standardized test scores in 5th grade, and standardized test scores in 5th grade squared.

We control for health problems because physical and mental health problems are associated with academic failure and behavioral problems (Needham, Crosnoe, & Muller, 2014). Visits to GP are obtained from the Control and Payment of Health Reimbursement (KUHR) database and includes all patient contact with their GP. From these data, we generate control variables that include the number of GP visits at age 9, the number of GP visits at age 9 squared, and five dummies that indicate the reason for GP visits.

**The difference-in-difference model**

Schools implementing interventions like SWPBS are potentially different from other schools. These schools may experience higher levels of problem behavior or have more proactive school management. We use a difference-in-difference (DiD) design to account for schools selecting into SWPBS on characteristics of schools or students that are constant across cohorts. This design compares changes in outcomes between subsequent cohorts of students within schools following the implementation of SWPBS. Changes within SWPBS schools are compared with changes in other (control) schools to account for trends in outcomes common to all schools.

We study the effects of program exposure in grades 4-7 on students’ later outcomes. The unit of observation is the student, and the basic model is:

\[
Y_{ics} = \beta_0 + \beta T_{cs} + \delta X_{ics} + \gamma_c + \mu_s + \epsilon_{ics}
\]

4 Appendix Figure A10.1 compares effect estimates from the DiD model with less comprehensive strategies.
$Y_{ics}$ is the outcome (e.g., end-of-compulsory school grades or criminal charges by age 17) of a student $i$ belonging to cohort $c$ that attended school $s$ in grades 4-7. Cohort refers to the year students exit primary school (and exposure to SWPBS ends). Since $T$ indicates whether a given cohort in a given school was enrolled after the implementation of SWPBS ($T = 1$), $\beta$ is the effect on student outcome of having attended an SWPBS school. $\beta_0$ is a constant term, $\gamma_c$ is the cohort fixed effect, $\mu_s$ is school fixed effect, and $X_{ics}$ are observed student characteristics (female, fathers’ and mothers’ education and earnings, immigrant background). Throughout, we use ordinary least squares for both continuous and binary outcome variables, and cluster residuals at the school level using the Huber-White (robust) sandwich estimator of variance in Stata 15.0 (Rogers, 1993).

A major advantage of DiD is that it accounts for all time-invariant differences between schools, such as stable school traits, teacher characteristics, and student characteristics, irrespective of proxies for these differences. Any such differences will be captured by the school fixed effects. The design may still give biased estimates, however, if there are unobserved differential changes in intervention and control schools that are concurrent with the introduction of the program, but not part of it.

The key identifying assumption that must be satisfied for the effect estimates to be unbiased is that outcomes across cohorts would change in parallel for students of SWPBS schools and control schools, in the absence of the SWPBS intervention (net of the effects of changes in observed student characteristics). This “parallel trends”-assumption is untestable,

---

5 Cluster-robust standard errors are used for nested data, e.g. students in schools, when there is a concern that unobserved characteristics or shocks make residuals, $\epsilon$, correlated within nests. Clustered standard errors adjust for these correlations, and tend to increase the estimated standard errors.
but we can evaluate its credibility indirectly by comparing trends in student outcomes of cohorts of program and non-program schools, where all students had left the SWPBS schools before the program was implemented. Before implementation, outcomes of students in (later) SWPBS schools may differ from those in other schools. However, if these differences are stable across cohorts, the evidence supports our identification strategy. We estimate program implementation leads and lags coefficients using

\[ Y_{ics} = \beta_0 + \sum_{p=-4}^{4} \beta_p T_{csp} + \delta X_{ics} + \gamma_c + \mu_s + \epsilon_{ics} \]

, where the \( \beta_p \) parameters identify any pre-program differentials (\( p < 0 \)) and post-implementation effects (\( p > 0 \)). Here, \( T_{csp} = 1 \) for students of a cohort with a time distance of \( p \) years since the implementation of the program. The cohort leaving 7th grade one year before implementation is chosen as a reference (i.e., \( \beta_1 = 0 \)). For example; \( \beta_3 \) measures the effect on \( Y \) for students exiting primary school three years after the initiation of SWPBS relative to the outcomes of students exiting primary school just before initiation, i.e., the effect of being exposed SWPBS for three years (grades 5-7).

Any indication of systematic differences across school cohorts by future SWPBS status would suggest a violation of the identifying assumption of similar trends absent the program. In our case, the pattern of estimated leads coefficients of equation (2) gives no reason to worry (reported in Appendix Figure A2.1). There are no apparent differences between cohorts within the same school before the implementation (net of differences explained by time-varying covariates and general time trends), and the pre-implementation estimates are close to zero. Moreover, a formal test backs this conclusion. For each of the outcomes studied, the joint F-test of parallel trends (i.e., \( \beta_{-4} = \beta_{-3} = \beta_{-2} = 0 \)) does not reject the null hypothesis of no
pre-intervention differences, with p-values ranging from 0.10 to 0.95 (see note in Appendix Figure A2.1).

All in all, we find that outcome differentials of the pre-implementation cohorts are stable. This pattern also supports a simplified specification of the main effect analyses where we compare the intervention effects ($p > 0$) with the pre-intervention cohorts (reference category):

\[ Y_{ics} = \beta_0 + \sum_{p=1}^{4} \beta_p T_{csp} + \delta X_{ics} + \gamma_c + \mu_z + \epsilon_{ics} \]

We further discuss the credibility of our design and the pre- and post-implementation differences in the discussion part, after presenting the main results.

**Linking students to schools**

Norway does not maintain a central registry of which primary school students have attended. Therefore, we impute the school attended from residential address using links constructed from test score data (see Appendix 2 for details). This imputation will cause some misclassification of the school attended, and thus of program exposure: Some students we classify as exposed to the program will not actually have been, while some we classify as not exposed will have been. Generally, this type of misclassification could bias the effect estimates either upwards or downwards; however, the specific misclassification in this study probably causes a slight attenuation bias in the effect estimates. Students in cities have a larger set of school options, including private schools, compared to those in rural areas. Thus, these students may have different characteristics and outcomes than other students, and have a different share of misclassification. However, because our research design accounts for
unobserved persistent differences in the outcomes of students in different schools as well as family background for each student, it is unlikely that misclassification is correlated with unobserved factors affecting outcomes.

Assuming it is conditionally random, the misclassification in this study will cause attenuation bias in the effect estimates (e.g., Lewbel, 2007): If a share $a$ of students classified as exposed were correctly classified ("true positive") and the share $b$ among those classified as not exposed were in fact exposed ("false negative"), the difference in actual program exposure between students classified as exposed or not would be $a-b<1$. This will bias our effect estimates towards zero. However, knowing $a$ and $b$ we can inflate our effect estimates with a factor $1/(a-b)$ to get an unbiased estimate of the true effect.

Although a central student register does not exist, we do observe the schools of students taking standardized tests in grade five from 2007 onwards (born 1997 and later). By extrapolating the misclassification from samples where we have school assignments from test score data to our cohorts, we can estimate the size of $a$ and $b$. 6 Results from such analyses indicate that $a=0.9$ and $b=0.0$ (see Appendix Figure A1.1). In other words, despite some misclassification, those classified as exposed to the program mostly are exposed, while very few of those classified as not exposed are actually exposed, reflecting that only a minority of schools are program schools. Consequently, students incorrectly assigned to a control school will mostly be attending some other control school. Thus, we can assume that the effect estimates are attenuated by a factor of about 0.9, and can inflate coefficients and standard

6 We have no evidence of large changes in school assignment in the relevant period, and private school attendance is stable and very low. Correlating predicted student numbers with student counts from register data we find consistently high correlations ($r=0.90$), and although the correlations decline slightly with increasing time difference between the cohorts studied and those used for constructing linkages, the correlation is still above .85 at the extremes of our sample (see Appendix Figure A1.2).
errors by about 10% (1/0.9=1.1) to adjust for the bias. This adjustment is relatively minor, and it is based on the subset of the sample for whom we observe standardized tests, and accordingly has some estimation uncertainty. Therefore, we will not explicitly implement this adjustment as part of our estimator, but refer to it in our discussion of the results instead.

**Identifying at-risk students**

To investigate the heterogeneous effects of SWPBS, we estimate the effects for students at risk of academic failure or behavioral problems, in addition to the average effect for all students. To identify program effects, the student population at risk must be defined similarly in both program schools and control schools. In this paper, we define the student population at risk by (pre-determined) characteristics of each student and his/her family. Because we do not have information on which students receive the more targeted tier II and tier III interventions, we do not know the extent to which the students we identify as at-risk overlap with the students the schools consider at high risk. The results for at-risk students cannot be interpreted directly as the effects of a more intensified intervention.

We estimate models to predict which students are at risk of academic failure and behavioral problems, building on the extensive literature on risk factors (Crews et al., 2007; Deater–Deckard, Dodge, Bates, & Pettit, 1998; Dubow, Huesmann, Boxer, & Smith, 2016; Lösel & Farrington, 2012; Murray & Farrington, 2010; Wasserman et al., 2003). Put simply, we use part of the students in the control schools to identify characteristics of at-risk students, calculate distributions of predicted outcomes among students in SWPBS schools, define at-risk students, and finally estimate a DiD model including only at-risk students using the remaining students in the control schools and the students in the SWPBS schools.

The available cohorts and variables differ by outcome variable, but the design is the
same. To explain the details, we use dropout as an example. First, we randomly cut the student sample of the control schools into three groups; a prediction sample, a hold-out sample, and a control group sample, consisting of 25, 25, and 50% of the students, respectively. Second, we estimate a logistic regression model with dropout as the outcome variable using the prediction sample. The logistic regression model includes standard background variables such as parental education and earnings, as well as information on whether the students have older siblings that have experienced dropout by 21 years of age, are charged for criminal activity by age 17, or have been in contact with child welfare services at age 18 (see control variables above for description). Third, we use the coefficients from the logistic regression model to predict the probability of dropout in the hold-out sample (used to test how well we identify at-risk students, see below) and the control group sample (used to estimate program effects for at-risk students). Finally, each student is classified as at-risk or not, based on what is defined as the cutoff in the predicted dropout distribution.

When deciding on the cutoff, there is a tradeoff between having a sample of at-risk students that include many of the actual dropouts and having a sample where many of those classified as dropouts, actually drop out. To discuss this tradeoff, we label students classified as at-risk to take value 1, and the others value 0. In each group, there are students that are correctly classified as 1 (c₁) and 0 (c₀), while others are misclassified (w₁ and w₀). Two relevant metrics are

\[
Sensitivity = \frac{c_1}{c_1 + w_0}
\]

\[
Precision = \frac{c_1}{c_1 + w_1}
\]

Sensitivity measures how many of the total dropouts that are correctly classified as dropouts
(completeness), while precision measures how many of the classified dropouts (correctly or not) actually drop out (relevance). A stricter cutoff will reduce sensitivity and increase precision. That is, by raising the required dropout probability to be considered at-risk, we will reduce the share of the total number of dropouts (lower sensitivity). Thus, a stricter cutoff will lower the number of students included in the analysis and reduction in statistical power. At the same time, a higher share of those classified as at-risk will actually have dropped out with a stricter cutoff (higher precision). Consequently, there is a tradeoff between precision and power. To avoid overfitting, we use the hold-out sample to test the predictions using the sensitivity and precision measure.

In the main analyses, we identify at-risk students as the students with 20% highest risk (i.e., cutoff=20%). The effect of SWPBS on individual student outcomes is then estimated using a DiD model that includes only students at-risk of dropout, in both intervention and control schools. The cutoff for classifying students as at-risk was partly based on prior Nordic studies indicating that about 20% of youth have behavioral problems such as truancy and 10% have behavioral problems such as delinquency and violence, while about 3.5% match a diagnosis of conduct disorders (Skogen & Torvik, 2013). However, the chosen cutoff for classifying students into low-risk and at-risk is somewhat arbitrary. To check the sensitivity of the results, we have estimated effects using cutoffs ranging from 30% to 10% in steps of 5 percentage points. These sensitivity results show that the results are robust to different cutoffs (Appendix Figure 8.1).

The approach we use to identify at-risk students allows for background variables such as parental criminal records to contribute more to high risk for some high-risk variables. The choice of at-risk variables is described in Appendix Table A5.1. However, the correlation between the predicted probability of dropout, criminal charge, and poor school behavior is high, as demonstrated in Appendix Figure A4.4. Thus, students at risk of criminal charges are
also at risk of dropout.

Results

Data description and fidelity of implementation

Table 1 shows descriptive statistics for outcomes (panel A) and control variables (panel B). On average, 15% have low test scores in standardized national tests in 8th grade and 17% dropout from upper secondary within one year. About 12% have poor school behavior on their leaving certificate from compulsory education (10th grade), and 7% have been charged for a criminal offense by age 17.

Compared to students in other Norwegian schools, the students in schools that later become SWPBS schools tend to have more dropout and poor school behavior, with significant differences in dropout remaining after controlling for student composition (Appendix Table A3.1). These differences highlight the need for a research design able to correct for stable differences in outcomes not related to observable proxies.

[Table 1]

For 204 of the 244 schools that implement SWPBS, we can observe whether the intervention is implemented in accordance with the model (fidelity) using the Effective Behavior Support Self-Assessment Survey (EBS, 46 items). All teachers and school staff in SWPBS schools perform this survey annually, and it measures the perceived fidelity at the school level (18 items), at the classroom level (11 items), in individual cases (8 items), and in common areas like hallways and the playground (9 items). There is a risk of reporting bias in

__________________________

7 We cannot compute interrater reliability because we only observe school averages of the responses and not the individual responses of teachers and school staff.
fidelity data based on self-reported measures; however, assessment of fidelity by an external observer may be erroneous because some features are overlooked or misinterpreted (Bradshaw, Debnam, Koth, & Leaf, 2009). In our study, the school staff self-report relatively low fidelity of implementation (see below), indicating that social desirability bias is likely not an issue.

The fidelity measures range from 0% to 100%, and a minimum 80% threshold score on the fidelity scale is usually considered necessary for SWPBS to be adequately implemented. In our data, only 18% of the SWPBS schools reached an average fidelity score above this 80% threshold within three years (Appendix Figure A9.2). While this figure is low, suggesting that few schools are adequately implementing the model, the average scores on the EBS survey in Norway are in line with the EBS score in a previous randomized controlled effectiveness trial from the United States (Bradshaw et al., 2010). The EBS score may accordingly not be as sensitive to changes in school practice as other fidelity measures. Appendix 10 shows the distribution (Figure A9.1) and the average (Table A9.1) of the fidelity measure in Norwegian schools. Further details on the fidelity measure used in this paper can be found in Sørlie and Ogden (2015) and Borgen et al. (2019).

Compared to other SWPBS schools and control schools, students at schools that later implement the model with fidelity tend to have worse school behavior, higher dropout, and lower examination grades before implementing the model (Appendix Table A3.2). Bearing in mind the limitations regarding self-reported data, this suggests that the schools that need the model the most are also the ones implementing it the best.

**Average effects of SWPBS**

As for most environmental factors, the effect of SWPBS on student outcomes may depend on
exposure time. We begin, however, by focusing on the average effect of the program based on the unweighted linear combination of the four post-intervention coefficients labeled “Average 1-4 years” in Table 2. For all six outcomes, the standard error exceeds the estimate itself. Excluding the cohort who were exposed to the first year of the program, where we expect program effects to be minor, hardly affects the estimated effects of the program (“Average 2-4 years” in Table 2). Moreover, the absence of significant average effects is confirmed by the outcome-specific F-tests where the null is that all exposure-specific $\beta_p$'s $(p = 1,2,3,4)$ are zero. The p-values of these F-tests range from 0.36 to 0.95 (presented in the bottom row of Panel A in Table 2), and, based on these F-tests, we cannot reject the null hypothesis of no effect for any of the outcomes.

While the tests discussed above provide reasonable statistical conclusions, they provide little information on potential effect sizes that may be consistent with our data. In other words, could there be substantial program-related differences between SWPBS and control schools that we miss to detect because of limited statistical precision? To further investigate this, we present a range of effect estimates, along with intervals showing the effect estimates plus/minus 1.96 estimated standard errors. These intervals should not be interpreted as confidence intervals because we study multiple outcomes for several groups of students (i.e., test multiple hypotheses). While we could have created 95% confidence intervals such that there was only a five percent probability of any false positive, these wider intervals would have been less informative about the precision of any single estimate.⁸

---

⁸ Without any correction, testing several hypotheses increases the chance of rejecting a correct null hypothesis. However, for a single pre-selected coefficient of interest, the calculated intervals can in fact be interpreted as a 95% confidence interval. Of course, if we had found estimates that were significant by single parameter tests, we would have considered whether they were robust to corrections for multiple testing.
Consider the average intervention effects for students exposed to the model in two, three, or four years (Panel A in Table 2). For the standardized tests in 8th grade, the 1.96 interval ranges from -0.069 to 0.021. Adjusting the intervals for misclassification (cf. our previous discussion of linking students to schools) widens these ranges slightly, to [-.075, .024]. This interval suggests that an improvement in average test scores greater than 2.4% of a standard deviation is unlikely, as is a reduction greater than 7.5% of a standard deviation. A reduction in the share of low-performers on the standardized test of more than 1.2 percentage points or an increase of more than 1.8 percentage points is similarly unlikely.\(^9\)

For the other outcomes, we similarly do not find any indications of effects. The effect estimates are insignificant and close to zero throughout. Focusing on the most desirable outcomes within the 1.96 intervals (as above, not adjusted for misclassification), we conclude that it is unlikely that the intervention improves examination grades by more than .023 SD or that it reduces the shares with poor school behavior, dropout, or charges by more than 2.1, 0.42, and 0.8 percentage points, respectively. As shown above, adjusting the intervals for misclassification only slightly increases the possible effect consistent with the data.

Finally, we study the effects of SWPBS on individual student outcomes by the length of exposure to the model (Table 2 and Figure 1). In some cases, the effect estimates are positive or negative, but they are consistently small and never statistically significant. Further, the estimates show no clear trend as neither of them increases or decreases with time since implementation.

\(^9\) For the share with low performance we get a 1.96 interval of [-.011, .017], which gives an interval of [-.012, .018] adjusting for misclassification.
SWBPS effects for at-risk students

To identify characteristics of students at risk of adverse outcomes, such as academic failure and behavioral problems, we have used a subset of the students at control schools to estimate a prediction model (see methods section above for details). Our baseline definition of at-risk is based on the 20% cutoff, where we include all students above the 80th percentile in the predicted distribution.\(^\text{10}\)

As for all students, we find no evidence of effect for at-risk students. To begin, the joint F-tests for any effects (across all exposures) give p-values ranging from 0.20 to 0.97 across outcomes, as shown in Panel B of Table 2. That said, at-risk students exposed to SWPBS for three years seem to have worse school behavior, and they also have a higher probability of dropout (Table 2, Figure 1). However, these estimated coefficients are not statistically significant at conventional levels if compared to critical values adjusted for the number of hypotheses tested (i.e., Bonferroni correction). Moreover, there are no effects on poor school behavior and dropout for students exposed either two or four years.

The sample size drops when we focus on at-risk students, and the precision is reduced relative to the results for all students. Looking at the average effects, the relative increase in the 1.96 intervals varies from a factor of 1.3 for test scores to 3.5 for the share with a low test score. Specifically, the largest desirable outcomes for the at-risk students that lay within the 1.96 intervals (not adjusting for misclassification) are increases in average test score of 0.054 SD and exam grades of 0.077 SD, and reductions in the shares with low test scores, poor school behavior, school dropout, or charged by respectively 2.8, 1.5, 3.0, and 4.2 percentage points.

\(^{10}\) For the outcome variables examination grades and standardized national tests we use a prediction model for a related outcomes, see Table A5.1 in the Online Appendix 5.
Based on these intervals, we are not able to make any strong claims. For dropout and crime, in particular, the effect estimates for at-risk students are less informative. For test scores and exam grades, moderately-sized effects are unlikely, even for this presumably most relevant group of treated students.

There is no fixed cutoff that defines students at risk. Instead, we face a tradeoff between precision and sensitivity. For example, consider the probability of low achievement on the standardized test in 8th grade. Among the students with the 20% highest risk, 45.5% get low test scores (precision), and the sample includes 74.7% of all students with low test scores (sensitivity). If we reduce the cutoff and focus on those with an even higher risk score, we would raise precision at the cost of smaller at-risk sample size and a smaller fraction included of those who are actually at risk (lower sensitivity). However, there is no indication of significant program effects for any of the definitions of students at risk we have investigated (see Figure A7.1). There is no indication that the effect estimates becomes more substantial when we focus on students with a higher probability of adverse outcomes, but the 1.96 intervals expand. These results illustrate that identifying effects for students with particularly high predicted risk, or heterogeneous effects in general, is difficult due to loss of power.

**Effects by fidelity**

Studies indicate that intervention effects are related to the degree of program fidelity (Weare & Nind, 2011). Thus, we have checked whether intervention effects are stronger for schools that implement SWPBS with fidelity (at least 80% overall fidelity) than for SWPBS schools that

---

11 In Table A4.1 and Figure A4.1 we report how this tradeoff differs across outcome measures. Appendix 4 also includes more detailed results, including the estimated density (Figure A4.1), the ROC-curve (Figure A4.3), and the predicted probability of adverse outcomes at different cutoffs (Figure A4.5).
have a poor implementation (60% or less overall fidelity). Unfortunately, these results show no clear pattern and, as discussed above, may be biased. In high-fidelity schools, we find better school behavior and less dropout when considering all students (Appendix Figure A9.2). In low-fidelity schools, we find worse school behavior for all students and worse outcomes on standardized national tests in 8th grade for at-risk students (Appendix Figure A9.3).

While these results suggest that good (poor) implementation of SWPBS may improve (hurt) student outcomes, the sample split implies that our identification strategy seems to fail. In our DiD design, any program effects before implementation are indicative of a violation of the parallel trends assumption. While this assumption seems to hold when including all intervention schools (Appendix Figure A2.1), it does not hold when including only high-fidelity schools or only low-fidelity schools. Therefore, we are reluctant to give the estimated intervention effects a causal interpretation.

Discussion

In this paper, we have used Norwegian population-wide register data to study the effects of SWPBS on short- and long-term individual outcomes. Like most previous research (Gage et al., 2015), we find no evidence that exposure to SWPBS in primary school improves short-term academic performance. The relatively small standard errors indicate that average effects on test scores greater than about 0.02 SD are unlikely. Moreover, we have examined effects on long-term academic performance and dropout, as well as differential effects for students at high risk of academic failure, again finding no effects of SWPBS. As for test scores, the maximum effects consistent with the estimated standard errors are small. Thus, the null-results are not because the precision is too low to detect meaningful effects.

Unlike academic performance, previous studies have found evidence for effects of
SWPBS on short-term behavioral outcomes (Mitchell et al., 2018), including in Norway (Sørlie et al., 2018; Sørlie & Ogden, 2015). However, no study has looked at behavioral outcomes for students after they leave an SWPBS school. The main purpose of this paper is to examine whether the SWPBS intervention has long-term effects on policy-relevant outcomes. We find no indication that SWPBS reduces criminal charges in Norway, nor that it affects school behavior as measured by marks in order and conduct on students’ school leaving certificate.

The present study is an example of a study that appears to be at odds with what is typical evidence in smaller-scales studies (Eisner & Malti, 2015). One explanation is a scale-up effectiveness penalty of our study that may also apply elsewhere. When interventions are evaluated under near-optimal conditions of delivery, which is typically the case in RCTs, the intervention effects are likely to be stronger than in the more uncontrolled real-world conditions (Flay et al., 2005; Hulmane & Cordray, 2009; Stuart et al., 2015). Importantly, implementation quality is likely a vital factor in the scale-up penalty (Welsh et al., 2010). Fidelity is high in the RCTs from the United States (Pas & Bradshaw, 2012), and it was also high in a study including 28 of the first SWPBS schools in Norway (Sørlie & Ogden, 2015). In this paper, we study the effects of SWPBS in an educational setting where 10 percent of the schools have implemented SWPBS, and we, therefore, expect lower fidelity. Indeed, using the staff-reported EBS measure, we find that only 18% of the schools had implemented with fidelity. Although we have reasons to question the sensitivity of the EBS measure (Bradshaw et al., 2010), poor program implementation is in line with previous research, where it is reported that only 2 out of 10 effectiveness studies in the United States report high fidelity among SWPBS schools (Chitiyo et al., 2012).

The ability to be effectively implemented in real-world settings is an essential factor of prevention programs, and there is accordingly an increasing focus on large-scale
implementation (Pas & Bradshaw, 2012). Measuring fidelity of implementation in SWPBS schools can, however, be a challenge due to the broad group of informants (the entire school staff) and the complex longitudinal and multi-level design. Moreover, different from more restricted intervention programs targeting students at risk with detailed instructions and procedures, the SWPBS model is a systemic yet flexible framework of guidelines that schools can use over time to implement preventive school-wide discipline interventions (G. Sugai et al., 2010). Studying the SWPBS model as implemented in Norway, our results indicate that any favorable short-term effects of SWPBS on students seem to fade out as they become older, rather than preventing high-risk youth from entering a downward spiral of increasing behavioral problems.

Of course, absent effects on later student outcomes may also be explained by limited short-run effects of SWPBS, for the average student in particular. Not all studies identify immediate effects of SWPBS, and the outcomes where the effects are found are typically suspension rates or office discipline referrals (Chitiyo et al., 2012; Mitchell et al., 2018). Further, studies almost exclusively find effects on teacher reports only. Borgen et al. (2019) is the only study that has found effects of the SWPBS on student-assessed outcomes, suggesting that SWPBS improved classroom order but not outcomes such as bullying and school wellbeing. Although improving classroom order and reducing disruptions improves the learning environment and may be of value in itself, it may very well be that such improvements have limited effects on long-term behavioral outcomes such as criminal charges and school dropout.

The lack of effects of SWPBS on long-term behavioral and academic outcomes can also be explained by contextual factors. Even if Norway is similar to other western countries when it comes to bullying (Due et al., 2009), upper secondary education dropout (OECD, 2017), and
classroom disciplinary climate (OECD, 2009), the prevalence of behavioral problems is lower than in the United States (Rescorla et al., 2012). Additionally, the variation in student outcomes across Norwegian schools is generally low (Hermansen, Borgen, & Mastekaasa, 2019). Several studies do not find effects of interventions and inputs in Norwegian schools, such as class size effects on short term grades (Leuven, Oosterbeek, & Rønning, 2008) and long-term educational attainment and earnings (Leuven & Løkken, 2018).

Unlike in the US, where studies have found effects of SWPBS both for high-risk (6.6% of the student population) and for at-risk (23.3%) students (Bradshaw et al., 2015), a previous study from Norway only found short-term effects on behavioral outcomes for a much smaller high-risk group (2.5%), characterized by persistently high levels of behavioral problems (Sørlie et al., 2018). This study defines at-risk as students with the 20 percent highest predicted risk of adverse outcomes, with the results not sensitive to the exact cutoff for risk. Unfortunately, we do not have statistical power to meaningfully identify (relevant) effects on a group smaller than 10% of the students.

If the behavior of only 1 out of 40 students in Norway are affected by the intervention (Sørlie et al., 2018), with potential effects fading over time, average intervention effects would be small. Even so, because of the substantial economic and social burden of dropout and criminal activity, even small improvements may be cost-beneficial from a public health perspective (Levin & Belfield, 2007). For short-term academic outcomes, we have the precision to suggest that effects that are of the magnitude of about 2% of a standard deviation or larger are unlikely. For other outcomes, however, there may be effects of policy relevance that we do not have the precision to rule out comfortably. For instance, our 1.96 intervals include reductions in criminal charges greater than 0.8 percentage point when looking at all students and greater than 4.2 percentage point when studying at-risk students.
Finally, the results are subject to several methodological limitations. Imputing school from residential address introduces measurement error in program exposure that biases the results towards zero by a factor of about 0.9. If the effects of SWPBS are limited to a small group of students with behavioral problems, register data alone will have limited power to identify such effects. Nor is our design sensitive to possible local effects, delimited to some schools or specific categories of schools.

Additionally, the effect estimates reflect all observed changes in outcomes in program schools beyond the concurrent changes in control schools. We assume that the changes in the control schools reflect the changes we could have expected in program schools in the absence of SWPBS. As noted earlier, we do not find evidence of pre-implementation “effects”; the individual year-by-outcome estimates show little indication of systematic trends before the implementation, alternating between positive and negative estimates. Among other things, these results indicate that SWPBS is not implemented in response to worsening outcomes within schools; if this was the case, we should expect to see a dip in the program schools shortly before the implementation (Ashenfelter, 1978). While not a definite proof, this absence of systematically different pre-implementation trends suggests that the SWPBS and control schools would have continued to evolve similarly absent SWPBS.

Nevertheless, effects of the SWPBS may be masked by simultaneous, unrelated changes of similar magnitude but opposite sign in the SWPBS school, or of same-sign changes in the control schools. For example, one concern could be that there are differential changes in student composition between SWPBS and control schools. Empirically, we approach this concern by using a placebo test where we re-estimate the DiD model replacing the 8th-grade test scores with the 5th-grade test scores. Any “effects” of the SWPBS on outcomes before the exposure would indicate that our DiD strategy fails. The very similar null effects on 5th and 8th
grade (see Appendix Figure A12.1) suggests that there are no concurrent unrelated changes in student composition that would invalidate the DiD design. In sum, while we cannot rule out treatment and control schools being on different trajectories or preceding or unrelated simultaneous changes masking an actual effect, we find no indications of either in the data.

All estimated effects are relative to what changes control schools achieve, not to some hypothetical setting without any intervention. Consequently, the lack of effects may be due to many control schools implementing other programs (Bradshaw et al., 2010), program contamination to nearby control schools, or the SWPBS program replacing other programs or practices. In the current study, data on other programs or school practices, in general, were not available. Thus, we do not know whether SWPBS replaces other, equally effective programs. More evidence is needed to confirm whether the lack of long-term effects in our study is due to methodological limitations or low fidelity of implementation, and whether our conclusions apply outside the Norwegian context.

Conclusion

While an extensive literature has found that SWPBS successfully address social and behavioral problems in schools, there is scarce evidence on potential persistent effects on student outcomes after they have graduated from an SWPBS school. Register data allows us to examine individual long-term effects of SWPBS that are scaled up and operating under normal conditions, rather than within a controlled environment under ideal implementation. We find no significant effects of SWPBS on a range of relevant long-term student outcomes. However,

---

12 Similarly, for short-term academic outcomes, we can compare effect estimates with and without additional controls for prior test scores and GP visits. The results suggest that removing these controls does not introduce bias in the effect estimates (Appendix Figure A11.1), but it does affect the prediction of at-risk students (Appendix Figure A4.4).
since precision is compromised when we study at-risk students, the results do not imply that effects are close to zero for all students. Still, we can conclude that (at best) only a few students experience a reduced risk of academic failure and antisocial behavior reflected in school behavior, dropout, and criminal charges after the implementation of SWPBS.

As a methodological point, our analysis underlines that there may be differences between program and control schools that are not caused by the program. Thus, valid effect estimates require a research design able to handle such differences. Our evidence suggests that a difference-in-difference strategy, using variation across cohorts within schools, is a useful design. Register data offers opportunities to study effects on outcomes measured independently of the program and can also be used to classify students by risk. However, the ability to predict student risk varies considerably by the outcome. While we can predict low test scores reasonably well, our prediction of criminal charges is less precise, and with the current data, we are not able to say how the different tiers of SWPBS treatment correlate with our risk measures. Register data would be even more useful for this type of research when combined with individual student information that went beyond academic achievement and included measures of emotional and behavioral problems.

References


positive behaviour support: Results from National Longitudinal Register Data. *International Journal of Psychology*.


# Tables and figures

**Table 1**: Descriptive statistics.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Outcomes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dropout high school (11th grade)</td>
<td>743885</td>
<td>.166</td>
<td>.372</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Examination grades (10th grade)</td>
<td>755485</td>
<td>.000818</td>
<td>1</td>
<td>-3.83</td>
<td>2.25</td>
</tr>
<tr>
<td>Low test scores (8th grade)</td>
<td>502321</td>
<td>.146</td>
<td>.353</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Standardized national tests (8th grade)</td>
<td>494632</td>
<td>.00788</td>
<td>.882</td>
<td>-3.39</td>
<td>2.3</td>
</tr>
<tr>
<td>Criminal charge (by age 17)</td>
<td>743885</td>
<td>.0654</td>
<td>.247</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Poor school behavior (10th grade)</td>
<td>332158</td>
<td>.118</td>
<td>.322</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Panel B: Control variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>964924</td>
<td>.487</td>
<td>.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>First-generation immigrants</td>
<td>964914</td>
<td>.00994</td>
<td>.0992</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Second-generation immigrants</td>
<td>964914</td>
<td>.0439</td>
<td>.205</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Father’s education (attainment level)</td>
<td>954355</td>
<td>4.29</td>
<td>1.71</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Mother’s education (attainment level)</td>
<td>964238</td>
<td>4.33</td>
<td>1.73</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Father’s earnings (1 000 NOK)</td>
<td>954925</td>
<td>387</td>
<td>214</td>
<td>0</td>
<td>3160</td>
</tr>
<tr>
<td>Mother’s earnings (1 000 NOK)</td>
<td>964380</td>
<td>182</td>
<td>135</td>
<td>0</td>
<td>2499</td>
</tr>
<tr>
<td>Father social welfare</td>
<td>964924</td>
<td>.0659</td>
<td>.248</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mother social welfare</td>
<td>964924</td>
<td>.0741</td>
<td>.262</td>
<td>0</td>
<td>1s</td>
</tr>
<tr>
<td>Father criminal charge</td>
<td>964924</td>
<td>.074</td>
<td>.262</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mother criminal charge</td>
<td>964924</td>
<td>.0173</td>
<td>.13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sibling criminal charge</td>
<td>930188</td>
<td>.066</td>
<td>.248</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sibling dropout</td>
<td>964388</td>
<td>.109</td>
<td>.312</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sibling child welfare service</td>
<td>964388</td>
<td>.0217</td>
<td>.146</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Differences in the number of observations for outcome variables reflect that the outcomes are measured at different ages and differences in availability in the registers (Appendix Table A5.1).
Figure 1: Effect estimates with intervals based on 1.96 estimated standard errors, based on Table 2.

Note: The point estimates show changes in outcomes for students exposed to SWPBS in 1, 2, 3, or 4 years compared to what would have happened in the absence of the intervention. Point estimates above (below) the dashed line suggest that SWPBS increases (decreases) the outcome. The corresponding interval bands indicate the uncertainty in the estimated intervention effects. Coefficients with bands that overlap the dashed line are not statistically significant different from zero at the 5 percent level. Outcome metrics: Standardized for national tests and examination grades. Observed share for low achievement, poor school behavior, dropout, and charged.
Table 2: Intervention effects of exposure to SWPBS in 4th-7th grade for all students (panel A) and at-risk students (panel B).

<table>
<thead>
<tr>
<th></th>
<th>8th grade</th>
<th>10th grade</th>
<th>11th grade</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low test</td>
<td>Standardized</td>
<td>Examination</td>
</tr>
<tr>
<td></td>
<td>score</td>
<td>national tests</td>
<td>grades</td>
</tr>
<tr>
<td><strong>Panel A: All students</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression coefficients</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed 1 year</td>
<td>0.0040</td>
<td>-0.0244</td>
<td>0.0177</td>
</tr>
<tr>
<td>(0.0072)</td>
<td>(0.0197)</td>
<td>(0.0191)</td>
<td>(0.0087)</td>
</tr>
<tr>
<td>Exposed 2 years</td>
<td>0.0029</td>
<td>-0.0326</td>
<td>-0.0062</td>
</tr>
<tr>
<td>(0.0072)</td>
<td>(0.0245)</td>
<td>(0.0199)</td>
<td>(0.0110)</td>
</tr>
<tr>
<td>Exposed 3 years</td>
<td>0.0030</td>
<td>-0.0141</td>
<td>-0.0171</td>
</tr>
<tr>
<td>(0.0086)</td>
<td>(0.0251)</td>
<td>(0.0237)</td>
<td>(0.0135)</td>
</tr>
<tr>
<td>Exposed 4 years</td>
<td>-0.0051</td>
<td>-0.0072</td>
<td>-0.0096</td>
</tr>
<tr>
<td></td>
<td>(0.0274)</td>
<td>(0.0128)</td>
<td>(0.0092)</td>
</tr>
<tr>
<td><strong>Average effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-4 years</td>
<td>0.0033</td>
<td>-0.0237</td>
<td>-0.0027</td>
</tr>
<tr>
<td>(0.0060)</td>
<td>(0.0197)</td>
<td>(0.0151)</td>
<td>(0.0076)</td>
</tr>
<tr>
<td>2-4 years</td>
<td>0.0030</td>
<td>-0.0233</td>
<td>-0.0094</td>
</tr>
<tr>
<td>(0.0070)</td>
<td>(0.0230)</td>
<td>(0.0166)</td>
<td>(0.0088)</td>
</tr>
<tr>
<td><strong>F-test</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joint p-value</td>
<td>0.9481</td>
<td>0.4473</td>
<td>0.6680</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>131898</td>
<td>131195</td>
<td>303177</td>
</tr>
</tbody>
</table>

**Panel B: At-risk students**

| Regression coefficients |           |             |            |            |         |         |
| Exposed 1 year          | 0.0294    | -0.0097     | 0.0015     | 0.0224     | 0.0079  | 0.0187  |
| (0.0267)               | (0.0312)  | (0.0404)    | (0.0219)   | (0.0193)   | (0.0169) |
| Exposed 2 years         | 0.0209    | -0.0101     | 0.0317     | -0.0021    | -0.0049 | -0.0011 |
| (0.0258)               | (0.0325)  | (0.0422)    | (0.0236)   | (0.0193)   | (0.0235) |
| Exposed 3 years         | 0.0201    | 0.0009      | -0.0464    | 0.0693*    | 0.0365  | -0.0154 |
| (0.0304)               | (0.0351)  | (0.0509)    | (0.0313)   | (0.0228)   | (0.0210) |
| Exposed 4 years         | 0.0438    | 0.0008      | -0.0272    | -0.0086    |         |         |
| (0.0564)               | (0.0289)  | (0.0250)    | (0.0301)   | (0.0301)   |         |         |
| Average effects         |           |             |            |            |         |         |
| 1-4 years               | 0.0235    | -0.0063     | 0.0077     | 0.0226     | 0.0031  | -0.0016 |
| (0.0222)               | (0.0268)  | (0.0317)    | (0.0164)   | (0.0148)   | (0.0138) |
| 2-4 years               | 0.0205    | -0.0046     | 0.0097     | 0.0226     | 0.0015  | -0.0084 |
| (0.0245)               | (0.0300)  | (0.0341)    | (0.0192)   | (0.0161)   | (0.0170) |
| **F-test**             |           |             |            |            |         |         |
| Joint p-value           | 0.6988    | 0.9744      | 0.5753     | 0.2011     | 0.1979  | 0.7010  |
| **N**                  | 26182     | 25739       | 56140      | 31631      | 61612   | 57466   |

*p < 0.05, **p < 0.01, ***p < 0.001
Note: The average effects are obtained using a linear combination of coefficients. For test scores and standardized national tests, we only observe the students for 1 to 3 years, and the average effects are calculated across 1-3 years and 2-3 years. In the F-test, the null hypothesis is that all the exposure specific coefficients are equal to 0 (i.e., \( \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \) in equation 3).
Supplementary Online Appendix

Heterogeneity in Short- and Long-term Impacts of School-Wide Positive Behavior Support (SWPBS) on academic outcomes, behavioral outcomes, and criminal activity

Nicolai Topstad Borgen¹,², Oddbjørn Raaum², Lars Johannessen Kirkebøen³, Mari-Anne Sørlie⁴, Terje Ogden⁴, and Ivar Frønes¹,⁴

¹ Department of Sociology and Human Geography, University of Oslo, Norway
² Ragnar Frisch Centre for Economic Research, Oslo, Norway
³ Statistics Norway, Oslo, Norway
⁴ Norwegian Center for Child Behavioral Development, Oslo, Norway

Content of supplementary online appendix:

Online Appendix 1: School and student identification
Online Appendix 2: Parallel trends assumption
Online Appendix 3: Pre-intervention differences in the outcome variables
Online Appendix 4: Prediction fit
Online Appendix 5: Available birth cohorts and control variables
Online Appendix 6: Correlation between outcome variables
Online Appendix 7: Different cutoffs
Online Appendix 8: Type of criminal charges
Online Appendix 9: Fidelity
Online Appendix 10: Effect size without the DiD design
Online Appendix 11: Effects on short-term academic outcomes without additional controls
Online Appendix 12: Placebo regression
Online Appendix 1: School and student identification

As there is no central register of compulsory school students in Norway, we generally do not know which school students attend. To assign students to schools, we used the fact that Norwegian primary school students overwhelmingly attend their local schools (less than 5 percent attend private schools) and impute school assignments based on residential addresses.

Register data identifies the residential location of students, in particular the basic statistical units (basic districts). There are about 14,000 such units, which according to Statistics Norway constitute “... small, stable geographical units which may form a flexible basis to work with and present regional statistics. (...) geographically coherent areas. (...) homogeneous, with respect to nature and basis for economic activities, conditions for communications, and structure of buildings.” The units have 0-6000 inhabitants (mean = 379).

From 2007 and onwards, students take standardized tests in grades 5 and 8 (later also in grade 9). Results are recorded at the student level and include school id and student id. Based on the standardized test data for 5th graders in 2007-2009, we found the most frequently attended school for students in each basic statistical unit. We then imputed school characteristics, including SWPBS participation, using these modal schools.

For students with standardized test data, we could compare actual and predicted school id. We found that about 85 percent of students in the cohorts used to construct residence school links attended the predicted school. For other cohorts, we have slightly worse identification of schools. Still, for each cohort within +/- four years of the cohorts used to impute schools, we found that more than 80 percent attended the predicted school. These shares were similar for SWPBS and control schools.

Furthermore, for the estimation of the average effects of SWPBS, we do not need to
correctly assign students to schools, but rather, to the correct SWPBS treatment status. This means that we need to assign students to a school with the correct year of SWPBS implementation.

Figure A1.1 Share of students correctly classified as SWPBS schools and control schools. Note: Based on cohorts completing standardized national tests in 5th grade between 2007 and 2012.

In Figure A1.1, we study the year of SWPBS implementation by year of predicted implementation for students with different predicted implementation years, as well as for
students predicted not to attend an SWPBS school. We notice that consistently about 90 percent of students predicted to attend an SWPBS school do indeed complete the grade 5 test at a school with the predicted implementation year. Notably, this share is similar for students taking the test in 2007-2009, the years used to link residential address and school, and students in later years (we do not have test score data for earlier cohorts). Furthermore, the share of students observed in a school with a different implementation year than predicted is essentially zero. Thus, both the students incorrectly predicted to attend an SWPBS school and the student predicted to attend a control school in general overwhelmingly attend control schools. To sum up, of the students predicted to attend an SWPBS school, about 90 percent are correctly classified in terms of the SWPBS treatment, and the remaining 10 percent attend a control school. Essentially all predicted control school students are correctly classified.

In the evaluation of SWPBS, we were mainly concerned about the school attended around the time of introduction of the program, which matches closely with the years used to impute schools. However, as we also study earlier and later cohorts, we want to know how our predictions perform away from the 2007-2009 period. As we only have test score data from 2007 onwards, for earlier cohorts (born before 1996), we cannot use test score data to evaluate the relevance of the predicted school attended directly.

However, data on student counts for each school exist for all years from 1992 onwards, allowing us to compare our predicted student counts with schools registered student numbers. A large majority of schools have existed since the start of the school register in 1992 (the schools present both in 1992 and 2007 have 96 percent of the students in 2007). The count data do not exactly correspond to the test score data. While both standardized tests and student counts are performed in the fall, the timing may differ slightly. More importantly, some students are exempted from the standardized tests, and whether and how these are
reported varies. Still, we expect our predicted number of students to be highly correlated with the count data. This is also what we find, with correlations of the number of students per school of about 0.90. Also, predicting the number of students by gender*grade*year*school and correlating this predicted student counts with observed counts, we found a correlation of 0.89.

Figure A1.2 shows how the correlation varies by year of observation, separately for all schools and SWPBS schools. We see that for both groups of schools, the correlation is about 0.90 and stable over time. This result indicates that even over long periods, we are mostly able to predict school assignments based on residence correctly.

![Figure A1.2: Correlations between the predicted number of students (based on residential address) and student counts from register data.](image)

**Figure A1.2** Correlations between the predicted number of students (based on residential address) and student counts from register data.
Figure A2.1: Effect estimates (after) and pre-program heterogeneity for all students and at-risk students with intervals based on 1.96 estimated standard errors.

Note: Outcome metrics: Standardized for national tests and examination grades. Observed share for low achievement, poor school behavior, dropout, and charged. Contrary to the bottom four panels, the top two panels include controls for prior academic achievement (5th grade) and visits to the GP, and are accordingly likely more protected from omitted variable bias than the other panels. Because of these additional controls,
the axis in the top two panels ranges from three years before three years after initiation of the SWPBS program (and not -4 to 4). The P-values of the joint F-tests of the pre-intervention leads are 0.8816 (low achievement), 0.9516 (standardized national tests), 0.0988 (examination grades), 0.4348 (poor school behavior), 0.4111 (dropout), and 0.2367 (charged by 17).
Online Appendix 3: Pre-intervention differences in the outcome variables

### Table A3.1 Pre-intervention differences between SWPBS schools and control schools.

<table>
<thead>
<tr>
<th></th>
<th>Low achievement</th>
<th>Standardized national tests</th>
<th>Dropout</th>
<th>Examination grades</th>
<th>Charged by 17</th>
<th>Poor school behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWPBS</td>
<td>0.000842</td>
<td>-0.000669</td>
<td>-0.0176</td>
<td>-0.00751</td>
<td>0.0152</td>
<td>0.00644</td>
</tr>
<tr>
<td></td>
<td>(0.00640)</td>
<td>(0.00386)</td>
<td>(0.0233)</td>
<td>(0.0120)</td>
<td>(0.00487)</td>
<td>(0.00252)</td>
</tr>
<tr>
<td>SWPBS 0-60%</td>
<td>-0.000286</td>
<td>-0.000289</td>
<td>-0.000799</td>
<td>-0.00187</td>
<td>0.0121</td>
<td>0.00179</td>
</tr>
<tr>
<td></td>
<td>(0.00057)</td>
<td>(0.000567)</td>
<td>(0.0325)</td>
<td>(0.0181)</td>
<td>(0.00712)</td>
<td>(0.00332)</td>
</tr>
<tr>
<td>SWPBS 61-79%</td>
<td>0.0238</td>
<td>0.00926</td>
<td>0.00942</td>
<td>0.0891**</td>
<td>0.00859</td>
<td>0.00703</td>
</tr>
<tr>
<td></td>
<td>(0.0255)</td>
<td>(0.0144)</td>
<td>(0.0961)</td>
<td>(0.0308)</td>
<td>(0.0135)</td>
<td>(0.00922)</td>
</tr>
<tr>
<td>SWPBS 80-100%</td>
<td>0.0127</td>
<td>0.00556</td>
<td>-0.0533</td>
<td>-0.0246</td>
<td>0.0332**</td>
<td>0.0171*</td>
</tr>
<tr>
<td></td>
<td>(0.0155)</td>
<td>(0.0128)</td>
<td>(0.0544)</td>
<td>(0.0318)</td>
<td>(0.0108)</td>
<td>(0.00745)</td>
</tr>
<tr>
<td>Cohort FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Student characteristics</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>465666</td>
<td>389673</td>
<td>458660</td>
<td>384503</td>
<td>695858</td>
<td>622794</td>
</tr>
</tbody>
</table>

Note: Cluster robust standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

### Table A3.2 Pre-intervention differences between SWPBS schools that implement with fidelity (80-100%), schools with medium implementation (61-79%), and schools with poor implementation (0-60%), all compared to control schools (reference category).

<table>
<thead>
<tr>
<th></th>
<th>Low achievement</th>
<th>Standardized national tests</th>
<th>Dropout</th>
<th>Examination grades</th>
<th>Charged by 17</th>
<th>Poor school behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWPBS</td>
<td>0.000842</td>
<td>-0.000669</td>
<td>-0.0176</td>
<td>-0.00751</td>
<td>0.0152</td>
<td>0.00644</td>
</tr>
<tr>
<td></td>
<td>(0.00640)</td>
<td>(0.00386)</td>
<td>(0.0233)</td>
<td>(0.0120)</td>
<td>(0.00487)</td>
<td>(0.00252)</td>
</tr>
<tr>
<td>SWPBS 0-60%</td>
<td>-0.000286</td>
<td>-0.000289</td>
<td>-0.000799</td>
<td>-0.00187</td>
<td>0.0121</td>
<td>0.00179</td>
</tr>
<tr>
<td></td>
<td>(0.00057)</td>
<td>(0.000567)</td>
<td>(0.0325)</td>
<td>(0.0181)</td>
<td>(0.00712)</td>
<td>(0.00332)</td>
</tr>
<tr>
<td>SWPBS 61-79%</td>
<td>0.0238</td>
<td>0.00926</td>
<td>0.00942</td>
<td>0.0891**</td>
<td>0.00859</td>
<td>0.00703</td>
</tr>
<tr>
<td></td>
<td>(0.0255)</td>
<td>(0.0144)</td>
<td>(0.0961)</td>
<td>(0.0308)</td>
<td>(0.0135)</td>
<td>(0.00922)</td>
</tr>
<tr>
<td>SWPBS 80-100%</td>
<td>0.0127</td>
<td>0.00556</td>
<td>-0.0533</td>
<td>-0.0246</td>
<td>0.0332**</td>
<td>0.0171*</td>
</tr>
<tr>
<td></td>
<td>(0.0155)</td>
<td>(0.0128)</td>
<td>(0.0544)</td>
<td>(0.0318)</td>
<td>(0.0108)</td>
<td>(0.00745)</td>
</tr>
<tr>
<td>Cohort fixed effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Student characteristics</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>456419</td>
<td>381469</td>
<td>449487</td>
<td>376358</td>
<td>673145</td>
<td>602409</td>
</tr>
</tbody>
</table>

Note: Cluster robust standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Online Appendix 4: Prediction fit

Table A4.1: Classification table. Sensitivity and precision in the hold-out sample.

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Low test scores (8th grade)</th>
<th>School behavior (10th grade)</th>
<th>Dropout VG1 (11th grade)</th>
<th>Charged (11th grade)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Precision</td>
<td>Sensitivity</td>
<td>Precision</td>
</tr>
<tr>
<td>70</td>
<td>87.3</td>
<td>35.6</td>
<td>62.0</td>
<td>20.3</td>
</tr>
<tr>
<td>75</td>
<td>82.2</td>
<td>40.3</td>
<td>55.3</td>
<td>21.7</td>
</tr>
<tr>
<td>80</td>
<td>74.7</td>
<td>45.5</td>
<td>48.7</td>
<td>23.8</td>
</tr>
<tr>
<td>85</td>
<td>64.7</td>
<td>52.9</td>
<td>39.9</td>
<td>26.2</td>
</tr>
<tr>
<td>90</td>
<td>50.2</td>
<td>62.1</td>
<td>29.6</td>
<td>29.4</td>
</tr>
</tbody>
</table>

Note: Cutoff is the percentile rank in the estimated probability of adverse outcomes.

Figure A4.1 Density estimates for the analysis sample.
Figure A4.2 Precision and sensitivity in the hold-out sample.

Figure A4.3 ROC curve in the hold-out sample.
Figure A4.4 Correlation between the predicted probability of dropout, poor school behavior, criminal charges, and low test scores.

Note: Moderate correlations between low test scores on the one hand and dropout, criminal charges, and poor school behavior on the other is because of the inclusion of prior test scores in the estimated risk of low test scores. Excluding prior test scores, the correlations are 0.790 (dropout), 0.731 (charged by 17), and 0.792 (poor school behavior).
Figure A4.5 Predicted probability of different adverse outcomes at different cutoffs.
### Online Appendix 5: Available birth cohorts and control variables

### Table A5.1: Birth cohorts observed with different outcome variables.

<table>
<thead>
<tr>
<th>Outcome variables</th>
<th>Birth cohorts</th>
<th>High-risk variable</th>
<th>Control variables(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short-term academic outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized test scores ((8^{th}) grade)</td>
<td>1997-2002</td>
<td>Low achievements ((8^{th}) grade)</td>
<td>Parental characteristics, individual characteristics, sibling characteristics, prior standardized tests ((5^{th}) grade), and GP visits.</td>
</tr>
<tr>
<td>Low achievements ((8^{th}) grade)</td>
<td>1997-2002</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Long-term academic outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Examination grades ((10^{th}) grade)</td>
<td>1986-1999</td>
<td>Dropout VG1 ((11^{th}) grade)</td>
<td>Parental characteristics, individual characteristics, and sibling characteristics.</td>
</tr>
<tr>
<td>Dropout VG1 ((11^{th}) grade)</td>
<td>1986-1998</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Long-term school behavior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor school behavior ((10^{th}) grade)</td>
<td>1992-1995, 1997-1998</td>
<td>Poor school behavior ((10^{th}) grade)</td>
<td>Parental characteristics, individual characteristics, and sibling characteristics.</td>
</tr>
<tr>
<td><strong>Long-term criminal behavior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criminal charge by age 17 ((11^{th}) grade)</td>
<td>1986-1997</td>
<td>Criminal charge by age 17 ((11^{th}) grade)</td>
<td>Parental characteristics, individual characteristics, and sibling characteristics.</td>
</tr>
</tbody>
</table>

Note: \(^{1}\) See appendix table A5.2 for a description of the groups of control variables.

### Table A5.2: Groups of control variables.

<table>
<thead>
<tr>
<th>Control variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parental characteristics</strong></td>
<td>Parent’s education (dummies), mother’s education (dummies), father’s earnings, father’s earnings squared, mother’s earnings, mother’s earnings squared, father social welfare benefits (dummy), mother social welfare benefits (dummy), father charged (dummy), mother charged (dummy), father’s marital status (dummies), mother’s marital status (dummies), mother’s year of birth, father’s year of birth squared,</td>
</tr>
<tr>
<td><strong>Individual characteristics</strong></td>
<td>Immigrant classification (dummies), year of immigration (dummies), gender, birth order, birth order squared, number of siblings, number of siblings squared, number of half siblings, number of half siblings squared, year of birth (dummies), child welfare service (dummy).</td>
</tr>
<tr>
<td><strong>Sibling characteristics</strong></td>
<td>Sibling charged (dummies), sibling completed upper secondary education (dummies), sibling in child welfare service (dummies).</td>
</tr>
<tr>
<td><strong>Prior standardized tests</strong></td>
<td>Low test score (5^{th}) grade, standardized test score (5^{th}) grade, standardized test score (5^{th}) grade squared</td>
</tr>
<tr>
<td><strong>GP visits</strong></td>
<td>Number of GP visits, number of GP visits squared, reason for GP visits (dummies)</td>
</tr>
</tbody>
</table>
Online Appendix 6: Correlation between outcome variables

Table A6.1: Pairwise correlations between outcome variables.

<table>
<thead>
<tr>
<th></th>
<th>Dropout</th>
<th>Examination grades</th>
<th>Low test scores</th>
<th>Standardized test scores</th>
<th>Charged 17</th>
<th>Poor school behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropout</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Examination grades</td>
<td>-0.305</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low test scores</td>
<td>0.275</td>
<td>-0.414</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized test scores</td>
<td>-0.257</td>
<td>0.670</td>
<td>-0.632</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charged 17</td>
<td>0.172</td>
<td>-0.207</td>
<td>0.107</td>
<td>-0.143</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Poor school behavior</td>
<td>0.278</td>
<td>-0.314</td>
<td>0.136</td>
<td>-0.192</td>
<td>0.264</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Online Appendix 7: Different cutoffs

**Figure A7.1:** Intervention effects for all students and students at different risk of adverse outcomes with intervals based on 1.96 estimated standard errors.

Note: Outcome metrics: Standardized for national tests and examination grades. Observed share for low achievement, poor school behavior, dropout, and charged. High-risk students are defined as the students with the highest estimated probability of academic failure (low achievement, standardized national tests), poor school behavior (poor school behavior), dropout (examination grades, dropout), and criminal charge (charged by 17).
Online Appendix 8: Type of criminal charges

Figure A8.1: Intervention effects for different types of charges with intervals based on 1.96 estimated standard errors.
Figure A8.2: Intervention effects for by the severity of charge with intervals based on 1.96 estimated standard errors.
Online Appendix 9: Fidelity

Table A9.1: Average score on the fidelity measures from the EBS survey by year of implementation.

<table>
<thead>
<tr>
<th>Year</th>
<th>Classroom level</th>
<th>Individual-level</th>
<th>Non-classroom level</th>
<th>School-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.48</td>
<td>0.26</td>
<td>0.46</td>
<td>0.52</td>
</tr>
<tr>
<td>1</td>
<td>0.54</td>
<td>0.28</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.36</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td>3</td>
<td>0.76</td>
<td>0.50</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td>4</td>
<td>0.78</td>
<td>0.57</td>
<td>0.79</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Figure A9.1: Percent in place by year after initiation of SWPBS.
Figure A9.2: Quantile plot that shows the fraction of schools that implement at a given fidelity level or higher by year since implementation. The dotted line at 0.8 represents the threshold of 80% implementation.
Figure A9.3: Effect estimates for high-fidelity schools (at least 80% implementation) with intervals based on 1.96 estimated standard errors.

Note: Outcome metrics: Standardized for national tests and examination grades. Observed share for low achievement, poor school behavior, dropout, and charged. Standard errors are clustered at the school level for all outcomes except standardized national tests, where we use the Huber-White (heteroskedasticity-robust) standard errors without adjustment for clustering. For standardized national tests, we have only three SWPBS schools in the pre-treatment period that later implement with high fidelity. In cases with few clusters, in total or treated, and in some cases with cluster-fixed effects (as we have), estimation error in the cluster adjustment
can lead to too small standard errors and over-rejection of the null hypothesis.

**Figure A9.4:** Effect estimates for low-fidelity schools (less than 60% implementation) with intervals based on 1.96 estimated standard errors.
Note: Outcome metrics: Standardized for national tests and examination grades. Observed share for low achievement, poor school behavior, dropout, and charged.
Online Appendix 10: Effect size without the DiD design

Figure A10.1: Intervention effects with controls for birth cohort (M1) and controls for birth cohort and observed student characteristics (M2) compared to the DiD model (M3).
Online Appendix 11: Effects on short-term academic outcomes without additional controls

Figure A11.1: Effect estimates (after) and pre-program heterogeneity with and without additional controls for prior test scores and GP visits for all students with intervals based on 1.96 estimated standard errors.

Note: Outcome metrics: Standardized for national tests and observed share for low achievement.
Online Appendix 12: Placebo regression

Figure A12.1: Effect estimates (after) and pre-program heterogeneity on academic achievement in 5th grade (placebo test) and 8th grade with intervals based on 1.96 estimated standard errors.

Note: Outcome metrics: Standardized for national tests and observed share for low achievement.