Schoner, Florian; Mergele, Lukas; Zierow, Larissa

Conference Paper

Grading Student Behavior

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2022: Big Data in Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Schoner, Florian; Mergele, Lukas; Zierow, Larissa (2022) : Grading Student Behavior, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2022: Big Data in Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/264140

CESifo WORKING PAPERS

9275
2021

Original Version: August 2021
This Version: October 2021

Grading Student Behavior

Florian Schoner, Lukas Mergele, Larissa Zierow

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9275*

Grading Student Behavior

Abstract

Numerous countries require teachers to assign comportment grades rating students' social and work behavior in the classroom. However, the impact of such policies on student outcomes remains unknown. We exploit the staggered introduction of comportment grading across German federal states to estimate its causal effect on students' school-to-work transitions as well as academic achievement and non-cognitive abilities. Analyzing census data, household surveys, and nationwide student assessments, we show that comportment grading does not meaningfully affect these outcomes and rule out large effect sizes. Exploring reasons for this finding, supplementary analyses suggest that comportment grades do not convey much information beyond students' grade point average.

JEL-Codes: D910, I210, I280, J240.

Keywords: school reforms, report cards, school-to-work transition, student achievement.

*Florian Schoner**
ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Munich / Germany
schoner@ifo.de

Lukas Mergele
ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Munich / Germany
mergele@ifo.de

Larissa Zierow
ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Munich / Germany
zierow@ifo.de

*corresponding author

October 22, 2021

We would like to thank Luca Facchinello, Joshua Goodman, Elisabeth Grewenig, Dan Hamermesh, Eric Hanushek, Rasmus Landersø, Sven Resnjanskij, Felix Rösel, Pedro Sant'Anna, Felix Weinhardt, and Ludger Woessmann for helpful comments and suggestions. We are grateful for financial support by the Leibniz Association under the competitive procedure for the project "Efficiency and Equity in Education: Quasi-Experimental Evidence from School Reforms across German States (EffEE)". Diva Barisone and Sophia Hueber provided excellent research assistance.

1 Introduction

Although comportment grading is used worldwide, there is no evidence as to its impact on student outcomes. The policy involves assigning a mark to student's social and work behavior in school and is commonplace in numerous European countries, including Italy, Germany, Poland, and Norway (see Table A.1 for an overview). Countries outside Europe such as Japan and Hong Kong follow similar practices, requiring teachers to rate students' behavior on school report cards (Urabe 2006; Cheung and Llu 2000). Comportment grading was also a mainstay in US schools (Maynard 1977; Currie 1995) until their shift towards objective measures of educational output and standards-based grading (Tyre 2010; Duckworth et al. 2012).[1]

The merits of comportment grading are fiercely debated. Proponents argue that the threat of receiving poor comportment grades might incentivize students to behave better in class. This would be in line with literature showing that grades, in general, can serve as incentives in the schooling context and that these incentives are important for students' educational investments (Hvidman and Sievertsen 2019). To the extent that comportment grades can also serve as indicators of non-cognitive abilities (Landersø and Heckman 2017), they might enable students to signal these abilities to employers, thereby reducing information asymmetry and facilitating students' transition into the labor market (Protsch and Solga 2015). Opponents of comportment grades point out that these grades are highly context-dependent, rendering them hardly comparable. This lack of standardization could lead students to feel they are being treated unfairly when receiving them, which might have demotivating effects on their learning as well as behavior (Close 2009). Furthermore, intrinsic motivation for good behavior could be crowded out by grade-driven extrinsic motivation (see Koch et al. 2015, for a comprehensive discussion of such motivational crowding-out in the context of educational interventions). Thus, the theoretical case for comportment grading is ambiguous. Empirical evidence on the causal effects of comportment grading does not exist.

This paper exploits a sequence of reforms across German federal states that introduced comportment grades in schools between 2001 and 2007 as a natural experiment. Due to its federal structure, the German setting offers a rare laboratory to examine education policies within a common political and economic framework. We exploit this policy variation using a staggered difference-in-differences design. After providing evidence that the main identifying assumption – parallel trends – is likely to hold, we adopt a two-way fixed effects (TWFE) estimation strategy. To avoid its pitfalls arising in the presence of heterogeneous or dynamic treatment effects (Goodman-Bacon 2021; de Chaisemartin and D'Haultfœuille 2020), we adopt the estimation routine put forward in Callaway and Sant'Anna (2020). We establish that heterogeneity in treatment effects are unlikely to matter in our application.

Based on German census data, we find that the effect of comportment grading on the probability of being employed or in training after school is not distinguishable from zero. We can reject effect sizes larger than 6 percentage points in absolute value. Next, we analyze

[1] Figures A.1 and A.2 in the appendix provide famous examples of report cards including comportment grading (also referred to as "deportment" or "conduct" grades) from former US presidents.

representative household surveys and nationally standardized student assessments to investigate potential mechanisms through which the school-to-work transition might be affected. Both analyses yield a concordant picture: Neither non-cognitive skills nor academic achievement are significantly affected by the reform.

In sum, our results confirm that the comportment grading reforms did not substantially alter student outcomes. To arrive at this conclusion, we adopt a careful approach to identification and estimation and rely on three different data sources, two of which include large numbers of observations. Moreover, we think that there is no reason to expect a null-effect of comportment grading ex-ante. Therefore, our findings are of high informational value in the sense that they shift beliefs about the causal effect of comportment grading reforms (Abadie 2020).

Investigating potential explanations for our results, we conduct a supplementary analysis using report card data on students' actual comportment grades as well as their subject grades. Positive correlations among grades suggest that subject grades and GPA partly contain the information in comportment grades. In fact, grades and GPA together explain a substantial share of the variation in comportment grades.

Our paper contributes to three strands of literature. First, we advance the knowledge on the factors within the schooling environment that facilitate a successful school-to-work transition (Ryan 2001). While there is a large literature studying these factors (e.g. Zimmermann 2013), we are the first to focus on the effect of grading students' behavior in school, a widely implemented policy. To our knowledge, there are only two other papers studying the extent to which receiving school grades generally matters for labor market outcomes. Facchinello (2020) examines a Swedish reform that postponed the introduction of grades in school by several years. He finds that while students from advantaged backgrounds are more likely to be unemployed early in their career, disadvantaged students see their incomes increase. However, effects do not persist in the long run. Tan (2020) compares the labor market outcomes of similarly able students who receive different letter grades at a Singaporean university. He shows that better letter grades translate into higher earnings. In contrast, we provide evidence on grades meant to measure behavior, not academic achievement.

Second, we contribute to the understanding of non-cognitive skill formation in school (Bowles and Gintis 2002) by investigating whether they can be fostered through grading comportment. Research on skill formation (e.g. Cunha and Heckman 2007) shows that these skills are malleable, for instance through mentoring programs in childhood and adolescence (Kautz et al. 2014; Kosse et al. 2020; Resnjanskij et al. 2021). There is also evidence for a robust link between these abilities and labor market outcomes (Heckman et al. 2006; Almlund et al. 2011; Heineck and Anger 2010). We are able to test whether students indeed adopt behaviors that are more compliant with conduct requirements in order to obtain positive feedback - as is often proposed as an argument in favor of comportment grades. In Germany, the requirements for good comportment grades include being companionable, hard-working, and honest. These concepts are closely related to the non-cognitive skills agreeableness and conscientiousness from the Big Five personality factors and to trust, which measures pro-social beliefs (Becker

et al. 2012).

Third, we investigate whether comportment grades foster cognitive skill formation and academic achievement more broadly. Both are associated with better labor market outcomes (Card 2001; Hanushek et al. 2015) as they constitute mechanisms through which the school-to-work transition might be affected. Comportment grades also enable teachers to sanction disruptive behaviors in a way that is also visible to parents, potentially incentivizing students to behave better in class. Since less disruptive classrooms enhance academic achievement, comportment grading might have beneficial effects on human capital formation (Lazear 2001; Angrist et al. 2013; Kristoffersen et al. 2015; Dobbie and Fryer 2020).

The remainder of this paper proceeds as follows: Section 2 details the institutional background underlying our work. Section 3 introduces the data sources we use. Section 4 outlines how we identify and estimate the causal effect of comportment grading. Section 5 presents our results and robustness checks. Section 6 offers potential explanations for our findings and Section 7 concludes.

2 Institutional Setting

In Germany each of the country's 16 federal states is solely responsible for its respective school system. This leads to policy differences across states although the general structure remains similar. Figure A.3 in the appendix provides a graphical overview of the school system. After four years in primary school, children are placed into one of three secondary school tracks: basic school (*Hauptschule*), middle school (*Realschule*), and academic track school (*Gymnasium*). Whereas academic track schools prepare students for studying at university, the other two tracks prepare students for entering the labor market through vocational training.

A common feature of these school types has been the evaluation of students' comportment. Students' biannual report card contains not only subject-specific grades as an assessment of their academic achievement but also grades related to working habits and social behavior as an assessment of their comportment. Only in the final two years of academic track schools do comportment grades not come into use. [2] When grading work and social behavior, teachers typically consider students' diligence, commitment, dependability, willingness to address conflict, readiness to help, and reflection capability.[3] Underscoring their significance, these grades are referred to as "head grades" (in German: "Kopfnoten") since they are placed at the top of the report card, above the subject grades. Comportment grades do not determine which secondary school track a student is able to attend. For school-to-work transitions, however, comportment grades signal important non-cognitive skills. Correspondence studies show that comportment grades are an important selection criteria in the apprenticeship market, the main

[2] An exception is the regulation of North Rhine-Westphalia after 2007 where comportment grades were mandatory even in the final years of academic track schools. However, this regulation falls outside of our sample period.

[3] As an example, Tables A.2 and A.3 in the appendix present teacher guidelines for the assessment of behavior in the states of Baden-Wuerttemberg and Saxony.

labor market entrance for students without tertiary-level education. Protsch and Solga (2015) show that employers may value comportment grades even more than regular subject grades.

After fierce public debates on the potential effects of comportment grades in the 1970s, some West German states dismissed comportment grading in schools (Helbig and Nikolai 2015). In East Germany comportment grades were the norm prior to reunification but were later abolished in several states. We exploit the second wave of reforms in East and West German states, which began reintroducing comportment grading in the early 2000s. Students in all German states received comportment grades by 2007.

As Figure 1 shows, comportment grading was adopted in four federal states during our period of study (1996 to 2007): Bremen (introduced in 2001), Brandenburg (2001), Saxony-Anhalt (2003), and North Rhine-Westphalia (2007). This policy cannot be clearly assigned to a specific political program as it was introduced by center-left governments (Bremen and Brandenburg) and center-right ones (Saxony-Anhalt and North Rhine-Westphalia) alike.

3 Data

To investigate whether the introduction of comportment grading affects students' school-to-work transition and skill formation, one would ideally draw on a single set of panel data with detailed information on individuals' schooling history, skill measures, and employment records. Given the lack of such a dataset, we compile repeated cross-section data drawn from three different sources. First, we use census data – the German Microcensus – to get information about individuals' school-to-work transition. Second, we use individual-level survey measures of respondents' non-cognitive skills from the German Socio-Economic Panel. Third, data on ninth-grade literacy test scores and track attendance are drawn from nation-wide student assessment studies. Applying the same set of sample restrictions across datasets makes this data well-suited to test our hypotheses. All of the data sources provide individuals' year of enrollment in school and their federal state of schooling, linking them with the respective date at which the reform was introduced. For these three sources, we add state-level information about whether schools grade comportment to derive treatment and control group assignments. Finally, we retrieve report card data from the National Educational Panel Study to deconstruct the relationship between subject and comportment grades, and make sense of our reform analysis.

3.1 Data on the school-to-work transition

The German Microcensus offers an administrative data source covering one percent of the German population in annual waves since 1970. We make use of the 2011–2016 waves and restrict the sample to individuals aged between 15 and 25 living in a state that introduced comportment grading. We exclude individuals who still attend secondary school and those studying towards a university entrance degree who have yet to transition into the labor market.

Thus, we focus on students who have completed secondary education, which enables them to start working directly after school or to begin vocational training ("Ausbildung").

We use information on individuals' employment status to derive a binary measure capturing successful school-to-work transitions. More specifically, we consider an individual to have successfully transitioned from school to work if she is in vocational training, completing secondary-schooling degree after finishing a lower one, or is employed at least part-time. Conversely, unsuccessful transitions include individuals that are marginally employed, looking for work, or temporarily out of the labor force.[4]

Table A.4 provides the descriptive statistics of this sample, which consists of 22,895 observations.

3.2 Non-cognitive skill measures

Survey measures of non-cognitive skills are taken from the German Socio-Economic Panel (SOEP, see Goebel et al. 2019), a survey data set representative of private households in Germany. From the SOEP, we build a cross-section of individuals aged 17 to 25 from different survey years ($2003 - 2018$) and born between 1990 and 2000.

We investigate the formation of non-cognitive skills that directly relate to criteria teachers are expected to consider when grading comportment (Table A.2 and A.3). We focus on agreeableness and conscientiousness from the "Big Five" personality factors, which overlap with the "Camaraderie" and "Work effort" criteria (Table A.2). We also investigate an individual's level of trust, which potentially affects one's willingness to be honest, and is therefore related with the "Honesty" dimension (Table A.2). Each of these latent concepts is measured using answers to three survey items on Likert-type scales. To generate a single measure for each concept, we average the items' scores for each individual. If measures from different survey years are available for a given individual, we take the earliest available measure.

Table A.5 provides the descriptive statistics of our SOEP sample which consists of 2,121 individuals.

3.3 Nationwide student assessments

Measures of students' academic achievement are taken from the German extension of the Programme for International Student Assessment (PISA-E), which is available with federal state identifiers for the years 2000, 2003, 2006, and 2012. For the years 2009 and 2015, we employ data from the National Assessment Study by the Institute for Educational Quality Improvement (IQB), which is collected in accordance with PISA. The data were made available by the Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB).

[4]This corresponds closely to the definition of "out-of-school joblessness" given in Ryan (2001), which includes those unemployed according to the ILO/OECD definition and those not enrolled in an educational course. We add the marginally employed to this group since we are interested in transitions from school into stable employment relationships.

All achievement testings target students in ninth grade and are always performed between May and July. As participating schools within each state are drawn at random, each wave constitutes a cross-section of ninth graders that is representative at the state level. Taken together, these waves form a quasi-panel of German states from 2000 to 2015, with observations occurring every three years. We impose identical sample restrictions as used for the census data wherever possible.

In additon to compulsory tests measuring students' reading skills, there are questionnaires given to schools, students, and parents that elicit a wide range of socio-demographic background characteristics. Test scores are standardized and comparable across waves. To capture different facets of student achievement, we focus on reading test scores in ninth grade and whether students attend an academic track school, the most demanding school track in Germany and the one leading to a university-entrance qualification. While the latter is an indicator variable, we standardize reading test scores to have mean zero and unit standard deviation. We do not consider math skills as they are only tested in every other wave of the National Assessment Study.[5]

Table A.6 provides the descriptive statistics of our student assessment data which consists of 42,415 observations.

3.4 Comportment grading reforms

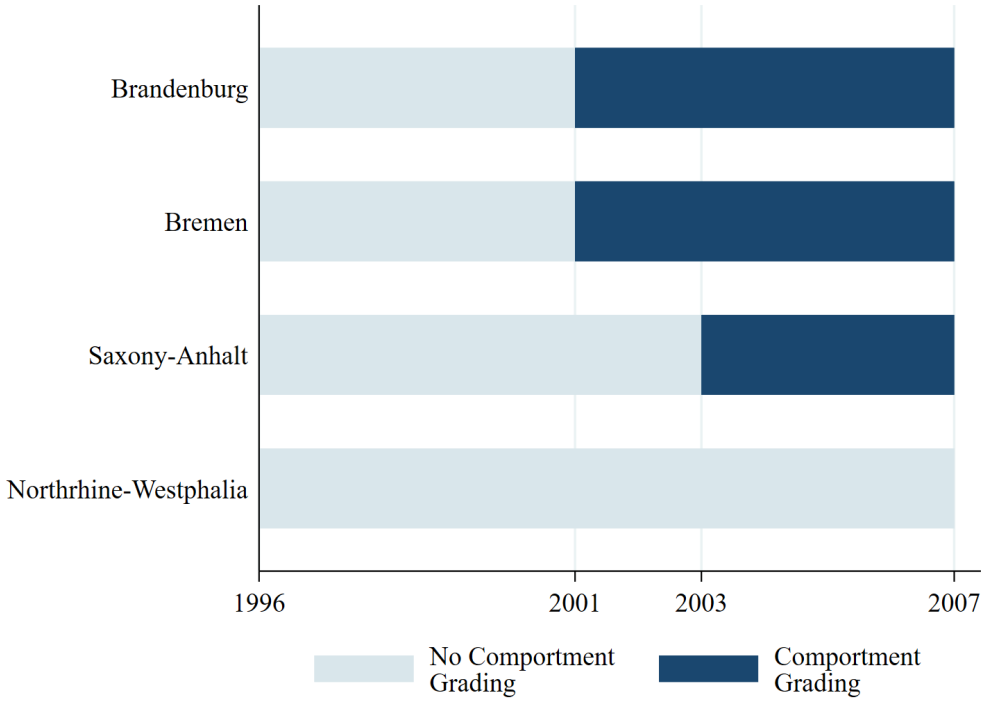
Data on state-level comportment grading policies were gathered from school reform coding based on the states' schooling legislation and collected by Helbig and Nikolai (2015). We classify state policies according to four categories: (1) no comportment grading, (2) optional verbal comportment grading, (3) mandatory verbal comportment grading, or (4) mandatory numerical comportment grading. Given that we are interested in the effect of comportment grading per se, we consider individuals experiencing any kind of mandatory grading of comportment as treated ((3) and (4)) and the others as non-treated ((1) and (2)).

As Figure 1 shows, four federal states adopted comportment grading during our sample period (1996 to 2007). More specifically, Bremen and Brandenburg constitute the first treatment group, which introduced comportment grading in 2001; Saxony-Anhalt followed suit in 2003 and therefore serves as the second treatment group. Finally, North Rhine-Westphalia did not adopt comportment grading until 2007 and consequently serves as our control group. The other federal states already had comportment grading schemes in place prior to our sample period. These states are excluded from our analysis since their untreated potential outcomes are never observed.

Individuals are considered treated if comportment grading was in place in the year they enrolled in school, i.e. treated students received these grades throughout their school career since none of the reforms were revoked. As a robustness check, we report results where we assign individuals to the treatment group if they have received comportment grades the year

[5]For more details on the data, refer to Baumert et al. (2002), Prenzel et al. (2007), Prenzel et al. (2010), Prenzel et al. (2019), Sachse et al. (2012), and Schipolowski et al. (2018).

FIGURE 1. The introduction of comportment grading over time and by state within the sample



Notes: Comportment grading is defined as the implementation of either mandatory verbal or mandatory numerical comportment grading. German states excluded from the overview introduced comportment grading ahead of our sample period. We exluce these states from our analysis. North Rhine-Westphalia did not introduce comportment grading until 2007.
Sources: Own representation based on Helbig and Nikolai (2015).

the transition into secondary school takes place. Since secondary school is a subset of one's entire school career, this implies a larger treatment group (see Tables A.4, A.6, and A.5).

3.5 Report card data

Finally, we use the Starting Cohort 3 from the National Educational Panel Study (NEPS SC3, version 10.0.0) to compare actual comportment grades with subject grades included in report cards (Blossfeld et al. 2011). The first wave from fall 2010 includes individual-level data for students in grade five. These individuals were resurveyed at regular intervals until wave 10, collected in fall 2018 when students were about 19 years old. Comportment grades were elicited in waves eight to ten, referring to students' respective final report card at graduation. We also retrieve the final grade-point average (GPA) as well as subject grades in Math and German. To harmonize the grading information, we round all grades to the next integer as reporting formats differ across grades. We also reverse the standard German grading scale to ease interpretation, such that higher numbers indicate better grades. This implies that our grades range from 1 ("insufficient") to 6 ("very good"). Moreover, we use data from wave 10 for information on agreeableness and conscientiousness from the "Big Five" personality factors. We focus on students who received comportment grades within their final report

cards. If individuals achieve more than one school degree, we keep the first one that contained comportment grades on the final report card.

Table A.7 provides the descriptive statistics of our grading data which consists of 812 students and their report cards.

4 Empirical strategy

Identifying the average effect of comportment grading on the treated relies on the staggered adoption of comportment grading across federal states, which gives rise to a generalized difference-in-differences approach.

Therefore, we are interested in the coefficient δ of the following regression

$$Y_{ist} = \gamma_s + \lambda_t + \delta \cdot \text{CG}_{st} + \mathbf{X}_{ist}^{\top}\boldsymbol{\beta} + \varepsilon_{ist}, \tag{1}$$

where Y_{ist} is an outcome for student i attending school in state s in cohort t. CG_{st} is a dummy variable equal to one if schools in this state graded comportment for this cohort and zero otherwise. $\mathbf{X}_{ist}$ contains an individual's sex and migration background to increase precision. Furthermore, we include a set of fixed effects capturing the federal state of schooling (γ_s), year of enrollment (λ_t), and the survey year since responses are taken from different survey years. ε_{ist} is an error term.

Standard errors are clustered at the level of treatment assignment, which is the level of federal states (Abadie et al. 2017). To account for the small number of clusters as a potential source of bias in the coefficients' variance estimates (e.g. Cameron et al. 2008), we apply the wild cluster bootstrap (WCB) procedure outlined in Roodman et al. (2019).

To identify the causal effect of comportment grading on student outcomes, we need to assume parallel trends. This means that, in absence of the reforms, the relationship of outcomes would have followed the same trajectory both treatment groups relative to the respective control group. Although fundamentally untestable, we corroborate this assumption by investigating pre-treatment trends in an event-study specification using our main dataset, the census data.[6] Note that we allow parallel trends to hold only after conditioning on student sex and migration background since the latter is unbalanced across groups and potentially affects the evolution of student outcomes (Abadie 2005; Heckman et al. 1997). Figure 2 shows that pre-treatment coefficients are not statistically distinguishable from zero for both treatment groups, suggesting that the parallel trends assumption is likely to hold.

Another threat to identification arises from different school reforms introduced at the same time as comportment grading. For this reason, we investigated the compendium of German school reforms since World War II by Helbig and Nikolai (2015) and did not find any concomitant school reform. The only exception is Saxony-Anhalt, where comportment grading was introduced in parallel to a shortening of the duration of primary school from six to four

[6]Note that our event-study results are based on the approach put forward in Sun and Abraham (2020) and therefore not biased by treatment effect heterogeneity as explained below.

years. We address this potential concern by analyzing the two groups of states that introduced comportment grades in different years separately (Figure 2), showing that the effects are highly similar in both groups.

Ordinary least squares estimates of δ using the TWFE specification above capture a causal effect only if treatment effects are homogeneous across time and units (de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021). This is because the TWFE estimator corresponds to a weighted average of all possible 2x2 difference-in-means estimates during the sample period. These include invalid comparisons of newly-treated to already-treated units. If treatment effects evolve over time, the estimated 2x2 effects from invalid comparisons might be weighted negatively, i.e. subtracted from the estimate when being aggregated to a single measure (Goodman-Bacon 2021). Although dynamics might be a lesser concern here, we want to ensure our estimates' robustness regarding the issues arising from treatment effect heterogeneity.[7] Therefore, we exclude states already using comportment grading schemes prior to our sample period and implement the estimator proposed by Callaway and Sant'Anna (2020), henceforth (C/S). It is robust against both forms of treatment effect heterogeneity and differs from the TWFE approach mainly by ensuring that newly-treated units are only compared to not-yet-treated units. Note that the C/S approach does not allow us to conduct cluster-robust inference due to the small number of clusters. We therefore report confidence intervals based on heteroskedasticity-robust confidence intervals. In contrast to usual practice, we report simultaneous instead of pointwise confidence intervals which are robust to multiple hypothesis testing.

We implement two measures to ensure comparability between the two approaches. First, we average group- and unit-specific 2x2 effects across both time periods and treatment groups to obtain a single estimate that can be interpreted as a multi-period and multi-group extension of the average treatment effect on the treated (ATT; Callaway and Sant'Anna 2020).[8] Second, we run both estimation routines on the exact same set of individuals by dropping units that have received treatment already before or at the start of the sample period. As expected, Table 1 shows that C/S and TWFE estimates hardly differ, irrespective of whether controls are included. This suggests that treatment effect heterogeneity is not an issue here. For this reason, we will adhere to the TWFE approach for the remaining analyses as it allows us to account for the small number of clusters when conducting inference.

5 Results

Our results show that comportment grading does not affect students' school-to-work transition. Potential intermediate outcomes, such as non-cognitive skills and student achievement, are also

[7] Since we use repeated cross-section data, dynamic treatment effects would be equivalent to assuming cross-cohort spillover effects. More specifically, treatment effects would need to be a function of the number of cohorts that had already been treated prior to the current cohort. This is because we do not observe individuals repeatedly, i.e. there is no way treatment effects can evolve for individuals.

[8] See Appendix section A.2 for a formal description of the C/S estimand and how we aggregate effects in our setting.

unaffected by the comportment grading reforms.

5.1 Comportment grading and the school-to-work transition

Panel A of Table 1 displays estimates of the aggregated ATT estimand in equation 2 (columns 1 and 2) and from estimating equation 1 (columns 3 and 4) using our preferred definitions of variables. Even-numbered columns add student sex and migration background as control variables. Point estimates obtained from C/S imply that the comportment reform-induced change in the probability of transitioning from school to work successfully are very close to zero, amounting to 0.13 and 0.14 percentage points in columns 1 and 2, respectively. Confidence intervals suggest that effect sizes larger than six percentage points in absolute value are highly unlikely. Plausible effect sizes are of small magnitude given that an average 86% of individuals transition from school to work successfully. The third and fourth column display the results from estimating equation 1. Estimated effect sizes hardly differ across estimation techniques, corroborating the notion that dynamic treatment effects and ensuing negative weights issues are a lesser concern in our setup. Again, the results do not change much when controlling for individual-level characteristics. Note that while confidence intervals based on cluster-robust standard errors are much smaller than those robust to multiple hypothesis testing, we expect the former to be too narrow given the small number of clusters. Therefore, we use the wild cluster bootstrap procedure outlined in Roodman et al. (2019). The third row of Table 1 shows the resulting p-values (0.8859 and 0.9209), reinforcing that we fail to reject the hypothesis of a zero-effect by a wide margin.

To test the robustness of this zero-effect of comportment grading on the school-to-work transition, we apply several changes to our main approach. The results are shown in the remaining panels of Table 1.

Panel B changes the assignment of treatment based on the year of secondary school enrollment. While point estimates are negative, effects are indistinguishable from zero at conventional levels of significance.[9]

Panel C defines "successful school-to-work transition" more strictly by considering employed individuals without a vocational qualification prior to their employment as unsuccessful. Estimated coefficients are negative but once again indistinguishable from zero except for column 3. There, we reject the hypothesis of a zero effect at the five percent level, but fail to do so at smaller levels. Panel D restricts the sample to individuals on the labor market by excluding those who have completed a further degree after secondary school to rule out that they are driving our results. This reduces the sample size to 19,263 individuals. Although point estimates are larger than those in panel A, they are indistinguishable from zero at any conventional level.

Finally, panel E combines both restrictions taken in the approaches shown in panels C and

[9] Although confidence intervals based on cluster-robust standard errors suggest a statistically significant effect in columns 3 and 4, we expect them to overreject and therefore ground our interpretation on the p-values obtained from the wild cluster bootstrap.

D. Point estimates are closer to zero than in panels C and D, corroborating our zero-effect finding.

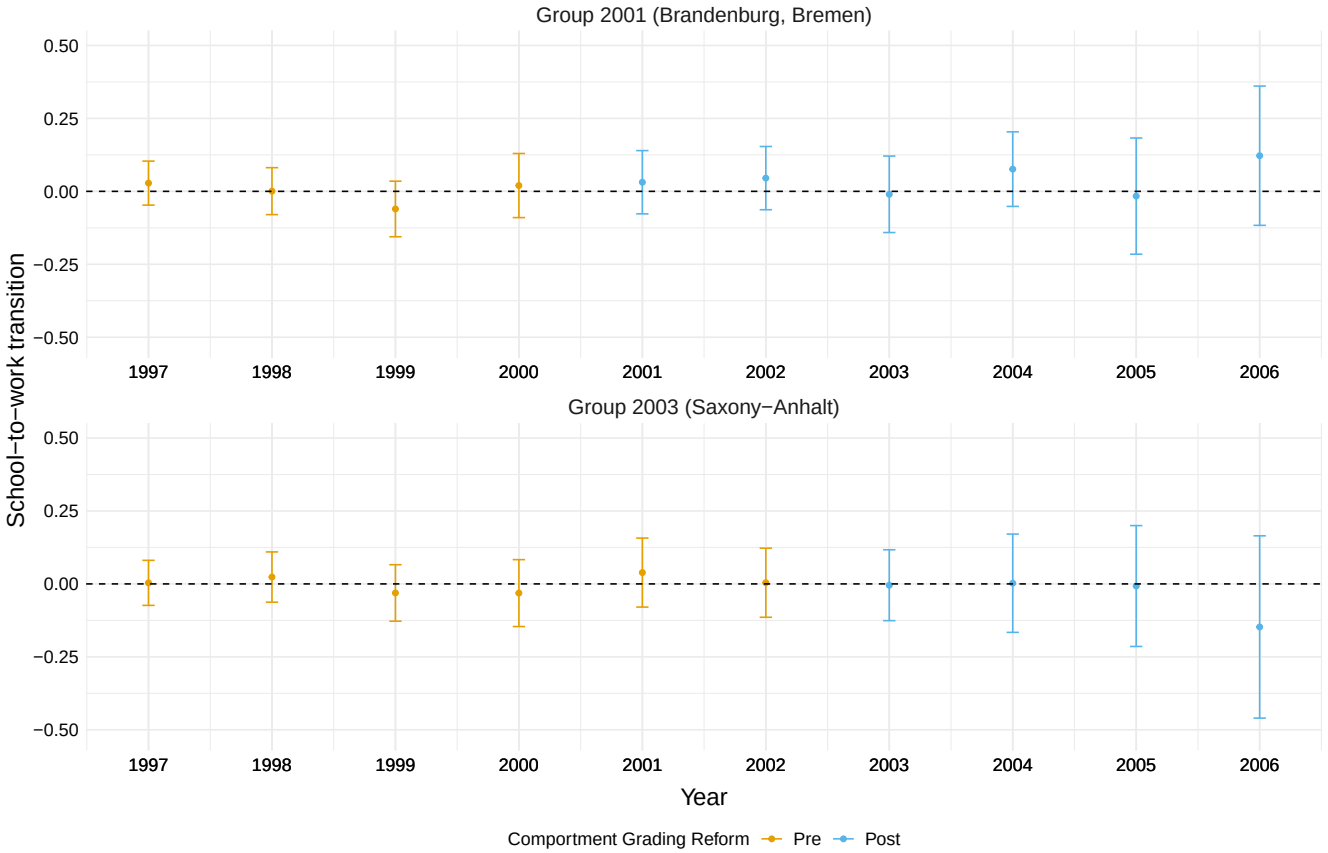
Three patterns emerge from these results. Most importantly, comportment grading neither enhances nor reduces the chances of a successful school-to-work transition. Second, effect sizes are similar across estimation techniques and throughout panels, showing that heterogeneous treatment effects are not an issue in our setting. The latter point is also vividly illustrated in Figure 1, which shows that post-reform point estimates differ neither across time periods nor groups. Finally, point estimates are robust to including individual-level controls.

TABLE 1. Effect of comportment grading on school-to-work transitions

	Successful School-to-work Transition			
	Callaway & Sant'Anna (C/S)		Two-Way-Fixed-Effects (TWFE)	
	(1)	(2)	(3)	(4)
Panel A: Main				
	0.0013	0.0014	0.0018	0.0021
	[−0.0602, 0.0628]	[−0.0549, 0.0577]	[−0.0237, 0.0274]	[−0.0188, 0.0231]
WCB p-val.	-	-	0.8859	0.9209
Panel B: Treatment assignment in grade 4				
	−0.0268	−0.0184	−0.0224	−0.0189
	[−0.0619, 0.0084]	[−0.0543, 0.0175]	[−0.0308, −0.0139]	[−0.0290, −0.0088]
WCB p-val.	-	-	0.2923	0.3183
Panel C: Stricter definition of success				
	−0.0211	−0.0159	−0.0410	−0.0406
	[−0.0859, 0.0437]	[−0.0737, 0.0419]	[−0.0490, −0.0330]	[−0.0549, −0.0263]
WCB p-val.	-	-	0.0460	0.1301
Panel D: Without those catching up				
	0.0440	0.0393	0.0338	0.0334
	[−0.0263, 0.1142]	[−0.0228, 0.1014]	[0.0068, 0.0609]	[0.0124, 0.0545]
WCB p-val.	-	-	0.3073	0.2052
Panel E: Combine C and D				
	0.0153	0.0186	−0.0089	−0.0100
	[−0.0583, 0.0890]	[−0.0494, 0.0866]	[−0.0173, −0.0006]	[−0.0253, 0.0053]
WCB p-val.	-	-	0.9219	0.6877
Mean Dep. Var. (A, B, D)	0.86	0.86	0.86	0.86
Mean Dep. Var. (C, E)	0.77	0.77	0.77	0.77
N (A – C)	22,895	22,895	22,895	22,895
N (D – E)	19,263	19,263	19,263	19,263
Controls	No	Yes	No	Yes
Std. Error	Robust	Robust	Cluster	Cluster

Notes: Estimates of the overall ATT (see equation 2) according to Callaway and Sant'Anna (2020) (columns 1 and 2) and from TWFE regressions using state, cohort and survey year fixed effects (columns 3 and 4). Columns 2 and 4 additionally include a female and migration background indicator as control variables. Columns 1 and 2 report simultaneous 95% confidence intervals robust to heteroskedasticity and multiple hypothesis testing. Columns 3 and 4 report 95% confidence intervals based on cluster-robust standard errors and p-values from the wild cluster bootstrap routine using weights from Webb's distribution (Roodman et al. 2019) and 999 iterations. Excluding individuals that catch up on a degree leads to a sample size of 19,263.
Source: Microcensus waves 2011–2016.

FIGURE 2. Dynamic effect of comportment grading on school-to-work transitions



Notes: Figure displays estimates of period- and group-specific ATTs for the two treatment groups. The dependent variable is binary and indicates a successful school-to-work transition (see Section 3). Specifications include indicators for students' sex and migration background. Error bars correspond to simultaneous 95% confidence bands based on robust standard errors.
Sources: Microcensus waves 2011–2016

5.2 Comportment grading and non-cognitive skills

Having established that grading social and work behavior does not alter school-to-work-transitions, we analyze whether intermediate outcomes such as non-cognitive skills are affected. Table 2 shows the results of estimating equation 1 using measures of non-cognitive skills as outcomes from our SOEP sample. In line with the zero-effect finding on the school-to-work transition, we do not detect statistically significant effects of comportment grading on any of the non-cognitive skill measures. Estimated effect sizes are small: They range from zero to four percent of a standard deviation in absolute value and fail to reject the null at conventional levels of significance. p-values from the wild cluster bootstrap procedure bolster this finding. Tables A.8 and A.9 in the appendix contain results of robustness checks. While Table A.8 displays results of regressions without any control variables, Table A.9 changes the assignment of treatment based on the year of secondary school enrollment as in Panel B of Table 1. Results in both cases corroborate our zero-effect finding.

TABLE 2. Effect of comportment grading on non-cognitive skills

	Trust	Conscientiousness	Agreeableness
ATT	−0.0277	0.0003	−0.0416
	[−0.1512, 0.0958]	[−0.0922, 0.0927]	[−0.1168, 0.0336]
WCB p-val.	0.7137	0.7968	0.6066
Adj. R-squared	0.0228	0.1114	0.0277
Observations	2,121	2,121	2,121
Std. Error	Cluster	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, cohort, and survey year fixed effects. All outcomes are standardized to have mean zero and unit standard deviation. Controls include student sex and a dummy for migration background. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values use weights from Webb's distribution and rely on 999 iterations (Roodman et al. 2019). 95% confidence intervals are in box brackets.
Sources: SOEP-Core v36

5.3 Comportment grading and student achievement

Next, we test an alternative intermediary for potential long-term effects of comportment grading and investigate whether comportment grading affects student achievement by the end of ninth grade. Table 3 reports estimates from equation 1 using OLS employing indicators of student achievement on the left-hand side. Reading test scores are z-scored and academic track attendance is an indicator variable. In line with results above, estimated effects on reading test scores and academic track school attendance are statistically indistinguishable from zero. Using wild cluster bootstrap routines leads to the same inferences.

These patterns are robust against a variety of concerns, as demonstrated by further analyses in the appendix. First, in Table A.10, we exclude any background characteristics, indicating that even raw differences do not suggest any changes in student achievement. In Table A.11, we control for an extended set of individual characteristics as the national assessment data contains the richest and most complete set of background characteristics at the student and school level. This leads to smaller point estimates, which are still not statistically different from zero. In Table A.12, we add controls at the school level, which does not change results but reduces our sample size. Our previous analyses simply used the first plausible value from the probability distribution of a student's reading skills. Although this approach should lead to unbiased coefficient estimates for sufficiently large samples, we assess whether using one of the other four plausible values offered by the data providers would lead to differing conclusions (see Table A.13). Comparing the five estimation results, it turns out that the first plausible value (PV1) as used in previous analyses spans the largest confidence interval, thus providing a particularly conservative approach. Using any plausible value as an outcome measure supports our zero-effects finding of comportment grading on reading skills. Finally, we perform the same robustness test as in Panel E of Table 1, that is, we change the assignment of treatment

TABLE 3. Effect of grading behavior on academic achievement in ninth grade

	Reading Skills	Academic Track School Attendance
ATT	0.2238	0.0002
	[−0.2039, 0.6515]	[−0.1818, 0.1821]
Outcome mean	0.01	0.34
WCB p-val.	0.364	0.995
Adj. R-squared	0.043	0.029
Observations	42,415	42,415
Std. Error	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, cohort, and survey year fixed effects. Controls include student sex and a dummy for migration background. Reading Skills are standardized to have mean zero and unit standard deviation while Gymnasium Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values use weights from Webb's distribution and rely on 999 iterations. 95% confidence intervals are in box brackets.
Sources: PISA 2000, PISA 2003, PISA 2006, IQB-LV 2008-9 (v2), PISA 2012, IQB-BT 2015 (v5).

year (year of enrollment in secondary school instead of enrollment in primary school). As shown in Table A.14, the effect on academic track school attendance is robust to this change, and the effect on reading skills is positive yet not significant when considering wild cluster bootstrap p-values.

6 Potential explanations

Having established that comportment grades do not meaningfully alter student outcomes raises the question why this is the case and, relatedly, what these grades actually measure. Although our analyis cannot distinguish between alternative hypotheses, we outline several explanations that are in line with auxiliary findings and the existing literature.

In terms of measurement, appendix Table A.15 documents cross-correlations among different types of grades and noncognitive skills based on the NEPS data. Positive correlations suggest that subject grades and GPA as well as noncognitive skill measures partly contain the information in comportment grades. This assertion is reinforced by the fact that grades and GPA together explain a substantial share of the variation in comportment grades (adjusted R^2 of 0.2, see Table A.16). Furthermore, while subject grades and GPA as well as conscientiousness are positively correlated among each other, higher agreeableness is only associated with better comportment grades, while there is no relationship with the other variables. This finding indicates that comportment grades indeed measure personality beyond what can already be inferred from subject grades and GPA.[10]

[10]Grades are found to capture various aspects of personality (Borghans et al. 2016). As an example, more conscientious individuals take assignments in schools more seriously.

Regarding the school-to-work transition, the previous results suggest that comportment grades might not convey relevant information about students' labor market potential beyond what is contained in subject grades and GPA. As is many other countries, subject grades and GPA in Germany are not only based on exam results of students but also contain an assessment of their oral participation in class. Information on students' non-cognitive skills that are potentially relevant for the labor market might be part of this component of the grade. If anything, comportment grades contain additional information about students' agreeableness, a skill that has been found to have little effect on labor market success (Borghans et al. 2008; Heineck and Anger 2010).

Additionally, comportment grades – as implemented in Germany and other countries – rather provide a low-stake incentive to behave better as they typically do not count towards tracking decisions and the promotion to the next grade. It is therefore reasonable to expect that students do not exert much effort to obtain better comportment grades (e.g. Schlosser et al. 2019). As a consequence, both non-cognitive skills as well as academic achievement should not be affected by the reform. This is in line with our findings.

Finally, the biannual release of report cards and comportment grades in Germany may not provide the timely feedback necessary in order to change students' behaviors. For instance, Levitt et al. (2016) show that students no longer respond to performance incentives once rewards are provided with a delay. Moreover, Jalava et al. (2015) demonstrate that giving numerical or letter grades may not be effective to incentivize students, whereas giving symbolic rewards or providing relative rank information could be more effective. Similarly, teachers might provide feedback regarding students' behavior through other means than grades, e.g. pedagogical disciplinary concepts such as reprimands. This substitution effect could only be investigated with data on teaching styles, which is not available in our setting.

7 Conclusion

Exploiting policy variation across German federal states, we document that grading students' comportment in school does not affect students' success in transitioning from school to work. The point estimates' confidence intervals allow us to derive bounds for the population effect of comportment grading on this transition. We can reject that receiving comportment grades is associated with an increase or decrease of more than six percentage points in the probability of successfully transitioning. In line with this finding, non-cognitive skills and academic achievement measured earlier in students' life are not affected either. Our robustness checks further bolster this zero-effect finding: Using alternative estimation strategies, including different sets of control variables, and applying other sample restrictions hardly affects our results. Finally, we explore potential explanations for our results and find that subject grades and GPA partly contain the information in comportment grades, rendering them less informative.

The findings suggest that the arguments of neither proponents nor opponents of comportment grading can be supported by causal evidence. A caveat of the study could be the

specific context of the German comportement grading reforms in the sense that other countries could experience different outcomes. Yet, given that the comportment grading policies are similar in other countries (see Table A.1), we remain confident in the external validity of our results, in the way that they are informative for other countries considering the introduction or abolition of comportment grading.

Considering the costs teachers incur through grading students' comportment has direct implications for policy. A lower bound for the cost of the reform per student per year amounts to approximately $7.[11] The same costs could be used, for instance, to finance virtual coaching programs for students (Oreopoulos et al. 2020) or to run information campaigns aimed at improving student behavior (see Peter et al. 2021, for a related campaign in the German context).

In sum, this paper shows that the introduction of comportment grading does not have an effect on student outcomes. Finding zero effects of educational reforms is not uncommon (e.g. Dale and Krueger 2002; Fryer 2011; Jerrim et al. 2017; Leuven and Løkken 2018; Bird et al. 2021). At the same time, this is highly informative from a policy perspective: It is crucial to know whether much-debated reforms affect student outcomes at all. In this sense, the results presented can shift beliefs about the causal effect of comportment grading reforms (Abadie 2020). Our finding of zero-effects suggests that policy efforts should focus on other domains to increase the efficiency of the education system.

[11] This figure is based on a teacher salary of $88,071 (OECD) per year and assuming that teachers work 40 hours a week and need five minutes per report card and per student, that is, 10 minutes per year, we arrive at an estimated cost of $7.11 per student per year. Given the roughly 11 million pupils in Germany, the annual cost amounts to $78 millions. Note that this estimate is conservative since we assume that only one teacher is involved in conducting the grading and we ignore social security contributions by the employer.

References

Abadie, A. (2005). "Semiparametric Difference-in-Differences Estimators". In: *The Review of Economic Studies* 72.1, pp. 1–19. DOI: 10.1111/0034-6527.00321.

— (2020). "Statistical Nonsignificance in Empirical Economics". In: *American Economic Review: Insights* 2.2, pp. 193–208. DOI: 10.1257/aeri.20190252.

Abadie, A., S. Athey, G. Imbens, and J. Wooldridge (2017). *When Should You Adjust Standard Errors for Clustering?*

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). "Personality Psychology and Economics". In: *Handbook of the Economics of Education*. Vol. 4. Elsevier, pp. 1–181. DOI: 10.1016/B978-0-444-53444-6.00001-8.

Angrist, J. D., P. A. Pathak, and C. R. Walters (2013). "Explaining Charter School Effectiveness". In: *American Economic Journal: Applied Economics* 5.4, pp. 1–27. DOI: 10.1257/app.5.4.1.

Baumert, J. et al. (2002). *Pisa 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012). "The Relationship Between Economic Preferences and Psychological Personality Measures". In: *Annual Review of Economics* 4.1, pp. 453–478. DOI: 10.1146/annurev-economics-080511-110922.

Bird, K. A. et al. (2021). "Nudging at Scale: Experimental Evidence from FAFSA Completion Campaigns". In: *Journal of Economic Behavior & Organization* 183, pp. 105–128. DOI: 10.1016/j.jebo.2020.12.022.

"Education as a Lifelong Process: The German National Educational Panel Study (NEPS)" (2011). In: *Zeitschrift für Erziehungswissenschaft*. Ed. by H.-P. Blossfeld, H. G. Roßbach, and J. von Maurice. Vol. 2. Special Issue 14. VS Verlag für Sozialwissenschaften. DOI: 10.1007/978-3-658-23162-0.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. Ter Weel (2008). "The Economics and Psychology of Personality Traits". In: *Journal of Human Resources* 43.4, pp. 972–1059.

Borghans, L., B. H. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2016). "What Grades and Achievement Tests Measure". In: *Proceedings of the National Academy of Sciences* 113.47, pp. 13354–13359. DOI: 10.1073/pnas.1601135113.

Bowles, S. and H. Gintis (2002). "Schooling in Capitalist America Revisited". In: *Sociology of Education* 75.1, pp. 1–18. DOI: 10.2307/3090251.

Callaway, B. and P. H. C. Sant'Anna (2020). "Difference-in-Differences with Multiple Time Periods". In: *Journal of Econometrics*. DOI: 10.1016/j.jeconom.2020.12.001.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors". In: *The Review of Economics and Statistics* 90.3, pp. 414–427. DOI: 10.1162/rest.90.3.414.

Card, D. (2001). "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems". In: *Econometrica* 69.5, pp. 1127–1160. DOI: 10.1111/1468-0262.00237.

Cheung, C.-k. and S.-c. Llu (2000). "Acculturation, Social Integration and School Achievement among Low-ability Seventh Graders' School Achievement in Hong Kong". In: *International*

Journal of Adolescence and Youth 8.1, pp. 81–108. DOI: 10.1080/02673843.2000.9747843. eprint: https://doi.org/10.1080/02673843.2000.9747843. URL: https://doi.org/10.1080/02673843.2000.9747843.

Close, D. (2009). "Fair Grades". In: *Teaching Philosophy* 32.4, pp. 361–398. DOI: 10.5840/teachphil200932439.

Cunha, F. and J. Heckman (2007). "The Technology of Skill Formation". In: 97.2, p. 17.

Currie, J. M. (1995). *Welfare and the Well-Being of Children*. Harwood Fundamentals of Pure and Applied Economics. Ed. by F. Welch. Harwood Academic Publishers.

Dale, S. B. and A. B. Krueger (2002). "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables*". In: *The Quarterly Journal of Economics* 117.4, pp. 1491–1527. DOI: 10.1162/003355302320935089.

De Chaisemartin, C. and X. D'Haultfœuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". In: *American Economic Review* 110.9, p. 33.

Dobbie, W. and R. G. Fryer (2020). "Charter Schools and Labor Market Outcomes". In: *Journal of Labor Economics* 38.4, pp. 915–957. DOI: 10.1086/706534.

Duckworth, A. L., P. D. Quinn, and E. Tsukayama (2012). "What No Child Left Behind Leaves Behind: The Roles of IQ and Self-Control in Predicting Standardized Achievement Test Scores and Report Card Grades". In: *Journal of Educational Psychology* 104.2, pp. 439–451. DOI: 10.1037/a0026280.

Facchinello, L. (2020). *Short- and Long-Run Effects of Early Grades*. SSRN Scholarly Paper ID 2966571. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.2966571.

Fryer, R. G. (2011). "Financial Incentives and Student Achievement: Evidence from Randomized Trials *". In: *The Quarterly Journal of Economics* 126.4, pp. 1755–1798. DOI: 10.1093/qje/qjr045.

Goebel, J. et al. (2019). "The German Socio-Economic Panel (SOEP)". In: *Jahrbücher für Nationalökonomie und Statistik* 239.2, pp. 345–360. DOI: 10.1515/jbnst-2018-0022.

Goodman-Bacon, A. (2021). "Difference-in-Differences with Variation in Treatment Timing". In: *Journal of Econometrics*. DOI: 10.1016/j.jeconom.2021.03.014.

Hanushek, E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2015). "Returns to Skills around the World: Evidence from PIAAC". In: *European Economic Review* 73, pp. 103–130. DOI: 10.1016/j.euroecorev.2014.10.006.

Heckman, J. J., J. Stixrud, and S. Urzua (2006). "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior". In: *Journal of Labor Economics* 24.3, pp. 411–482.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". In: *The Review of Economic Studies* 64.4, pp. 605–654. DOI: 10.2307/2971733.

Heineck, G. and S. Anger (2010). "The Returns to Cognitive Abilities and Personality Traits in Germany". In: *Labour Economics* 17.3, pp. 535–546. DOI: 10.1016/j.labeco.2009.06.001.

Helbig, M. and R. Nikolai (2015). "Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949". In: p. 383.

Hvidman, U. and H. H. Sievertsen (2019). "High-Stakes Grades and Student Behavior". In: *Journal of Human Resources*, 0718–9620R2. DOI: 10.3368/jhr.56.3.0718-9620R2.

Jalava, N., J. S. Joensen, and E. Pellas (2015). "Grades and Rank: Impacts of Non-Financial Incentives on Test Performance". In: *Journal of Economic Behavior & Organization* 115, pp. 161–196. DOI: 10.1016/j.jebo.2014.12.004.

Jerrim, J., L. Macmillan, J. Micklewright, M. Sawtell, and M. Wiggins (2017). "Does Teaching Children How to Play Cognitively Demanding Games Improve Their Educational Attainment? Evidence from a Randomised Controlled Trial of Chess Instruction in England". In: *Journal of Human Resources*, p. 0516. DOI: 10.3368/jhr.53.4.0516.7952R.

Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014). *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*. Working Paper 20749. National Bureau of Economic Research. DOI: 10.3386/w20749.

Koch, A., J. Nafziger, and H. S. Nielsen (2015). "Behavioral Economics of Education". In: *Journal of Economic Behavior & Organization*. Behavioral Economics of Education 115, pp. 3–17. DOI: 10.1016/j.jebo.2014.09.005.

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk (2020). "The Formation of Prosociality: Causal Evidence on the Role of Social Environment". In: *Journal of Political Economy* 128.2, pp. 434–467. DOI: 10.1086/704386.

Kristoffersen, J. H. G., M. V. Krægpøth, H. S. Nielsen, and M. Simonsen (2015). "Disruptive School Peers and Student Outcomes". In: *Economics of Education Review* 45, pp. 1–13. DOI: 10.1016/j.econedurev.2015.01.004.

Landersø, R. and J. J. Heckman (2017). "The Scandinavian Fantasy: The Sources of Intergenerational Mobility in Denmark and the US". In: *The Scandinavian Journal of Economics* 119.1, pp. 178–230. DOI: 10.1111/sjoe.12219.

Lazear, E. P. (2001). "Educational Production". In: *The Quarterly Journal of Economics* 116.3, pp. 777–803.

Leuven, E. and S. A. Løkken (2018). "Long-Term Impacts of Class Size in Compulsory School". In: *Journal of Human Resources*, p. 0217. DOI: 10.3368/jhr.55.2.0217.8574R2.

Levitt, S. D., J. A. List, S. Neckermann, and S. Sadoff (2016). "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance". In: *American Economic Journal: Economic Policy* 8.4, pp. 183–219. DOI: 10.1257/pol.20130358.

Maynard, R. A. (1977). "The effects of the rural income maintenance experiment on the school performance of children". In: *The American Economic Review* 67.1, pp. 370–375.

Oreopoulos, P., U. Petronijevic, C. Logel, and G. Beattie (2020). "Improving non-academic student outcomes using online and text-message coaching". In: *Journal of Economic Behavior & Organization* 171, pp. 342–360.

Peter, F., C. K. Spiess, and V. Zambre (2021). "Informing students about college: Increasing enrollment using a behavioral intervention?" In: *Journal of Economic Behavior & Organization* 190, pp. 524–549.

Prenzel, M. et al. (2007). *Programme for International Student Assessment 2003 (PISA 2003)Programme for International Student Assessment 2003 (PISA 2003)*. Version 1. IQB - Institute for Educational Quality Improvement. DOI: 10.5159/IQB_PISA_2003_V1.

Prenzel, M. et al. (2010). *Programme for International Student Assessment 2006 (PISA 2006)Programme for International Student Assessment 2006 (PISA 2006)*. Version 1. IQB - Institute for Educational Quality Improvement. DOI: 10.5159/IQB_PISA_2006_V1.

Prenzel, M. et al. (2019). *Programme for International Student Assessment 2012 (PISA 2012)Programme for International Student Assessment 2012 (PISA 2012)*. Version 5. IQB - Institute for Educational Quality Improvement. DOI: 10.5159/IQB_PISA_2012_V5.

Protsch, P. and H. Solga (2015). "How Employers Use Signals of Cognitive and Noncognitive Skills at Labour Market Entry: Insights from Field Experiments". In: *European Sociological Review* 31.5, pp. 521–532. DOI: 10.1093/esr/jcv056.

Resnjanskij, S., J. Ruhose, S. Wiederhold, and L. Woessmann (2021). "Can Mentoring Alleviate Family Disadvantage in Adolescence? A Field Experiment to Improve Labor-Market Prospects". In: *CESifo Working Paper no. 8870*, p. 130.

Roodman, D., M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb (2019). "Fast and Wild: Bootstrap Inference in Stata Using Boottest". In: *The Stata Journal: Promoting communications on statistics and Stata* 19.1, pp. 4–60. DOI: 10.1177/1536867X19830877.

Ryan, P. (2001). "The School-to-Work Transition: A Cross-National Perspective". In: *Journal of Economic Literature* 39.1, pp. 34–92.

Sachse, K. A. et al. (2012). "IQB-Ländervergleich 2008/2009". In: DOI: 10.18452/3126.

Schipolowski, S., N. Haag, F. Milles, S. Pietz, and P. Stanat (2018). *IQB-Bildungstrend 2015*. Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. DOI: 10.18452/19997.

Schlosser, A., Z. Neeman, and Y. Attali (2019). "Differential Performance in High Versus Low Stakes Tests: Evidence from the Gre Test". In: *The Economic Journal* 129.623, pp. 2916–2948. DOI: 10.1093/ej/uez015.

Sun, L. and S. Abraham (2020). "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects". In: *Journal of Econometrics*. DOI: 10.1016/j.jeconom.2020.09.006.

Tan, B. (2020). *Grades as Noisy Signals*. SSRN Scholarly Paper ID 3544407. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.3544407.

Tyre, P. (2010). "A's for Good Behavior". In: *The New York Times. Week in Review*.

Urabe, M. (2006). "Cultural Barriers in Educational Evaluation: A Comparative Study on School Report Cards in Japan and Germany." In: *International Education Journal* 7.3, pp. 273–283.

Zimmermann, K. F. (2013). "Youth Unemployment and Vocational Training". In: *Foundations and Trends® in Microeconomics* 9.1-2, pp. 1–157. DOI: 10.1561/0700000058.

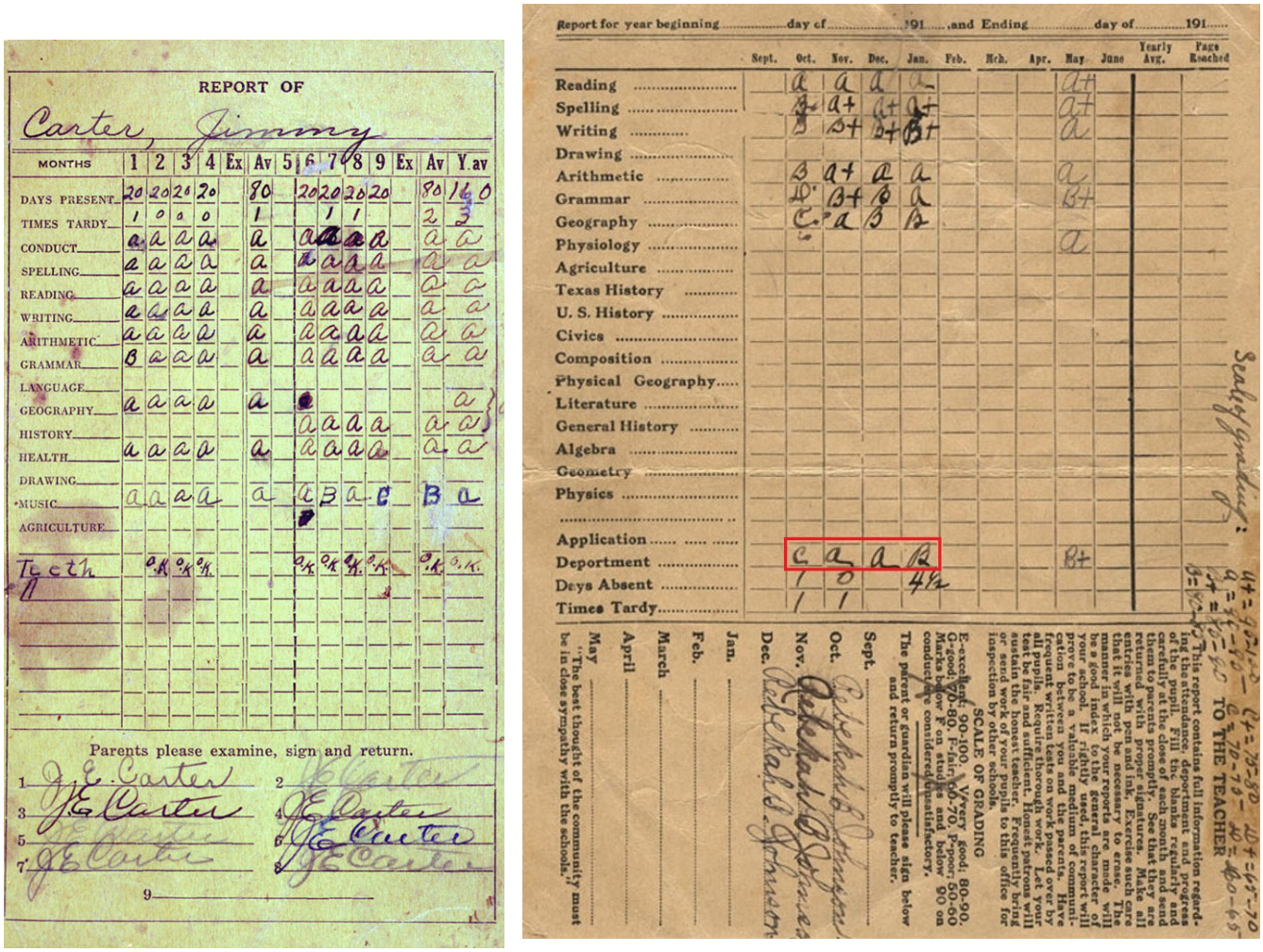
Appendix

(For Online Publication)

A Appendix

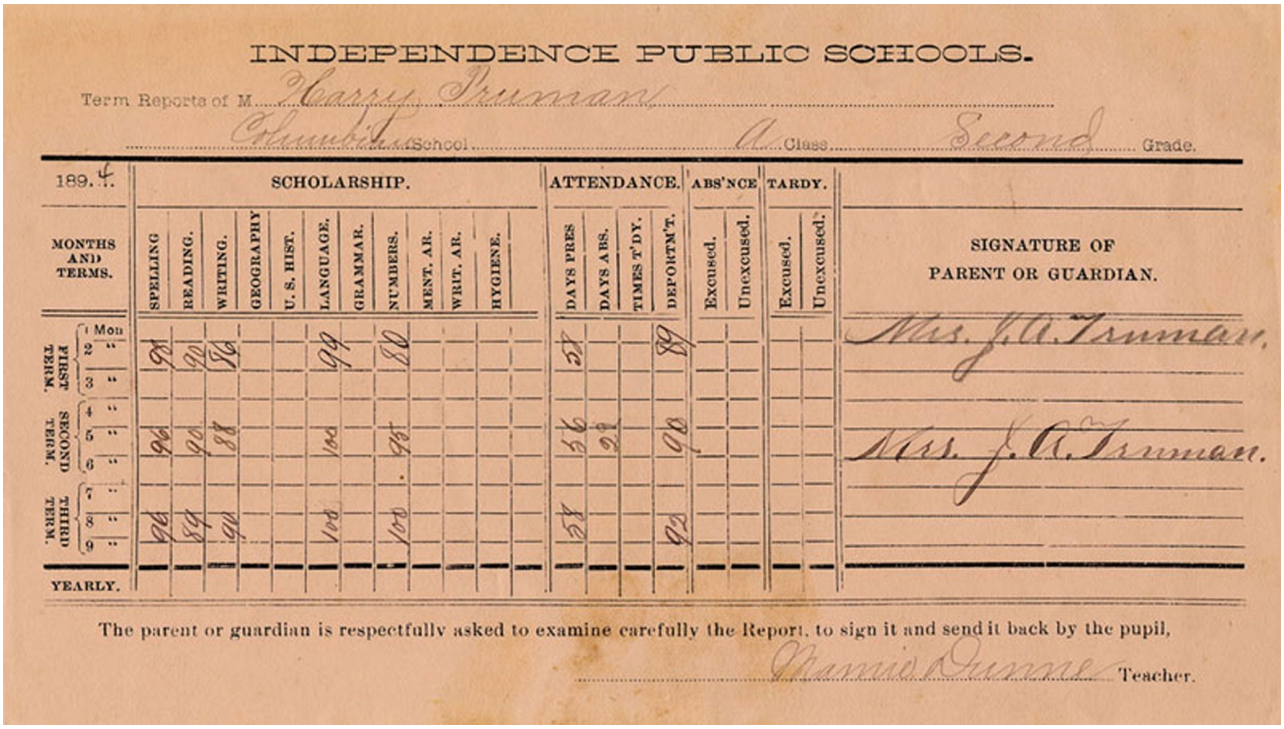
A.1 Policy background

FIGURE A.1. Report cards by Jimmy Carter (left) and Lyndon B. Johnson (right)



Notes: Carter (*1924): sixth-grade report card, includes grade for "conduct" (third item). Johnson (1908-1973): third-grade report card, includes grade for "deportment" (third but last item, synonym for "comportment"). *Source:* Carter Library and Johnson City Foundation.

FIGURE A.2. Harry S. Truman's school report card



INDEPENDENCE PUBLIC SCHOOLS.

Term Reports of M. Harry Truman

Church School A Class Second Grade.

189.4.

MONTHS AND TERMS.	SCHOLARSHIP. SPELLING	READING.	WRITING.	GEOGRAPHY	U. S. HIST.	LANGUAGE.	GRAMMAR.	NUMBERS.	MENT. AR.	WRIT. AR.	HYGIENE.	ATTENDANCE. DAYS PRES	DAYS ABS.	TIMES T'DY.	DEPORTM'T.	ABS'NCE Excused.	ABS'NCE Unexcused.	TARDY. Excused.	TARDY. Unexcused.	SIGNATURE OF PARENT OR GUARDIAN.
FIRST TERM. 1 Mon	95	90	90			90		80				54			90					Mrs. J. A. Truman.
2 "																				
3 "																				
SECOND TERM. 4 "	96	90	88			100		96				56	2		90					Mrs. J. A. Truman.
5 "																				
6 "																				
THIRD TERM. 7 "	96	99	90			100		100				54			95					
8 "																				
9 "																				
YEARLY.																				

The parent or guardian is respectfully asked to examine carefully the Report, to sign it and send it back by the pupil,

Minnie Dunne Teacher.

Notes: Truman (1884-1972): second grade report card, includes grade for "deportment" (synonym for "comportment", last item within "attendandance" category).
Source: Harry Truman Library.

FIGURE A.3. Stylized Overview of the German school system

Years of schooling				Age	
	Vocational training/ further schooling/ labor market	Vocational training/ technical college/ labor market	University entrance	19	Secondary school
13			Academic track (Gymnasium)	18	
12				17	
11				16	
10		Middle school (Realschule)		15	
9	Basic school (Hauptschule)			14	
8				13	
7				12	
6				11	
5				10	
4	Primary school (Grundschule)			9	Primary school
3				8	
2				7	
1				6	

Source: Own representation based on Helbig and Nikolai (2015).

TABLE A.1. Grading of social and work behavior in selected European countries

Country	Grading of work and social behavior
Austria	Behavioral grades exist for all school types and grade behavior in the middle school years. In 2014, parents' associations tried to abolish these grades (*Die Presse*, Sept. 18, 2014).
Czech Republic	Students' behavior is assessed as (1) very good, (2) satisfactory, or (3) unsatisfactory.
Denmark	Until 2013, students received grades on the orderliness/organization/neatness of their written exams in Danish and mathematics (Landersø and Heckman 2017).
France	A grade for comportment ("note de vie scolaire") was abolished in 2014. The grade considered punctuality, respect for rules, participation in the school's social life, and attaining a road safety education certificate. It was abolished following criticism regarding its subjectivity (*Avis du Conseil supérieur des programmes sur la note de vie scolaire*, Nov. 21, 2013).
Greece	At the end of each quarter and when grades have been finalized and recorded, parents receive an individual progress report and are informed about student performance, diligence, attendance and behavior.
Hungary	Behavior and effort/diligence are evaluated on a four-grade scale: exemplary (5), good (4), varying (3), or poor (2).
Italy	The assessment of students' conduct refers to the development of citizenship competences, in accordance with what is established by each school's regulations and the 'Joint responsibility agreement' signed by students and parents. Students with a mark below 6/10 in conduct cannot progress to the following grade.
Norway	The students are assessed in conduct.
Poland	A grade for behavior exists and does not influence the promotion to a higher grade or graduation. Yet, receiving an inadmissible grade for behavior in two consecutive years student cannot be promoted to the next grade or finish school .
Sweden	A proposal to reintroduce comportment grading in schools caused a long debate in 2019. A majority of members of the Riskdag upheld the proposal with the aim of reducing disruptive behavior in schools. The Swedish Teachers' Association is critical and fears that grading conduct might even be counterproductive (*Göteborgs-Posten*, Apr. 2, 2019).
Switzerland	Social conduct and attitude to work may be assessed depending on canton. In 2016, the canton of Zurich also decreed that these grades count towards students' promotions to high-track schools (*Tages-Anzeiger*, Dec. 19, 2016).

Source: European Commission (2021). *Eurydice: Better knowledge for better education policies.* National Education Systems. Individual country reports retrieved from `https://eacea.ec.europa.eu/national-policies/eurydice/national-description_en` (as of July 5, 2021).

TABLE A.2. Teacher guidelines for the evaluation of behavior (excerpt) in the state of Baden-Wuerttemberg

Criterion	Commendable behavior	Gross misconduct
General conduct	Polite, friendly, controlled, calm, placid	Naughty, defiant, malicious, uncontrolled, quick-tempered
Camaraderie	Companionable, helpful, compassionate, compatible	Non-companionable, ruthless, unbearable, spiteful
Honesty	Sincere, honest, candid	Insincere, dishonest, lying
Restraint	Modest, restrained, discreet	Immodest, boastful, presumptuous, arrogant
Work effort	Takes over community tasks willingly	Refuses to take over community tasks
Acceptance of rules	Recognition of principles of order, sense of order, willingness to comply, reliable, punctual, regular participation in class, compliant	Negligently or intentionally violates principles of order, disorganized, belligerent, unreliable, frequently arrives late, frequently misses class without sufficient justification, continually disrupts class

Notes: This table was suggested as a teacher aide to assess student behavior in the state of Baden-Wuerttemberg. Most commonly, students receive the grade "good". If the student's behavior is particularly cooperative, the grade "very good" might be assigned. If the student's behavior frequently meets the description given by the columns *Misconduct* (not shown in this excerpt) or *Gross misconduct*, the student might receive a "satisfactory" or "insufficient" grade.
Source: Hausmann, Johanna (2010). *Beeinflussungstendenzen bei Kopfnoten: welche Faktoren fließen in die Noten unserer Kinder ein?* Hamburg, Diplomica.

TABLE A.3. Teacher guidelines for the evaluation of behavior in the state of Saxony

Criterion	Behaviors to be considered
Order	Care, punctuality, reliability, compliance with rules, having teaching materials ready
Cooperation	Initiative, willingness to cooperate, ability to work in a team, independence, creativity, responsibility
Conduct	Attentiveness, helpfulness, civic courage and appropriate handling of conflicts, considerateness, tolerance, sociability, self-perception
Diligence	Willingness to learn, determination, endurance, regularity in fulfilling task.

Notes: This table represents the concept of comportment grading in the state of Saxony. Students will be assigned a grade between *1* ("exemplary") and *5* ("insufficient").
Source: Bohl, Thorsten (2010). "Aktuelle Regelungen zur Leistungsbeurteilung und zu Zeugnissen an deutschen Sekundarschulen". In: *Zeitschrift für Pädagogik* 49.4, p. 558.

A.2 Treatment Effect Estimands

Following the expositon by Callaway and Sant'Anna (2020), this section details how the ATTs we report in columns 1 and 2 of Table 1 are obtained. Let G_i be the time period when unit i becomes treated and $t = 1, \ldots, T$ denote time periods. $Y_{it}(g)$ is unit i's potential outcome in time period t if they become treated in period g.

Under (conditional) parallel trends and for all $t \geqslant g$, they show that the following group- and period-specific average treatment effect on the treated is identified using modified differences in expectations

$$\text{ATT}(g,t) := \text{E}(Y_t(g) - Y_t(0)|G = g).$$

Effects with $t < g$ can be used for pre-testing. In the canonical 2x2 design, $\text{ATT}(g = 2, t = 2)$ is the estimand of interest, corresponding to the instantaneous treatment effect for the group receiving treatment in the second period. In general staggered designs with many more ATTs, aggregates of these can be used to get an idea of the overall treatment effect.

In our setup, units correspond to German federal states, i.e. $i \in \{\text{Brandenburg, Bremen, Saxony-Anhalt, Northrhine-Westphalia}\}$. We restrict the sample period to $t = 1996, \ldots, 2006$. There are two treatment groups receiving treatment in 2001 and 2003, respectively ($g \in \{2001, 2003\}$). This means that we have 10 ATTs for each group, 4 (6) pretreatment and 6 (4) post-treatment effects for the group with $g = 2001$ ($g = 2003$). In a first step, we average over post-treatment effects for each group:

$$\begin{aligned}\theta_S(g = 2001) &:= \frac{1}{2006 - 2001 + 1} \sum_{t=1997}^{2006} \mathbb{1}\{2001 \leqslant t\} \, \text{ATT}(g = 2001, t) \\ &= \frac{1}{6} \sum_{t=2001}^{2006} \text{ATT}(g = 2001, t)\end{aligned}$$

$$\theta_S(g = 2003) := \frac{1}{4} \sum_{t=2003}^{2006} \text{ATT}(g = 2003, t).$$

To arrive at a single measure that resembles a multi-group multi-period extension of the ATT in the 2x2 design, we further average across treatment groups to obtain

$$\begin{aligned}\text{ATT} &:= \sum_{g=1997}^{2006} \theta_S(g) \cdot \Pr(G = g) \\ &= \theta_S(g = 2001) \cdot \Pr(G = 2001) + \theta_S(g = 2003) \cdot \Pr(G = 2003).\end{aligned} \tag{2}$$

Table 1 shows estimates of the estimand in equation 2 under different scenarios.

A.3 Descriptive statistics

Table A.4. Descriptive statistics Microcensus

	Mean	SD	Min	Max	N
Successful school-to-work transition	0.86	0.35	0	1	22,895
Successful school-to-work transition, strict	0.77	0.42	0	1	22,895
Comportment grading (Enrollment)	0.04	0.19	0	1	22,895
Comportment grading (4th grade)	0.15	0.36	0	1	22,895
Female	0.41	0.49	0	1	22,895
First-generation migrant	0.28	0.45	0	1	22,895

Notes: Sample includes students from the federal states of Bremen, Brandenburg, Saxony-Anhalt, North Rhine-Westphalia. Comportment group indicators are defined as whether there is comportment grading when the student is enrolled or in 4th grade, respectively. Success strict is an alternative measure of successful school-to-work transition, excluding employed individuals who have not earned any vocational qualification prior to their employment.
Sources: Microcensus waves 2011–2016.

Table A.5. Descriptive statistics SOEP

	Mean	SD	Min	Max	N
Trust	−0.00	1.00	−2.50	2.88	2,121
Conscientiousness	−0.00	1.00	−3.38	1.66	2,121
Agreeableness	−0.00	1.00	−3.84	1.66	2,121
Comportment grading (enrollment)	0.11	0.31	0	1	2,121
Comportment grading (4th grade)	0.40	0.49	0	1	2,121
Female	0.49	0.50	0	1	2,121
First-generation migrant	0.35	0.48	0	1	2,121

Notes: Sample includes students from the federal states of Bremen, Brandenburg, Saxony-Anhalt and North Rhine-Westphalia. Comportment group indicators are defined as whether there is comportment grading when the student is enrolled or in 4th grade, respectively.
Sources: SOEP-Core v36.

TABLE A.6. Descriptive statistics nationwide student assessments

	Mean	SD	Min	Max	N
Reading skills	0.01	1.00	−4.91	5.21	42,415
Academic track school attendance	0.34	0.47	0.00	1.00	42,415
Comportment grading (enrolment)	0.14	0.35	0.00	1.00	42,415
Comportment grading (4th grade)	0.26	0.44	0.00	1.00	42,415
Female	0.49	0.50	0.00	1.00	42,415
First generation migrant	0.12	0.32	0.00	1.00	42,415
Age (months)	187.39	6.56	148.96	230.01	42,415
Low SES	0.30	0.46	0.00	1.00	42,415
School size (students)	630.29	303.82	32.00	1825.00	33,272
Public school	0.95	0.22	0.00	1.00	33,272
Town (<15,000 inhabitants)	0.20	0.40	0.00	1.00	33,272
Large town (15,000-100,000 inhabitants)	0.42	0.49	0.00	1.00	33,272
City (100,000-1,000,00 inhabitants)	0.39	0.49	0.00	1.00	33,272
Observations	42,415				

Notes: Sample includes repeated cross-sections of students in ninth grade from the federal states of Bremen, Brandenburg, Saxony-Anhalt, North Rhine-Westphalia. Low SES defined as parents having obtained the education level ISCED Level 3B/C at most.
Sources: PISA 2000, PISA 2003, PISA 2006, IQB-LV 2008-9 (v2), PISA 2012, IQB-BT 2015 (v5).

TABLE A.7. Descriptive statistics NEPS

	Mean	SD	Min	Max	N
German Grade	4.31	0.79	2.00	6.00	812
Math Grade	4.15	0.95	2.00	6.00	812
GPA	4.51	0.64	3.00	6.00	812
Standardized values of Conscientiousness	0.02	1.00	-2.77	2.08	812
Standardized values of Agreeableness	-0.00	1.00	-3.24	2.57	812
Comportment Grade	4.95	0.71	1.00	6.00	812
Observations	812				

Notes: Sample includes German students that were interviewed in fifth grade in autumn/ winter 2010 for the first time and re-surveyed in an approximately annual interval until autumn/ winter 2018. Comportment grades are those that are available in a student's graduation report. The subject grades represent half-year grades and, like the GPA, are taken from the graduation year. Subject grades and GPA are rounded to integers. Non-cognitive skills are standardised and taken from the survey in autumn/winter 2018. All variables assume that higher values are better. If students have taken more than one degree, the first one is included in the sample. Excluding students with incomplete information leads to a sample size of 812.
Sources: NEPS SC3 10.0.0

A.4 Robustness checks non-cognitive skills (Socio-Economic Panel)

A.4.1 Without controls

TABLE A.8. Effect of grading comportment on non-cognitive skills - without controls

	Trust	Conscientiousness	Agreeableness
ATT	−0.0369	0.0108	−0.0369
	[−0.1683, 0.0946]	[−0.0758, 0.0975]	[−0.1102, 0.0364]
WCB p-val.	0.6446	0.6957	0.6667
Adj.R.squared	0.0072	0.0754	0.0186
Observations	2,121	2,121	2,121
Std.Error	Cluster	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, cohort, and survey year fixed effects. All outcomes are standardized to have mean zero and unit standard deviation. Specifications do not include further covariates. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values use weights from Webb's distribution and rely on 999 iterations (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: SOEP-Core v36.

A.4.2 Treatment assignment in grade 4

TABLE A.9. Effect of grading comportment on non-cognitive skills - Treatment assignment in grade 4

	Trust	Conscientiousness	Agreeableness
ATT	0.0324	0.0054	0.1325
	[-0.0736, 0.1383]	[-0.1048, 0.1156]	[-0.0817, 0.3467]
WCB p-val.	0.9229	0.8629	0.6396
Adj.R.squared	0.0228	0.1114	0.0284
Observations	2,121	2,121	2,121
Std.Error	Cluster	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, cohort, and survey year fixed effects. All outcomes are standardized to have mean zero and unit standard deviation. Controls include student sex and a dummy for migration background. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values use weights from Webb's distribution and rely on 999 iterations (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: SOEP-Core v36.

A.5 Robustness checks student achievement (nationwide student assessments)

TABLE A.10. Effect of grading behavior on academic achievement - without controls

	(1) Reading Skills	(2) Academic Track School Attendance
CG_{st}	0.2238 [−0.2325, 0.6801]	0.0001 [−0.1874, 0.1877]
Outcome mean	0.01	0.34
WCB P-Value	0.388	0.992
R-squared	0.019	0.023
Observations	42,415	42,415
St. Error	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, cohort, and survey year fixed effects. Specifications do not include further covariates. Reading Skills are standardized to have mean zero and unit standard deviation while Gymnasium Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values rely on 999 bootstrap iterations using weights from Webb's distribution (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: PISA 2000, PISA 2003, PISA 2006, IQB-LV 2008-9 (v2), PISA 2012, IQB-BT 2015 (v5).

TABLE A.11. Effect of grading behavior on academic achievement - extended individual-level controls

	(1) Reading Skills	(2) Academic Track School Attendance
CG_{st}	0.0917	−0.0429
	[−0.3275, 0.5108]	[−0.2153, 0.1295]
Outcome mean	0.01	0.34
WCB P-Value	0.554	0.570
R-squared	0.181	0.128
Observations	42,415	42,415
St. Error	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, cohort, and survey year fixed effects. Controls include student sex, migration background, age in months, and an indicator for parental SES. Reading Skills are standardized to have mean zero and unit standard deviation while Gymnasium Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values rely on 999 bootstrap iterations using weights from Webb's distribution (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: PISA 2000, PISA 2003, PISA 2006, IQB-LV 2008-9 (v2), PISA 2012, IQB-BT 2015 (v5).

TABLE A.12. Effect of grading behavior on academic achievement - school-level controls

	(1) Reading Skills	(2) Academic Track School Attendance
CG_{st}	0.2601	0.0153
	[−0.1670,0.6873]	[−0.1769,0.2074]
Outcome mean	0.01	0.34
WCB P-Value	0.333	0.864
R-squared	0.051	0.027
Observations	33,272	33,272
St. Error	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, cohort, and survey year fixed effects. Controls include student sex, migration background, age in months, and an indicator for parental SES. Additional school-level controls include school size, school type (public vs. private), and city size. Reading Skills are standardized to have mean zero and unit standard deviation while Gymnasium Attendance is an indicator variable. Sample size lower due to limited availability of school-level controls. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values rely on 999 bootstrap iterations using weights from Webb's distribution (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: PISA 2000, PISA 2003, PISA 2006, IQB-LV 2008-9 (v2), PISA 2012, IQB-BT 2015 (v5).

TABLE A.13. Effect of grading behavior on academic achievement - plausible values

	(1) Reading PV1	(2) Reading PV2	(3) Reading PV3	(4) Reading PV4	(5) Reading PV5
CG_{st}	0.2238	0.2311	0.2101	0.2134	0.2183
	[−0.2039, 0.6515]	[−0.1703, 0.6326]	[−0.1936, 0.6137]	[−0.1973, 0.6241]	[−0.1982, 0.6348]
Outcome mean	0.01	0.01	0.01	0.01	0.01
WCB P-Value	0.364	0.332	0.364	0.366	0.365
R-squared	0.043	0.044	0.044	0.042	0.043
Observations	42,415	42,415	42,415	42,415	42,415
St. Error	Cluster	Cluster	Cluster	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates for different plausible values (PV) from the probability distribution of a student's reading skills. PV1 is the plausible value used in our baseline analyses. All specifications include federal state, cohort, and survey year fixed effects. Controls include student sex and migration background. Reading Skills are standardized to have mean zero and unit standard deviation. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values rely on 999 bootstrap iterations using weights from Webb's distribution (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: PISA 2000, PISA 2003, PISA 2006, IQB-LV 2008-9 (v2), PISA 2012, IQB-BT 2015 (v5).

TABLE A.14. Effect of grading behavior on academic achievement - Treatment defined in 4th grade instead of enrollment

	(1) Reading Skills	(2) Gymnasium Attendance
CG_{st}	0.1775	0.0359
	[0.0170, 0.3379]	[−0.0599, 0.1317]
Outcome mean	0.01	0.34
WCB P-Value	0.204	0.377
R-squared	0.043	0.029
Observations	42,415	42,415
St. Error	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state, time, and survey year fixed effects. Controls include student sex and migration background. Reading Skills are standardized to have mean zero and unit standard deviation while Gymnasium Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level; wild cluster bootstrap p-values rely on 999 bootstrap iterations using weights from Webb's distribution (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: PISA 2000, PISA 2003, PISA 2006, IQB-LV 2008-9 (v2), PISA 2012, IQB-BT 2015 (v5).

A.6 Supplementary Analyses

Table A.15. Correlation among grades and non-cognitive skills

	Comportment Grade	Math Grade	German Grade	GPA	Conscientiousness	Agreeableness
Comportment Grade	1.000					
Math Grade	0.220	1.000				
	(0.000)					
German Grade	0.333	0.250	1.000			
	(0.000)	(0.000)				
GPA	0.411	0.515	0.520	1.000		
	(0.000)	(0.000)	(0.000)			
Conscientiousness	0.212	0.050	0.097	0.098	1.000	
	(0.000)	(0.145)	(0.004)	(0.005)		
Agreeableness	0.099	-0.036	-0.006	0.009	0.124	1.000
	(0.004)	(0.287)	(0.865)	(0.804)	(0.000)	

Notes: Correlation matrix of comportment grades, subject grades, GPA, and non-cognitive skills. Comportment grades are taken from a students' graduation report. The subject grades represent half-year grades and, like GPA, are taken from the final school year. Subject grades and GPA are rounded to integers. Non-cognitive skills are taken from the survey in autumn/winter 2018. This matrix is based on the assumption that for all variables a higher value is considered better. P-values in parentheses. Excluding all individuals with incomplete information leads to a sample size of 812.
Sources: NEPS SC3 10.0.0.

Table A.16. Explanatory power of subject grades regarding comportment grades

	Comportment Grade							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
German Grade	0.311	0.292					0.168	0.145
	(11.725)	(10.443)					(5.679)	(4.776)
Math Grade			0.168	0.176			0.006	0.017
			(7.406)	(7.871)			(0.251)	(0.718)
GPA					0.463	0.446	0.352	0.345
					(14.709)	(14.024)	(8.683)	(8.519)
Constant	3.567	3.595	4.203	4.057	2.828	2.824	2.585	2.606
	(30.664)	(9.036)	(43.180)	(9.553)	(19.499)	(6.711)	(17.086)	(5.469)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
R-squared	0.11	0.12	0.05	0.08	0.17	0.18	0.19	0.20
Adj. R-squared	0.11	0.11	0.05	0.08	0.17	0.18	0.19	0.20
Observations	1320	1320	1320	1320	1240	1240	1240	1240

Notes: Each column presents separate OLS coefficient estimates with *t*-statistics in brackets. Controls include student age and gender. Comportment Grade, German Grade, Math Grade, and GPA are rounded to take on integers from 1 (worst) to 6 (best).
Sources: NEPS SC3 10.0.0.