CESifo WORKING PAPERS

9275
2021

Original Version: August 2021
This Version: March 2023

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 skills. Analyzing administrative data, household surveys, and nationwide student assessments, we show that comportment grading does not meaningfully affect these outcomes, and reject large effect sizes. Teachers likely offset potential effects by using alternative methods for providing student feedback and managing classroom discipline in place of comportment grading.

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
Reutlingen University
ESB Business School
Reutlingen / Germany
larissa.zierow@reutlingen-university.de

*corresponding author

March 15, 2023
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, Ludger Woessmann and participants of the EffEE PhD Workshop on Causal Analyses of School Reforms and the annual meeting of the Verein für Socialpolitik's standing committee in education economics 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, Lisa Eitinger, 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 grade to students' 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 2004) until their shift towards objective measures of educational output and standards-based grading (Tyre 2010; Duckworth et al. 2012).[1] Receiving feedback on their comportment in the classroom might encourage students to behave better. However, unlike the "No Excuses" approach adopted by some US charter schools (Angrist et al. 2013), the comportment grading policy focuses solely on grades. Still, it may empower teachers to address disruptive in-class behaviors in a manner that is also noticeable to parents.

Empirical evidence on the causal effects of comportment grading does not exist, and the theoretical case for comportment grading is ambiguous. The merits of comportment grading are thus often debated based on gut-feelings rather than on empirical evidence. Proponents argue that the threat of receiving poor comportment grades might incentivize students to behave better in class and also help them with their transition into the labor market to signal their social skills to employers (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 feeling they are being treated unfairly when receiving them, which might have demotivating effects on their learning as well as their behavior (Close 2009).

This paper estimates the causal effect of comportment grading on student outcomes exploiting 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 run two-way fixed effects (TWFE) regressions. 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 (2021). We establish that heterogeneity in treatment effects are unlikely to matter in our application.

Our primary objective in this study is to investigate the effect of comportment grading on the probability of school leavers' employment or training status. To this end, we use administrative data from the German Mikrozensus. Our results indicate that the effect of comportment grading on the probability of being employed or in training after school is not distinguishable from

[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.

zero. To better understand the mechanisms behind this outcome, we also analyze students' non-cognitive skill formation and academic achievement as two channels through which an effect on the school-to-work transitions could operate. Both analyses using representative household surveys and nationally standardized student assessments yield a concordant picture, showing that the reform has no significant impact on either non-cognitive skills or academic achievement. For all of the outcomes under consideration, we conduct benchmarking exercises to demonstrate that non-rejectable effect sizes are mostly small in size.

Investigating potential explanations for our results, we draw upon three further sources of data. First, we conduct descriptive analyses using data on students' comportment and subject grades as well as measures of their non-cognitive skills. We find that the informational value-added of comportment grades about students' non-cognitive skills is limited once subject grades are known. Therefore, it becomes unlikely that comportment grades can have an impact through their ability to signal significant non-cognitive skills. Second, we draw upon cross-sectional variation in the PISA data to show that comportment grading is not linked to a more conducive classroom environment for learning. Better classroom behavior usually leads to higher academic achievement, but our finding shows that comportment grades do not correlate with classroom behavior. Hence, it is unlikely that they have an impact on academic achievement through this channel. Third, we ran a survey among teachers in Germany. The results show that teachers substitute comportment grades with alternative pedagogical disciplinary measures in the absence of comportment grading. Teachers therefore can maintain classroom discipline and students still receive feedback on their behavior regardless of whether comportment grades are provided. This finding offers a possible explanation for a lack of effects on students' skill formation.

Our paper contributes to three strands of literature that investigate inputs to producing student outcomes in the framework of an education production function (e.g., Hanushek 2020). 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 et al. 2013), we are the first to focus on the effect of grading students' behavior in school, a widely implemented policy. The closest paper to ours in this domain is Protsch and Solga (2015). They use a correspondence study design and find that behavioral reports of students are even more important than GPA for callback rates in the first round of the application process in the German apprenticeship market. However, their contribution is limited to investigating whether there is a direct signalling effect of comportment grades. By contrast, our analysis considers students' academic achievement and non-cognitive skill formation, both of which can have effects on the school-to-work transition.

There are several papers studying the relationship between grades and labor market outcomes. A higher GPA in tertiary education is generally found to increase earnings both in a variety of empirical settings and correspondence studies (Jones and Jackson 1990; Freier et al. 2015; Feng and Graetz 2017; Piopiunik et al. 2020; Tan 2022). Recently, Hansen et al. (2023) have shown that the GPA-related earnings premium quickly fades after labor market entry. By

contrast, the counterfactual scenarios in Facchinello (2020) are whether or not students receive grades at all, and therefore more similar to our setting. The author 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. In contrast to that literature, 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. 2022). There is also evidence for a robust link between these abilities and labor market outcomes (Heckman et al. 2006; Almlund et al. 2011; Deming 2017; Deming and Kahn 2018), suggesting that an effect on the school-to-work transition could be mediated by improved non-cognitive skills. 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 favour of comportment grades. In Germany, the requirements for good comportment grades include being companionable, diligent, and honest (see Tables A.3 and A.2 for more criteria). 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 performance on standardized tests and academic achievement more broadly. Both are associated with better labor market outcomes (Hanushek et al. 2015; Card 2001), and thus, they constitute mechanisms through which students' school-to-work transition might be affected. Comportment grades might increase academic achievement by enabling teachers to sanction disruptive behaviors, potentially incentivizing students to behave better in class.[2] Since less disruptive classrooms enhance academic achievement, comportment grading might benefit 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 and depicts our theoretical considerations. 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, Section 7 presents an analysis of the costs comportment grades cause, and Section 8 concludes.

[2]If comportment grades indeed measure behavior, there could be another way how they enhance academic achievement. Ferman and Fontes (2022) show that teachers inflate grades of well-behaved students.

2 Institutional Setting and Theoretical Considerations

2.1 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*). While academic track schools prepare students for studying at university, the other two tracks prepare students for entering the labor market through vocational training.

Evaluation of students' comportment is a common practice in these various school types, where students' biannual report card includes grades on working habits and social behavior in addition to subject-specific grades.[3] Yet, comportment grades have not always been uniformly used across all states and school types. When grading work and social behavior, teachers typically consider students' camaraderie or willingness to cooperate, diligence or work effort, orderliness, and honesty.[4] 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. The way in which these grades are presented in the report card can vary from state to state, however. As described in Helbig and Nikolai (2015), some states utilize a five- or six-point scale similar to the normal grading system, while others make standardized written statements about the work and social behavior of students. It is worth noting that the legal basis for grading in the German federal states is primarily criterion-referenced, with individual or collective reference norms rarely used. Under this system, grades reflect the extent to which a student has met the defined requirements, rather than how they compare to their peers or their own previous performance over a certain period of time (Kostorz 2016). As a result, a grading on the curve approach is generally not provided for by the legal framework in Germany, and it is unlikely to be applied when teachers assign comportment 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 labor market entrance for students without tertiary-level education. Protsch and Solga (2015) demonstrate that employers may value comportment grades even more than regular subject grades.

Comportment grades have a long-standing tradition in German schools, dating back to the inclusion of such grades in scholarship certificates for underprivileged families. Following

[3] Comportment grades are typically not utilized during the last two years of schooling in academic track schools. 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.

[4] 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.

the Second World War, all German states had incorporated comportment grades into their curriculum (Arnold and Vollstädt 2001). In the 1970s, public debates on the potential effects of comportment grading in schools led some West German states to dismiss it (Helbig and Nikolai 2015). Prior to reunification, comportment grades were the norm in East Germany, but they were later abolished in several states. We examine the second wave of reforms in East and West German states, which reintroduced comportment grading in the early 2000s. By 2007, all German states had reintroduced comportment grades.

One factor that led to the reintroduction of comportment grades was the increasing pressure from companies and organizations. For instance, Bremen, one of the states examined in this study, implemented comportment grades in response to the success of a project conducted in schools across the state with the help of companies such as DaimlerChrysler.[5] Additionally, North Rhine-Westphalia implemented comportment grades as a recognition of the increasing importance of ”soft skills” in the professional environment.

Figure 1 illustrates the adoption of comportment grading in four federal states, namely Bremen (introduced in 2001), Brandenburg (2001), Saxony-Anhalt (2003), and North Rhine-Westphalia (2007), during the period of study (1996 to 2007). It is noteworthy that this policy was implemented by governments of different political affiliations, including center-left and center-right ones, which suggests a non-partisan approach to the issue.

Furthermore, at the federal level, the trend towards reintroducing comportment grades is evident. In 2002, the German parliament established a commission of inquiry on the future of civic engagement, which emphasized the significance of fostering civic engagement in childhood and adolescence. In this context, the use of report cards as a means of recognizing and promoting essential civic traits was explicitly recommended (Enquete-Kommission „Zukunft des Bürgerschaftlichen Engagements“ des Deutschen Bundestages 2002).

Despite these developments, the issue of comportment grading remains contentious. After introducing this grading system in 2007, the new center-left coalition in North Rhine-Westphalia abolished it again in December 2010, following public demonstrations against it. Conversely, in 2013, Mecklenburg–Western Pomerania switched from a written assessment to a classic German grading system using a six-point scale ranging from 1 to 6 for comportment grades, while also making it mandatory for primary school children. In Bavaria, comportment grades in primary schools became compulsory from 2008 onwards.

[5]The project involved the development and evaluation of appropriate assessment methods for working habits and social behavior over a period of more than two years. The implementation of the project was based on various factors, including the results of workshops and surveys. Additionally, the regulation was informed by associated school research projects on key qualifications and assessment methods, as well as the dialogue with representatives from the business sector, and the example of corresponding forms of assessment of work and social behavior in the training system of DaimlerChrysler. The regulation also took into account similar regulations and documented experiences in other federal states, especially in Niedersachsen, Brandenburg, and Thüringen, as well as the evaluation of relevant assessment methods used in primary schools and comprehensive schools in Bremen. Finally, the proposed regulation underwent a ”trial evaluation” of the presented drafts. Information available at:https://www.bildung.bremen.de/sixcms/media.php/13/l129.pdf, last access on March 12, 2023.

2.2 Theoretical Considerations

From a theoretical perspective, comportment grading in school could affect student outcomes via the following channels.

(i) If students receive feedback on non-cognitive dimensions of their skills, they may be motivated to invest into these skills in order to get more positive feedback. 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 2021). Being companionable, diligent, and honest, for example, are among the criteria teachers ought to consider when grading comportment in Germany (see Table A.2). At the same time, they are closely related to the non-cognitive skills agreeableness and conscientiousness from the "Big Five" personality factors, and trust, which measures prosocial beliefs (Becker et al. 2012).

Research has shown that these non-cognitive skills can be easily influenced and shaped (e.g. through mentoring programs) in childhood and adolescence (Kautz et al. 2014; Kosse et al. 2020; Resnjanskij et al. 2022) and that these skills and later labor market outcomes are strongly linked (Heckman et al. 2006; Almlund et al. 2011; Deming 2017; Deming and Kahn 2018). Consequently, if students indeed invested more into their non-cognitive skills when they receive comportment grades, we would expect more favorable labor market outcomes, i.e. in our setting, a more successful transition into the labor market after school. Furthermore, we would expect to see an increase in non-cognitive skill measures as well as in student achievement.

(ii) If students receive feedback on their comportment in the classroom, this might have a disciplinary effect. In contrast to the "No Excuses" approach (e.g. Angrist et al. 2013) adopted by charter schools in the US, the comportment grading policy stands in isolation, that is, it does not affect the school environment beyond grades. Still, it enables teachers to sanction disruptive in-class behaviors in a way that is also visible to parents. If classroom discipline improves due to this rather unobtrusive policy, something teachers appear to affirm (see Figure D.4), this would have beneficial repercussions on human capital formation (Lazear 2001). Research has shown that disruptive classroom behavior decreases student achievement significantly (Kristoffersen et al. 2015; Ahn and Trogdon 2017). Therefore, comportment grading might enhance academic achievement, which would also lead to a higher probability to enter the labor market successfully after school.

(iii) To the extent that comportment grades can also serve as indicators of non-cognitive skills (Landersø and Heckman 2017), as supported by respondents in our teacher survey D.3, they might enable students to signal these skills to employers, thereby reducing information asymmetry and facilitating students' transition into the labor market (Protsch and Solga 2015).

(iv) However, critics of comportment grades argue that the presence of different settings when establishing these grades renders them hardly comparable, which is contested by teachers (see Figures D.2 and D.5). Due to a lack of standardization, students might question the fairness behind the grading process, which could in turn 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). In this scenario, we would expect negative effects on student outcomes.

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 administrative data – the German Mikrozensus – 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 birth and their federal state of schooling or federal state of residence at the time the survey was conducted. We impute their year of enrolment and federal state of schooling from this information. Then, we link the individual date via the year-of-enrolment and state identifier with the reform status of their state of schooling. By doing so, individuals are assigned as being affected by the reform (treated) if the reform was in place when they entered school.[6] For these three sources, we add state-level information about whether schools grade comportment to derive treatment and control group assignments.

To explore potential explanations for our findings, we draw on two further data sources. We retrieve report card data from the National Educational Panel Study (NEPS) to investigate the relationship between subject and comportment grades, and make sense of our reform analysis. Finally, we ran a survey among teachers in Germany eliciting their assessment of comportment grades.

Data on the school-to-work transition The German Mikrozensus offers an administrative data source covering one percent of the German population in annual waves since 1970. We make use of the 2011–2018 waves and restrict the sample to individuals aged between 15 and 20[7] 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

[6]Information on the federal state of schooling is available in the nation student assessment data as well as partially in the SOEP data. However, the Mikrozensus lacks this information. We use the SOEP data and information from the German statistical office to gauge what share of individuals is assigned the wrong federal state of schooling if one instead uses current federal state at ages 15 to 20. We thereby achieve misclassifcation rates of 2% for the SOEP and 4.2% for the Mikrozensus data. See Appendix H for more detailed information.

[7]Pinquart et al. (2003) argue that one should consider up to five years after regular school-leaving age for assessing students' school-to-work transition. Regular school-leaving age for the school tracks we consider is 15 and 16 for the low and medium track, respectively (see Figure A.3).

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").[8]

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 a 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.[9]

Table C.1 provides the descriptive statistics of this sample, which consists of 16,982 observations.

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 15 to 20 from different survey years ($2003 - 2020$) 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 (Tables 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" or "Diligence" criteria. 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. 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 score of each item for each individual. If measures from different survey years are available for a given individual, we take the earliest available measure.

Table C.2 provides the descriptive statistics of our SOEP sample, which consists of 5,547 individuals.

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). All achievement tests target students in ninth grade and are always

[8]To avoid potential sample selection issues caused by changes in student tracking resulting from the reform, we investigate whether the reform had any effect on tracking (see Table 3, Column 2). Fortunately, our analysis does not find any significant impact of the reform on tracking. As a result, we are confident that our estimation approach is not affected by sample selection issues.

[9]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.

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 pseudo-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 addition to compulsory tests that measure students' reading skills, questionnaires are given to schools, students, and parents, which 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. Reading test scores were generated as follows: To keep the length of student achievement tests tractable, test providers estimate individual test scores based on a random subset of the full questionnaire (Jerrim et al. 2017). Hence, five different estimates of reading test scores ('plausible values') are available throughout all waves to represent the distribution of *true* reading skills. We do not consider math skills as they are only tested in every other wave of the National Assessment Study.[10]

Table C.3 provides the descriptive statistics of our student assessment data, which consists of 128,249 observations. For an exploratory analysis of classroom discipline in Section 6, we also use PISA 2000 as a cross section because it includes item batteries on classroom discipline that were separately answered by students and school principals.

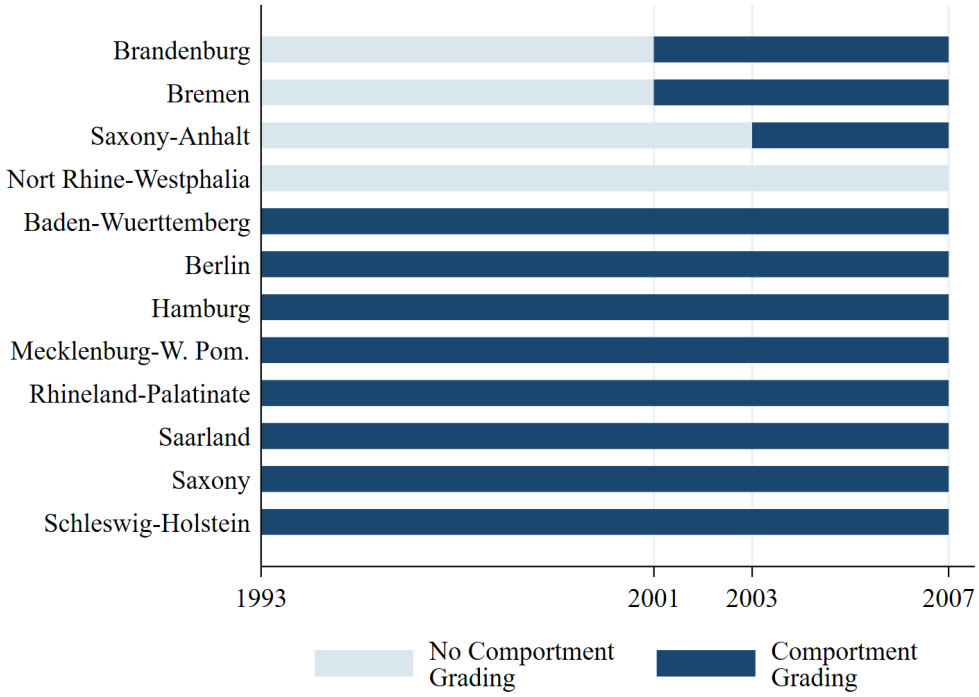
Data on comportment grading reforms Data on state-level comportment grading policies were gathered from school reform coding based on the states' schooling legislation and collected by Helbig and Nikolai (2015). We classify state policies according to four categories: (1) no comportment grading, (2) optional written comportment grading, (3) mandatory written 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 eight other federal states shown already had comportment grading schemes in place prior to our sample period.[11] After having established that heterogeneity in treatment effects is unlikely to

[10] 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. (2019).

[11] We cannot include all 16 federal states since the remaining 4 had either introduced or abolished comportment

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 written or mandatory numerical comportment grading. German states that reformed comportment grading shortly ahead of our sample period were excluded from the overview. We also exclude these states from our analysis. North Rhine-Westphalia did not introduce comportment grading until 2007.
Sources: Own representation based on Helbig and Nikolai (2015).

matter in our case, we use them to increase statistical power.

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. Based on our analysis of legal documents related to the reform, we chose this assignment method because the definitive treatment was only introduced for new student cohorts.[12] According to the legal rules, we found no evidence that the reform had to be applied retroactively to students who were already enrolled in school. To address any remaining concerns, we conduct a robustness test by assigning individuals to the treatment group if they had already been enrolled in school for some years at the time of the reform's implementation.

Report card data We use the Starting Cohort 3 from the National Educational Panel Study (NEPS SC3, version 11.0.0) to compare actual comportment grades with subject grades included in report cards (Blossfeld and Roßbach 2019). The first wave from fall 2010 includes individual-level data for students in grade five. These individuals were resurveyed at regular intervals

grading shortly before our sample period starts. This could confound our results.

[12] In the first and second grades, comportment grading was usually done in writing, while from the third grade onwards, numerical grading became mandatory. See for example here, last access on March 13, 2023.

until wave 10, which was conducted 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 C.4 provides the descriptive statistics of our grading data, which consists of 886 students and their report cards.

Teacher survey on comportment grading As a last step, we conduct an expert survey on the specifics of comportment grading among a convenience sample of 246 teachers in Germany. The respondents are recruited by distributing the survey on online teacher platforms, via teacher groups on social media, and by directly contacting schools. Overall, this online survey enables us to gain a more nuanced impression of how comportment grading takes place in practice. We are particularly interested in the teachers' assessment of the effectiveness of comportment grading, the time spent per student and report card as well as the number of teachers typically involved, and teachers' grading standards.

Table C.5 compares respondents' demographic characteristics to national averages from official statistics. It provides assurance that our sample – although not representative in nature – does not stray too far from the true distribution of age, gender, place of work, and school type among teachers in Germany.

4 Empirical Strategy

Identifying the average effect of comportment grading on the students who received comportment grades 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

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

where Y_{ist} is an outcome for student i attending school in state s in cohort t. CG_{st} is a dummy variable indicating that schools in this state graded comportment for this cohort (from the year of enrollment onwards) 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) and year of enrollment (λ_t). ε_{ist} is an error term.

Standard errors are clustered at the level of treatment assignment, that is, federal states (Abadie et al. 2022). In our main specification, we use 4 and 12 federal states, respectively, of which three introduce comportment grades at different points in time (see Figure 1). 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 obtain valid p-values.

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, that is, they are 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 the robustness of our estimates regarding the issues arising from treatment effect heterogeneity.[13] Therefore, we implement the estimator proposed by Callaway and Sant'Anna (2021), henceforth (C/S), which excludes states that already used comportment grading schemes prior to our sample period. 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. As detailed below, the C/S routine also estimates a weighted average of 2x2 effects.[14] Each 2x2 effect is an estimate of the average treatment effect on the treated (ATT):

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

where g denotes the year group G first receives the treatment and t are time periods. In our setup, there are two treatment groups receiving treatment in 2001 and 2003, respectively ($g \in \{2001, 2003\}$). We restrict the sample period to $t = 1992, \ldots, 2006$ since North Rhine-Westphalia introduces comportment grading in 2007. This means that we have 14 ATTs for each group, 8 (10) pre-treatment and 6 (4) post-treatment effects for the group with $g = 2001$ ($g = 2003$). To obtain a single estimate that can be interpreted as a multi-period and multi-group extension of the ATT in the canonical 2x2 design, we first average over post-treatment effects for each group and then across treatment groups to obtain

$$\text{ATT} := \frac{1}{6}\sum_{t=2001}^{2006} \text{ATT}(g = 2001, t) \cdot \Pr(G = 2001) + \frac{1}{4}\sum_{t=2003}^{2006} \text{ATT}(g = 2003, t) \cdot \Pr(G = 2003),$$

whose estimates can be compared directly to estimates of ATT obtained from TWFE

[13] 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, that is, there is no way treatment effects can evolve for individuals.

[14] See the Appendix B for a more detailed exposition.

specifications. If estimates on the same sample differ substantially, there must be heterogeneity in treatment effects either across time, across units, or both.

We run both estimation routines on the exact same set of individuals in the Mikrozensus sample by dropping units that have received treatment already before or at the start of the sample period. In addition, we show results from adding further federal states that did not change their comportment grading policy throughout our sample period for the analysis of non-cognitive skills and student achievement (see Figure 1). This increases statistical power and alleviates the inference issues arising from few clusters.

Note that the C/S approach does not allow us to conduct cluster-robust inference due to the small number of clusters (Callaway and Sant'Anna 2021, p.25). We therefore report heteroskedasticity-robust simultaneous confidence bands. In contrast to usual practice, simultaneous instead of pointwise confidence intervals capture the estimation uncertainty arising from estimating the whole sequence the group-time average treatment effects that go into the aggregation to obtain a single estimate of the treatment effect (see Figure 2). This automatically implies robustness against multiple hypothesis testing, which would be a concern if we used pointwise confidence intervals since each pre- and post-treatment coefficient in Figure 2 corresponds to one hypothesis test.

C/S confidence intervals grow in the number of estimated effects, that is, in the number of time periods. To maximize power, one can reduce the number of estimated effects in the following way. Our setup gives rise to four time periods: before the first group gets treated (1992 to 2000), after the first group got and before the second group gets treated (2000 to 2002), after the second group got treated and before North Rhine-Westphalia adopts the treatment (2003 to 2006), and 2007 and later. Following the same logic as above, this gives 0 (1) pre-treatment and 2 (1) post-treatment effects for the group with g = 2001 (g = 2003). This setup still allows for pre-testing and, crucially, leads to much narrower confidence intervals (see Figure E.2 and Panel B of Table 1). Since estimates of the overall ATT rely on averaging of post-treatment effects in Table 1 (see Appendix B), point estimates barely change between Panels A and B of Table 1.

There are two main threats to identifying the causal effect of comportment grading on treated students' outcomes. First, we need to assume parallel trends. This means that, in absence of the reforms, student outcomes would have followed the same trajectory over time for 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 Mikrozensus data.[15] 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 none of the individual pre-treatment coefficients are statistically distinguishable from zero for both treatment groups, suggesting that the parallel trends assumption is likely to hold.

[15]Note that our event-study results are based on the approach put forward in Sun and Abraham (2021) and therefore not biased by treatment effect heterogeneity.

Another threat to identification arises from different school reforms being introduced at the same time as comportment grading. 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 years. We address this potential concern by analyzing the two treatment groups that introduced comportment grades in different years separately (Figure 2), showing that the post-treatment effects have similar sizes in both groups. Another way to see this is by averaging the post-treatment effects for both groups separately, which shows that, reassuringly, both groups experience null effects (Figure E.1).

5 Results

5.1 Comportment Grading and the School-to-Work Transition

Panel A of Table 1 displays C/S estimates of the aggregated ATT estimand in equation 2 (columns 1 and 2)[16] and from estimating equation 1 (columns 3 and 4) using TWFE. 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 0.28 and -0.22 percentage points in columns 1 and 2, respectively.

Columns 3 and 4 display the results from estimating equation 1. Estimated effect sizes hardly differ across estimation techniques, corroborating the notion that heterogeneous treatment effects and ensuing negative weights issues are a lesser concern in our setup. Note that while confidence intervals based on cluster-robust standard errors are smaller than those robust to multiple hypothesis testing, we expect the former to be too narrow given the small number of clusters. Wild cluster bootstrap p-values alleviate this issue and similarly indicate that we fail to reject the hypothesis of a null effect by a wide margin.

Panel B reduces the number of time periods as outlined in Section 4.[17] The width of confidence intervals in columns 1 and 2 decreases substantially by more than 30% compared to Panel A, while point estimates remain very close to zero. Confidence intervals corresponding to TWFE estimations barely change since there is no multiple hypothesis testing correction. These results sharpen our inference in the sense that we can reject effect sizes beyond 3.3 percentage points in absolute value. Put differently, non-rejectable effect sizes are larger than one tenth of a standard deviation (the standard deviation of the variable successful school-to-work transition is 0.34; see Table C.1). Non-rejectable effect sizes are also small compared to other studies investigating effects on the school-to-work transition in Germany as an outcome. For example, Resnjanskij et al. (2022) find that a year-long mentoring program for disadvantaged adolescents

[16] Note that these numbers stem from the same estimation conducted for Figure 2. They can be obtained by averaging the post-treatment point estimates across both groups and time periods as detailed in Appendix B.

[17] This approach is visualized in Figure E.2.

increased the likelihood of doing an apprenticeship for students with low socio-economic status by 29.3 percentage points or .66 of a SD.

Robustness checks To alleviate concerns about the exact treatment timing as explained in Section 2, we report results using delayed treatment assignment. More specifically, we define the treatment as applying to all students who were in second, third, or fourth grade and higher instead of first grade, respectively, at the time of the reform. Appendix Table E.2 shows that our null result persists in Panels B and D. If we define treatment assignment as applying to third graders and above in Panel C, we find a negative effect that is borderline significant. Yet, considering the few-cluster adjusted inference from the WCB, we do not reject the null.

Further robustness checks in Appendix Table E.1 use different definitions of our outcome variable and sample restrictions. Panel A defines "successful school-to-work transition" more strictly by considering employed individuals without a vocational qualification prior to their employment as unsuccessful. Panel B 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 this is driving our results. Finally, Panel C combines both restrictions taken in the approaches shown in panels A and B. Crucially, all point estimates remain relatively close to zero, and nowhere can we reject the null hypothesis of a null effect.

Three patterns emerge from these results. Most importantly, comportment grading neither enhances nor reduces the chances of a successful school-to-work transition. Second, C/S and TWFE estimates hardly differ, irrespective of whether controls are included. This suggests that treatment effect heterogeneity is not an issue here. This is also vividly illustrated in Figure 1, which shows that post-reform point estimates differ neither across time periods nor groups. 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. Furthermore, for the analysis of non-cognitive skills and student achievement, we use eight further federal states to increase statistical power.[18] Ultimately, conclusions from estimated effects are robust to different definitions of the outcome variable and sample restrictions.

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 non-cognitive skill formation as an intermediate outcome is affected. Table 2 shows the results of estimating equation 1 using z-scored measures of non-cognitive skills as outcomes from our SOEP sample. In line with the null-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. p-values from the wild cluster bootstrap procedure bolster this finding.

[18]The analysis for the school-to-work transition using 12 states is provided in Appendix Table E.3

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				
	0.0028	−0.0022	−0.0055	−0.0062
	[−0.0413, 0.0469]	[−0.0472, 0.0428]	[−0.0400, 0.0290]	[−0.0370, 0.0246]
WCB p-val.	-	-	0.6446	0.6236
Panel B: Four period setup				
	−0.0020	−0.0033	−0.0062	−0.0069
	[−0.0331, 0.0291]	[−0.0327, 0.0262]	[−0.0402, 0.0278]	[−0.0372, 0.0234]
WCB p-val.	-	-	0.6326	0.6116
Mean Dep. Var.	0.88	0.88	0.88	0.88
N (A – B)	16,982	16,982	16,982	16,982
Controls	No	Yes	No	Yes
Std. Error	Robust	Robust	Cluster	Cluster

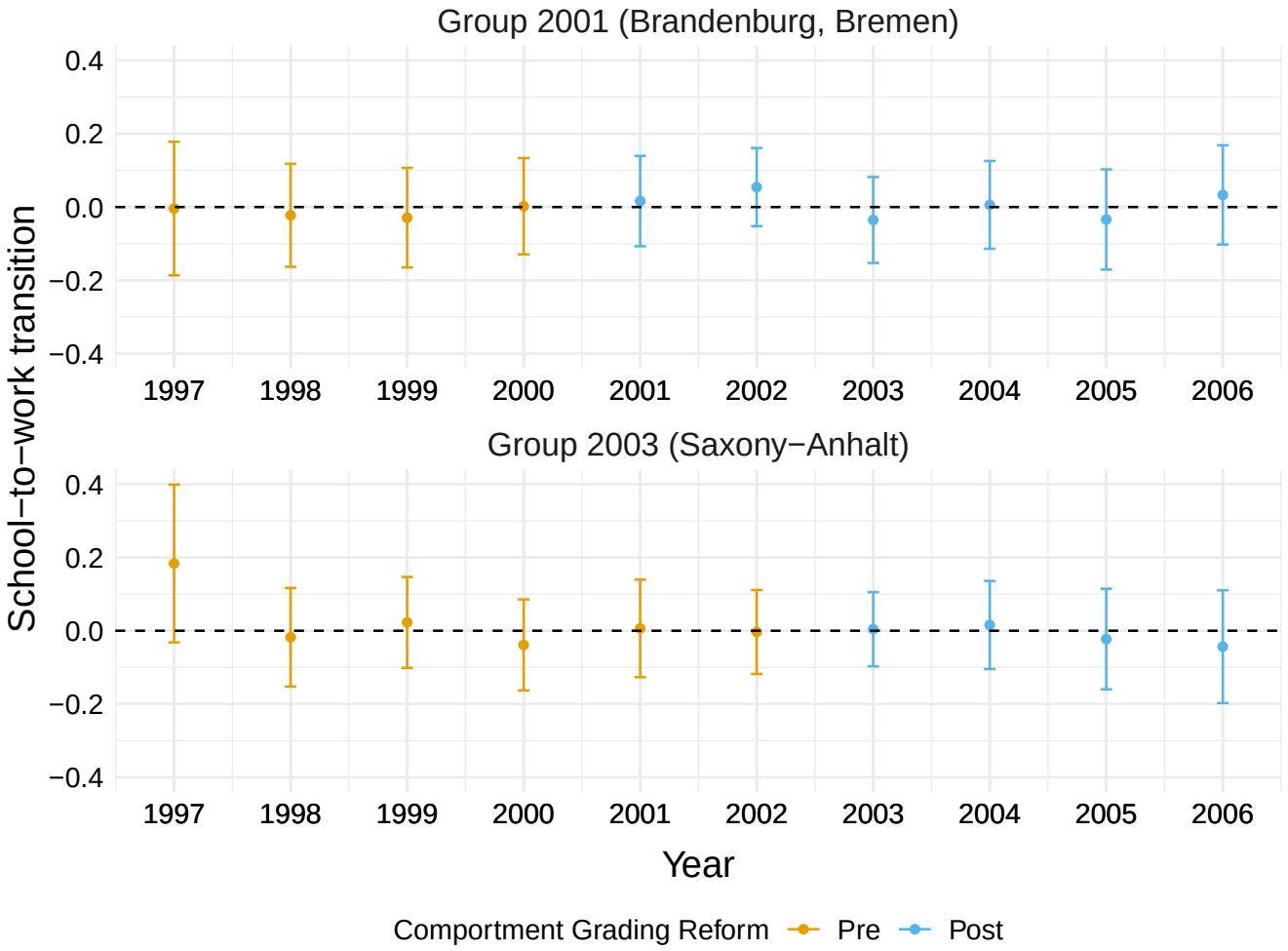
Notes: Estimates of the overall ATT (see equation 2 and 3) according to Callaway and Sant'Anna (2021) (columns 1 and 2) and from TWFE regressions using state and cohort fixed effects (columns 3 and 4). Columns 2 and 4 additionally include a female and migration background indicator as control variables. Columns 1 and 2 report simultaneous 95% confidence intervals robust to heteroskedasticity. 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.
Source: Mikrozensus waves 2011–2018.

Estimated effect sizes are small: They range from one to below three percent of a standard deviation in absolute value. In the case of trust as an outcome, we can benchmark our estimated effect against the effect other treatments studied in the literature had. Kosse et al. (2020) find that a year-long mentoring program increased a measure of primary school childrens' trust by .235 SDs (standard deviations). This effect is more than twice as large as the upper bound of our 95% confidence interval (.111 SDs). Similarly, Alan et al. (2019) find that students' grit, a trait capturing perseverance that is closely related to conscientiousness, increased by between .29 and .35 SDs through a teacher-training program.[19] Their estimated effects fall outside our confidence interval by a wide margin, too. We cannot present similarly rigorous evidence on agreeableness as an outcome but expect potential effect sizes to be of similar magnitude.

Tables E.4 and E.5 in the appendix contain results of robustness checks. While Table E.4 displays results of regressions without any control variables, Table E.5 changes the assignment of treatment based on the year of secondary school enrollment as in Panel D of Table E.2. Results in all three cases corroborate our null-effect finding.

[19]This program focused on three interrelated ideas underlying grit: students' growth mindset, perseverance through failures, and goal setting and involved videos, case studies, and classroom activities (Alan et al. 2019, p. 1123)

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: Mikrozensus waves 2011–2018

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 have five different plausible values available. We use the procedure from Macdonald (2008) that relies on Rubin (1987). Effectively, it combines estimates from five separate regressions with the respective plausible values and also considers the imputation error emerging from the stochastic nature of plausible values. Academic track attendance is an indicator variable, and is analyzed using a linear probability model. In line with results from our other outcomes, estimated effects on reading test scores and academic track school attendance are statistically indistinguishable from zero.

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

	Trust	Conscientiousness	Agreeableness
ATT	0.0151	−0.0241	0.0269
	[−0.0808, 0.1110]	[−0.1078, 0.0596]	[−0.0541, 0.1078]
WCB p-val.	0.7746	0.5737	0.5149
Observations	5547	5547	5547
Adj.R.squared	0.0204	0.0421	0.0118
Std.Error	Cluster	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state and cohort 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 and confidence intervals use weights from Webb's distribution and rely on 9999 iterations (Roodman et al. 2019). 95% confidence intervals are in box brackets.
Sources: SOEP-Core v37

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

	(1) Reading Skills	(2) Academic Track School Attendance
ATT	−0.0159	−0.0783
	[−0.2779, 0.2461]	[−0.2028, 0.0462]
Outcome mean	0.01	0.33
Adj. R-squared	0.188	0.126
Observations	128,249	128,249
St. Error	Cluster	Cluster

Notes: Each column presents 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. We report the average estimator across five regressions using separate plausible values of individual reading test scores as implemented by Macdonald (2008). Academic Track School Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level. 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).

Our confidence intervals allow us to reject positive effect sizes on reading test scores beyond roughly .25 SDs, which is admittedly still large. Using the rule of thumb presented in Woessmann (2016) that average student learning in a year is equal to about a .3 SDs test score increase, we can reject effect sizes that can be translated into learning of 83% of a school year or more. Yet, there are papers showing effects on students' test scores that fall outside our confidence interval. Carlana and La Ferrara (2021) find that an online tutoring program increased students' scores by .26 SDs on a test that covered math, Italian and English. For academic track school attendance, positive effects larger than one tenth of a SD can be rejected.

These patterns are robust against a variety of concerns, as demonstrated by further analyses in the appendix. First, in Table E.6, we exclude any background characteristics, indicating that even raw differences do not suggest any changes in student achievement. In Table E.7, we only control for individual characteristics also available in our previous two data sources. This leads to larger point estimates, which are still not statistically different from zero. As the national assessment data contains the richest and most complete set of background characteristics at the student and school level, in Table E.8, we add controls at the school level, which does not change results but reduces our sample size. Next, we perform the same robustness test as in Panel D of Table E.2, that is, we change the assignment of treatment year (year of enrollment in secondary school instead of enrollment in primary school). As shown in Table E.9, our conclusion remains unchanged.

5.4 Discussion of Main Results

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

We conduct benchmarking exercises for the outcomes we investigate to demonstrate that effect sizes we are unable to reject are mostly of small magnitude. To this end, we rely on effect sizes from randomized controlled trials that were considered impactful.

Based on the framework outlined in Abadie (2020), we argue that our results carry significant informational value despite being underpowered to reject meaningful effect sizes for some of the outcomes we investigate.[20] Prior beliefs of practitioners led us to put substantial prior probability on finding a positive effect of comportment grading on student outcomes. As soon as the prior probability of rejecting the null hypothesis of comportment grading having no effect on outcomes is larger than .5, the informational value of non-rejection exceeds that of rejection. Our findings therefore provide a major shift of posterior as compared to prior beliefs about the effectiveness of comportment grading policies. This emphasizes the important contribution this paper makes to the literature on school inputs.

6 Potential Explanations

Our analysis has established that comportment grades do not meaningfully alter student outcomes, but leaves open the question as to why this is the case. Section 2 delineates why the effect of comportment grading on student outcomes is ambiguous in theory. Similarly, practitioners disagree about the usefulness of comportment grades. On the one hand, teachers suggest that comportment grading has at best a small effect on various behavioral dimensions

[20] Abadie even goes so far as to point out that "[...] the informational value of nonsignificance relative to significance does not necessarily emanate from the estimation of "precise zeros."" More precisely, "even in empirical settings where power/precision is low relative to conventional requirements (e.g., power exceeding 0.80 is a usual benchmark in the design of experiments), nonsignificance may substantially outperform significance in terms of the informational value of the result [...]" (Abadie 2020, p. 195).

and academic achievement. For instance, Figure D.6 shows that more than 75% of respondents see no effect on students' thirst for knowledge, while respondents are split on the question whether there is a positive or no effect on academic performance. On the other hand, employers' fierce insistence on their usefulness to screen applicants (Tuch 2000), which has also been shown in a correspondence study by Protsch and Solga (2015), suggests that receiving comportment grades could indeed affect student outcomes. This section therefore explores the potential channels through which comportment grades could affect student outcomes using teacher survey data as well as additional analyses using NEPS and IQB data.

Direct signaling effect Even in the absence of an effect on academic achievement and non-cognitive skills, comportment grades could still help students signal non-cognitive abilities in the application process and therefore ease their school-to-work transition (Protsch and Solga 2015). However, finding such an effect rests on the assumption that comportment grades provide additional information to employers beyond other parts of the application. Analyzing variation in comportment grades using simple linear regressions, we find that once subject grades and overall GPA are accounted for, including measures of the two non-cognitive skills conscientiousness and agreeableness adds little explanatory power. More specifically, the adjusted R^2 increases by a mere 16% (2.7 percentage points).[21] Put differently, the informational value-added of comportment grades about non-cognitive skills relevant for labor market outcomes is limited once subject grades are known. Therefore, we should not expect a large effect on students' school-to-work transition operating through this signaling mechanism. Our finding regarding the informational content of subject grades is in line with the literature. In particular, subject grades are found to capture various aspects of personality in addition to IQ that should similarly be captured in comportment grades, too. For example, more conscientious individuals take assignments more seriously, which leads to better grades (Borghans et al. 2016). For the same reason, subject grades are more predictive of labor market outcomes than IQ. Furthermore, Ferman and Fontes (2022, p.1) show that teachers "inflate grades of well-behaved students and deduct points from worse-behaved ones" on high-stakes achievement tests in Brazil, supporting the notion that subject grades contain information about student behavior. Finally, teachers in our survey are split regarding the informational content of comportment grades: 30% of teachers (strongly) agree that comportment grades are already contained in subject grades, while roughly 50% (strongly) disagree.

Indirect effect mediated through academic achievement and non-cognitive skills A possible effect of comportment grades on the school-to-work transition could also be mediated by academic achievement and non-cognitive skills. However, there are two major theoretical arguments as to why it is reasonable to expect null effects. First, comportment grades as implemented in Germany and other countries provide a low-stake incentive to behave better as

[21] Note that most of the variance in comportment grades remains unexplained. This could be due to measurement error or because there are many important determinants of comportment grades that are missing in our data.

they typically do not count towards tracking decisions and the promotion to the next grade. It is therefore expected that students do not exert much effort to obtain better comportment grades (e.g., Schlosser et al. 2019). Yet, student effort is found to be an important input in the education production function (Stinebrickner and Stinebrickner 2004; De Fraja et al. 2010; Gneezy et al. 2019; Carrell and Rury 2021). Therefore, student outcomes should not be affected through this channel.

Second, the biannual – or even annual (see Figure D.7) – release of report cards with comportment grades in Germany may not provide feedback that is appropriate to change students' behavior. Previous research by Levitt et al. (2016) shows 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 comportment grades, e.g., pedagogical disciplinary concepts such as reprimands.

Empirically, we can provide indirect evidence to bolster these arguments. An effect of comportment grades on academic achievement might operate through a less disruptive classroom environment (Lazear 2001; Kristoffersen et al. 2015; Ahn and Trogdon 2017). We use the frequency of classroom disruptions as a proxy for how conducive the classroom environment is to foster learning. Tables G.1 and G.2 show that there is no correlation between a states' contemporaneous comportment grading policy and the number of classroom disruptions as assessed by students (Table G.1) and school principals (Table G.2). If anything, the signs of the estimates point towards a reduction in classroom disruptions and the amount of disciplinary problems if there is comportment grading. Teachers report that they rely less on other disciplinary measures when using comportment grading, such as one-to-one conversations with the disruptive student (see Figure D.4 in the appendix). This substitution between disciplinary approaches and comportment grading suggests that the feedback students receive about their behavior is similar regardless of whether comportment grades are given. Therefore, the introduction of comportment grades is unlikely to have a positive effect on students' classroom behavior. However, improved classroom behavior could potentially have direct positive effects on both non-cognitive skills and academic achievement. Consequently, neither outcome is likely to be impacted by the presence or absence of comportment grades.

Summary and limitations In sum, both theoretical and empirical arguments suggest that our null findings are meaningful. We provide evidence that comportment grades on report cards do not provide significant additional information relevant for labor market outcomes, making a direct signalling effect unlikely. In addition, we show that comportment grades are not associated with a better learning environment in the classroom, which is evidence against a direct effect on academic achievement. Finally, teachers report that they provide students with similar feedback on their behavior in the classroom irrespective of whether a comportment

grading policy is in place, implying that students probably do not change their behaviors in different ways under the two scenarios.

Although we lack the data and variation to empirically test the theoretical arguments surrounding the low-stakes nature of the incentives provided by comportment grades and the importance of timely feedback in behavior modification, both of them suggest that comportment grades, as implemented in German schools, are unlikely to affect student outcomes.

7 Costs of Comportment Grading

The costs of the reform can be substantial due to the time invested by the teachers: The expert survey suggested that in most cases (80%), more than one teacher is involved in comportment grading (see Figure D.8). When asked to provide exact numbers, most respondents stated that for each class, there are 6 to 10 other teachers involved in grading in addition to the class head teacher. Moreover, our expert survey highlighted considerable heterogeneity in the time invested in comportment grading. While most respondents take fewer than 30 minutes per student and report card, some need 60 minutes or more. Rationalizing such large numbers, one of our respondents explains that documentation of students' behavior may be necessary throughout the entire school year as a preparation for the final grade. Based on these answers, an exemplary cost calculation may reasonably assume that per student and report card, seven teachers are involved who each take the median of 15 minutes to grade the respective student's comportment (see Figures D.9 and D.10). Given that comportment grading is usually conducted for both half-term and end-of-year certificates, they invest about 30 minutes each year. Assuming a teacher typically works 40 hours per week and taking teachers' average salary of $88,071 as given by the OECD, we arrive at an estimated cost of $21.17 per student and year. For the roughly 11 million students in Germany, this adds up to a total cost of about $233 million. This amount is sufficient to, for instance, 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).

8 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 confidence intervals of the point estimates 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 3.3 percentage points in the probability of successfully transitioning into the labor market. In line with this finding, non-cognitive skills and academic achievement as potential mediators of such an effect are not affected either. We provide benchmarking exercises for the outcomes we investigate to demonstrate our ability to reject effect sizes from interventions that were deemed impactful.

Our robustness checks further bolster our findings: 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 the null result and find both theoretical and empirical evidence that comportment grades lack the effectiveness needed to significantly affect student outcomes. To arrive at this conclusion, we adopt a careful approach to identification and estimation and rely on numerous data sources.

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

In sum, this paper shows that the introduction of comportment grading does not have an effect on student outcomes. Finding null effects of educational reforms is not uncommon (e.g. Dale and Krueger 2002; Fryer 2011; Jerrim et al. 2018; Leuven and Løkken 2020; 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, especially when they incur significant costs. In this sense, the results presented can shift beliefs about the causal effect of comportment grading reforms (Abadie 2020). Our finding of null effects suggests that policy efforts should focus on other domains to increase the efficiency of the education system.

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. W. Imbens, and J. M. Wooldridge (2022). "When Should You Adjust Standard Errors for Clustering?" In: *The Quarterly Journal of Economics* 138.1, pp. 1–35. DOI: 10.1093/qje/qjac038.

Ahn, T. and J. G. Trogdon (2017). "Peer Delinquency and Student Achievement in Middle School". In: *Labour Economics* 44, pp. 192–217. DOI: 10.1016/j.labeco.2017.01.006.

Alan, S., T. Boneva, and S. Ertac (2019). "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit". In: *The Quarterly Journal of Economics* 134.3, pp. 1121–1162. DOI: 10.1093/qje/qjz006.

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). "Personality Psychology and Economics". In: *Handbook of the Economics of Education* 4, 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.

Arnold, K.-H. and W. Vollstädt (2001). "Arbeits- Und Sozialverhalten in Der Schule. Möglichkeiten Und Grenzen Ihrer Beurteilung Durch "Kopfnoten"." In: *Die deutsche Schule* 93.2, pp. 199–209.

Baumert, J. et al., eds. (2002). *PISA 2000 — Die Länder der Bundesrepublik Deutschland im Vergleich*. Wiesbaden: VS Verlag für Sozialwissenschaften. DOI: 10.1007/978-3-663-11042-2.

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.

Blossfeld, H.-P. and H.-G. Roßbach, eds. (2019). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)*. Vol. 3. Edition ZfE. Wiesbaden: Springer Fachmedien. DOI: 10.1007/978-3-658-23162-0.

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. JSTOR: 3090251.

Callaway, B. and P. H. C. Sant'Anna (2021). "Difference-in-Differences with Multiple Time Periods". In: *Journal of Econometrics*. Themed Issue: Treatment Effect 1 225.2, pp. 200–230. 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.

Card, D. and A. B. Krueger (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". In: *The American Economic Review* 84.4, pp. 772–793. JSTOR: 2118030.

Carlana, M. and E. La Ferrara (2021). "Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3777556.

Carrell, S. and D. Rury (2021). "Knowing What It Takes: The Effect of Information About Returns to Studying on Study Effort and Achievement". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3970822.

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.

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: *American Economic Review* 97.2, pp. 31–47. DOI: 10.1257/aer.97.2.31.

Currie, J. M. (2004). *Welfare and the Well-Being of Children*. London: Routledge. DOI: 10.4324/9780203987575.

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, pp. 2964–2996. DOI: 10.1257/aer.20181169.

De Fraja, G., T. Oliveira, and L. Zanchi (2010). "Must Try Harder: Evaluating the Role of Effort in Educational Attainment". In: *The Review of Economics and Statistics* 92.3, pp. 577–597. DOI: 10.1162/REST_a_00013.

Deming, D. and L. B. Kahn (2018). "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals". In: *Journal of Labor Economics* 36.S1, S337–S369. DOI: 10.1086/694106.

Deming, D. J. (2017). "The Growing Importance of Social Skills in the Labor Market*". In: *The Quarterly Journal of Economics* 132.4, pp. 1593–1640. DOI: 10.1093/qje/qjx022.

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.

Enquete-Kommission „Zukunft des Bürgerschaftlichen Engagements" des Deutschen Bundestages (2002). "Bürgerschaftliches Engagement: auf dem Weg in eine zukunftsfähige Bürgergesellschaft". In: *Bericht. Bürgerschaftliches Engagement: auf dem Weg in eine zukunftsfähige Bürgergesellschaft*. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 55–154. DOI: 10.1007/978-3-322-92328-8_1.

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.

Feng, A. and G. Graetz (2017). "A Question of Degree: The Effects of Degree Class on Labor Market Outcomes". In: *Economics of Education Review* 61, pp. 140–161. DOI: 10.1016/j.econedurev.2017.07.003.

Ferman, B. and L. F. Fontes (2022). "Assessing Knowledge or Classroom Behavior? Evidence of Teachers' Grading Bias". In: *Journal of Public Economics* 216, p. 104773. DOI: 10.1016/j.jpubeco.2022.104773.

Freier, R., M. Schumann, and T. Siedler (2015). "The Earnings Returns to Graduating with Honors — Evidence from Law Graduates". In: *Labour Economics*. European Association of Labour Economists 26th Annual Conference 34, pp. 39–50. DOI: 10.1016/j.labeco.2015.03.001.

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.

Gneezy, U. et al. (2019). "Measuring Success in Education: The Role of Effort on the Test Itself". In: *American Economic Review: Insights* 1.3, pp. 291–308. DOI: 10.1257/aeri.20180633.

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*. Themed Issue: Treatment Effect 1 225.2, pp. 254–277. DOI: 10.1016/j.jeconom.2021.03.014.

Hansen, A. T., U. Hvidman, and H. H. Sievertsen (2023). "Grades and Employer Learning". In: *Journal of Labor Economics* forthcoming. DOI: 10.1086/724048.

Hanushek, E. A. (2020). "Education Production Functions". In: *The Economics of Education (Second Edition) - A Comprehensive Overview*. Ed. by S. Bradley and C. Green, pp. 161–170. DOI: 10.1016/B978-0-12-815391-8.00013-6.

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.

Helbig, M. and R. Nikolai (2015). "Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949". In: *Book, Verlag Julius Klinkhardt*, pp. 1–383.

Hvidman, U. and H. H. Sievertsen (2021). "High-Stakes Grades and Student Behavior". In: *Journal of Human Resources* 56.3, pp. 821–849. 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. A. Lopez-Agudo, O. D. Marcenaro-Gutierrez, and N. Shure (2017). "What Happens When Econometrics and Psychometrics Collide? An Example Using the PISA Data". In: *Economics of Education Review* 61, pp. 51–58. DOI: 10.1016/j.econedurev.2017.09.007.

Jerrim, J., L. Macmillan, J. Micklewright, M. Sawtell, and M. Wiggins (2018). "Does Teaching Children How to Play Cognitively Demanding Games Improve Their Educational Attainment? Evidence from a Randomized Controlled Trial of Chess Instruction in England". In: *Journal of Human Resources* 53.4, pp. 993–1021. DOI: 10.3368/jhr.53.4.0516.7952R.

Jones, E. B. and J. D. Jackson (1990). "College Grades and Labor Market Rewards". In: *Journal of Human Resources* 25.2, pp. 253–266.

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.

Kostorz, P. (2016). "Bewertungsmaßstäbe und Bezugsnormen bei der Notenvergabe unter der Lupe des Schulrechts – Was ist pädagogisch sinnvoll, was juristisch möglich?" In: *RdJB Recht der Jugend und des Bildungswesens* 64.2, pp. 270–289. DOI: 10.5771/0034-1312-2016-2-270.

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. JSTOR: 2696418.

Leuven, E. and S. A. Løkken (2020). "Long-Term Impacts of Class Size in Compulsory School". In: *Journal of Human Resources* 55.1, pp. 309–348. 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.

Macdonald, K. (2008). *PV: Stata Module to Perform Estimation with Plausible Values*.

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. JSTOR: 1815932.

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. DOI: 10.1016/j.jebo.2020.01.009.

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. DOI: 10.1016/j.jebo.2021.07.032.

Pinquart, M., L. P. Juang, and R. K. Silbereisen (2003). "Self-Efficacy and Successful School-to-Work Transition: A Longitudinal Study". In: *Journal of Vocational Behavior* 63.3, pp. 329–346. DOI: 10.1016/S0001-8791(02)00031-3.

Piopiunik, M., G. Schwerdt, L. Simon, and L. Woessmann (2020). "Skills, Signals, and Employability: An Experimental Investigation". In: *European Economic Review* 123, p. 103374. DOI: 10.1016/j.euroecorev.2020.103374.

Prenzel, M. et al. (2007). *Programme for International Student Assessment 2003 (PISA 2003)Programme for International Student Assessment 2003 (PISA 2003)*. 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)*. 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)*. 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, L. Woessmann, and K. Wedel (2022). "Can Mentoring Alleviate Family Disadvantage in Adolescence? A Field Experiment to Improve Labor-Market Prospects". In: *Working Paper*, pp. 1–130.

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.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Ltd. DOI: 10.1002/9780470316696.indauth.

Ryan, P. (2001). "The School-to-Work Transition: A Cross-National Perspective". In: *Journal of Economic Literature* 39.1, pp. 34–92. JSTOR: 2698454.

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 (2019). *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.

Stinebrickner, R. and T. R. Stinebrickner (2004). "Time-Use and College Outcomes". In: *Journal of Econometrics*. Higher Education (Annals Issue) 121.1, pp. 243–269. DOI: 10.1016/j.jeconom.2003.10.013.

Sun, L. and S. Abraham (2021). "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects". In: *Journal of Econometrics*. Themed Issue: Treatment Effect 1 225.2, pp. 175–199. DOI: 10.1016/j.jeconom.2020.09.006.

Tan, B. J. (2022). "The Consequences of Letter Grades on Labor Market Outcomes and Student Behavior". In: *Journal of Labor Economics* forthcoming. DOI: 10.1086/719994.

Tuch, P. (2000). "Sozialverhalten im Zeugnis – Betragen: Sehr Gut". In: *Deutsches Ärzteblatt*.

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

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.

Woessmann, L. (2016). "The Importance of School Systems: Evidence from International Differences in Student Achievement". In: *Journal of Economic Perspectives* 30.3, pp. 3–32. DOI: 10.1257/jep.30.3.3.

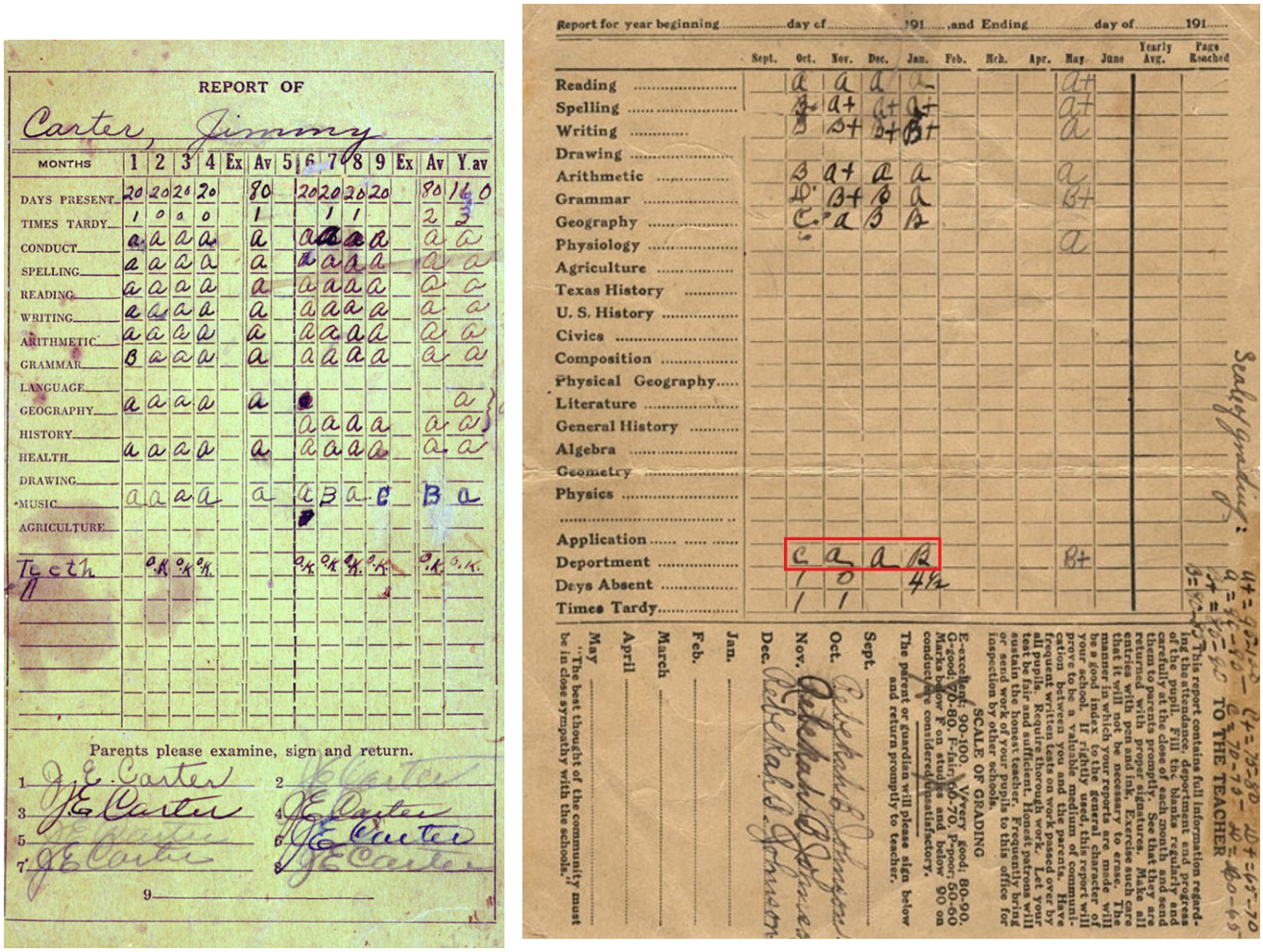
Zimmermann, K. F. et al. (2013). "Youth Unemployment and Vocational Training". In: *Foundations and Trends® in Microeconomics* 9.1-2, pp. 1–157. DOI: 10.1561/0700000058.

Appendix

(For Online Publication)

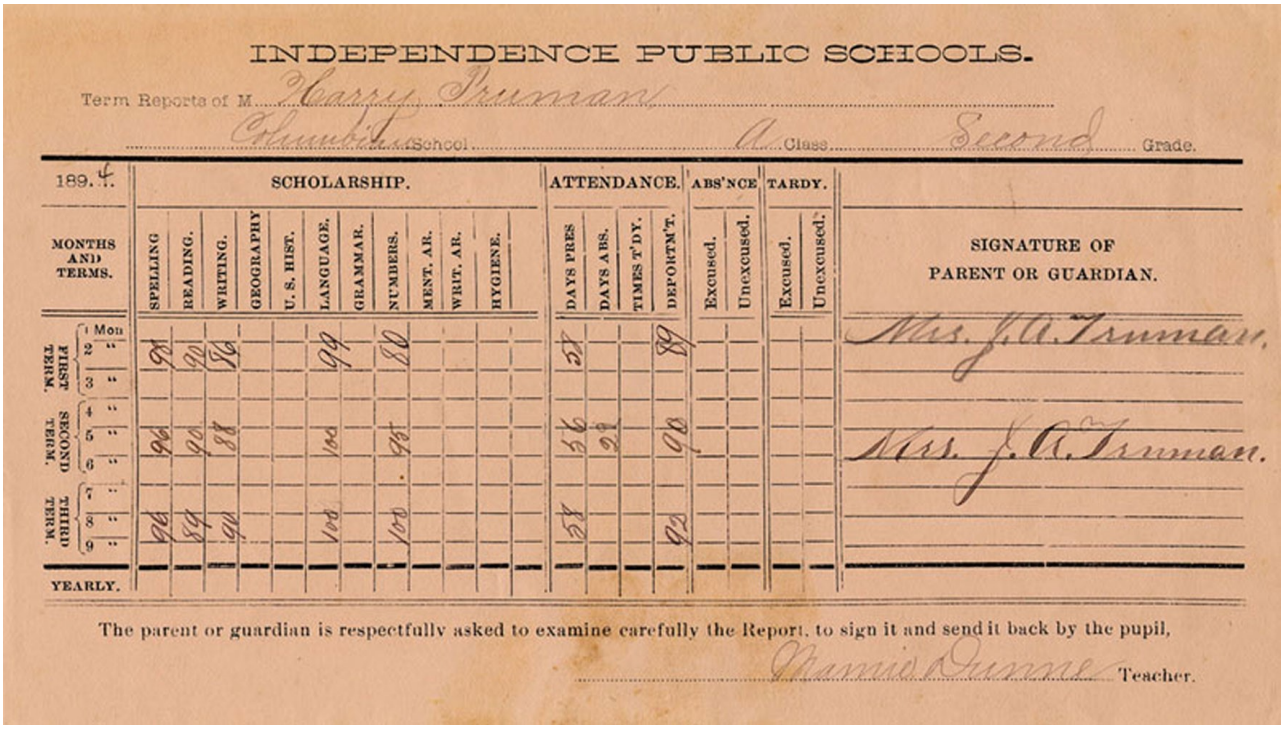
A Policy Background

FIGURE A.1. Report cards for 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. School report card for Harry S. Truman



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. Mon 1	95	90	90			90		80				54			90					Mrs. J. A. Truman.
2	"	"	"			"		"				"			"					
3	"	"	"			"		"				"			"					
SECOND TERM. 4	96	90	88			100		90				56	2		90					Mrs. J. A. Truman.
5	"	"	"			"		"				"			"					
6	"	"	"			"		"				"			"					
THIRD TERM. 7	96	95	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.

B Treatment Effect Estimands

Following the expositon by Callaway and Sant'Anna (2021), 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. $i \in \{\text{Brandenburg, Bremen, Saxony-Anhalt, North Rhine-Westphalia}\}$. We restrict the sample period to $t = 1992, \ldots, 2009$. Note that we lack a control group in 2007 and later. There are two treatment groups receiving treatment in 2001 and 2003, respectively ($g \in \{2001, 2003\}$). This means that we have 14 ATTs for each group, 8 (10) 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:

$$\begin{aligned}\overline{\text{ATT}}_S(g = 2001) &:= \frac{1}{2006 - 2001 + 1} \sum_{t=1993}^{2009} \mathbb{1}\{2001 \leqslant t \leqslant 2006\} \, \text{ATT}(g = 2001, t) \\ &= \frac{1}{6} \sum_{t=2001}^{2006} \text{ATT}(g = 2001, t)\end{aligned}$$

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

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

$$\begin{aligned}\text{ATT} &:= \sum_{g \in \{2001, 2003\}} \overline{\text{ATT}}_S(g) \cdot \Pr(G = g) \\ &= \overline{\text{ATT}}_S(g = 2001) \cdot \Pr(G = 2001) + \overline{\text{ATT}}_S(g = 2003) \cdot \Pr(G = 2003).\end{aligned} \tag{2}$$

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

As outlined in the empirical strategy (Section 4), the simultaneous confidence bands account for the estimation uncertainty along the entire path of group-time effects. This implies that confidence intervals widen in the number of estimated group-time effects. To maximize power

when estimating the ATT in equation 2, one can reduce the number of estimated effects in the following way without loss of generality. The reforms and our sample periods give rise to four time periods: before the first group gets treated (1992 to 2000), after the first and before the second group gets treated (2000 to 2002), after the second group got treated and before North Rhine-Westphalia adopts the treatment (2003 to 2006), and 2007 and later. They are denoted by $t^* = 1, \ldots, 4$. Note that we lack a control group in period 4. Following the same logic as above, this gives 0 (1) pre-treatment and 2 (1) post-treatment effects for the group with $g = 2001$ ($g = 2003$). Although this setup is more akin to classical difference-in-differences (e.g., Card and Krueger 1994), having a second treatment group allows us to use one period to pre-test for parallel trends. The averaging of effects then takes the following form:

$$\begin{aligned}\overline{\text{ATT}}_S(g = 2) &:= \frac{1}{2} \sum_{t^*=1}^{4} \mathbb{1}\{2 \leqslant t^* \leqslant 3\} \, \text{ATT}(g = 2, t^*) \\ &= \frac{1}{2} \sum_{t^*=2}^{3} \text{ATT}(g = 2, t^*)\end{aligned}$$

$$\overline{\text{ATT}}_S(g = 3) := \text{ATT}(g = 3, t^* = 3).$$

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

$$\begin{aligned}\text{ATT} &:= \sum_{g \in \{2,3\}} \overline{\text{ATT}}_S(g) \cdot \Pr(G = g) \\ &= \overline{\text{ATT}}_S(g = 2) \cdot \Pr(G = 2) + \overline{\text{ATT}}_S(g = 3) \cdot \Pr(G = 3).\end{aligned} \tag{3}$$

C Descriptive Statistics

TABLE C.1. Descriptive statistics Mikrozensus

	Mean	SD	Min	Max	N
Success	0.86	0.34	0	1	16982
Success strict	0.80	0.40	0	1	16982
CG (Enrolment)	0.09	0.29	0	1	16982
CG (2nd grade)	0.07	0.26	0	1	16982
CG (3rd grade)	0.05	0.22	0	1	16982
CG (4th grade)	0.03	0.17	0	1	16982
Female	0.42	0.49	0	1	16982
Migrant	0.30	0.46	0	1	16982

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: Mikrozensus waves 2011–2018.

TABLE C.2. Descriptive statistics SOEP

	Mean	SD	Min	Max	N
Trust	−0.00	1.00	−2.51	2.81	5547
Conscientiousness	−0.00	1.00	−3.41	1.80	5547
Agreeableness	0.00	1.00	−4.56	1.71	5547
Openness	−0.00	1.00	−3.17	1.96	5546
Extraversion	−0.00	1.00	−3.02	1.68	5547
Neuroticism	0.00	1.00	−2.41	2.50	5545
Risk_love	0.00	1.00	−2.64	1.87	5536
CG (Enrolment)	0.64	0.48	0.00	1.00	5547
CG (4th grade)	0.74	0.44	0.00	1.00	5547
Female	0.50	0.50	0.00	1.00	5547
Migrant	0.29	0.45	0.00	1.00	5547

Notes: Sample includes students from the federal states of Bremen, Brandenburg, Saxony-Anhalt, North Rhine-Westphalia, Baden-Wuerttemberg, Berlin, Hamburg, Mecklenburg-Vorpommern, Rhineland-Palatinate, Saarland, Saxony, and Schleswig-Holstein. Comportment group indicators are defined as whether there is comportment grading when the student is enrolled or in 4th grade, respectively.
Sources: SOEP-Core v37.

TABLE C.3. Descriptive statistics nationwide student assessments

	Mean	SD	Min	Max	N
Reading skills	0.01	1.00	−6.00	5.11	128249
Academic track school attendance	0.33	0.47	0.00	1.00	128,249
Comportment grading (enrolment)	0.72	0.45	0.00	1.00	128,249
Comportment grading (4th grade)	0.76	0.43	0.00	1.00	128,249
Comportment grading (current grade)	0.83	0.38	0.00	1.00	128,249
Female	0.49	0.50	0.00	1.00	128249
First generation migrant	0.12	0.33	0.00	1.00	128,249
Age (months)	187.26	6.39	148.96	242.00	128,249
Low SES	0.30	0.46	0.00	1.00	128,249
School size (students)	627.59	284.94	3.00	2344.00	101,320
Public school	0.95	0.22	0.00	1.00	101320
Town (<15,000 inhabitants)	0.25	0.43	0.00	1.00	101,320
Large town (15,000-100,000 inhabitants)	0.35	0.48	0.00	1.00	101,320
City (100,000-1,000,00 inhabitants)	0.39	0.49	0.00	1.00	101,320
Observations	128,249				

Notes: Sample includes repeated cross-sections of students in ninth grade from the federal states of Bremen, Brandenburg, Saxony-Anhalt, North Rhine-Westphalia, Baden-Wuerttemberg, Berlin, Hamburg, Mecklenburg-W. Pom., Rhineland-Palatinate, Saarland, Saxony, Schleswig-Holstein. 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 C.4. Descriptive statistics NEPS

	Mean	SD	Min	Max	N
German Grade	4.33	0.80	2.00	6.00	891
Math Grade	4.17	0.95	2.00	6.00	891
GPA	4.52	0.64	3.00	6.00	891
Conscientiousness	0.006	1.01	−3.15	2.10	891
Agreeableness	−0.006	1.01	−3.17	2.57	891
Comportment Grade	4.97	0.70	1.00	6.00	886

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 standardized 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 886.
Sources: NEPS SC3 11.0.0

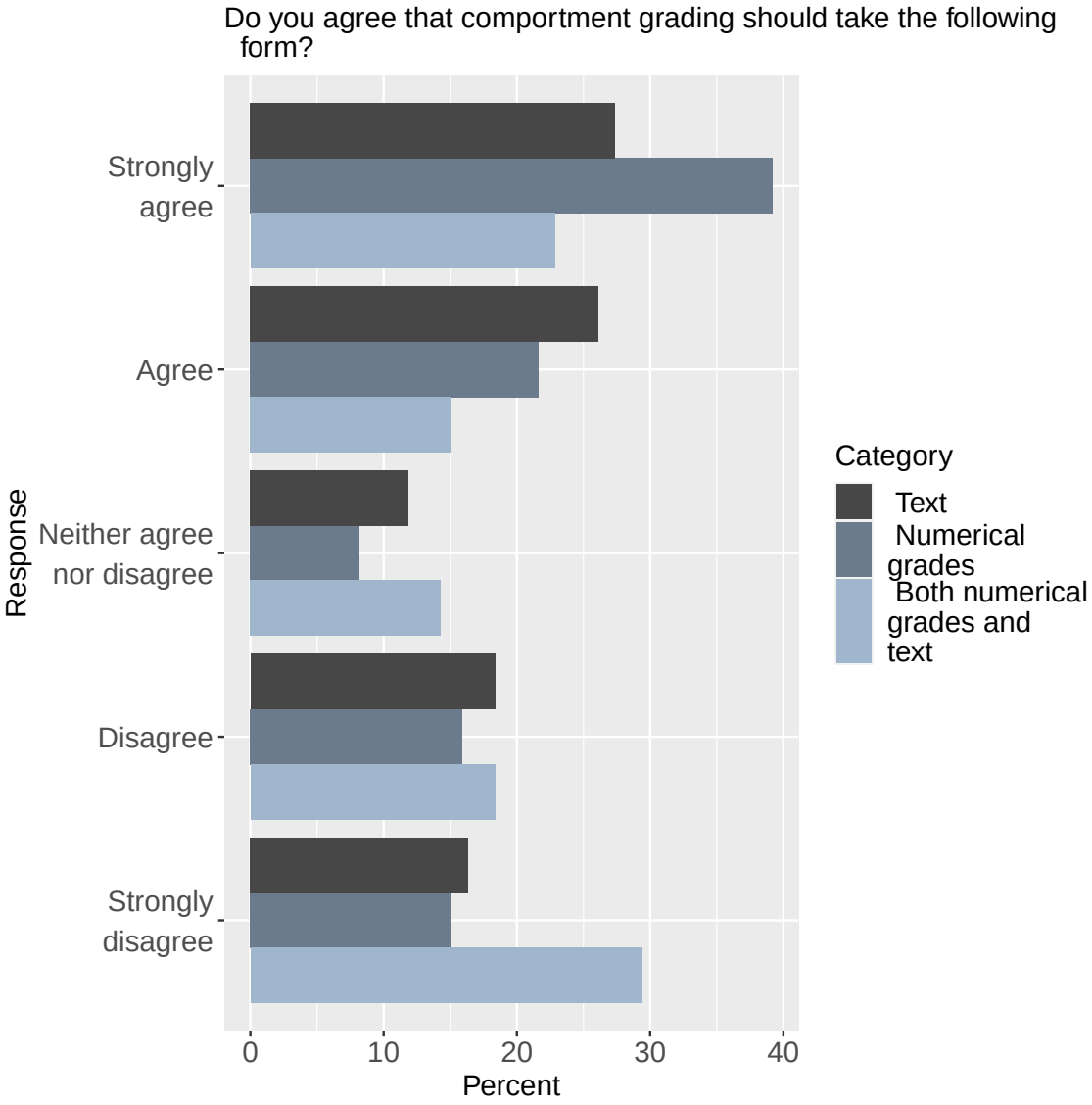
TABLE C.5. Descriptive statistics teacher survey: Teacher characteristics

	Sample percentage	National percentage
School type		
Primary school	25.31	31.2
Basic school (Hauptschule)	6.12	4.0
Middle school (Realschule)	23.67	8.1
Academic track (Gymnasium)	22.45	26.4
Integrated compreh. school	14.69	13.8
Others (e.g., vocationnal schools)	7.76	NA
Place of Work		
West Germany	61.22	73.0
East Germany	38.37	25.3
No answer	0.41	1.7
Age		
29 years or younger	3.27	7.0
30-39 years	26.53	29.0
40-49 years	22.45	26.0
50-59 years	36.73	26.0
60 years or older	10.61	11.0
No answer	0.41	0.0
Gender		
Female	68.57	73.4
Male	29.39	26.6
Diverse	0.41	NA
No answer	1.63	NA
Observations	246	799314

Notes: Sample includes teachers that participated in our online survey. *NA* indicates that a certain characteristic was not available from official statistics.
Sources: The national percentage is taken from the German Statistical Office's 2021/2022 school report.

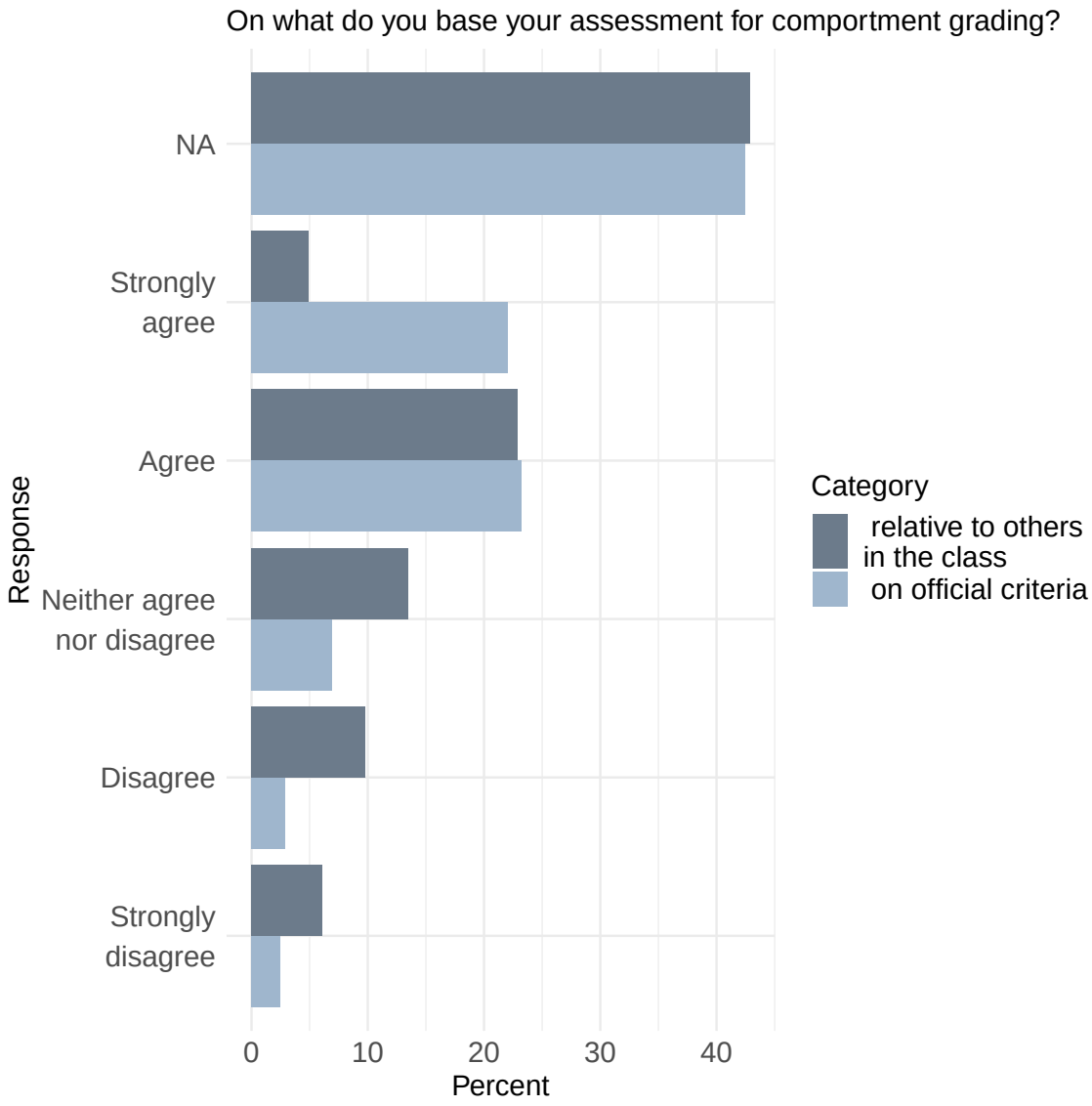
D Results of the Teacher Survey

FIGURE D.1. Form of comportment grading



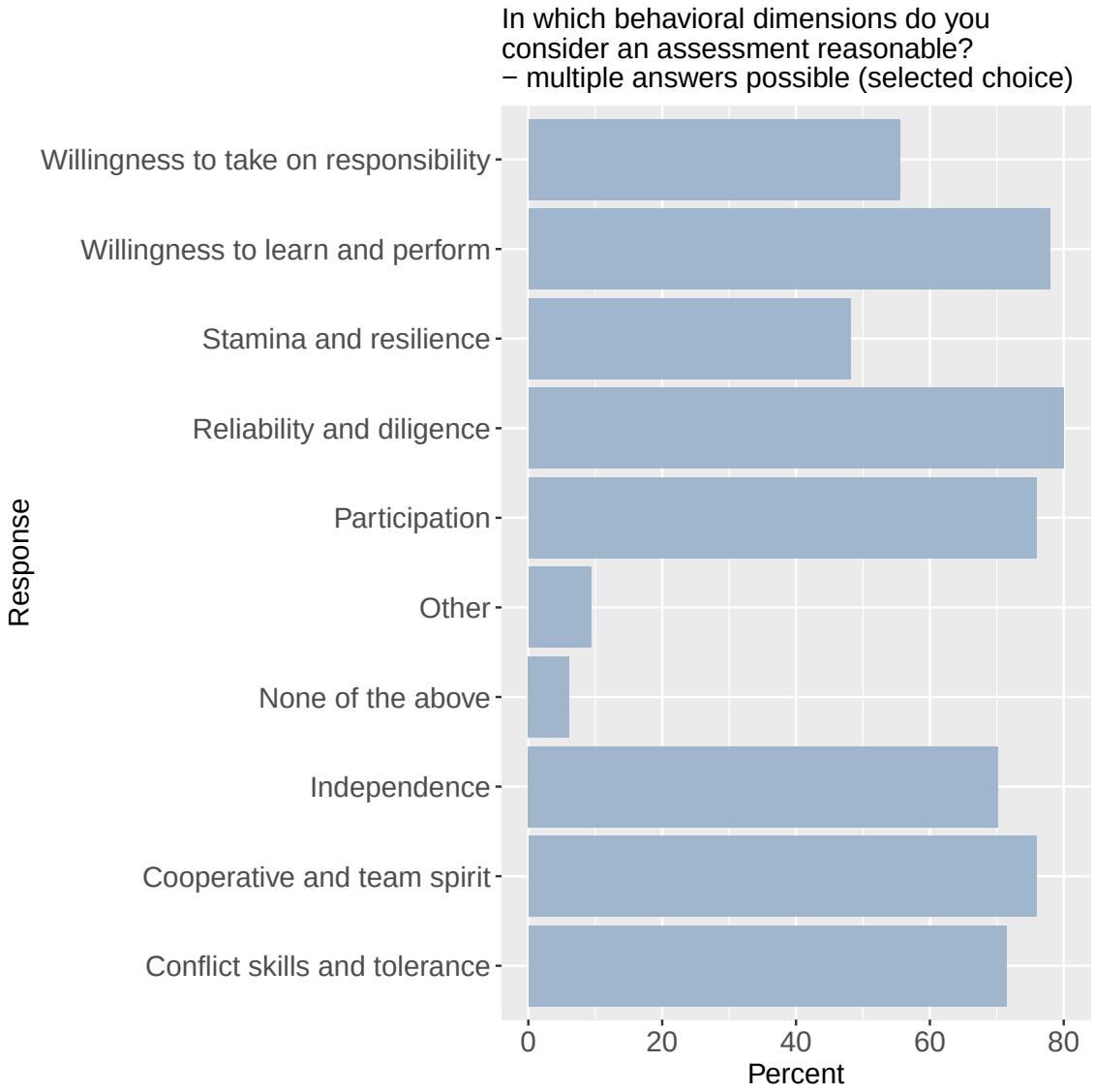
Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

FIGURE D.2. Conduct of comportment assessment



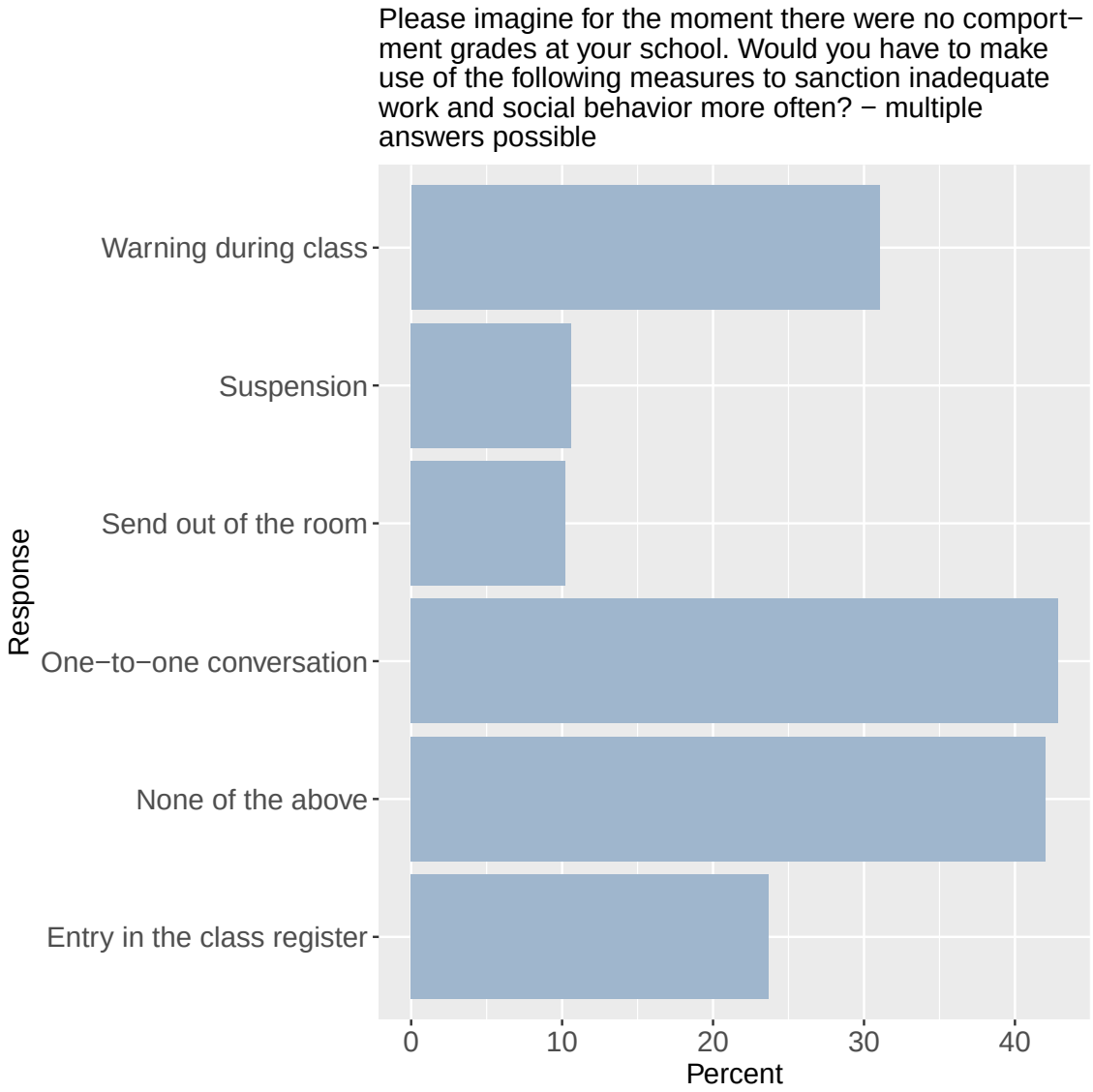
Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

FIGURE D.3. Behavioral dimensions assessed for comportment grading



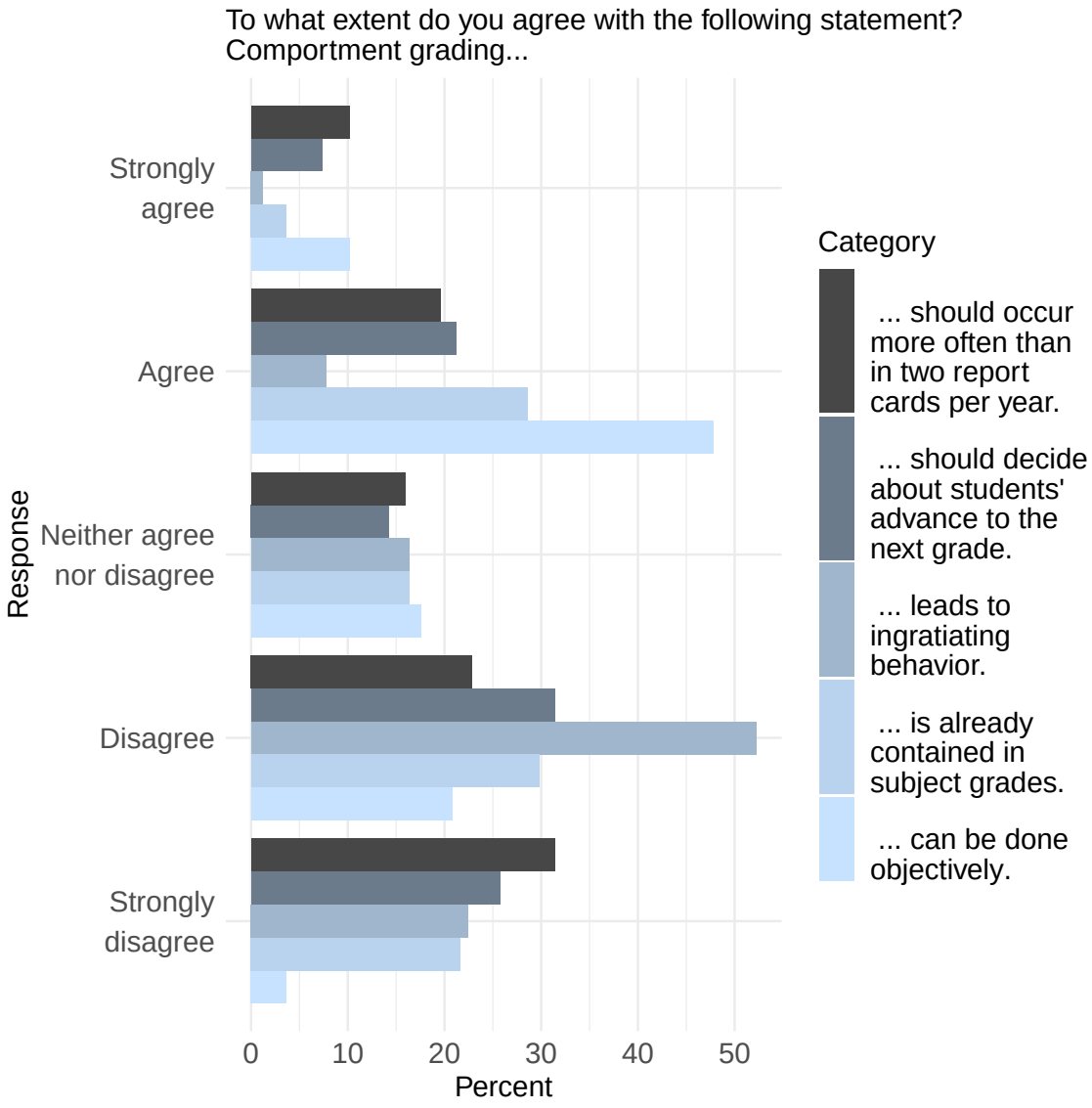
Notes: The above statistics are based on 245 responses. As multiple answers were possible for this question, percentages use this overall number of responses as reference.
Source: Own survey among German teachers.

FIGURE D.4. Disciplinary measures without comportment grading



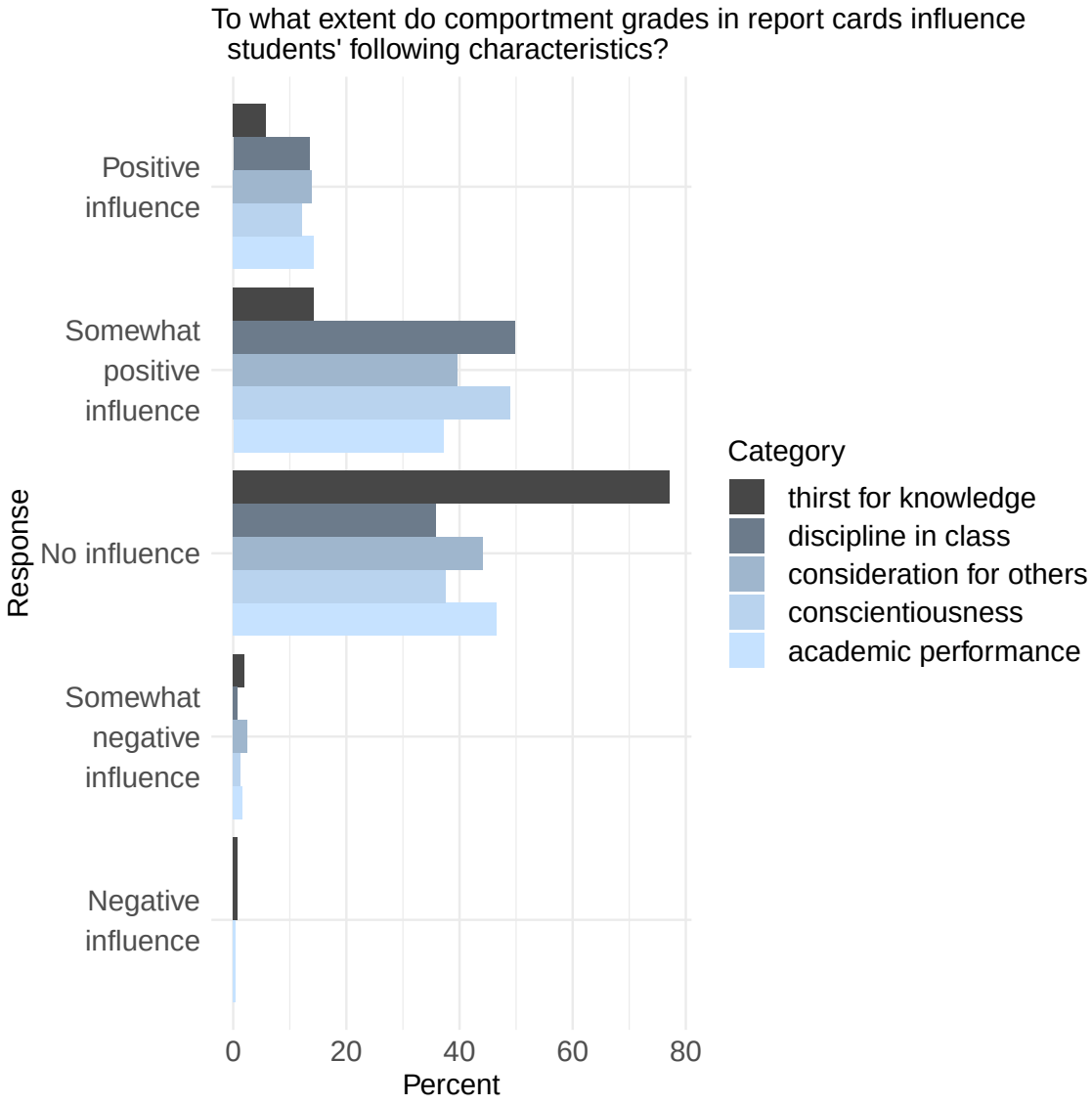
Notes: The above statistics are based on 245 responses. As multiple answers were possible for this question, percentages use this overall number of responses as reference.
Source: Own survey among German teachers.

FIGURE D.5. Teachers' opinions on the implementation and consequences of comportment grading



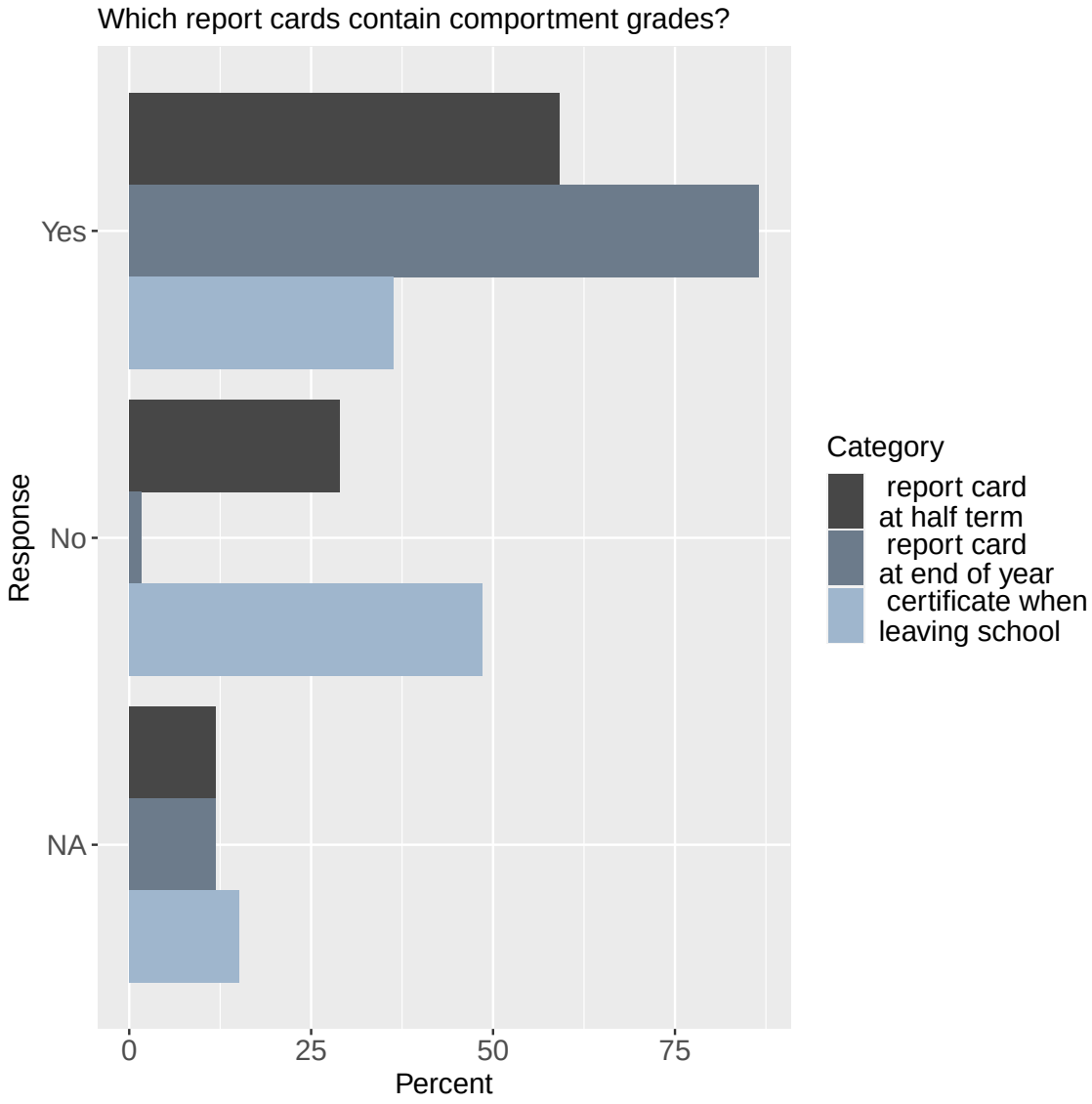
Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

FIGURE D.6. Influence of comportment grading on students



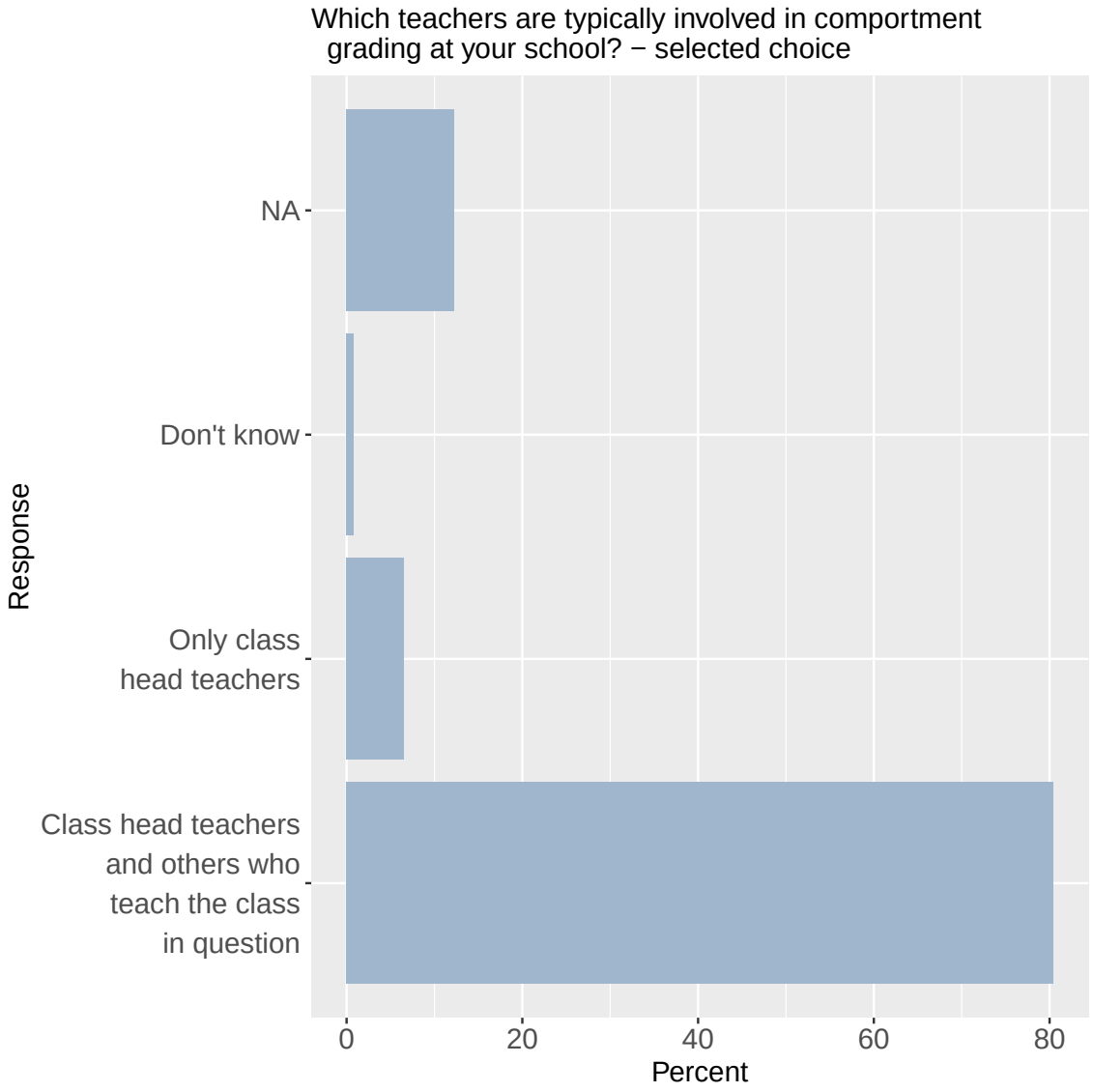
Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

FIGURE D.7. Report cards containing comportment grades



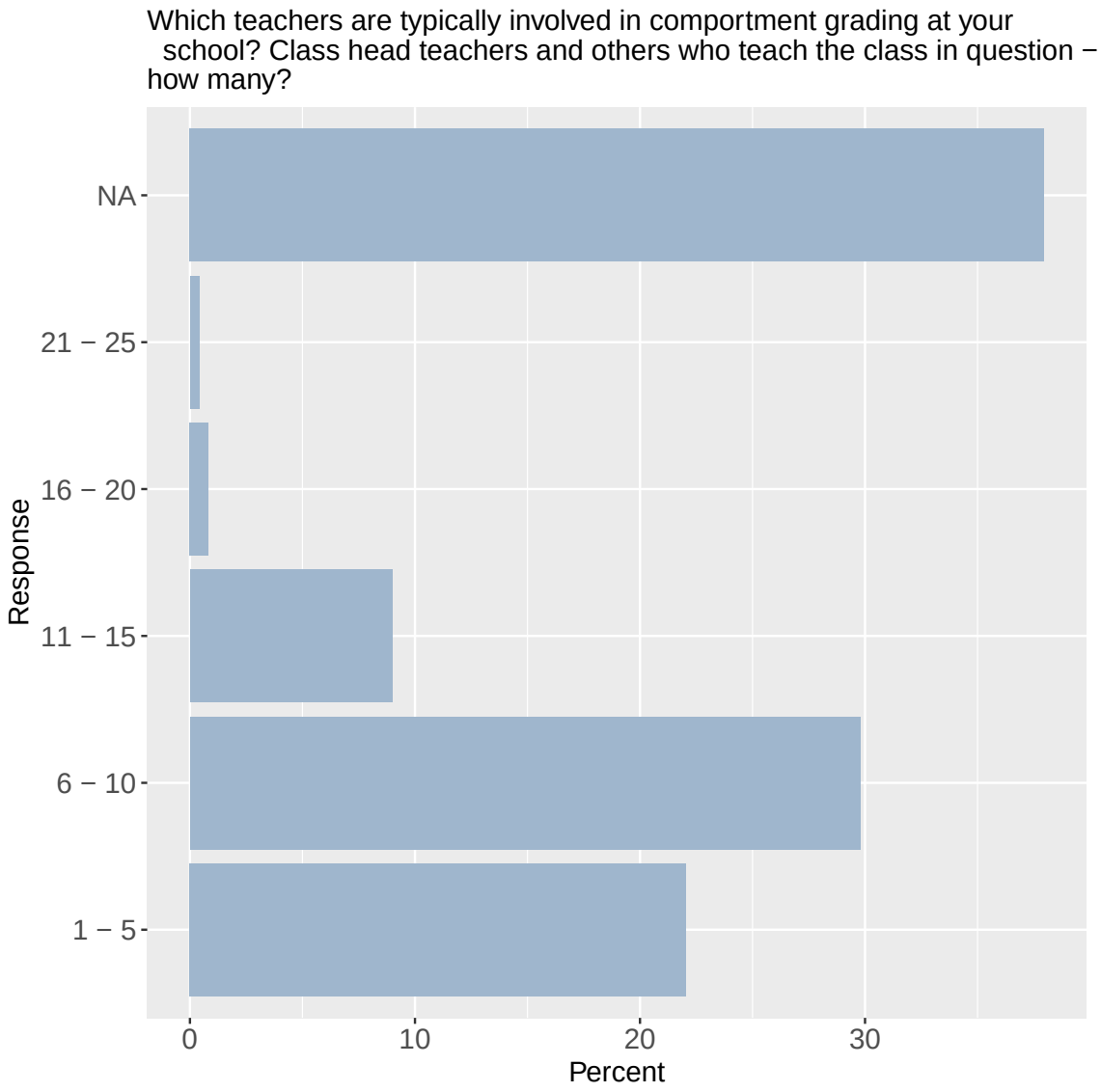
Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

FIGURE D.8. Teachers involved in comportment grading



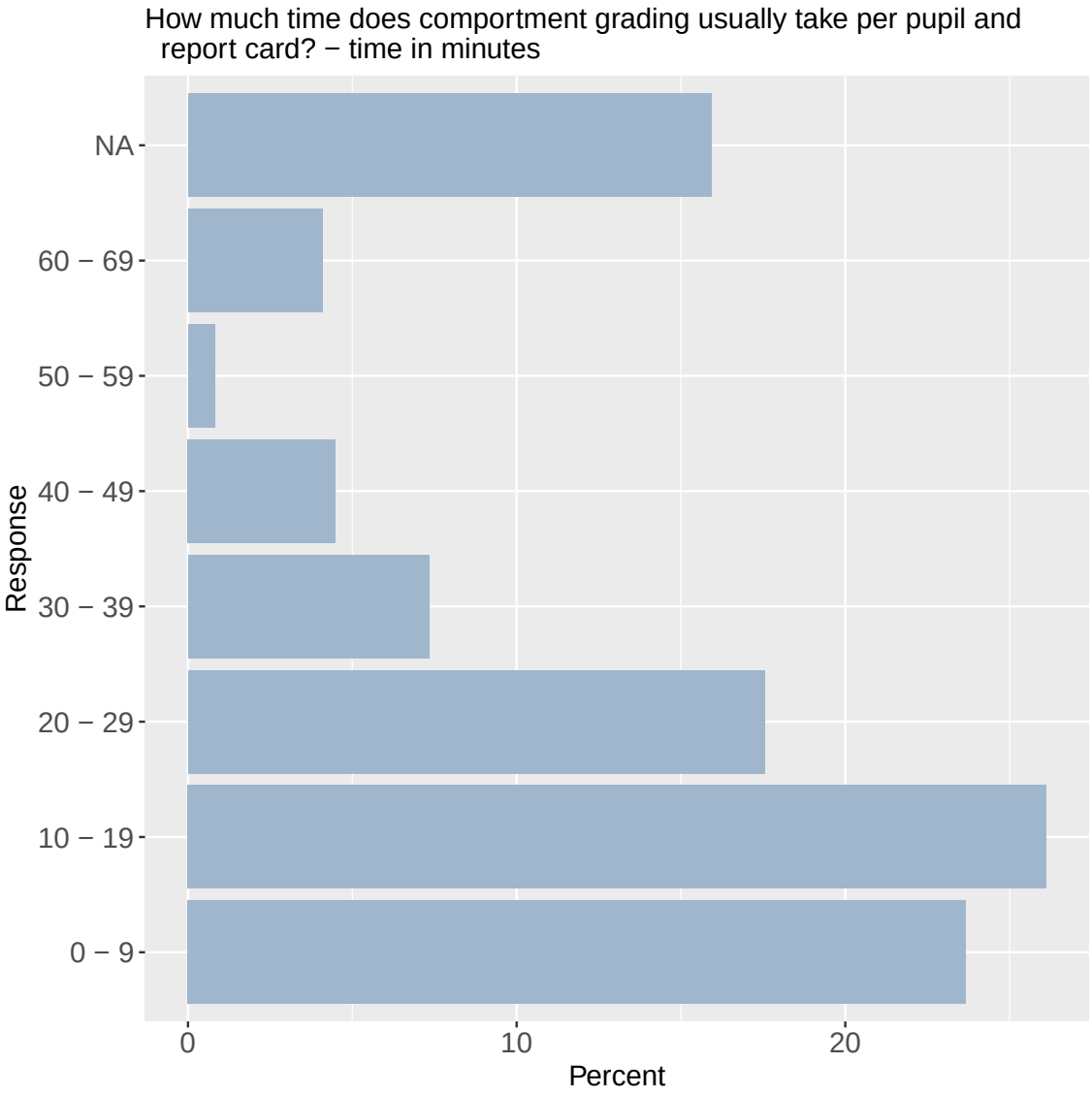
Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

FIGURE D.9. Other teachers involved in comportment grading



Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

FIGURE D.10. Minutes per student and report card

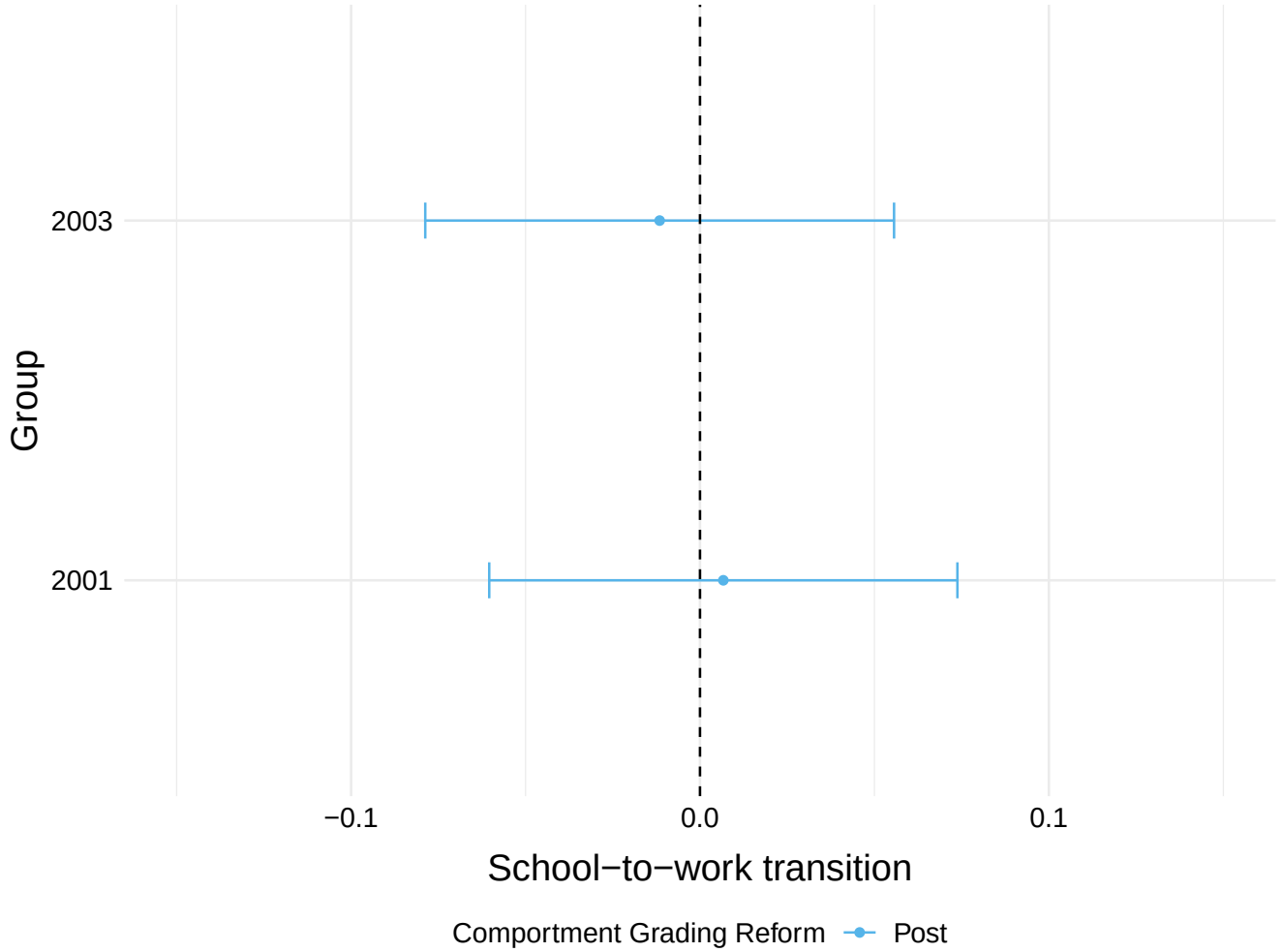


Notes: The above statistics are based on 245 responses.
Source: Own survey among German teachers.

E Robustness Checks

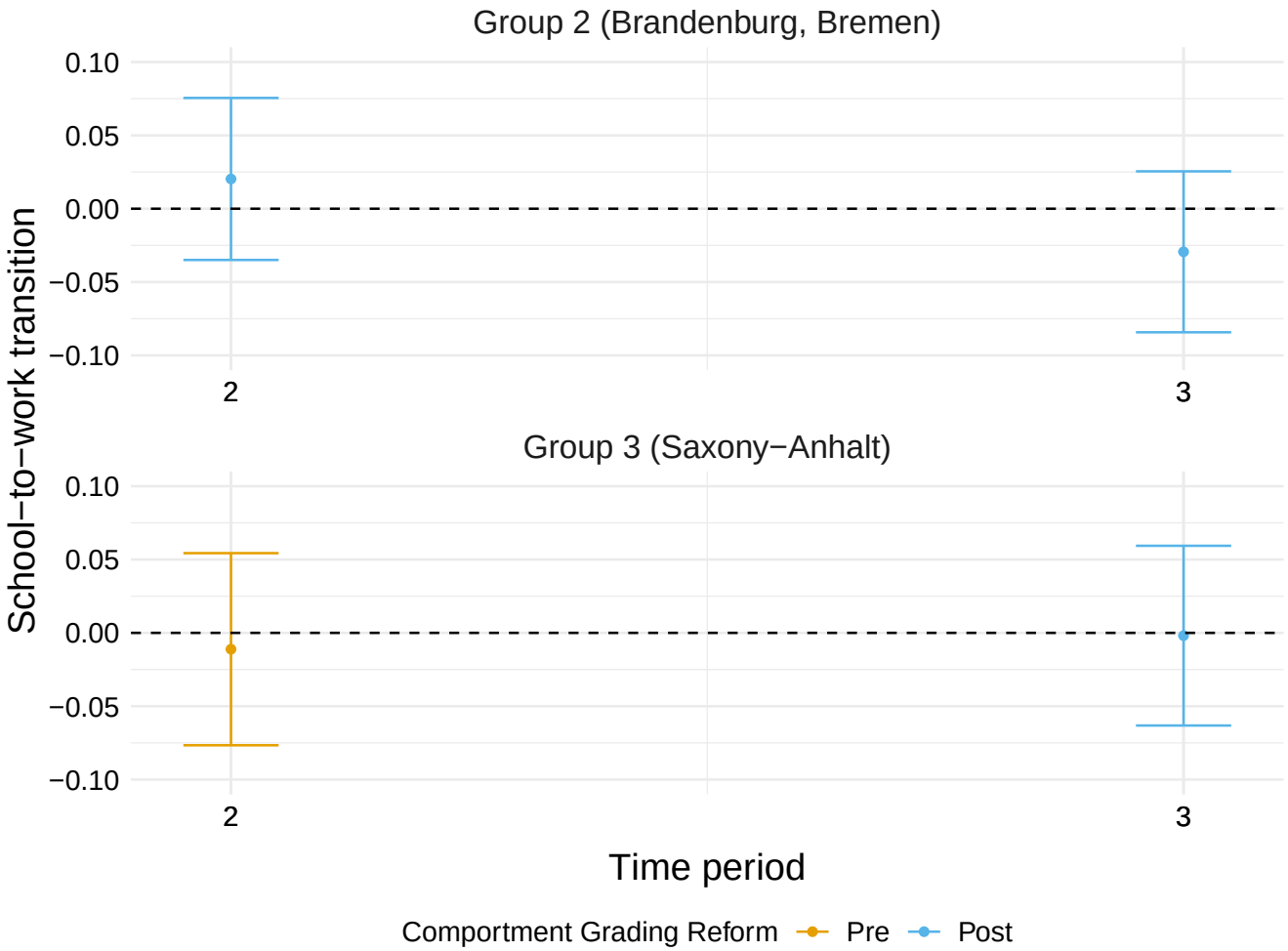
E.1 School-to-Work Transition

FIGURE E.1. By group treatment effect of comportment grading on school-to-work transitions



Notes: Figure displays estimates of treatment-group-specific effects. 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: Mikrozensus waves 2011–2018

FIGURE E.2. Four period setup: Dynamic effect of comportment grading on school-to-work transitions



Notes: Figure displays estimates of treatment-group-specific effects. 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: Mikrozensus waves 2011–2018

TABLE E.1. Effect of comportment grading on school-to-work transitions: Robustness checks

	Successful School-to-work Transition			
	Callaway & Sant'Anna (C/S)		Two-Way-Fixed-Effects (TWFE)	
	(1)	(2)	(3)	(4)
Panel A: Stricter definition of success				
	−0.0017	−0.0036	−0.0160	−0.0169
	[−0.0500, 0.0467]	[−0.0563, 0.0491]	[−0.0459, 0.0138]	[−0.0427, 0.0088]
WCB p-val.	-	-	0.2763	0.1782
Panel B: Without those catching up				
	0.0242	0.0225	0.0107	0.0081
	[−0.0270, 0.0754]	[−0.0285, 0.0735]	[−0.0251, 0.0464]	[−0.0220, 0.0382]
WCB p-val.	-	-	0.5746	0.6446
Panel C: Panel A and B combined				
	0.0254	0.0305	0.0041	0.0009
	[−0.0298, 0.0806]	[−0.0250, 0.0860]	[−0.0289, 0.0372]	[−0.0251, 0.0269]
WCB p-val.	-	-	0.7137	0.9329
Mean Dep. Var.	0.88	0.88	0.88	0.88
N (A)	16,982	16,982	16,982	16,982
N (B – C)	12,867	12,867	12,867	12,867
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 (2021) (columns 1 and 2) and from TWFE regressions using state and cohort 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. 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.
Source: Mikrozensus waves 2011–2018.

TABLE E.2. Effect of comportment grading on school-to-work transitions: Robustness checks using treatment assignment in different grades

	Two-Way-Fixed-Effects (TWFE)	
	(1)	(2)
Panel A: First grade treatment assignment (Main)		
	−0.0055	−0.0062
	[−0.0400, 0.0290]	[−0.0370, 0.0246]
WCB p-val.	0.6446	0.6236
N	16,982	16,982
Panel B: Second grade treatment assignment		
	−0.0144	−0.0165
	[−0.0453, 0.0166]	[−0.0450, 0.0120]
WCB p-val.	0.3754	0.3113
N	17,385	17,385
Panel C: Third grade treatment assignment		
	−0.0432	−0.0454
	[−0.0834, −0.0030]	[−0.0852, −0.0056]
WCB p-val.	0.0591	0.0611
N	17,477	17,477
Panel D: Fourth grade treatment assignment		
	−0.0374	−0.0388
	[−0.0848, 0.0099]	[−0.0833, 0.0056]
WCB p-val.	0.0571	0.0631
N	17,486	17,486
SE	Cluster	Cluster
Controls	No	Yes

Notes: Estimates of the overall ATT (see equation 2) from TWFE regressions using state and cohort fixed effects. Column 2 additionally includes a female and migration background indicator as control variables. Columns 1 and 2 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.
Source: Mikrozensus waves 2011–2018.

TABLE E.3. Effect of comportment grading on school-to-work transitions: Robustness checks - 12 instead of 4 federal states

	Successful School-to-work Transition			
	Callaway & Sant'Anna (C/S)		Two-Way-Fixed-Effects (TWFE)	
	(1)	(2)	(3)	(4)
	0.0028	−0.0022	0.0273	0.0277
	[−0.0413, 0.0469]	[−0.0472, 0.0428]	[−0.0045, 0.0591]	[−0.0042, 0.0597]
WCB p-val.	-	-	0.2032	0.1712
N	16,982	16,982	43,851	43,851
SE	Robust	Robust	Cluster	Cluster
Controls	No	Yes	No	Yes

Notes: Estimates of the overall ATT (see equation 2) from TWFE regressions using state and cohort fixed effects. Column 2 additionally includes a female and migration background indicator as control variables. Columns 1 and 2 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.
Source: Mikrozensus waves 2011–2018.

E.2 Non-cognitive Skills (Socio-Economic Panel)

E.2.1 Without controls

TABLE E.4. Effect of comportment grading on non-cognitive skills - without controls

	Trust	Conscientiousness	Agreeableness
ATT	−0.0011	−0.0132	0.0272
	[−0.1071, 0.1048]	[−0.1017, 0.0753]	[−0.0586, 0.1130]
WCB p-val.	0.9846	0.7762	0.5525
Observations	5547	5547	5547
Adj.R.squared	0.0045	0.0085	0.0035
Std.Error	Cluster	Cluster	Cluster

Notes: Each column presents separate OLS coefficient estimates with federal state and cohort 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 and confidence intervals use weights from Webb's distribution and rely on 9999 iterations (Roodman et al. 2019). 95% confidence intervals in box brackets.
Sources: SOEP-Core v37.

E.2.2 Treatment assignment in grade 4

TABLE E.5. Effect of comportment grading on non-cognitive skills - Treatment assignment in grade 4

	Trust	Conscientiousness	Agreeableness
ATT	0.0113	−0.0163	0.0175
	[−0.0717, 0.0944]	[−0.1394, 0.1069]	[−0.1067, 0.1418]
WCB p-val.	0.7957	0.8171	0.7788
Observations	5547	5547	5547
Adj.R.squared	0.0204	0.0421	0.0118
Std.Error	Cluster	Cluster	Cluster

Notes: This specification checks the robustness of the effects assuming that comportment grading was introduced in the fourth rather than the first year. 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.

E.3 Student Achievement (Nationwide Student Assessments)

TABLE E.6. Effect of comportment grading on academic achievement - without controls

	(1) Reading Skills	(2) Academic Track School Attendance
ATT	0.0033	−0.0767
	[−0.2909,0.2975]	[−0.2120,0.0587]
Outcome mean	0.01	0.33
R-squared	0.023	0.021
Observations	128,249	128,249
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. We report the average estimator across five regressions using separate plausible values of individual reading test scores as implemented by Macdonald (2008). Academic Track School Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level. 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).

TABLE E.7. Effect of comportment grading on academic achievement in ninth grade - harmonized controls

	Reading Skills	Academic Track School Attendance
ATT	0.0165	−0.0741
	[−0.2605,0.2935]	[−0.2062,0.0581]
Outcome mean	0.01	0.33
R-squared	0.046	0.025
Observations	128,249	128,249
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 and a dummy for migration background. Reading skills are standardized to have mean zero and unit standard deviation. We report the average estimator across five regressions using separate plausible values of individual reading test scores as implemented by Macdonald (2008). Academic Track School Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level. 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).

TABLE E.8. Effect of comportment grading on academic achievement - school-level controls

	(1) Reading Skills	(2) Academic Track School Attendance
ATT	0.0410	−0.0486
	[−0.2432,0.3251]	[−0.1941,0.0969]
Outcome mean	0.01	0.33
R-squared	0.250	0.262
Observations	101,320	101,320
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. We report the average estimator across five regressions using separate plausible values of individual reading test scores as implemented by Macdonald (2008). Academic Track School Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level. 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).

TABLE E.9. Effect of comportment grading on academic achievement - Treatment defined in 4th grade instead of enrollment

	(1) Reading Skills	(2) Academic Track School Attendance
ATT	0.0414	−0.0277
	[−0.1162,0.1989]	[−0.1093,0.0540]
Outcome mean	0.01	0.33
R-squared	0.046	0.133
Observations	128,249	128,249
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, migration background, age in months, and an indicator for parental SES. Reading skills are standardized to have mean zero and unit standard deviation. We report the average estimator across five regressions using separate plausible values of individual reading test scores as implemented by Macdonald (2008). Academic Track School Attendance is an indicator variable. Robust standard errors allow for clustering at the federal state level. 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).

F Relationship between Comportment and Subject Grades

TABLE F.1. Explanatory power of subject grades regarding comportment grades

	(1)	(2)	(3)	(4)
German Grade	0.124	0.129	0.119	0.123
	(0.034)	(0.034)	(0.034)	(0.034)
Math Grade	−0.006	−0.005	−0.008	−0.007
	(0.029)	(0.029)	(0.029)	(0.029)
GPA	0.328	0.326	0.324	0.322
	(0.046)	(0.046)	(0.046)	(0.046)
Agreeableness		0.052		0.040
		(0.023)		(0.023)
Conscientiousness			0.113	0.108
			(0.022)	(0.022)
Constant	1.957	1.991	2.081	2.103
	(0.373)	(0.369)	(0.380)	(0.377)
Controls	Yes	Yes	Yes	Yes
N.	886	886	886	886
R2	0.172	0.177	0.197	0.201
R2 Adj.	0.167	0.171	0.192	0.194

Notes: Each column presents separate OLS coefficient estimates with standard errors 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). Measures of Conscientiousness and Agreeableness are z-scored.
Sources: NEPS SC3 11.0.0.

G Relationship between Comportment Grades and Classroom Discipline

TABLE G.1. Contemporaneous comportment grading and classroom disruptions as rated by students

	Classroom disruptions (higher values - more frequent)			
	(1)	(2)	(3)	(4)
SG_s	-0.0317	-0.0343	-0.0414	-0.0307
	[-0.0866,0.0231]	[-0.0891,0.0205]	[-0.0965,0.0137]	[-0.0950,0.0336]
Student controls	no	yes	yes	yes
Classroom controls	no	no	yes	yes
School controls	no	no	no	yes
Outcome mean	0.00	0.00	0.00	-0.01
R-squared	0.000	0.002	0.008	0.013
Observations	28,165	28,165	28,165	22,761

Notes: Each column presents separate OLS coefficient estimate from a cross-sectional regression of the contemporaneous comportment grading policy on the freqeuency of learning impairments at school. SG_s is binary and equals 1 if state s had comportment grading in place. The outcome is standardized to have mean zero and unit standard deviation, averaging the six items on disciplinary problems in German classes: "teacher has to wait until students become quiet", "students cannot work undisturbed", "students do not listen to teacher", "students only start working long after beginning of class", "it is noisy in class", "nothing happens in the first five minutes". Each item can be answered by the students with "never" (1), "in few lessons" (2), "in most lessons" (3), "in all lessons" (4). Student controls include gender, age, migration background, and parental education. Classroom controls includes the student body composition regarding the same variables. School controls include school type, school size, public school dummy, and school location. 95% confidence intervals in box brackets. Heteroskedasticity-robust standard errors used. Sample includes all 16 states to maximize cross-sectional variation.
Source: PISA 2000.

TABLE G.2. Contemporaneous comportment grading and disciplinary problems as rated by school principal

	Disciplinary problems in class (higher values - more problems)		
	(1)	(2)	(3)
SG_s	-0.0637	-0.0531	-0.0834
	[-0.2038,0.0765]	[-0.2009,0.0947]	[-0.2199,0.0531]
Principal controls	no	yes	yes
School controls	no	no	yes
Outcome mean	0.01	0.02	0.02
R-squared	0.001	0.010	0.193
Observations	1,112	1,034	1,013

Notes: Each column presents separate OLS coefficient estimate from a cross-sectional regression of the contemporaneous comportment grading policy on the existence of learning impairments at school. SG_s is binary and equals 1 if state s had comportment grading in place. The outcome is standardized to have mean zero and unit standard deviation, averaging the six items on learning impairments at school: "classroom disruptions by students", "student truancy", "students lacking respect for teachers", "bullying of students by classmates", "frequent absence of student", "lacking parental support when learning at home". Each item can be answered by the school principal with "not impaired" (1), "little impaired" (2), "somewhat impaired" (3), "very impaired" (4). Principal controls include principal's gender, age, and years of experience as a teacher. School controls include school type, school size, public school dummy, and school location. 95% confidence intervals in box brackets. Heteroskedasticity-robust standard errors used. Sample includes all 16 states to maximize cross-sectional variation.
Source: PISA 2000.

H Federal State of Schooling

To alleviate concerns regarding measurement error in treatment assignment, we restricted our sample to minimize the number of wrongly assigned individuals. By doing so, we only misclassify an average of 2.03% of individuals in the German Socio-Economic Panel (SOEP) sample and of 4.19% in the Mikrozensus sample.[22] In the SOEP, we decided to mainly include individuals who participated in the youth surveys. These target adolescents aged 16 to 17 in the survey households, treating them as first-time adult respondents. We add further individuals from the adult surveys aged between 15 and 20. Similarly, the Mikrozensus sample is now restricted to young adults aged between 15 and 20.

For the SOEP, we have three different routes to imputing the federal state in which the individuals enrolled in primary school at our disposal. The first is to simply use the information on respondents' federal state of enrolment or, alternatively, on their state of transition into secondary school, which is available for a subset of respondents. This first strategy allows us to obtain the federal state at enrolment in primary school for about 40% of the sample.

For the remainder, for whom this is not possible, our "second-best" strategy exploits the fact that the households in which the youth respondents live participate in the SOEP themselves. This allows us to retrieve the federal state of respondents' households in those survey years in

[22] Average misclassification rates are computed separately for each age and then weighted by the age distribution in each sample.

which respondents were between 6 and 10 years old, i.e., attended primary school. The federal state of enrolment for another 10% of our sample is thus added.

Lastly, for the latter half of our sample, whose federal state of enrolment we did not grasp by any of the previous means, we use the federal state given in the youth questionnaire, that is, at age 17. Reassuringly, the misclassification rate remains low at roughly 3.2% when we compare the federal state at age 17 with that during primary school as obtained in the previous two routes. This is little surprising given that about 98% of 17-year-olds in Germany still live with their parents according to Mikrozensus data for 2019 (see here).

This third route is also how we proceed in the Mikrozensus sample, that is, we use the federal state at the time the survey was conducted. Comparing the federal state at ages 15 to 20, i.e., the age range of our Mikrozensus sample, to that during primary school in the SOEP sample produces a misclassification rate of 5.9% at maximum for those aged 20 and an average rate of 4.19% for the entire sample.