

Another Empty Promise of Grade Retention? Peer Effects of Classmate Retention

Samuel Jalalian^{1,2}
(corresponding author)
s.jalalian@lmu.de

Fabian T. Pfeffer^{1,2}
fabian.pfeffer@lmu.de

¹ Department of Sociology, LMU Munich, Munich, Germany

² Munich International Stone Center for Inequality Research (ISI), LMU Munich, Munich, Germany

Acknowledgments

We thank Steffen Schindler for his conceptual input on the research gap and theoretical framing during the early stages of this project, and Corinna Kleinert for her advice on coding and working with NEPS data. We also thank Natalie Backes, Lea Dittebrand, and Theresa Scheck for their helpful feedback and discussions during the early stages of the project. We further thank participants at the 9th NEPS Conference and the 2024 joint doctoral colloquium of the Brüderl and Auspurg research groups at LMU Munich for their valuable feedback. We are also grateful to Katja Scharenberg and Paul Fabian for their helpful comments.

Abstract

Grade retention is a common but costly educational practice and studies have repeatedly shown that, in many contexts, it offers no academic benefit to retained students. A remaining justification for its continued use is the presumed benefit for the promoted peers from whose class the retained student was removed. This study tests that assumption using longitudinal data from the German National Educational Panel Study (NEPS) and a difference-in-differences approach, focusing on grades and competencies from Grades 1 to 9. Results predominantly yield null findings, with a small positive effect on mathematics grades in lower secondary education (0.08–0.09 SD) as the only exception. Overall, our findings do not provide support for presumed peer benefits as a justification for grade retention.

Introduction

Grade retention or grade repetition, i.e., the practice of having a student repeat a grade level, is a common and costly measure in educational systems worldwide. In the United States, 2.3% of approximately 50 million students are retained in a single year, creating additional school expenditures exceeding \$12 billion annually (West, 2012). Similarly, in Germany—the focus country of this study—around 250,000 students (2.6%) were retained in 2007/08, resulting in direct costs of approximately €931 million (Klemm, 2009).¹ Of course, the high price of this policy could be justified if it created substantial benefits. Research assessing the impact of grade retention on the further academic development of retained students, however, has long been skeptical about such benefits (e.g., Jimerson, 2001). Recent meta-analyses and systematic reviews conclude that, on average, grade retention does not improve the academic performance of grade repeaters, with any positive effects confined to the short run and to very specific institutional contexts (Goos et al., 2021; Valbuena et al., 2021).

Despite limited effectiveness and substantial costs, grade retention persists. A common additional rationale offered in support of this practice is the belief that it benefits other students, namely those who are promoted. Some policy-makers, teachers, and administrators assume that low-achieving classmates impede the learning of higher-achieving students and that, therefore, removing these classmates should improve learning conditions for the remaining class by increasing classroom homogeneity (see Bellenberg & im Brahm, 2010, pp. 524–525; Klemm, 2009, p. 6; U.S. Department of Education, 1999, p. 7). This line of reasoning follows a similar logic to that applied in defending strong tracking or institutional segregation of students: even if the sorting out of low-achieving students does not support their own academic achievement, students who remain in higher-achieving tracks or schools are assumed to benefit from more homogeneous learning

environments. Under this logic, increased educational inequality is counterbalanced by increased educational quality for some. The empirical evidence, however, rejects this assumed trade-off, at least for the case of institutional differentiation (Heisig & Matthewes, 2022; Pfeffer, 2015). For grade retention, by contrast, there is very limited empirical evidence on whether promoted students benefit when one of their classmates is retained. This study addresses this gap by estimating the effect of retaining classmates on the grades and competencies of promoted students.

The remainder of this article is organized as follows. The next section discusses how the prior literature may help build expectations about the potential benefits or drawbacks of classmate retention for promoted students. The subsequent section describes the data, variables, and methods. After presenting our findings, we conclude that there is little evidence to support the assumption that classmate retention benefits promoted peers. Taken together with prior research showing that the practice fails to benefit retained students in many contexts (Goos et al., 2021; Valbuena et al., 2021) and reinforces educational inequality (National Research Council, 1999, p. 122; Pipa et al., 2025, p. 15), these findings call for a critical reassessment of grade retention as an educational policy.

Background

Prior Research on the Effects of Grade Retention

Much prior research has examined the effects of grade retention on the “repeater” and is largely skeptical about its benefits (Allen et al., 2009; Hattie, 2009, pp. 97–99; Holmes, 1989; Jimerson, 2001). More recently, Valbuena et al. (2021) reviewed 42 causal-inference studies published between 2001 and 2020 across various countries. They conclude that early-grade retention can have positive effects, but only in the short term and mostly within very specific institutional contexts.

By contrast, retention in later grades tends to produce null or even negative effects on student outcomes. Taken together, they conclude that the measure is overall ineffective for retained students (Valbuena et al., 2021). Similarly, another recent meta-analysis of 84 methodologically sound studies by Goos, Pipa and Peixoto (2021) indicates that grade retention has, on average, no effect on the further academic development of repeaters. Notably, the effectiveness of the measure varies by country: it appears most beneficial in systems with less intense tracking, such as the United States, and least effective in countries like Germany, Belgium, Luxembourg, and Uganda, where tracking, in the form of strict institutional differentiation, is extensively used as a response to student heterogeneity (Goos et al., 2021, p. 9). Research from Germany aligns with this comparative literature: While Klapproth et al. (2016) observe short-term academic benefits of grade retention for repeaters, their study, along with others (Ehmke et al., 2010, 2017; Fabian, 2020), finds no medium- or long-term improvements in grades or competencies. Additionally, research from both the United States (Locke & Sparks, 2019; National Research Council, 1999, p. 122) and Europe (Pipa et al., 2025, p. 15) indicates that grade retention reinforces educational inequality.

While there is thus ample and convincing evidence on the limited benefits of grade retention for retained students (e.g., Valbuena et al., 2021), including for Germany (e.g., Fabian, 2020), the literature on how this practice affects the academic performance of promoted classmates is very thin. A recent U.S. working paper by Hwang and Koedel (2025) is the only study to date that directly examines this question, leveraging plausibly exogenous variation generated by a third-grade retention policy change in the state of Indiana. They find that promoted fourth-grade cohorts exposed to higher levels of classmate retention had fewer disciplinary incidents (concentrated among boys), while girls showed higher achievement in English language arts. However, these advantages were substantively small and faded out after one year, and no consistent effects were

found for math achievement or unexcused absences. Overall, their findings thus suggest that grade retention does not improve the long-term outcomes of promoted students. Limitations of their analysis include the reliance on school-by-grade data without classroom identifiers, which may lead to an understatement of classroom-level effects (Hwang & Koedel, 2025, p. 2). In addition, although they can follow promoted fourth-grade cohorts over multiple years, they cannot account for other instances of grade retention occurring in earlier or later grades. We contribute to this literature by providing evidence from a different institutional context (Germany) using data that allow us to identify classrooms and to analyze school grades and competencies separately for Grades 1–9, while accounting for instances of grade repetition across all observed grade levels.

To better understand potential effects on promoted students, we next consider the small, related literature on how students perform in classrooms that receive a retained peer. In the United States, a few studies suggest that the addition of a grade repeater negatively affects classmates' academic outcomes (Bietenbeck, 2020; Gottfried, 2013b; A. J. Hill, 2014) and increases their rates of unexcused absences (Gottfried, 2013a), whereas Hwang and Koedel (2025) find no such evidence for third graders. At the same time, Bietenbeck (2020) finds that repeaters positively impact their classmates in the long run by improving their non-cognitive skills, such as their discipline and effort, and increasing their likelihood of graduating from high school and taking a college entrance exam. Using data from China, Xu et al. (2022) report that low-achieving peers—defined as those retained in primary school—negatively affect both cognitive and non-cognitive outcomes of their classmates. Given these mostly negative effects in receiving classes, one might expect grade retention to benefit the former classmates in the promoted class. However, the effects of adding versus losing a repeater may not be symmetrical. One reason is that repeaters may be more

closely aligned with the achievement level of students in the receiving class compared to those in the promoted class, as they have already been exposed to the material.

To build expectations about how retaining classmates may affect promoted students, we present two competing theoretical perspectives from prior educational research in the next sections: the “bad apple” and “rainbow” perspectives.

Potential Benefits of Removing the “Bad Apple”

Why might classmate retention benefit promoted students? Lazear's (2001) “bad-apple” model of peer effects posits that students’ learning progress is influenced by the academic abilities and behaviors of their peers (Lazear, 2001, pp. 779–780). In this view, academically low-performing students can negatively influence their peers' academic outcomes, as they require more teacher attention, which reduces the time available to support others, and collaborations with low-achieving students may be less productive (Sacerdote, 2011, p. 255). The “bad apple” view also worries about misbehaving classmates who may hinder the learning progress of all students as they disrupt others' tasks, instigate disruptive behavior among their peers, and potentially add psychological stress for teachers (Sacerdote, 2011, p. 255). This is consistent with research showing that exposure to low-achieving (e.g., Lavy, Paserman, et al., 2012; Lavy, Silva, et al., 2012) and disruptive students (e.g., Carrell & Hoekstra, 2010; Neidell & Waldfogel, 2010), as well as lower average peer achievement (e.g., Hanushek et al., 2003; Zimmer & Toma, 2000), negatively influences individual students’ academic achievement. Since low academic achievement and misbehavior are among the strongest predictors of grade retention (Pipa et al., 2025, p. 15), the “bad apple” perspective

would thus assume that removing such students from the classroom improves the academic performance of their peers.

In addition to these peer-composition mechanisms, the mere possibility of retention could incentivize students to increase their effort, suggesting that retention might benefit all students. Such incentive effects may become particularly salient for promoted students when a classmate is retained, as this experience could make the consequences of insufficient academic performance more tangible and thereby strengthen students' motivation to avoid a similar outcome. Yet, the limited existing research on this mechanism provides no or only weak support for such incentive effects (e.g., Belot & Vandenberghe, 2014).

Potential Benefits of Low-Achievers for their Classmates

In contrast, students may also benefit from the presence of low-achieving classmates, meaning that their removal could in fact harm the outcomes of promoted students. For example, questions from low-performing students and the additional explanations teachers provide in response to them may strengthen the foundational understanding of all students. In addition, students could benefit from collaborating with lower-achieving classmates, as explaining content to others enhances one's own understanding. In line with this, the “Rainbow Model” suggests that greater ability heterogeneity within a class benefits all students (Sacerdote, 2011, p. 256). For instance, exposure to a broader range of perspectives can deepen comprehension, while interactions with diverse (including neurodiverse) peers may foster non-cognitive development. This last point aligns with Bietenbeck's (2020) findings, which suggest that having a repeater in the classroom can improve classmates' non-cognitive skills. More broadly, some research has also documented benefits of heterogeneous

learning settings (Lehmann, 2006; Murphy et al., 2017; Scharenberg, 2012b, pp. 251–252). Nevertheless, empirical evidence on whether classroom heterogeneity or homogeneity is more beneficial for students' academic outcomes remains inconsistent (Scharenberg, 2012a, p. 3).

Beyond these compositional mechanisms, social comparison processes may also play a role. From the perspective of the “big-fish-little-pond” effect (Marsh, 1987), the presence of a low-achieving student may benefit the academic performance of their peers. The idea is that students compare their abilities with those of their classmates, and perceiving themselves as more capable can boost their academic self-concept. This, in turn, may enhance motivation and ultimately improve academic performance (e.g., Marsh & Martin, 2011). Therefore, losing a low-achieving peer may harm the academic outcomes of promoted students. Comparable reference-group effects may also arise from the teacher's perspective, with the lowest-achieving student serving as a baseline for evaluating academic performance in the classroom. This matters because teachers often consciously or unconsciously grade in relation to the overall achievement level of the class. Removing a low-achieving student may cause the same level of performance to be perceived as weaker in the new, higher-achieving context (Ingenkamp, 1989; Neumann et al., 2010). As a result, the grades of promoted students may decline following the removal of the low-achieving student.

Research Design

The German Context

In Germany, students typically transition to lower secondary education (*Sekundarstufe I*) after completing Grade 4 in elementary school, with exceptions in the federal states of Berlin and Brandenburg, where the transition usually occurs after Grade 6 (Helbig & Nikolai, 2015, p. 83). The

dominant secondary school tracks, ranked by academic rigor, are *Gymnasium* (higher), *Realschule* (intermediate), and *Hauptschule* (lower secondary track). Grade repetition generally occurs when students fail to meet the academic requirements for promotion to the next grade. It does not usually imply an immediate track change, although repeated non-promotion may eventually lead to a change to a lower track. This study encompasses primary education—attended by approximately 2.71 million students in 2013/2014—and most major school types within Germany’s secondary education system, including *Gymnasium* (2.33 million students), *Realschule* (1.02 million), *Hauptschule* (554,000), and schools with multiple educational tracks (454,000) (Destatis, 2024).²

Data

The data for this study are drawn from the German National Educational Panel Study (NEPS), a longitudinal study that provides information about educational processes and the competence development of students in Germany (Blossfeld & Roßbach, 2019). Each of the NEPS survey cohorts was drawn to approximate national representativeness of the respective student population in Germany.

To examine the effect of retaining classmates on promoted students’ academic performance, it is necessary to observe classes over at least two years to compare academic performance at two distinct points: when the student who is later retained is still part of the class and the following year after their retention. We therefore rely on a longitudinal class identifier (Flores et al., 2019), which is only available for starting cohorts 2³ (Grades 1–4; years: 2013–2015/16) and 3⁴ (Grades 5–9; years: 2010/11–2015). Consequently, our analysis uses starting cohorts 2 (SC2) and 3 (SC3) of the NEPS and covers Grades 1 through 9 (NEPS Network, 2024, 2025). Grades 10 and

above are omitted because the longitudinal class identifier does not cover these grade levels (Flores et al., 2019, p. 8). However, this is less critical for Grades 11 and above, which are not of analytical interest, as upper secondary education in Germany generally follows a course-based structure rather than fixed class groups (KMK, 2026, p. 120). The longitudinal class identifier allows us to follow students' classroom affiliation over time as long as they remain on regular educational trajectories. Once students deviate from such a trajectory—through grade repetition, class or school changes, grade skipping, or early or delayed school enrollment—their subsequent classroom affiliation can no longer be identified. Accordingly, our analyses of promoted students are restricted to those with regular educational trajectories. A related data constraint is that classroom coverage is often incomplete in the NEPS data. Based on teacher-reported class sizes, an average of 62% of students per class in SC2 and 46% in SC3 are not observed, limiting our ability to identify all instances of classmate retention. We assess the implications of this constraint in two ways: first, using minimum detectable effect analyses that incorporate plausible levels of unobserved retention, and second, using robustness checks with more fully observed classrooms. We summarize the results of both approaches and their implications for our estimates in the “Potential Threats to Identification” section.

Combining both starting cohorts, the initial sample includes 17,654 students (SC2: 9,337; SC3: 8,317). Because the longitudinal class identifier does not reliably cover irregular educational trajectories, we excluded 5,110 students for which classroom affiliation could not be identified. Furthermore, 453 students attending integrated comprehensive schools (ICS) were excluded (as explained in endnote 2). The resulting analytic sample comprises 12,091 students (SC2: 5,157; SC3: 6,934) with a total of 40,062 observations, most from the lower secondary level ($n = 22,463$), followed by elementary schools ($n = 17,599$). Within the former, observations are concentrated in

the higher track ($n = 11,294$), followed by the intermediate ($n = 5,514$), lower ($n = 3,060$), and multiple-track schools ($n = 2,595$). Finally, observations with missing information on the outcome variable⁵ as well as singletons—classes in which only one student was observed—were excluded from the analyses. Final sample sizes by analysis are reported in Appendix A, Table A.1, as they vary across the outcomes described next.

Outcome Measures

Our outcome variables are student grades and competencies. Grade point averages (GPA) were calculated using math and German grades as reported by parents (SC2; elementary school) or students (SC3; lower secondary level) based on their respective end-of-year report cards. Where grade information was missing, we used student-reported grades in SC2 and parent-reported grades in SC3 to supplement the data. No grades were collected for Grade 1, and for Grade 9, half-year report cards were used because NEPS did not collect end-of-year grades at that level. GPA was computed only when both math and German grades were available; in addition, we also analyze math and German grades separately. All grades and GPA were coded such that higher values indicate better performance.

Our measures of competencies are based on Item Response Theory (IRT) scales based on weighted maximum likelihood estimation. These scores represent the best point estimates of individual competencies, are comparable across time and cohorts, and are scaled so that higher values reflect better performance (Pohl & Carstensen, 2012, p. 16). This makes them well suited to capturing changes in competence at the individual level (Pohl & Carstensen, 2012, p. 9). Accordingly, our analyses estimate the average effect of classmate retention on a non-retained student, rather

than on the entire promoted class, which is an appropriate approach given that classes are not fully observed (see “Data”). That said, a key challenge with the competence measures in both starting cohorts is that they were rarely collected in consecutive years, which could dilute the results. To address this, only competency measurements taken no more than two school years apart were included in the analysis. With this restriction, the following competencies were considered: math measured in Grades 1, 2, and 4 (SC2) and Grades 5, 7, and 9 (SC3); vocabulary and science measured in Grades 1 and 3 (SC2); and reading measured in Grades 5, 7, and 9 (SC3).

To facilitate comparisons across outcomes and cohorts and aid the interpretation of effect magnitudes, we report all effect estimates in standard deviation units. For each outcome and cohort, we use the standard deviation among students in the analytic sample who had not yet been exposed to classmate retention, measured in the earliest grade in which the outcome is observed. This rescaling changes only the scale of the estimates and does not affect statistical significance. Table 1 reports descriptive statistics for all outcomes on their original scales.

Table 1: Descriptive Statistics of Outcome Measures (Main Analysis Sample)

	N	Mean	SD	Min	Max
GPA					
SC2	7,462	4.95	0.67	1.00	6.00
SC3	18,733	4.29	0.76	1.00	6.00
Math Grade					
SC2	7,503	4.98	0.77	1.00	6.00
SC3	18,790	4.25	0.97	1.00	6.00
German Grade					
SC2	7,490	4.92	0.75	1.00	6.00
SC3	18,789	4.33	0.82	1.00	6.00
Math Competence					
SC2	11,803	2.79	1.63	-3.56	9.50
SC3	10,603	0.76	1.25	-4.65	6.29
Vocabulary/Reading Competence					
SC2	7,190	2.11	0.98	-2.30	6.66
SC3	10,150	0.79	1.34	-3.54	6.33
Science Competence					

SC2	7,292	2.15	1.10	-1.35	6.60
SC3	-	-	-	-	-

Note: SC2 denotes Starting Cohort 2 (elementary school), and SC3 denotes Starting Cohort 3 (lower secondary level). Dashes indicate outcomes not observed for the respective survey cohort. N denotes the number of student-grade observations; individual students may be observed in multiple grades.
Source: NEPS data; authors' calculations.

Treatment and Control Groups

Our treatment group consists of students who had one or more classmates retained following the previous school year, meaning that those classmates are no longer part of their current class. In contrast, the control group includes students who did not (yet) experience any classmates being retained. To identify the treatment and control groups, it is crucial to reliably determine which students repeated a grade and when. For this purpose, we construct a grade level variable by combining reports from parents, students, and survey information, as the original grade level variable in the dataset showed certain limitations with respect to our analytical aims (see Appendix B for details). Based on this self-generated variable, students observed in the same grade for two consecutive years are coded as class repeaters. Given the construction of our self-generated grade-level variable, all identified cases represent actual repeaters (see Appendix B). Using the longitudinal class identifier from Florean et al. (2019), we then identify the classes to which these repeaters belonged, enabling the identification of treatment and control groups.

A few additional restrictions apply. Non-retained students who left the promoted class, such as those who switched to another class, were excluded from the treatment and control groups once they entered a new class, as their subsequent class membership could not be identified (Florean et al., 2019). Consequently, our analyses are restricted to promoted students with regular educational trajectories (see “Data” section). Additionally, students who had just entered elementary school (1st grade) or the lower secondary level (5th or 7th grade) cannot, by design, be part of the

treatment group, as they could not yet have experienced the retention of a classmate in their new class. Lastly, students who repeated more than once could not contribute additional treatment events after their first repetition because their subsequent class affiliation could not be identified. This concerns about 6% of all observed repeaters, but their first repetition remains included in the treatment definition. We discuss potential bias from unobserved retention events in the “Potential Threats to Identification” section.

Overall, 551 instances of grade retention could be used to define treatment exposure. Appendix C provides further details on the distribution of treatment and control observations across grade levels, school types, and model specifications, which are described in the following Analytical Strategy section.

Analytical Strategy

The estimand of our study is the effect of classmate retention on the academic performance of promoted students. To illustrate our estimation strategy, consider the canonical two-period difference-in-differences (DiD) setup, in which the average treatment effect on the treated (ATT) can be written as

$$ATT = (Y_{post}^T - Y_{pre}^T) - (Y_{post}^C - Y_{pre}^C), \quad (1)$$

where Y denotes students’ academic performance. The superscript T refers to promoted students whose class experiences a classmate’s retention between the pre- and post-periods (treatment group), so that the post-period is their first observation after treatment. The superscript C refers to promoted students whose class does not experience a classmate’s retention in either the pre- or the

post-period (control group), although they may experience such a retention in a later period. This canonical two-period DiD setup illustrates the core idea of our estimation strategy.

In our empirical setting, however, students are observed across multiple school grades and may experience classmate retention at different points in time, including more than once. To account for this staggered and potentially repeated treatment, we rely on recent advances in DiD methodology and use heterogeneity-robust estimators (Callaway & Sant’Anna, 2021; de Chaisemartin & D’Haultfœuille, 2026), as described below. Using these estimators, promoted students who experience the retention of a classmate between two consecutive grades are compared to promoted students who have not yet experienced such an event over the same period. The resulting treatment-effect estimates are aggregated separately within elementary school (SC2; Grades 1–4) and lower secondary education (SC3; Grades 5–9) as follows:

$$\hat{\theta}_s = \sum_k w_{k,s} \hat{\tau}_k. \quad (2)$$

Here, s denotes the educational stage (SC2 or SC3), k indexes the underlying treatment-effect estimates, $\hat{\tau}_k$ denotes an underlying treatment-effect estimate, and $w_{k,s}$ the estimator-specific weight assigned to that estimate. Thus, $\hat{\theta}_s$ denotes the aggregated treatment-effect estimate for educational stage s .

To obtain the underlying DiD estimates that enter this aggregation and to assess robustness, we apply two model specifications. In the cumulative specification, each instance in which a student experiences classmate retention is treated as a separate exposure; for this specification, we apply the approach by de Chaisemartin and D’Haultfœuille (2026), implemented in Stata 18 via *did_multiplt_dyn*, and report the average cumulative (total) effect per unit of treatment (de Chaisemartin & D’Haultfœuille, 2026, pp. 871–872). In the non-cumulative specification, students

are permanently classified as treated once a classmate from their class is retained; here, we use the estimator by Callaway and Sant’Anna (2021), implemented in Stata 18 via *csdid*, and report the overall ATT based on treatment-group aggregation (Callaway & Sant’Anna, 2021, p. 211, Eq. 3.11). Across both specifications, individual- and time-specific heterogeneity (e.g., stable student characteristics or grade-level trends) are differenced out, while avoiding the well-documented shortcomings of classical two-way fixed effects estimators in staggered settings (Roth et al., 2023). In addition, time-invariant class- and school-specific heterogeneity are differenced out by design, as the analytic sample excludes students who change classes or schools (see “Data”). We include no additional covariates in the main analysis, and standard errors are clustered at the class level. Finally, all analyses are unweighted. Given that there is no universally preferred approach to survey weighting for causal effect estimation (Solon et al., 2015), we re-estimate the full set of analyses using different NEPS survey weights in Appendix F; the weighted estimates are substantively similar.

To clarify the identification logic underlying our approach, it is important to note that observational studies on grade retention that compare retained and promoted students typically face endogeneity problems (Valbuena et al., 2021, pp. 413–414). Our design avoids this particular source of bias by comparing only promoted students: those who lose a classmate due to retention (treatment group) are compared with promoted students who have not or not yet experienced such a loss (control group). Identification therefore relies on the assumption that, absent classmate retention, the two groups would have followed parallel outcome trends. This assumption is arguably more plausible in our setting because a classmate’s retention is not directly determined by the focal student’s own academic performance. Because students in the analytic sample remain in the same

class and school (see “Data”), within-student differencing also eliminates time-invariant differences across classes and schools, including persistent differences in retention cultures. However, time-varying classroom-level factors could still jointly affect classmate retention and promoted students’ outcomes, for example through changes in classroom composition or teacher assignment. We examine such threats in multiple sensitivity analyses (see “Threats to Identification” section).

Empirically, the parallel trends assumption of the DiD approach cannot be directly tested because it concerns an unobserved counterfactual; therefore, pre-treatment trends can only provide indirect evidence. In our case, these diagnostics are further limited by the availability of only one to three pre-treatment event-time periods, depending on the outcome. That said, we find no differential trends for most outcomes, except for mathematics competence in SC2 and GPA in SC3 (see Appendix A, Table A.2). Overall, we consider the parallel trends assumption plausible primarily on theoretical and design-based grounds and take the two suggestive pre-treatment differences into account when interpreting the corresponding estimates in the Results section.

Results

Descriptive Results

We begin by describing differences in academic performance between students who experienced the retention of a classmate (treatment group) and those who had not yet experienced such an event (control group). Figure 1 summarizes average group differences across grade levels, calculated as zero-order differences in outcomes (treatment minus control). Across most outcomes and grades, students in the treatment group perform similarly to or slightly worse than students in the control

group. These differences tend to be more pronounced at the lower secondary level (SC3). In elementary school (SC2), most differences are not statistically distinguishable from zero; the sole exception is mathematics grades in Grade 2, for which treatment-group students perform better than control-group students. GPA differences in Grade 2 point in the same direction but are not statistically significant. Importantly, these comparisons are purely descriptive and do not account for any unobserved heterogeneity across students or over time. We therefore turn to our DiD estimates to assess the effect of classmate retention on the academic performance of promoted students.

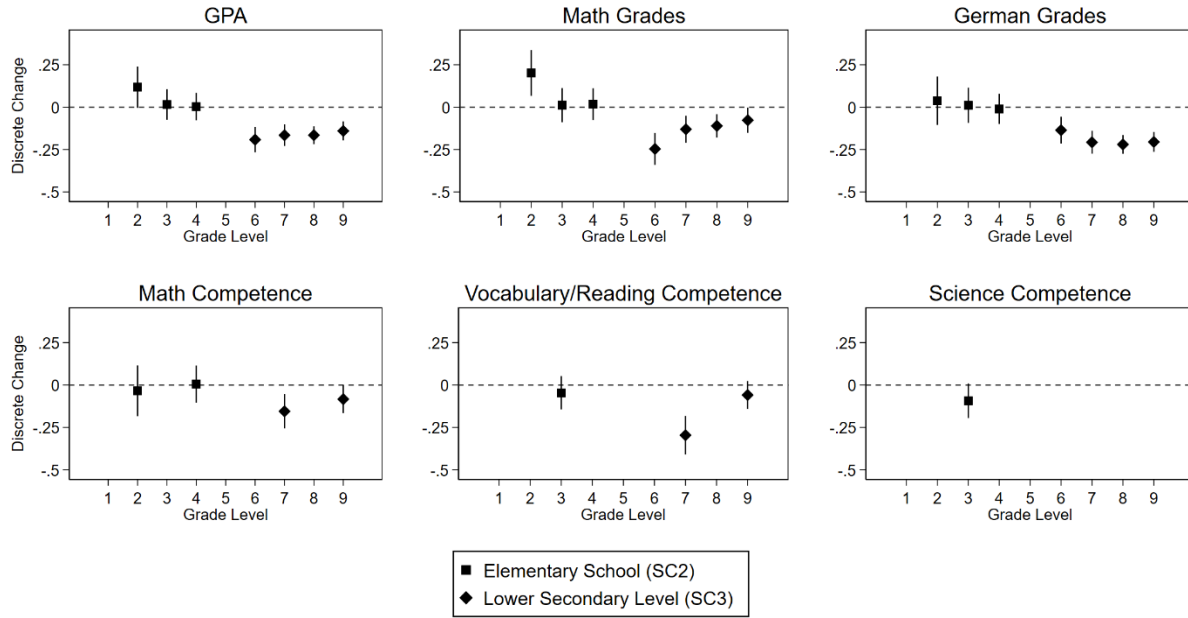


Figure 1: Average Group Differences (Treatment – Control) by Grade Level

Note: Negative values indicate worse performance of the treatment group relative to the control group. Point estimates and 95% confidence intervals are based on unadjusted group-level differences and do not account for any unobserved heterogeneity across students or over time; standard errors are not clustered.

Source: NEPS data; authors' illustration.

Difference-in-Differences Results

Figure 2 presents the aggregated treatment-effect estimates from the cumulative and non-cumulative specifications described in the Analytical Strategy section. Positive estimates indicate improved performance among promoted students following the retention of a classmate relative to the comparison group, whereas negative estimates indicate worse performance. In elementary school (SC2), all estimated effects are small, with eight of the twelve point estimates being negative. The largest estimates indicate a decline in German grades of approximately 0.03 standard deviations in the cumulative and 0.04 standard deviations in the non-cumulative specification following the retention of a classmate; however, none of the SC2 estimates are statistically distinguishable from zero at the 5% level. In lower secondary education (SC3), the estimates for German grades and both competency measures are likewise not statistically significant. By contrast, mathematics-grade estimates are small, positive, and statistically significant in both the cumulative (0.09 SD; $p < .01$) and non-cumulative (0.08 SD; $p < .01$) specifications. Given the multiple outcomes and specifications examined, this isolated positive finding should be interpreted cautiously. GPA estimates in SC3 are also positive (cumulative: 0.05 SD; non-cumulative: 0.06 SD) but do not reach the conventional 5% significance level ($p = .054$ and $p = .053$, respectively). As GPA is constructed from mathematics and German grades, these estimates are driven by the positive mathematics-grade results. The pre-treatment diagnostics suggest differential pre-treatment patterns for GPA in SC3 and mathematics competence in SC2 (see Appendix A, Table A.2), although these diagnostics are limited by the small number of available pre-treatment periods. Interpreting the estimates for these two outcomes cautiously does not alter the overall pattern of results. Even when these outcomes are set aside, 16 of the remaining 18 estimates are not statistically distinguishable from zero at the 5% level.

Taken together, the main analyses predominantly yield null findings and fail to provide sufficient evidence that classmate retention improves promoted students' academic performance. Mathematics grades in SC3 are the only outcome showing statistically significant positive effects, amounting to 0.09 SD in the cumulative and 0.08 SD in the non-cumulative specification. Exact estimates on their original outcome scales are reported in Appendix A, Table A.3.

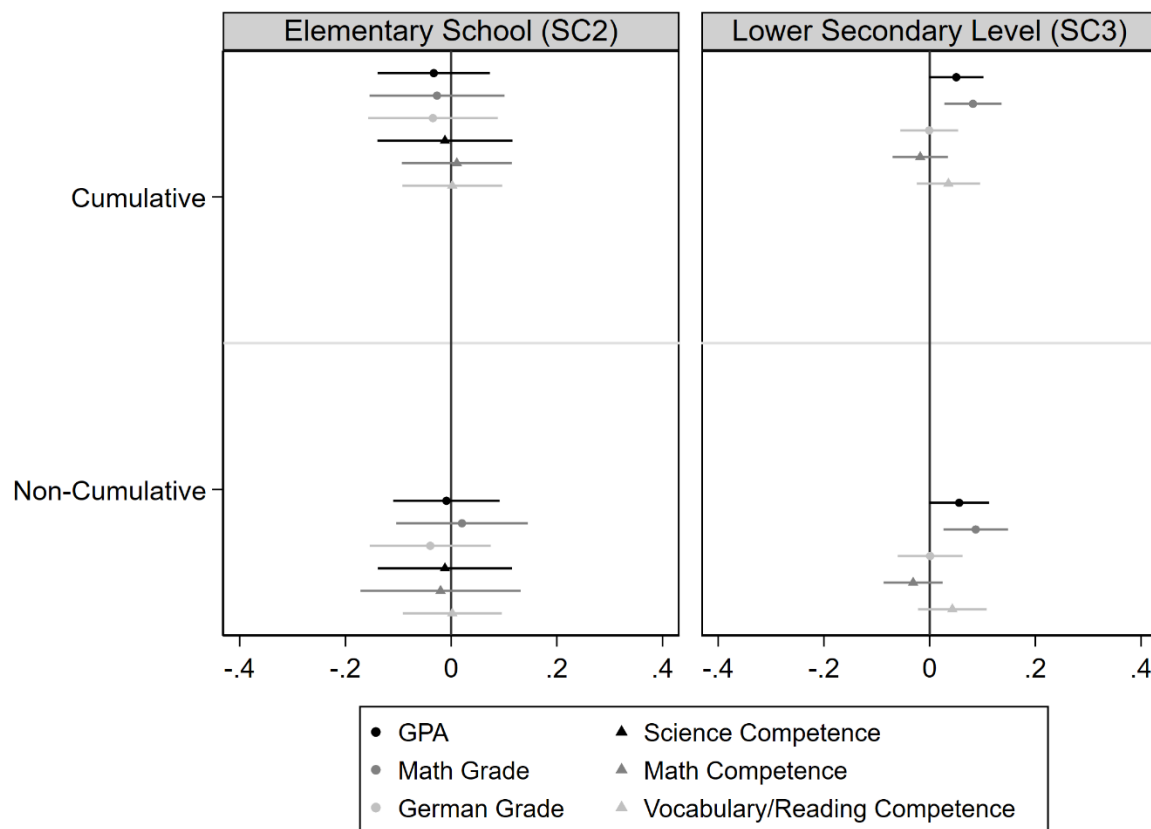


Figure 2: Effects of Classmate Retention on Promoted Students' Academic Performance

Note: Each marker shows an aggregated treatment-effect estimate. Horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

As an exploratory analysis of potential effect heterogeneity, we re-estimated the SC3 models separately for the academic and non-academic secondary tracks (see Appendix D for details). Most estimates are not statistically distinguishable from zero, but the overall pattern suggests less favorable effects in the non-academic tracks, particularly for mathematics competence (-0.11 to -0.14 SD; $p < .01$). By contrast, mathematics-grade estimates are positive and similar in magnitude across track groups but reach statistical significance only in the academic track. Given the smaller subgroup samples, we interpret these patterns as suggestive rather than strong evidence of effect heterogeneity. Overall, the track-specific analyses reproduce the predominantly null pattern of the main analysis and therefore do not alter its substantive conclusion, although the more favorable pattern of point estimates in the academic track provides some suggestive evidence of possible effect heterogeneity.

Potential Threats to Identification

While we believe that the DiD approach applied in the main analyses provides the most appropriate strategy available with our data for estimating the effects of classmate retention on promoted students, we have also considered additional sources of potential bias that may threaten identification. Specifically, we have conducted robustness checks to address concerns related to changes in classroom composition unrelated to grade retention (see Appendix E.1), variation in teacher assignment (see Appendix E.2), and potential confounding due to class size (see Appendix E.3). Across these robustness checks, point estimates vary somewhat, but mostly remain statistically indistinguishable from zero at the 5% level. The robustness analyses therefore reproduce the main substantive

pattern of predominantly null findings and provide little indication that the overall results are driven by these potential confounders.

A separate issue concerns the reassignment of retained students. The NEPS data do not allow us to identify the classes that repeaters enter after retention. It is therefore possible that a promoted class that loses one retained student subsequently receives another repeater, potentially offsetting some of the effect of the former student's removal. The incoming repeater, however, has already completed the respective grade once and may therefore differ academically from the student who was retained from the promoted class. More importantly, such reassignment is an inherent consequence of grade retention: retained students must enter another class. Our estimates should therefore be interpreted as capturing the net peer effect of classmate retention in realized school settings, rather than the isolated effect of removing a particular student while holding the rest of the classroom composition constant. If any benefits from removing one classmate are systematically offset by the arrival of another repeater, this itself limits the extent to which grade retention benefits promoted students in practice.

The most consequential remaining limitation is bias toward zero due to incomplete classroom observation. On average, approximately 62% of students per classroom in SC2 and 46% in SC3 are not observed in NEPS, meaning that some classes classified as untreated may in fact have experienced the retention of an unobserved classmate. This treatment misclassification is one-sided: observed retention events allow us to identify treated classes, whereas unobserved events can cause truly treated classes to enter the control group. Such contamination attenuates the treatment–control contrast toward zero. We assess this concern in two complementary ways. First, we progressively restrict the analyses to more fully observed classrooms (see Appendix E.4.1). Across our three increasingly restrictive specifications, none of the estimates reach statistical significance

at the 5% level, and no consistent pattern of beneficial effects emerges, although the restrictions substantially reduce sample size and precision. Second, we quantify how plausible levels of unobserved retention affect the statistical sensitivity of the main analyses (see Appendix E.4.2). Based on teacher-reported class sizes and official retention rates, the sensitivity analysis implies that approximately 17% of control observations in SC2 and 16–25% in SC3 may in fact be treated. Accounting for this contamination increases the outcome-specific minimum detectable effect sizes from 0.13–0.22 SD to 0.16–0.26 SD in elementary school (SC2) and from 0.08–0.09 SD to 0.10–0.12 SD in lower secondary education (SC3). Thus, incomplete classroom observation reduces our ability to detect smaller effects, particularly in elementary school, but the analyses remain capable of detecting effects above the adjusted outcome-specific thresholds. For the most precisely estimated outcomes, true effects of about 0.10 SD in SC3 and 0.16 SD in SC2 would still have an 80% chance of being detected as statistically significant at the 5% level despite the estimated attenuation. In SC3, this threshold is close in magnitude to the positive mathematics-grade estimates of 0.08–0.09 SD observed in the main analysis. At the same time, mathematics grades are the only outcome showing statistically significant positive effects across the 22 main estimates, so this isolated finding may also be a chance finding. Taken together, the overall pattern of estimates and the sensitivity analyses suggest that the true effects are either null or, if present, generally small and difficult to detect reliably with our data. The implications of this pattern for policy evaluation are discussed in the next section.

Discussion and Conclusion

This contribution examines whether retaining students affects the academic performance of their non-retained classmates who are promoted to the next grade, as measured by grades and competencies. Given the empirically well-founded skepticism about the benefits of grade retention for retained students themselves (Valbuena et al., 2021), particularly in heavily tracked education systems (Goos et al., 2021), and considering the high financial costs associated with this practice (Klemm, 2009; West, 2012), potential benefits for promoted, non-retained classmates would need to serve as justification for its continued use in such contexts. Our analyses provide no broad evidence for such benefits. The main analyses predominantly yield null findings, with small positive effects on mathematics grades in lower secondary education as the only statistically significant exception. The exploratory track-specific analyses likewise reproduce this predominantly null pattern, although point estimates are generally more favorable in the academic than in the non-academic tracks, suggesting possible effect heterogeneity. Overall, our findings provide little support for presumed peer benefits as a justification for grade retention.

By extension, our findings also offer little support for the “bad-apple” hypothesis (Lazear, 2001), according to which removing low-achieving classmates should improve the academic outcomes of promoted students. Our study instead challenges claims that student performance improves when the share of low-achieving classmates is reduced (e.g., Lavy, Paserman, et al. 2012), or average peer ability increases (e.g., Hanushek et al. 2003). Instead, our results align more closely with studies reporting no measurable peer effects of low-achieving classmates (e.g., Auestad 2018). A complementary perspective, partly derived from the “rainbow model” (Sacerdote, 2011, p. 256), holds that low-achieving students can positively contribute to their classmates’ learning and may lower teachers’ grading standards. Our results provide limited support for this view as

well, as we mostly find null effects. One possible explanation for our findings is that removing one or a few low-achieving students is too small an intervention to meaningfully change classroom dynamics. Another is that such students are less detrimental to their peers than some prior studies suggest (e.g., Lavy et al., 2012). While our data cannot disentangle these mechanisms, both are consistent with our overall finding of predominantly null effects and no consistent evidence of academic benefits for promoted peers.

At first sight, our conclusions for Germany may appear to diverge from those based on studies from the United States (Bietenbeck, 2020; Gottfried, 2013a, 2013b; A. J. Hill, 2014) and China (Xu et al., 2022). While we mostly observe null effects of losing a classmate due to grade retention, these studies predominantly report negative spillover effects when a class gains a repeating student. However, these diverging findings may simply reflect asymmetrical effects between losing and receiving a repeater. Like us, in ongoing work, Hwang and Koedel (2025) find no lasting effects of grade retention on the academic outcomes of either promoted cohorts (rising fourth graders) or lower cohorts that gain repeaters (rising third graders). These results are thus consistent with our findings for Germany. Nonetheless, as our study and Hwang and Koedel (2025) are the only ones to examine the effects of classmate retention on promoted students, further research is needed, including from other institutional contexts, to test the robustness of these findings.

As with any observational study, our findings are subject to limitations (discussed in detail in the Data and Potential Threats to Identification sections). We highlight two key limitations here. First, due to constraints in the class identifier used in this research, our analyses are restricted to children who started school at the typical age and followed regular educational trajectories (e.g., without track changes, class switches, or grade skipping). This limits the generalizability of our findings to non-retained students with uninterrupted and “typical” educational trajectories. Second,

unlike Hwang and Koedel (2025), who observe full grade cohorts but lack classroom identifiers, our data include classroom assignments but not complete classrooms. As a result, both designs face different identification constraints that may make it more difficult to detect smaller peer effects. For our study, we assess this concern through robustness checks using more fully observed classrooms, which likewise do not reveal a consistent pattern of beneficial effects, although these analyses are substantially less precise. In addition, sensitivity analyses of our main models indicate that incomplete classroom observation increases the minimum detectable effects to approximately 0.16–0.26 SD in elementary school (Grades 1–4) and 0.10–0.12 SD in lower secondary education (Grades 5–9). Thus, while our analyses remain informative about effects of these and larger magnitudes, smaller effects—particularly in elementary school—cannot be ruled out. Taken together with the predominantly null pattern of our estimates, these results suggest that the true effects are either null or, if present, likely below the magnitudes that our data can reliably detect. Importantly, effects below these detection thresholds could be either positive or negative. Whether such small positive effects on promoted students, if present, would be sufficient to justify grade retention is ultimately a normative policy judgment that must be considered alongside the policy’s costs and the lack of consistent academic benefits for retained students. Future studies should aim to combine the respective strengths of our study and that of Hwang and Koedel (2025) by examining class-level effects using data that include both complete classrooms and classroom identifiers where available (such as in large-scale administrative data sources).

Our study contributes to the long-standing critique of grade retention policies in educational research. Already in the late 20th century, the National Research Council (1999, pp. 128–132), along with other scholars (e.g., Jimerson, 2001), cautioned that grade retention is ineffective in improving the academic outcomes of repeaters and may exacerbate educational inequalities

(e.g., Locke & Sparks, 2019). Our findings contribute to this debate from a less-often studied perspective, providing little evidence of academic benefits for other students when a classmate is retained. This finding, combined with the substantial financial costs (Klemm, 2009; West, 2012) and the absence of consistent academic benefits for retained students in many highly stratified education systems (Goos et al., 2021; Valbuena et al., 2021), suggests that the practice of grade retention warrants critical reconsideration. Rather than retaining students for an entire school year, it may be more constructive to provide targeted, subject-specific remedial instruction and psychological support to struggling students. Such individualized interventions could help close learning gaps and foster motivation without the stigmatizing and socially selective consequences associated with grade retention. The fact that several countries, such as Iceland and Norway, practice automatic promotion rather than grade retention illustrates that retaining students is not an educational necessity but one of several possible policy choices.

Research Ethics Statement

The NEPS study is conducted under the supervision of the German Federal Commissioner for Data Protection and Freedom of Information (BfDI) and in coordination with the German Standing Conference of the Ministers of Education and Cultural Affairs (KMK) and—in the case of surveys at schools—the Educational Ministries of the respective Federal States. All data collection procedures, instruments and documents were checked by the data protection unit of the Leibniz Institute for Educational Trajectories (LIfBi). The necessary steps are taken to protect participants' confidentiality according to national and international regulations of data security. Participation in the

NEPS Study is voluntary and based on the informed consent of participants. This consent to participate in the NEPS study can be revoked at any time.

Endnotes

¹ The substantially higher cost estimate for the United States (\$12 billion) compared to Germany (€931 million) reflects both the larger absolute number of retained students and a methodological difference: While West (2012) assumes that each retained student generates the full cost of an additional year of schooling, Klemm (2009) adjusts for the fact that not every additional student results in higher staffing or class formation costs. Notably, Klemm (2009) remains the most recent national estimate of direct retention costs available for Germany.

² Enrollment figures from the 2013/2014 school year are reported because this year falls within the period covered by the data used in this study. As of the 2023/2024 school year, enrollment figures in Germany include 3.10 million students in elementary schools, 2.28 million in *Gymnasium*, 771,000 in *Realschule*, 334,000 in *Hauptschule*, 550,000 in schools offering multiple educational tracks, and 1.15 million in integrated comprehensive schools (ICS) (Destatis, 2024). We include schools offering multiple tracks because, although they house multiple tracks within a single institution, their students continue to be formally assigned to a specific track and are taught in relatively stable class groups across most subjects. Integrated comprehensive schools (ICS), in contrast, differentiate students by subject-specific performance rather than assigning them to fixed educational tracks (Helbig & Nikolai, 2015, pp. 89–96). We exclude ICS from this analysis, since their students generally do not repeat a grade but are instead reassigned to a lower performance

level in the respective subject, and because the internal structuring of ICS by subject-specific performance groups limits their comparability to school types with fixed class-based progression.

³ SC2 (version 11.0.0) includes children first sampled in spring 2011, who were expected to start school in 2012/2013. The sampling strategy involved a two-stage process: first, a random sample of elementary schools was drawn from a complete list and invited to participate; then, a random sample of the kindergartens providing students to these schools was also invited. Children from these kindergartens were tracked as they transitioned to elementary school, with the sample further augmented by including all first-grade classmates and the recruitment of additional elementary schools (FDZ-LIfBi, 2024, pp. 13–15).

⁴ The sampling strategy for SC3 (version 13.0.0) employed a stratified two-stage process: initially, schools were randomly selected from a complete list, followed by the random selection of two 5th-grade classes within each school during the 2010/11 school year (FDZ-LIfBi, 2025, pp. 12–14).

⁵ Regarding missing outcome values, we do not discard the entire case but still include the student in the analyses, provided that at least two valid observations are available. GPA is used here as an example: if a student has a missing value in Grade 6 but valid values in Grades 5, 7, 8, and 9, the valid observations are used in the analysis. In the analytic sample of the lower secondary level (SC3), most students had no missing GPA value (83.13%) or one missing value (14.74%), while cases with two (1.84%) or three (0.28%) missing values were rare. Within the analytic sample for the elementary school level (SC2), students had either one (81.08%) or two (18.92%) missing values, as GPA was not measured in Grade 1 and some students had an additional missing value in Grades 2, 3, or 4. We decided against imputing missing outcome data, as this would only add noise to the estimates (von Hippel, 2007, pp. 84–85).

References

- Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of Research Design Moderates Effects of Grade Retention on Achievement: A Meta-Analytic, Multilevel Analysis. *Educational Evaluation and Policy Analysis*, 31(4), 480–499. <https://doi.org/10.3102/0162373709352239>
- Auestad, M. L. (2018). The effect of low-achieving peers. *Labour Economics*, 55, 178–214. <https://doi.org/10.1016/j.labeco.2018.10.004>
- Bellenberg, G., & im Brahm, G. (2010). Reduzierung von Selektion und Übergangsschwellen. In G. Quenzel & K. Hurrelmann (Eds.), *Bildungsverlierer: Neue Ungleichheiten* (pp. 517–535). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92576-9_24
- Belot, M., & Vandenberghe, V. (2014). Evaluating the ‘threat’ effects of grade repetition: Exploiting the 2001 reform by the French-Speaking Community of Belgium. *Education Economics*, 22(1), 73–89. <https://doi.org/10.1080/09645292.2011.607266>
- Bietenbeck, J. (2020). The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR. *Journal of the European Economic Association*, 18(1), 392–426. <https://doi.org/10.1093/jeea/jvy049>
- Bloom, H. S. (1995). Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs. *Evaluation Review*, 19(5), 547–556. <https://doi.org/10.1177/0193841X9501900504>
- Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process. The German National Educational Panel Study (NEPS)*. Edition ZfE (2nd ed.). Springer VS.

- Callaway, B., & Sant’Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics, Themed Issue: Treatment Effect 1*, 225(2), 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>
- Carrell, S. E., & Hoekstra, M. L. (2010). Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone’s Kids. *American Economic Journal: Applied Economics*, 2(1), 211–228. <https://doi.org/10.1257/app.2.1.211>
- de Chaisemartin, C., & D’Haultfœuille, X. (2026). Difference-in-Differences Estimators of Inter-temporal Treatment Effects. *The Review of Economics and Statistics*, 108(4), 863–880. https://doi.org/10.1162/rest_a_01414
- Destatis. (2024). *Schülerinnen und Schüler an allgemeinbildenden Schulen*. Statistisches Bundesamt (Destatis). <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsindikatoren/schueler-allgemeinbildende-schulen-tabelle.html>
- Ehmke, T., Drechsel, B., & Carstensen, C. H. (2010). Effects of grade retention on achievement and self-concept in science and mathematics. *Studies in Educational Evaluation*, 36(1), 27–35. <https://doi.org/10.1016/j.stueduc.2010.10.003>
- Ehmke, T., Sälzer, C., Pietsch, M., Drechsel, B., & Müller, K. (2017). Kompetenzentwicklung im Schuljahr nach PISA 2012: Effekte von Klassenwiederholungen. *Zeitschrift für Erziehungswissenschaft*, 20(S2), 99–124. <https://doi.org/10.1007/s11618-017-0752-4>
- Fabian, P. (2020). *Leistungskonsolidierung, Leistungssteigerung – oder etwas ganz anderes? Die Effekte einer Klassenwiederholung auf die Leistungsentwicklung* (1st, New eds.). Waxmann.
- FDZ-LIfBi. (2024). *NEPS Starting Cohort 2—Kindergarten. From Kindergarten to Elementary School. Scientific Use File Version 11.0.0* (pp. 1–210). Leibniz Institute for Educational

- Trajectories (LIfBi). https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/11-0-0/NEPS_SC2_DataManual_11-0-0_en.pdf
- FDZ-LIfBi. (2025). *NEPS Starting Cohort 3 – Grade 5 (2010). Paths Through Lower Secondary School. Scientific Use File Version 13.1.0* (pp. 1–210). Leibniz Institute for Educational Trajectories (LIfBi). https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/13-1-0/NEPS_SC3_DataManual_13-1-0_en.pdf
- Flores, D., Hofmann, J., & Kleinert, C. (2019). Longitudinal class identifiers for NEPS Starting Cohort 3: Generation process and application examples. *NEPS Survey Papers No. 52*. <https://doi.org/10.5157/NEPS:SP52:1.0>
- Goos, M., Pipa, J., & Peixoto, F. (2021). Effectiveness of grade retention: A systematic review and meta-analysis. *Educational Research Review*, 34, 100401. <https://doi.org/10.1016/j.edurev.2021.100401>
- Gottfried, M. A. (2013a). Retained Students and Classmates' Absences in Urban Schools. *American Educational Research Journal*, 50(6), 1392–1423. <https://doi.org/10.3102/0002831213498810>
- Gottfried, M. A. (2013b). The Spillover Effects of Grade-Retained Classmates: Evidence from Urban Elementary Schools. *American Journal of Education*, 119(3), 405–444. <https://doi.org/10.1086/669851>
- Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5), 527–544. <https://doi.org/10.1002/jae.741>
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

- Heisig, J. P., & Matthewes, S. H. (2022). No Evidence that Strict Educational Tracking Improves Student Performance through Classroom Homogeneity: A Critical Reanalysis of Esser and Seuring (2020). *Zeitschrift Für Soziologie*, 51(1), 99–111. <https://doi.org/10.1515/zfsoz-2022-0001>
- Helbig, M., & Nikolai, R. (2015). *Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949*. Verlag Julius Klinkhardt. <https://doi.org/10.25656/01:11095>
- Hill, A. J. (2014). The costs of failure: Negative externalities in high school course repetition. *Economics of Education Review*, 43, 91–105. <https://doi.org/10.1016/j.econedurev.2014.10.002>
- Hill, C., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). *Empirical Benchmarks for Interpreting Effect Sizes in Research*. MDRC. https://www.mdrc.org/sites/default/files/full_84.pdf
- Holmes, C. T. (1989). Grade Level Retention Effects: A Meta-Analysis of Research Studies. In L. A. Shepard & M. L. Smith (Eds.), *Flunking Grades: Research and Policies on Retention. Education Policy Perspectives*. (pp. 16–33). Falmer.
- Hwang, N., & Koedel, C. (2025). *The Peer Effects of Grade Retention* [EdWorkingPaper: 25 - 1146]. <https://doi.org/10.26300/be3d-6t30>
- Ingenkamp, K. (1989). Sind Zensuren aus verschiedenen Klassen vergleichbar? In K. Ingenkamp (Ed.), *Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte* (8th ed., pp. 194–201). Beltz.
- Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. In *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed*

- Reviews [Internet].* Centre for Reviews and Dissemination (UK).
<https://www.ncbi.nlm.nih.gov/books/NBK68711/>
- Klapproth, F., Schaltz, P., Brunner, M., Keller, U., Fischbach, A., Ugen, S., & Martin, R. (2016). Short-term and medium-term effects of grade retention in secondary school on academic achievement and psychosocial outcome variables. *Learning and Individual Differences*, 50, 182–194. <https://doi.org/10.1016/j.lindif.2016.08.014>
- Klemm, K. (2009). *Klassenwiederholungen – teuer und unwirksam. Eine Studie zu den Ausgaben für Klassenwiederholungen in Deutschland*. Bertelsmann Stiftung. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/GP_Klassenwiederholungen_teuer_und_unwirksam.pdf
- KMK. (2026). *The Education System in the Federal Republic of Germany 2024/2025: A Description of the Responsibilities, Structures and Developments in Education Policy for the Exchange of Information in Europe*. Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/EN/dossier_en_ebook.pdf
- Lavy, V., Paserman, M. D., & Schlosser, A. (2012). Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers In the Classroom. *The Economic Journal*, 122(559), 208–237. <https://doi.org/10.1111/j.1468-0297.2011.02463.x>
- Lavy, V., Silva, O., & Weinhardt, F. (2012). The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools. *Journal of Labor Economics*, 30(2), 367–414. <https://doi.org/10.1086/663592>
- Lazear, E. P. (2001). Educational Production. *The Quarterly Journal of Economics*, 116(3), 777–803. <https://doi.org/10.1162/00335530152466232>

- Lehmann, R. H. (2006). Zur Bedeutung der kognitiven Heterogenität von Schulklassen für den Lernstand am Ende der Klassenstufe 4. In A. Schröder-Lenzen (Ed.), *Risikofaktoren kindlicher Entwicklung. Migration, Leistungsangst und Schulübergang* (pp. 109–121). VS, Verl. f. Sozialwiss.
- Locke, V. N., & Sparks, P. J. (2019). Who Gets Held Back? An Analysis of Grade Retention Using Stratified Frailty Models. *Population Research and Policy Review*, 38(5), 695–731. <https://doi.org/10.1007/s11113-019-09524-3>
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280–295. <https://doi.org/10.1037/0022-0663.79.3.280>
- Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *The British Journal of Educational Psychology*, 81(1), 59–77. <https://doi.org/10.1348/000709910X503501>
- Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology*, 51, 336–355. <https://doi.org/10.1016/j.cedpsych.2017.09.003>
- National Research Council. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation* (J. P. Heubert & R. M. Hauser, Eds.). The National Academies Press. <https://doi.org/10.17226/6336>
- Neidell, M., & Waldfogel, J. (2010). Cognitive and Noncognitive Peer Effects in Early Education. *The Review of Economics and Statistics*, 92(3), 562–576.

- NEPS Network. (2024). *National Educational Panel Study, Scientific Use File of Starting Cohort Kindergarten* [Data set]. Leibniz Institute for Educational Trajectories (LIfBi). <https://doi.org/10.5157/NEPS:SC2:11.0.0>
- NEPS Network. (2025). *National Educational Panel Study, Scientific Use File of Starting Cohort Grade 5 (2010)*. (Version 13.1.0) [Data set]. Leibniz Institute for Educational Trajectories (LIfBi). <https://doi.org/https://doi.org/10.5157/NEPS:SC3:13.1.0>
- Neumann, M., Milek, A., Maaz, K., & Gresch, C. (2010). Zum Einfluss der Klassenzusammensetzung auf den Übergang von der Grundschule in die weiterführenden Schulen. In K. Maaz, J. Baumert, C. Gresch, & N. McElvany (Eds.), *Der Übergang von der Grundschule in die weiterführende Schule: Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten* (pp. 229–252). Bundesministerium für Bildung und Forschung (BMBF), Referat Bildungsforschung. <https://www.fachportal-paedagogik.de/literatur/vollanzeige.html?FId=906930>
- Pfeffer, F. T. (2015). Equality and quality in education. A comparative study of 19 countries. *Social Science Research*, 51, 350–368. <https://doi.org/10.1016/j.ssresearch.2014.09.004>
- Pipa, J., Pepels, J., Gaitas, S., Goos, M., Peixoto, F., Belfi, B., Meissner, F., & Klapproth, F. (2025). *Summary of predictors and effects of grade retention in Europe*. Maastricht University Press. <https://doi.org/10.26481/mup.rep.alter.2502.en>
- Pohl, S., & Carstensen, C. H. (2012). NEPS Technical Report – Scaling the Data of the Competence Tests. *NEPS Working Paper, No. 14*. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf

- Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2), 2218–2244. <https://doi.org/10.1016/j.jeconom.2023.03.008>
- Sacerdote, B. (2011). Chapter 4 – Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 249–277). Elsevier. <https://doi.org/10.1016/B978-0-444-53429-3.00004-1>
- Scharenberg, K. (2012a). Do secondary students learn more in homogeneous or heterogeneous classes? The importance of classroom composition for the development of Reading achievement in secondary school. *Online Educational Research Journal*, 3(12). <https://doi.org/https://oerj.webspace.durham.ac.uk/wp-content/uploads/sites/249/2021/07/51.pdf>
- Scharenberg, K. (2012b). *Leistungsheterogenität und Kompetenzentwicklung: Zur Relevanz klassenbezogener Kompositionsmerkmale im Rahmen der KESS-Studie*. Waxmann.
- Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What Are We Weighting For? *Journal of Human Resources*, 50(2), 301–316. <https://doi.org/10.3368/jhr.50.2.301>
- Traini, C., Kleinert, C., & Bittmann, F. (2021). How does exposure to a different school track influence learning progress? Explaining scissor effects by track in Germany. *Research in Social Stratification and Mobility, Mechanisms of Educational Stratification*, 76, 100625. <https://doi.org/10.1016/j.rssm.2021.100625>
- U.S. Department of Education. (1999). *Taking Responsibility for Ending Social Promotion: A Guide for Educators and State and Local Leaders* (No. ED430319). U.S. Department of Education. <https://files.eric.ed.gov/fulltext/ED430319.pdf>

- Valbuena, J., Mediavilla, M., Choi, Á., & Gil, M. (2021). Effects of Grade Retention Policies: A Literature Review of Empirical Studies Applying Causal Inference. *Journal of Economic Surveys*, 35(2), 408–451. <https://doi.org/10.1111/joes.12406>
- von Hippel, P. T. (2007). Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology*, 37(1), 83–117. <https://doi.org/10.1111/j.1467-9531.2007.00180.x>
- West, M. R. (2012). Is Retaining Students in the Early Grades Self-Defeating? *Brookings CCF Brief*, No. 49. <https://www.brookings.edu/wp-content/uploads/2016/06/16-student-retention-west.pdf>
- Xu, D., Zhang, Q., & Zhou, X. (2022). The Impact of Low-Ability Peers on Cognitive and Non-cognitive Outcomes: Random Assignment Evidence on the Effects and Operating Channels. *Journal of Human Resources*, 57(2), 555–596. <https://doi.org/10.3368/jhr.57.2.0718-9637R2>
- Zimmer, R. W., & Toma, E. F. (2000). Peer Effects in Private and Public Schools across Countries. *Journal of Policy Analysis and Management*, 19(1), 75–92.

Appendix A: Supplementary Tables

Table A.1: Sample Sizes by Outcome Variable for Main and Supplementary Analyses

	GPA	Math Grade	German Grade	Math Comp.	Vocab./ Reading Comp.	Science Comp.
SC2 (Main Analysis; Main Text, Figure 2)						
Observations	7,462	7,503	7,490	11,803	7,190	7,292
Individuals (total)	2,745	2,755	2,752	4,404	3,595	3,646
Ever-treated individuals	167	170	167	239	322	332
Classes	699	699	700	931	793	794
SC3 (Main Analysis; Main Text, Figure 2)						
Observations	18,733	18,790	18,789	10,603	10,150	-
Individuals (total)	5,319	5,332	5,326	4,323	4,140	-
Ever-treated individuals	1,753	1,761	1,760	1,602	1,499	-
Classes	492	492	492	446	434	-
SC3 (Academic Track Only; Appendix D, Figure D.1)						
Observations	9,892	9,915	9,914	5,702	5,415	-
Individuals (total)	2,724	2,729	2,727	2,311	2,206	-
Ever-treated individuals	810	813	814	771	706	-
Classes	209	209	209	198	191	-
SC3 (Non-Academic Tracks Only; Appendix D, Figure D.1)						
Observations	6,748	6,777	6,776	3,795	3,658	-
Individuals (total)	1,900	1,907	1,902	1,523	1,458	-
Ever-treated individuals	813	818	815	710	671	-
Classes	207	207	207	184	180	-
SC2 ("Matching" Class Sizes Only; Appendix E, Figure E.1)						
Observations	284	286	286	1,806	-	-
Individuals (total)	142	143	143	903	-	-
Ever-treated individuals	7	7	7	11	-	-
Classes	37	37	38	170	-	-
SC3 ("Matching" Class Sizes Only; Appendix E, Figure E.1)						
Observations	1,857	1,863	1,869	-	-	-
Individuals (total)	892	895	897	-	-	-
Ever-treated individuals	94	94	95	-	-	-
Classes	77	77	78	-	-	-
SC2 (Same Teacher Only; Appendix E, Figure E.2)						
Observations	2,286	2,292	2,293	5,182	2,330	2,354
Individuals (total)	868	869	870	2,164	1,165	1,177
Ever-treated individuals	63	63	63	164	125	127
Classes	208	208	208	491	239	242
SC2 (Additional Control for Class Size; Appendix E, Figure E.3)						
Observations	2,360	2,388	2,374	4,482	-	-
Individuals (total)	1,180	1,194	1,187	2,241	-	-
Ever-treated individuals	35	35	35	205	-	-
Classes	355	356	359	492	-	-

SC3 (Additional Control for Class Size; Appendix E, Figure E.3)						
Observations	5,044	5,064	5,077	1,218	1,162	-
Individuals (total)	1,903	1,909	1,913	609	581	-
Ever-treated individuals	347	347	348	181	162	-
Classes	196	196	196	77	75	-
SC2 (Incomplete Classroom Observation Robustness Check 1; Appendix E, Figure E.4)						
Observations	582	582	582	861	418	422
Individuals (total)	244	244	244	364	209	211
Ever-treated individuals	27	27	27	26	26	26
Classes	36	36	36	40	22	22
SC3 (Incomplete Classroom Observation Robustness Check 1; Appendix E, Figure E.4)						
Observations	1,823	1,824	1,830	912	869	-
Individuals (total)	647	647	650	407	386	-
Ever-treated individuals	112	112	113	71	69	-
Classes	30	30	30	19	18	-
SC2 (Incomplete Classroom Observation Robustness Check 2; Appendix E, Figure E.5)						
Observations	344	344	344	509	418	422
Individuals (total)	125	125	125	188	209	211
Ever-treated individuals	21	21	21	12	26	26
Classes	17	17	17	20	22	22
SC3 (Incomplete Classroom Observation Robustness Check 2; Appendix E, Figure E.5)						
Observations	496	498	498	512	475	-
Individuals (total)	180	181	181	207	189	-
Ever-treated individuals	22	22	22	40	38	-
Classes	8	8	8	9	8	-
SC2 (Incomplete Classroom Observation Robustness Check 3; Appendix E, Figure E.6)						
Observations	97	97	97	146	192	188
Individuals (total)	35	35	35	51	96	94
Ever-treated individuals	21	21	21	26	13	13
Classes	4	4	4	5	9	9
SC3 (Incomplete Classroom Observation Robustness Check 3; Appendix E, Figure E.6)						
Observations	389	389	389	318	292	-
Individuals (total)	141	141	141	159	146	-
Ever-treated individuals	24	24	24	24	24	-
Classes	7	7	7	8	7	-

Note: SC2 denotes Starting Cohort 2 (elementary school sample), and SC3 denotes Starting Cohort 3 (lower secondary school sample). Classes are counted uniquely, even when observed across multiple years. Ever-treated individuals are promoted students who experience at least one instance of classmate retention within the respective analysis sample. Sample sizes for the survey-weighted robustness analyses (see Appendix F) are not reported because the full set of analyses is re-estimated using two alternative survey weights; reporting these sample sizes would therefore substantially expand the table.

Source: NEPS data; authors' calculations.

Table A.2: Tests for Differential Pre-Treatment Trends

	Elementary School (SC2)	Lower Secondary Level (SC3)
GPA	p = 0.77 [1]	p = 0.04 [3]
Math Grade	p = 0.28 [1]	p = 0.17 [3]
German Grade	p = 0.79 [1]	p = 0.59 [3]
Math Competence	p = 0.02 [1]	p = 0.11 [1]
Vocabulary/Reading Competence	- [0]	p = 0.98 [1]
Science Competence	- [0]	-

Note: P-values are from tests of the null hypothesis that pre-treatment trends equal zero, estimated using *estat pretrend* in *csdid* (Stata 18). Values are rounded to two decimal places. Numbers in brackets indicate the number of available pre-treatment event-time periods, corresponding to grade-to-grade transitions in which the respective treatment group is still untreated. Dashes indicate outcomes for which no pre-treatment test can be conducted; for vocabulary/reading and science competence (SC2), only two outcome measurements are available, leaving no pre-treatment event-time period for testing differential trends.

Source: NEPS data; authors' calculations.

Table A.3: Difference-in-Differences Estimates – Main Analysis

Outcome	Elementary School (SC2)		Lower Secondary Level (SC3)	
	Cumulative	Non-Cumulative	Cumulative	Non-Cumulative
GPA	-0.021 (0.035)	-0.006 (0.034)	0.037 (0.019)	0.041 (0.021)
Math Grade	-0.020 (0.048)	0.015 (0.047)	0.075** (0.025)	0.080** (0.028)
German Grade	-0.025 (0.046)	-0.029 (0.043)	-0.001 (0.023)	0.001 (0.025)
Science Competence	-0.010 (0.057)	-0.010 (0.057)	- -	- -
Math Competence	0.012 (0.058)	-0.022 (0.085)	-0.020 (0.031)	-0.036 (0.033)
Vocabulary/Reading Competence	0.002 (0.037)	0.002 (0.037)	0.043 (0.037)	0.052 (0.040)

Note: Standard errors are reported in parentheses and are clustered at the class level. Point estimates are reported on the original outcome scales and have not been rescaled to standard-deviation units. Dashes indicate outcomes that are not observed for the respective cohort. All numbers are rounded to three decimal places.

*p < .05. **p < .01. ***p < .001.

Source: NEPS data; authors' calculations.

Appendix B: Further Information on the Identification of Grade Repeaters

In principle, whether a student repeats a class can be inferred from the dataset by using a variable that indicates grade level: if a student remains in the same grade for two consecutive years, they are assumed to have repeated that grade. However, for SC3 (Grades 5–9), a comparison based on the full sample (prior to any case exclusions) reveals that this method substantially underestimates the true number of repeaters. While it identified 502 repetitions, 678 cases of retrospectively self-reported grade repetitions were recorded in later survey waves—despite reduced participation at that point—indicating limitations of the original grade level variable in accurately identifying repeaters. For SC2 (Grades 1–4), the same approach identified 1,037 repeaters, 903 of whom allegedly repeated Grade 2, suggesting potential data errors.

These discrepancies highlight that the grade level information in the dataset has limitations in identifying class repeaters. For this reason, a new grade level variable was constructed for the main analyses to more accurately identify grade repeaters and, based on this information, to define the treatment and control groups. For SC3, this variable was based on retrospective information on grade repetitions and skipping from students, collected from survey wave 8 onward (i.e., when students with regular educational careers are in Grade 11 and higher), and corresponding information from parents, collected closer in time, as well as information on the survey wave. If students or their parents indicated that the student had repeated or skipped a grade level, they were believed and the new variable on the grade level of students was constructed accordingly. If neither repetitions nor skipping were reported, we assumed that students were in the grade level targeted by the respective survey wave. This assumption is justified by the fact that NEPS surveys students primarily in their class context, and only those with irregular educational trajectories (e.g., repeaters, skippers) are individually tracked. For example, in SC3, students surveyed in wave 4 are expected

to be in grade 8 if they follow a regular educational trajectory, as this wave targets eighth-graders (FDZ-LIfBi, 2025, p. 26). The same approach was applied to SC2. However, here the information on grade repetitions and skipping is based solely on parental reports, as no information was available from the children in this starting cohort. Finally, students in both SC2 and SC3 who remained in the same grade for two consecutive school years were coded as grade repeaters. Note that all instances of grade repetition identified in this way represent actual repeaters rather than students observed in the same grade for consecutive years for other reasons (e.g., due to changing school tracks), as the grade level variable itself is based on parent or student reports of grade repetition and grade skipping, as explained above.

Table B.1 compares the original grade level variable from the dataset with the self-generated version (in the full sample before any case exclusions) and data from the German Federal Statistical Office. The self-generated variable identifies a greater number of grade levels than the original (72,612 vs. 56,195). In terms of grade repetition, the average number of repeaters in Grades 3 to 9—the range covered by official statistics—shows that the self-generated variable ($M = 1.33\%$) is marginally closer to the official figures ($M = 2.11\%$) than the original variable ($M = 1.28\%$). Nevertheless, both variables underestimate the actual number of grade repetitions as reported in official statistics, reflecting incompletely observed classrooms in our data. We discuss the implications of this limitation in the “Potential Threats to Identification” section.

Table B.1: Comparison of Grade Level Variables and Repetition Rates in NEPS and Official Statistics

Grades:	Total	1	2	3	4	5	6	7	8	9	Average (G1–9)	Average (G3–9)
Original Grade Level Variable												
N	56,195	5,909	7,240	5,320	4,949	6,164	5,905	7,695	6,989	6,024		
Repetitions	1,539	70	903	48	16	52	84	149	30	187		
% per year	2.74	1.18	12.47	0.90	0.32	0.84	1.42	1.94	0.43	3.10	2.51	1.28
Self-generated Grade Level Variable												
N	72,612	8,715	8,726	8,690	8,669	6,161	6,187	8,536	8,474	8,454		
Repetitions	906	75	95	46	12	57	78	222	169	152		
% per year	1.25	0.86	1.09	0.53	0.14	0.93	1.26	2.60	1.99	1.80	1.25	1.33
Official Statistics on Repetitions per School Year												
Repetitions				8,098	4,076	15,574	12,606	18,433	23,691	30,446		
% per year	-	-	-	1.2	0.6	2.1	1.6	2.4	3.0	3.9	-	2.11

Note: Average (G1–9) includes all grade levels; Average (G3–9) restricts the calculation to Grades 3–9 to ensure comparability with official statistics from the German Federal Statistical Office, which reports repetition rates starting in Grade 3. For official statistics, repetition rates from the corresponding NEPS survey years were used to ensure temporal comparability. NEPS figures are based on the sample prior to any case exclusions.

Source: NEPS, authors' calculations; German Federal Statistical Office (Destatis), Fachserie 11, Reihe 1: Allgemeinbildende Schulen, https://www.statistischebibliothek.de/mir/receive/DESerie_mods_00000110

Appendix C: Distribution of Grade Retentions and of Treatment and Control Groups by Grade Level, Model Specification, and School Type

This section describes the distribution of students in the treatment and control groups across grade levels, model specifications, and school types. The treatment group consists of promoted students who lose a classmate due to grade retention, while the control group includes promoted students who have not (yet) experienced such a loss. It is therefore informative to first examine the distribution of grade retention in the analytic sample. As shown in Table C.1, more than half of the 551 instances occur in Grades 7 and 8 (53.9%), with Grade 7 alone accounting for 32.5%. While this concentration in the later grades is pronounced across all secondary school types, repetitions in elementary school are concentrated in the earlier grades, with Grade 3 accounting for only 19.01% of repetitions.

The 551 instances reported in Table C.1 are fewer than the 906 instances reported in Table B.1 (Appendix B), which includes all repetitions identified when constructing our grade-level variable. For treatment assignment, we exclude repetitions in Grade 4 ($n = 12$) and Grade 9 ($n = 152$), because their potential effects on promoted classmates would only be observed outside the corresponding analytic samples (Grade 5 in SC2 and Grade 10 in SC3; see “Data” section). We further exclude repetitions in integrated comprehensive schools ($n = 5$), cases in which SC3 students were still enrolled in elementary school in Grades 5 or 6 ($n = 5$) (see “The German Context” section), and other repetitions for which classroom assignment could not be tracked ($n = 181$). We discuss the implications of incomplete classroom identification for our estimates in the “Threats to Identification” section.

Turning to the distribution of treatment and control groups, the left panel of Figure C.1, based on the cumulative model specification, shows that students remain in the treatment group

after their initial exposure and may experience multiple exposures if additional classmates are retained in later years. Consequently, the size of the treatment group increases over time, spanning Grades 1 to 4 in the elementary level and Grades 5 to 9 in the lower secondary level. As expected, first-time exposures are most common, followed by second-time exposures, while third and fourth exposures occur only rarely. The non-cumulative model specification (right panel of Figure C.1) yields identical treatment counts, as all exposures are classified as “first”. Figures C.2 and C.3 display the distribution of treatment and control group students across grade levels and model specifications, disaggregated by school type. In lower secondary education, grade repetitions are most frequent in the higher track ($n = 166$), followed by the intermediate ($n = 152$) and lower tracks ($n = 55$), with the fewest cases observed in schools offering multiple tracks ($n = 36$) (see Table C.1). Accordingly, the number of treated students is highest in the higher track and lowest in schools with multiple tracks. Because these figures describe the overall analytic sample prior to outcome-specific exclusions, the number of ever-treated individuals in each main and supplementary analysis is reported separately in Appendix A, Table A.1.

Table C.1: Grade Repetitions Used for Treatment and Control Group Definition, by Grade Level and School Type

Grade Level	1	2	3	5	6	7	8	Total
All School Types								
N	54	61	27	49	63	179	118	551
%	9.80	11.07	4.90	8.89	11.43	32.49	21.42	100
Elementary School								
N	54	61	27	-	-	-	-	142
%	38.03	42.96	19.01	-	-	-	-	100
Schools with Multiple Educational Tracks								
N	-	-	-	2	8	11	15	36
%	-	-	-	5.56	22.22	30.56	41.67	100
Higher/Academic Track (<i>Gymnasium</i>)								
N	-	-	-	14	31	64	57	166
%	-	-	-	8.43	18.67	38.55	34.34	100
Intermediate Track (<i>Realschule</i>)								
N	-	-	-	18	15	86	33	152
%	-	-	-	11.84	9.87	56.58	21.71	100
Lower Track (<i>Hauptschule</i>)								
N	-	-	-	15	9	18	13	55
%	-	-	-	27.27	16.36	32.73	23.64	100

Note: This table reports the number and percentage of grade-retention instances used to define the treatment and control groups, by grade level and school type. Percentages are row percentages and are rounded to two decimal places.

Source: NEPS, authors' calculations.

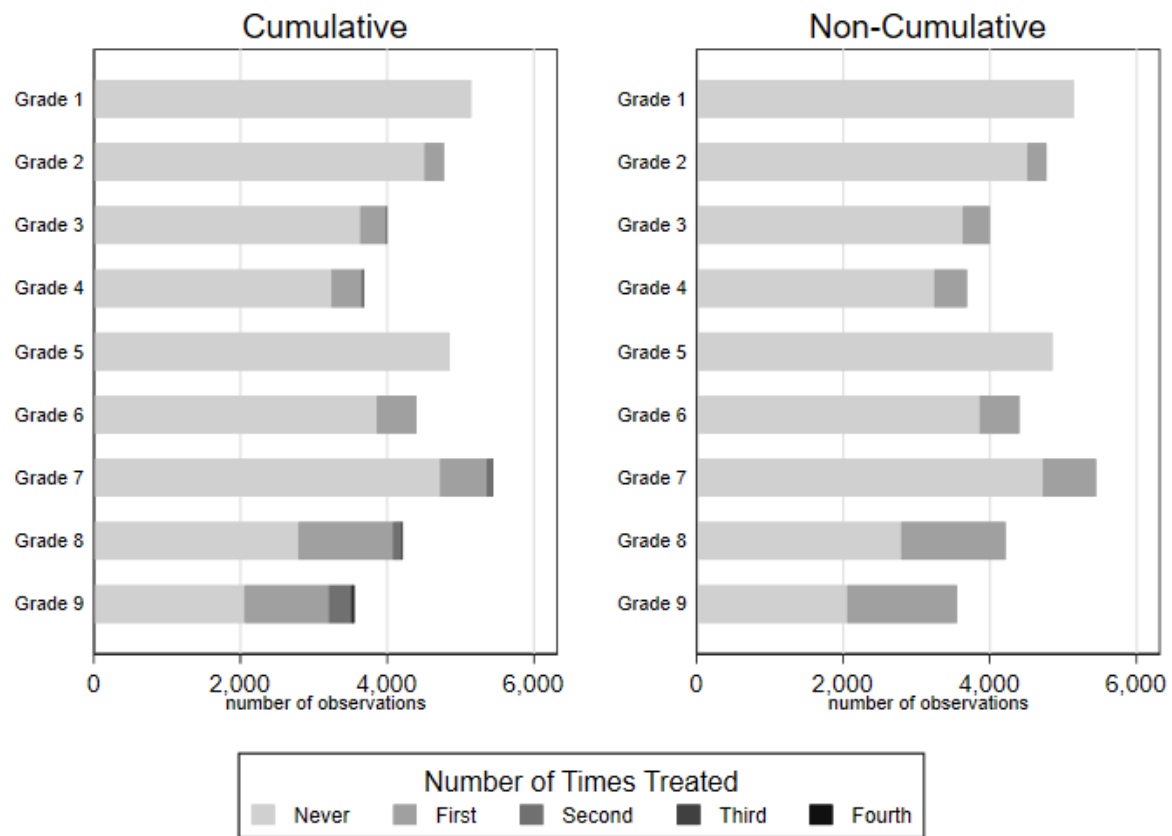


Figure C.1: Distribution of Treatment Exposure by Grade Level and Model Specification

Note: Bars show the distribution of observations by grade level and treatment status under cumulative and non-cumulative specifications, based on the analytic sample prior to excluding singletons and observations with missing outcome data.

Source: NEPS data; authors' illustration.

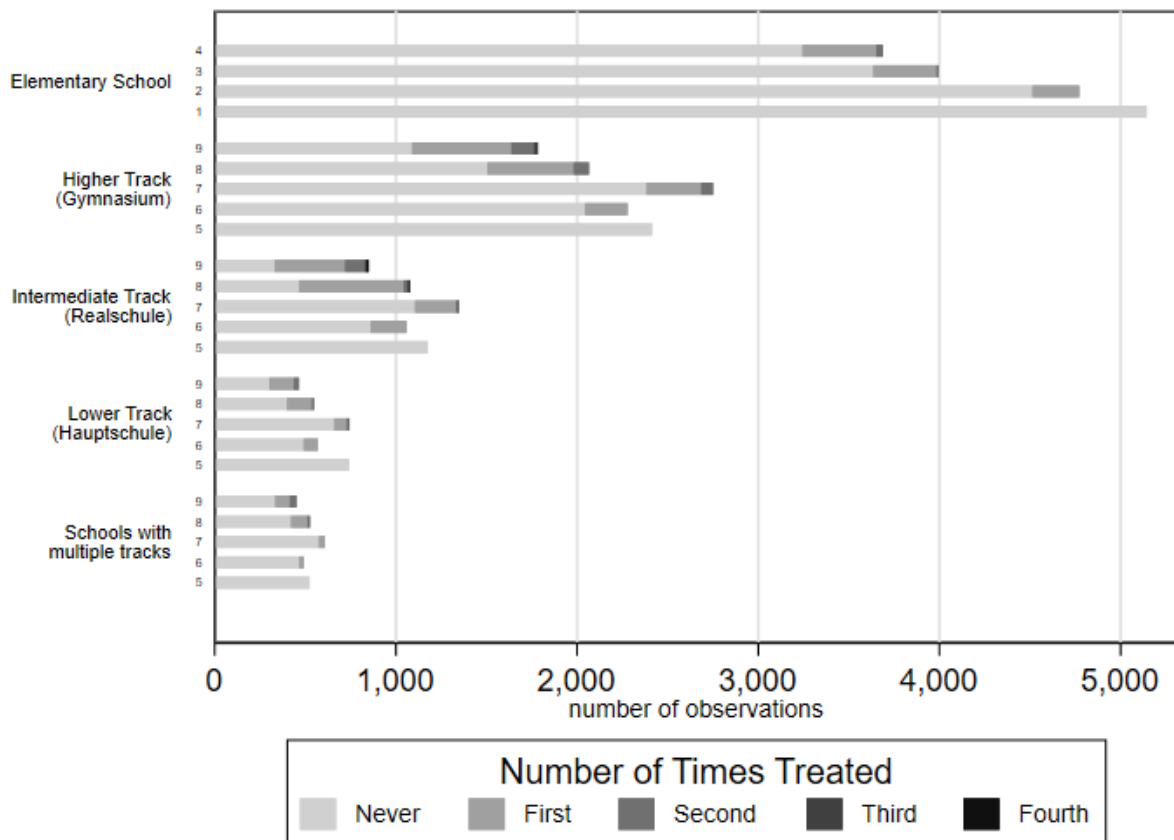


Figure C.2: Distribution of Treatment Exposure by Grade Level and School Type (Cumulative Specification)

Note: Bars indicate the number of observations by grade level and school type, disaggregated by treatment exposure (number of times treated), under the cumulative specification. The figure is based on the analytic sample prior to excluding singletons and observations with missing outcome data.

Source: NEPS data; authors' illustration.

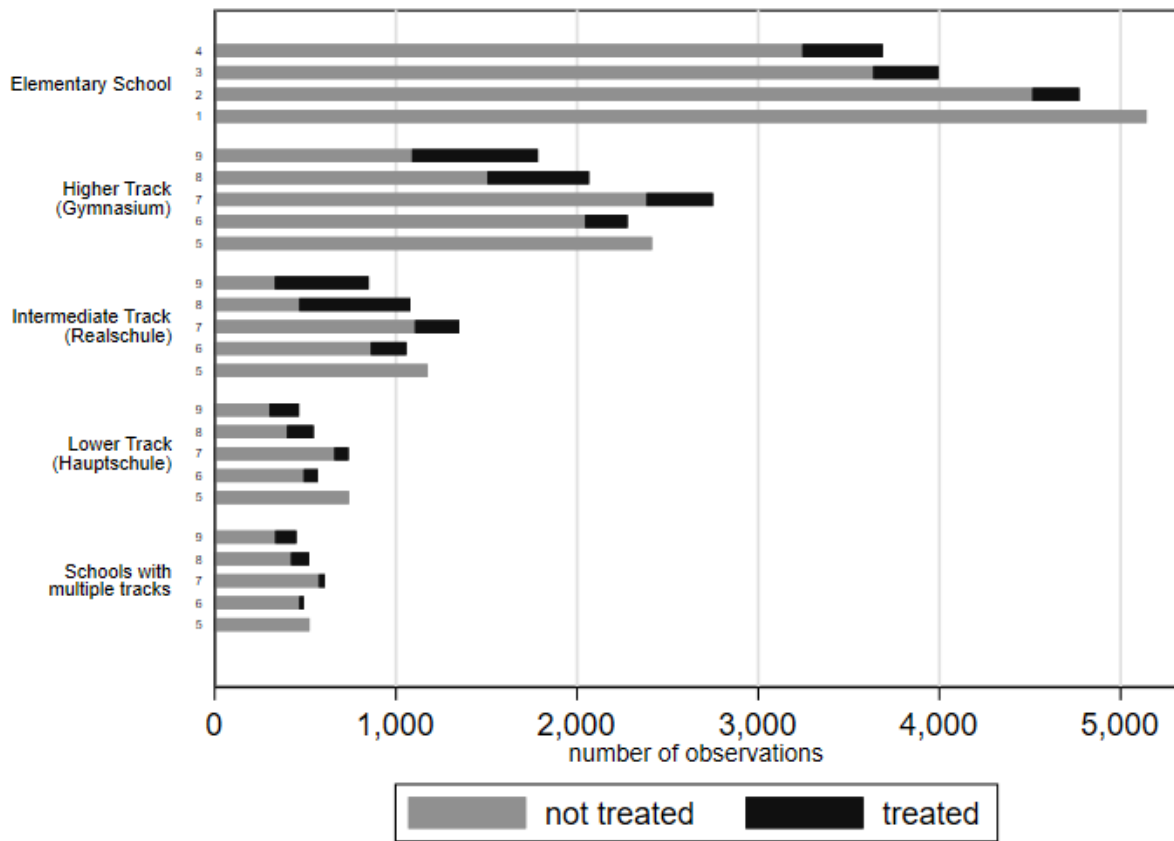


Figure C.3: Distribution of Treatment Exposure by Grade Level and School Type (Non-Cumulative Specification)

Note: Bars indicate the number of observations by grade level and school type, disaggregated by treatment status, under the non-cumulative specification. The figure is based on the analytic sample prior to excluding singletons and observations with missing outcome data.

Source: NEPS data; authors' illustration.

Appendix D: Effect Heterogeneity by Secondary School Track

In the German context, tracking is often argued to create distinct learning environments (e.g., Traini et al., 2021). As an exploratory analysis of potential effect heterogeneity, we therefore re-estimated our main models separately for the academic (higher) and non-academic (intermediate and lower) secondary tracks. Schools with multiple educational tracks were excluded because students' specific track assignment cannot be identified in our data. In addition, splitting the lower-secondary sample reduces sample sizes and statistical precision (see Appendix A, Table A.1); these results should therefore be interpreted more cautiously than those from the main analysis.

Figure D.1 reports the results. Positive estimates indicate improved performance among promoted students following the retention of a classmate relative to the comparison group, whereas negative estimates indicate worse performance. Across both track groups and specifications, 15 of the 20 estimates are statistically indistinguishable from zero at the 5% level. In the academic track, mathematics-grade estimates are positive and statistically significant in both the cumulative (0.09 SD; $p < .01$) and non-cumulative (0.08 SD; $p < .05$) specifications. The cumulative GPA estimate is also positive and statistically significant (0.07 SD; $p < .05$), whereas the corresponding non-cumulative estimate, the German-grade estimates, and all competency estimates are not statistically significant. All point estimates in the academic track are nevertheless positive. In the non-academic tracks, mathematics-grade estimates are similar in magnitude to those in the academic track (cumulative: 0.07 SD; non-cumulative: 0.08 SD) but do not reach conventional levels of statistical significance ($p = .162$ and $p = .154$, respectively). Mathematics competence, by contrast, shows negative effects in both the cumulative (-0.11 SD; $p < .01$) and non-cumulative (-0.14 SD; $p < .01$) specifications. Although the German-grade and reading estimates in the non-academic

tracks are not statistically distinguishable from zero, their point estimates are also negative, in contrast to the positive point estimates for all outcomes in the academic track.

Taken together, the pattern is suggestive of potential effect heterogeneity by secondary school track, with generally less favorable estimates in the non-academic tracks, particularly for mathematics competence. These findings should nevertheless be interpreted cautiously. The track-specific analyses involve substantially smaller samples and do not formally test differences between tracks. Moreover, we have no clear theoretical reason to expect mathematics competence in particular to exhibit stronger heterogeneity than the other outcomes. The difference in statistical significance of the mathematics-grade estimates across track groups should likewise not be interpreted as evidence that the effects themselves differ, given their very similar magnitudes. Finally, Figure D.1 reports 20 estimates across two related specifications, so statistically significant findings should be interpreted cautiously in light of multiple testing. Overall, the results therefore do not provide strong evidence of systematic effect heterogeneity across the full set of grade and competence outcomes, although both the mathematics-competence estimates and the broader pattern of point estimates suggest less favorable effects of classmate retention in the non-academic tracks.

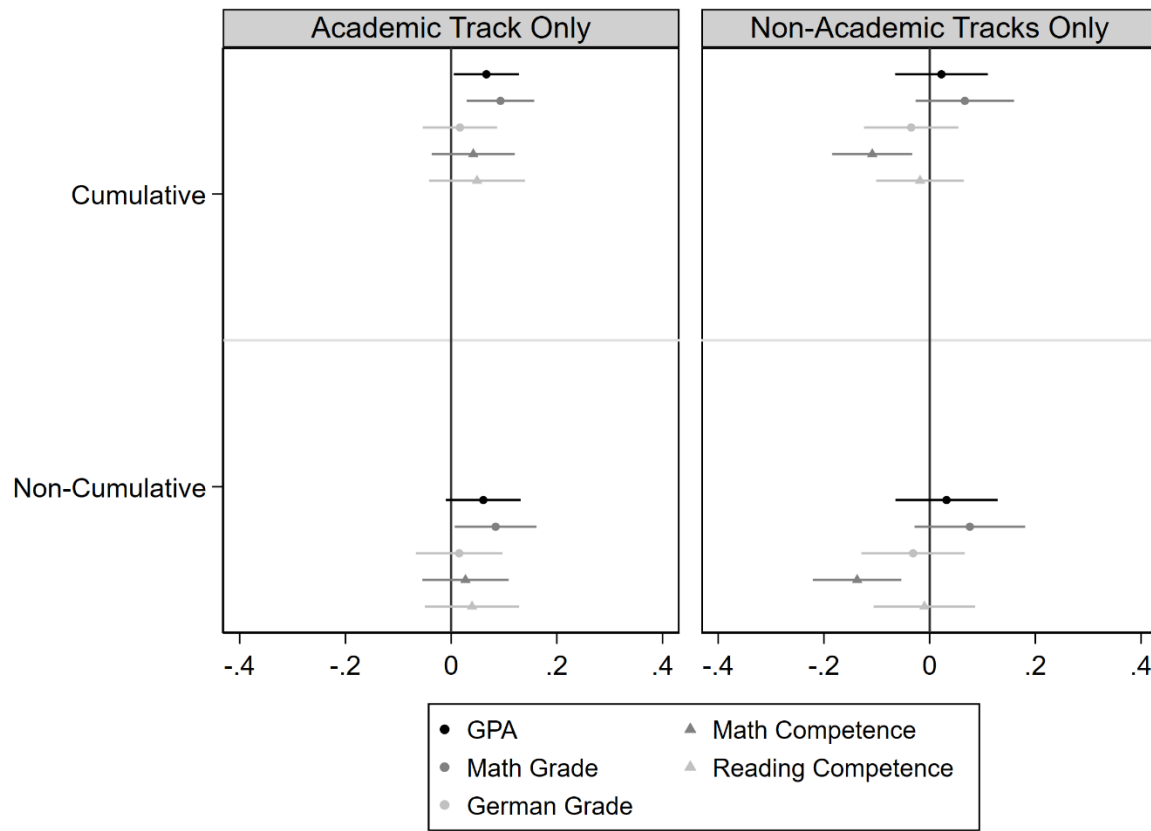


Figure D.1: Effects of Classmate Retention on Promoted Students' Academic Performance by Secondary School Track

Note: Each marker shows an aggregated treatment-effect estimate. Horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

Appendix E: Details on Threats to Identification

Here, we provide a more detailed discussion of the identification threats we report in the main text, along with the robustness checks conducted to assess their potential impact on our results.

E.1 Changes in Classroom Composition Unrelated to Grade Retention

A first concern is that changes in classroom composition unrelated to grade retention may occur—for example, when students switch classes, change school tracks, or join a class due to the merging of same-grade classes. Such changes may bias treatment effect estimates if they influence both the likelihood of a classmate being retained (i.e., treatment assignment) and the academic performance of the remaining students (i.e., the outcome). To address this concern, we conducted a robustness check using a restricted sample in which the number of retained students had to match the change in teacher-reported class size across adjacent grades—including cases with no retention and no change. We rely on teachers' reports of class sizes rather than on the number of observed students per class, as the latter systematically underestimates actual class sizes in our data. The logic of the restriction is as follows: if a Grade 2 class initially had 20 students and one student was retained, the reported class size in Grade 3 should be 19. If no student was retained, the class size should remain at 20. Limiting the sample to cases with such retention-consistent class size changes substantially reduced the number of observations and, consequently, the statistical power of the models (see Appendix A, Table A.1). Competency outcomes were excluded from this analysis because their measurements were taken over a year apart, and enforcing class-size matching across that

time span would have reduced the sample too much. An exception is math competencies in elementary school, which were assessed in both Grades 1 and 2, while Grade 4 measurements were excluded from the analysis.

The results of this robustness check are shown in Figure E.1. Most estimated effects are not statistically significant at the 5% level. One exception is mathematics competence in elementary school, for which both specifications yield a positive effect of approximately 0.45 SD ($p < .001$). This effect is substantively large, corresponding to roughly half of the annual mathematics achievement gain reported by C. Hill et al. (2007, p. 3) between Grades 1 and 2. At the same time, mathematics-grade estimates in elementary school are negative and also substantively sizable, although neither is statistically significant at the 5% level (cumulative: -0.31 SD, $p = .17$; non-cumulative: -0.31 SD, $p = .06$). Although our theoretical discussion allows for the possibility that competencies and grades respond differently—for example, if the removal of a low-achieving student raises the classroom reference standard for grading—the large, opposite-signed estimates for mathematics competence and grades do not constitute a coherent pattern. Given the substantial reduction in sample size and statistical precision in this restricted analysis, we therefore interpret these estimates cautiously and place limited weight on the positive mathematics-competence finding. Taken together, these results are consistent with our overall assessment that there is no consistent evidence of academic benefits from classmate retention for promoted students.

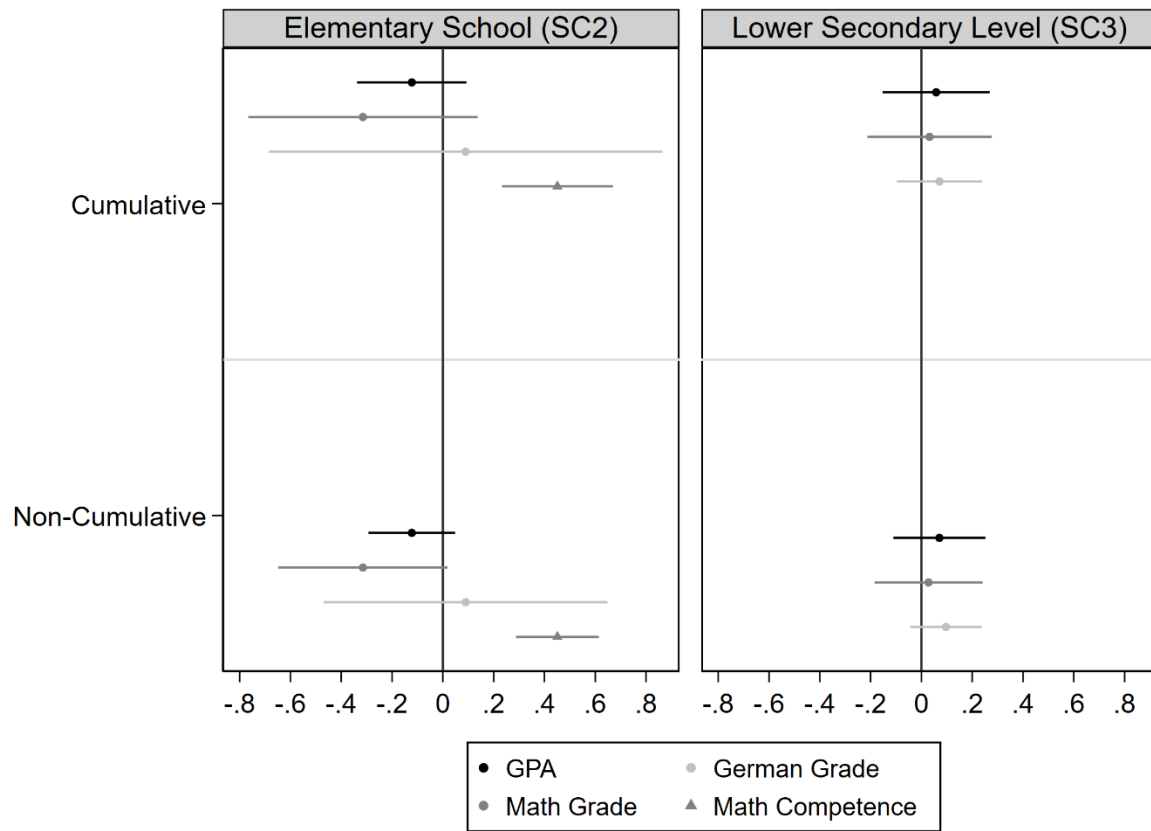


Figure E.1: Effects of Classmate Retention on Promoted Students' Academic Performance Using "Matching" Class Sizes

Note: Each marker shows an aggregated treatment-effect estimate; horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units. The sample is restricted to cases in which changes in teacher-reported class size across adjacent grades are consistent with the number of retained students.

Source: NEPS data; authors' illustration.

E.2 Potential Confounding by Teacher Assignment

Second, teacher assignments may be a potential confounder. Assume that a “bad teacher” increases retention rates and an observant school administrator compensates the class by assigning a more skilled teacher in the next year. That teacher may increase average student performance and therefore lead to a positive bias in our estimates. At the same time, the assignment of a “bad teacher” following a good one may produce the opposite bias. To minimize any potential bias due to teacher assignments following the retention of a classmate, we repeated the main analysis using a restricted sample of students who were taught by the same teacher throughout elementary school (see Figure E.2). This restriction also led to a substantial reduction in sample size (see Appendix A, Table A.1). We limited the robustness check to the elementary level, as students in Germany are typically taught by one teacher for most subjects in elementary school, whereas in secondary school, different teachers are often assigned to each subject.

None of the estimated effects is statistically significant at the 5% level. Although the point estimates are generally positive, they are imprecisely estimated and therefore provide at most weak indications of beneficial effects. Overall, this robustness check is consistent with our assessment that there is no clear evidence of academic benefits of classmate retention among promoted students.

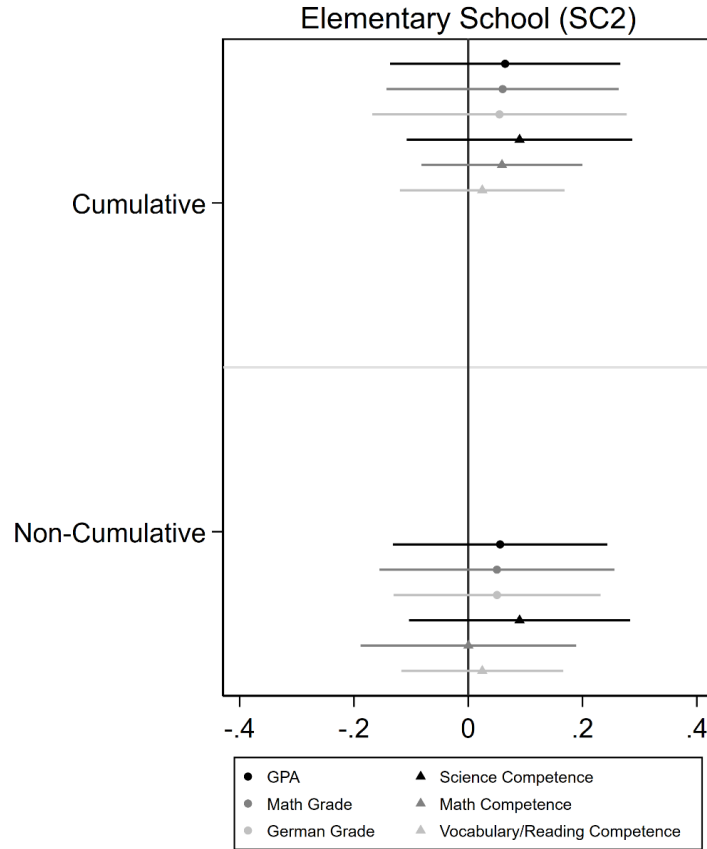


Figure E.2: Effects of Classmate Retention on Promoted Students' Academic Performance Using Classes with No Teacher Change

Note: Each marker shows an aggregated treatment-effect estimate; horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units. The sample is restricted to elementary school classes with no teacher change between grades.

Source: NEPS data; authors' illustration.

E.3 Potential Confounding by Class Size

Third, class size may act as a confounder, as it could influence both the likelihood of a classmate being retained—since struggling students in larger classes may be more easily overlooked—and individual student performance. To address this concern, we conducted a robustness check that controlled for class size measured in the year before the retention event, when the retained student was still part of the class. Because substantial changes in class size between grades may reflect broader shifts in classroom composition that could independently affect academic outcomes, we excluded classes in which the number of students changed by more than six from one year to the next (or more than ten students when competencies were measured two years apart). We relied on teacher-reported class sizes, as NEPS often does not capture all students within each class and observed student counts therefore underestimate true class sizes. While more accurate, teacher-reported class sizes contain many missing values, which substantially reduces statistical power when included as controls (see Appendix A, Table A.1). To evaluate the impact of the sample restriction, we plotted the results of the main analysis using the reduced sample (see Figure E.3, Panel a), alongside results from models estimated on the same sample that additionally controlled for class size (see Figure E.3, Panel b). The results show only marginal differences, suggesting that the observed changes are primarily driven by the reduced sample rather than by the inclusion of class size controls. This implies that the absence of class size controls in the main models is unlikely to have introduced meaningful bias. Regarding statistical significance, the reduced sample yielded no significant effects at the 5% level. Accordingly, this robustness check reproduces the predominantly null pattern of the main analysis.

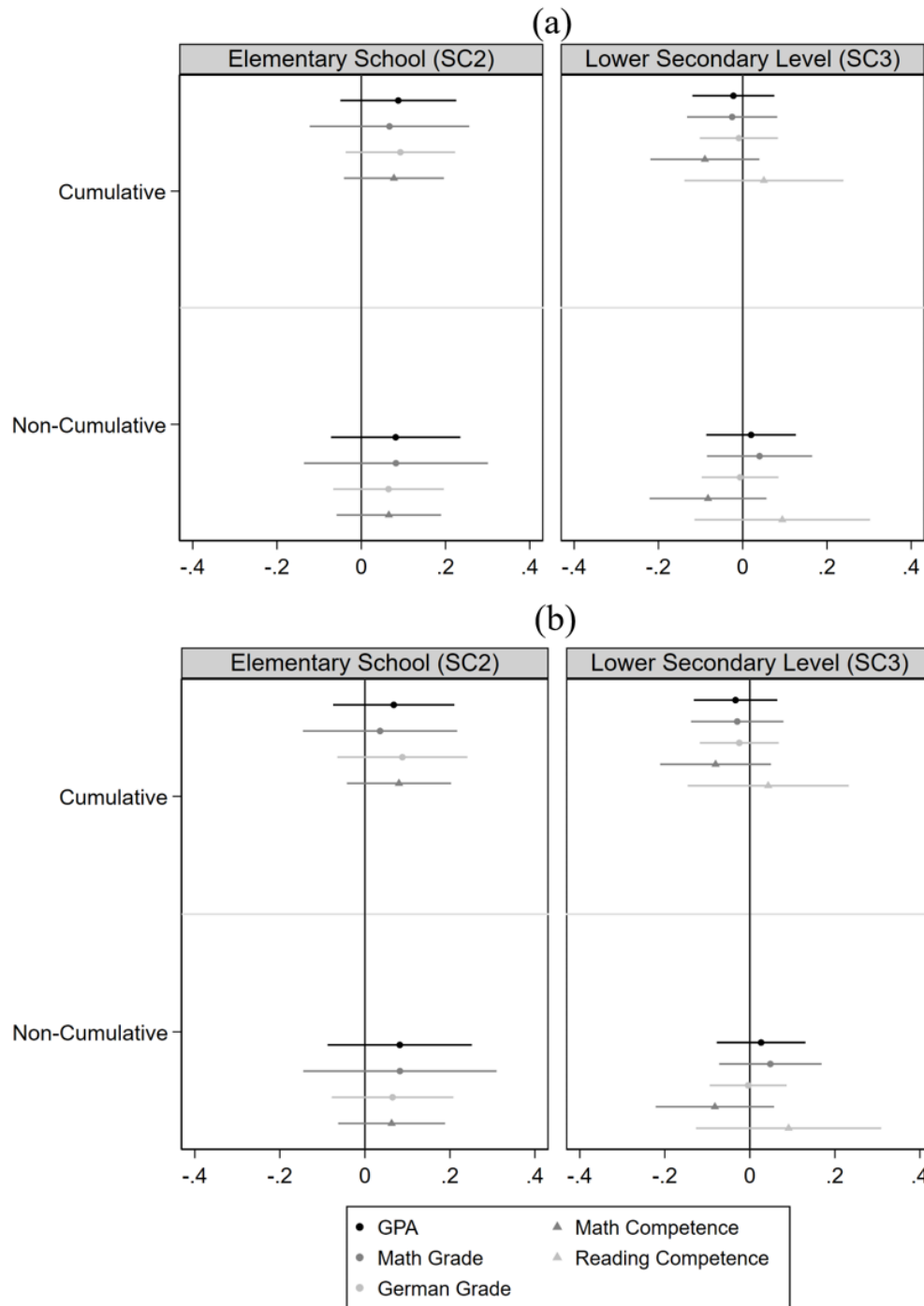


Figure E.3: Effects of Classmate Retention on Promoted Students' Academic Performance with Class Size Controls
Note: Each marker shows an aggregated treatment-effect estimate; horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units. Panel (a) replicates the main analysis using a reduced sample that excludes observations with missing class size data. Panel (b) uses the same reduced sample and additionally controls for class size.
Source: NEPS data; authors' illustration.

E.4 Control-Group Misclassification Due to Incomplete Classroom Observation

A limitation of the NEPS data is that classrooms are not fully observed (see “Data”). As a result, some instances of grade retention may go unobserved, causing truly treated classrooms to be misclassified as controls and thereby attenuating the estimated effects toward zero. We address this concern in two complementary ways. First, we replicate the main analyses using progressively more fully observed classrooms (Appendix E.4.1). Second, we conduct minimum detectable effect analyses that incorporate plausible levels of unobserved retention to quantify the implications of such misclassification for the effects our analyses can detect (Appendix E.4.2).

E.4.1 Analyses Using More Fully Observed Classrooms

First, we assess the sensitivity of our findings to this limitation by replicating the main analyses while progressively restricting the allowable deviation between the number of observed students per class and the teacher-reported class size, which serves as the reference measure. These restrictions increase confidence that classrooms in the control group are truly untreated, though at the cost of sample size (see Appendix A, Table A.1). Accordingly, confidence intervals are wider, and the coefficient plots therefore use a larger x-axis scale than in the main analysis (see Figures E.4–E.6). We first replicate the main analyses restricting the sample to classrooms in which the number of observed students deviates by no more than six students at the elementary level and three students at the lower secondary level from the teacher-reported class size, which serves as the reference value. For classrooms without a teacher-reported class size—approximately 20% at the elementary level and 40% at the lower secondary level—we retain only classes with at least 10 (elementary) or 20 (lower secondary) observed students. Figure E.4 presents the corresponding estimates, none of which is statistically significant at the 5% level.

A limitation of this specification is that the restrictions are applied at the grade–year level, such that only individual observations violating the deviation threshold are excluded, while the remaining observations of the same classroom are retained. To address this concern, we re-estimate the main analyses imposing the same restrictions as above but now requiring that they be satisfied for the entire observation period. Under this more stringent restriction, the estimated effects again overall yield null effects (see Figure E.5).

As a final robustness check, we restrict the sample to classrooms with a teacher-reported class size and exclude all classrooms without this information. To maintain sufficient sample size, we require that the number of observed students deviates by no more than five students from the teacher-reported class size over the entire observation period for both cohorts. Even under this most restrictive specification, the estimated effects yield overall null effects (see Figure E.6).

In sum, across all three increasingly restrictive robustness checks (Figures E.4–E.6), none of the estimated effects is statistically significant at the 5% level, and the estimates do not reveal a consistent pattern of beneficial effects. Thus, restricting the analyses to more fully observed classrooms does not uncover positive effects that may have been masked by control-group misclassification in the main analyses. Although these restrictions cannot fully eliminate the possibility that some grade-retention events remain unobserved and substantially reduce statistical precision, the results are broadly consistent with our overall assessment that there is no consistent evidence of academic benefits of classmate retention among promoted students.

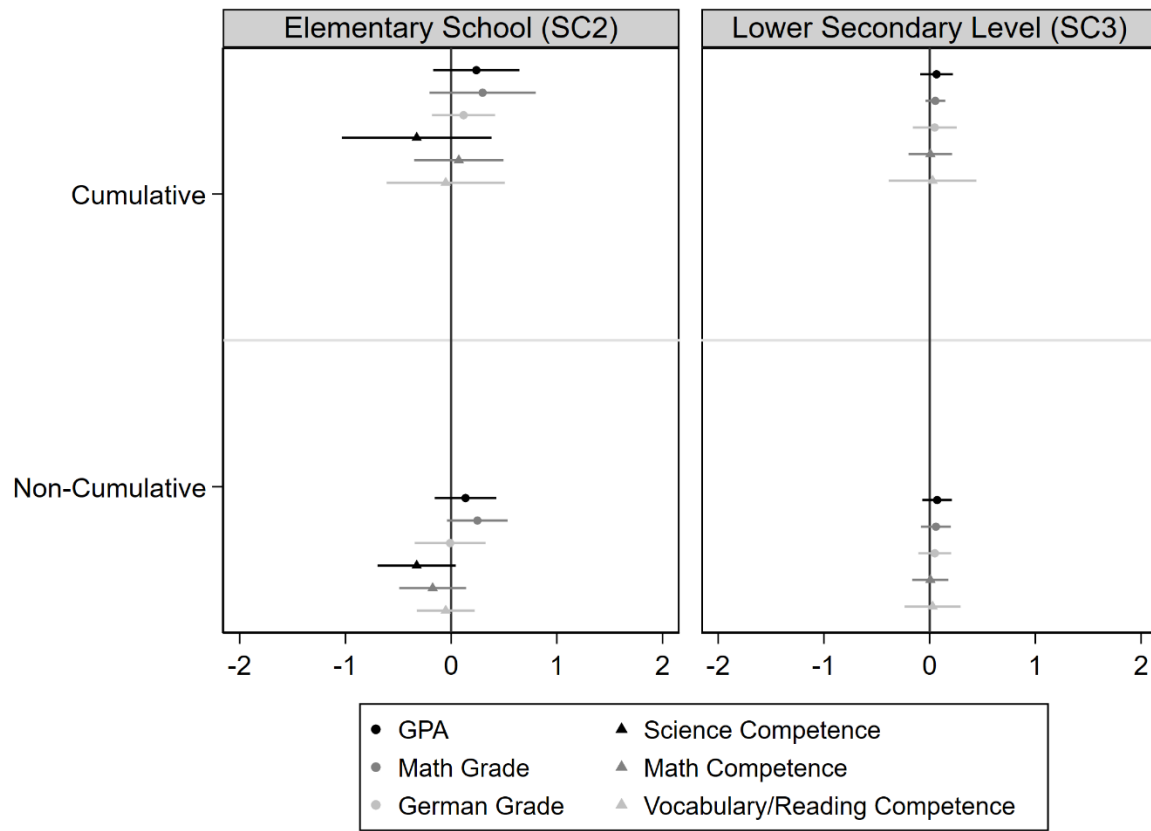


Figure E.4. Effects of Classmate Retention on Promoted Students' Academic Performance in a Class Size-Restricted Sample

Note: Each marker shows an aggregated treatment-effect estimate; horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units. The sample is restricted to classrooms in which the number of observed students deviates by no more than six students at the elementary level and three students at the lower secondary level from the teacher-reported class size. For classrooms without teacher-reported class size, only classes with at least 10 (elementary) or 20 (lower secondary) observed students are retained. *Source:* NEPS data; authors' illustration.

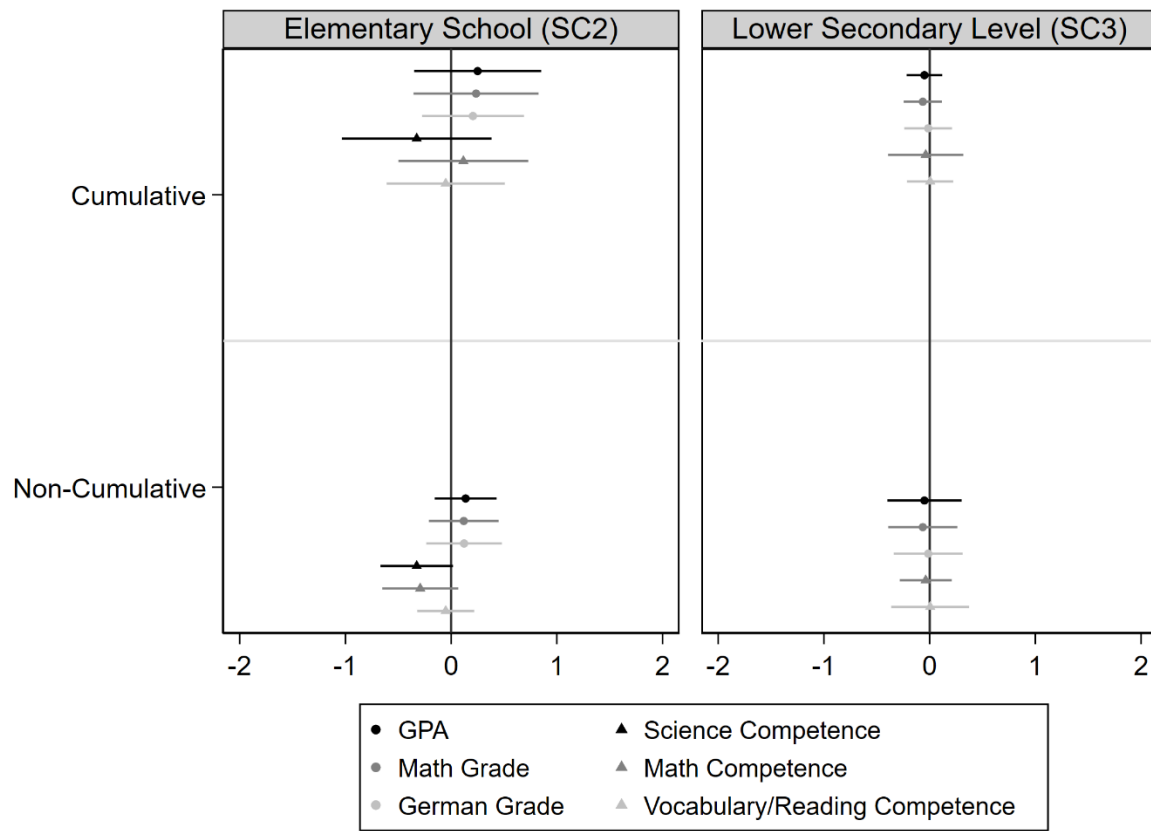


Figure E.5. Effects of Classmate Retention on Promoted Students' Academic Performance in a Class Size–Restricted Sample over the Entire Observation Period

Note: Each marker shows an aggregated treatment-effect estimate; horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units. The sample is restricted to classrooms that satisfy the same deviation thresholds as in Figure E.4, now required to hold over the entire observation period of each classroom.

Source: NEPS data; authors' illustration.

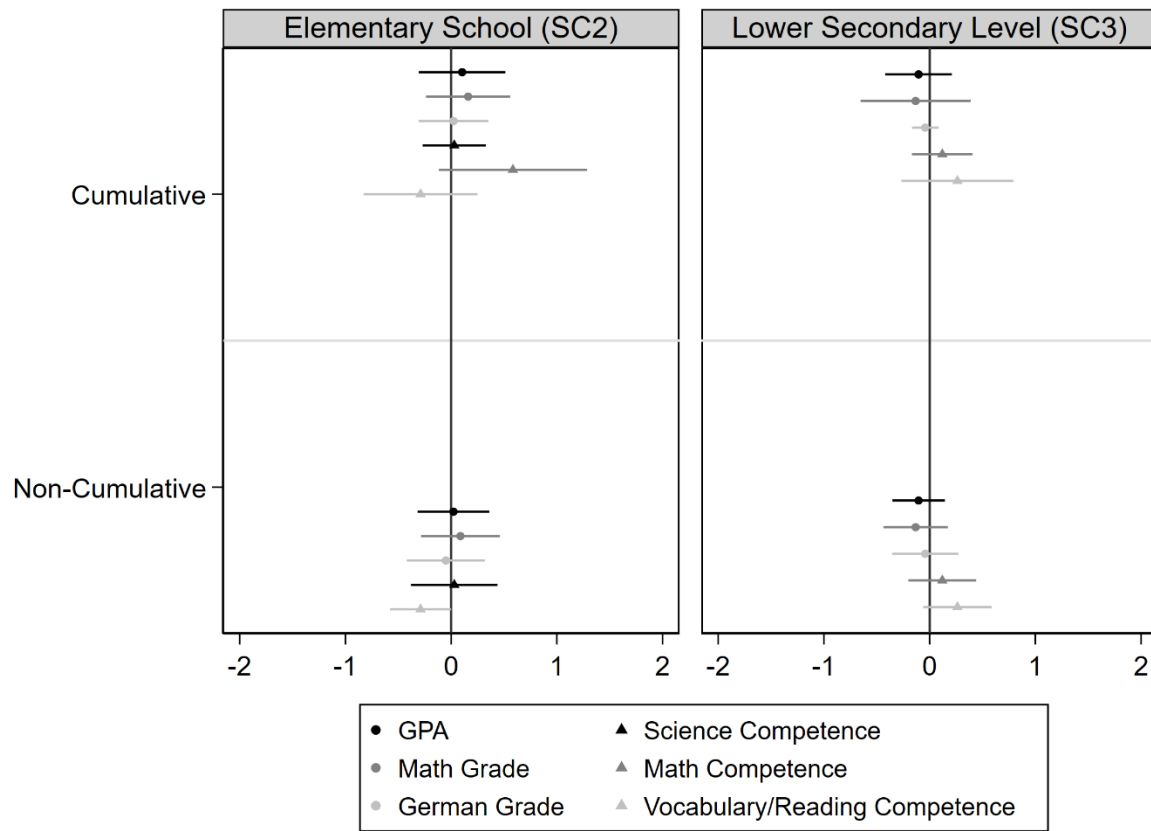


Figure E.6. Effects of Classmate Retention on Promoted Students' Academic Performance in a Class Size–Restricted Sample over the Entire Observation Period When Valid Teacher-Reported Class Size Information Is Available

Note: Each marker shows an aggregated treatment-effect estimate; horizontal lines indicate 95% confidence intervals. All outcomes are rescaled so that effect estimates are expressed in standard-deviation units. The sample is restricted to classrooms with teacher-reported class size information and requires that the number of observed students deviates by no more than five students from the teacher-reported class size over the entire observation period.

Source: NEPS data; authors' illustration.

E.4.2 Minimum Detectable Effects under Plausible Levels of Unobserved Retention

Second, we quantify how potential control-group misclassifications described above affect the magnitude of effects that our main analyses can detect. Because an unobserved retention event can only turn a truly treated class into an apparent control class, and never the reverse, this misclassification is one-sided. In our DiD models, this contamination of the control group absorbs part of the treatment–control contrast and therefore attenuates the estimated effect toward zero. Rather than restricting the sample, as in Section E.4.1, this section addresses the issue by asking how large a true effect would need to be for our analyses to detect it under plausible levels of unobserved retention. Because the following framework maps directly onto the binary, absorbing treatment definition of the non-cumulative specification, we conduct the minimum detectable effect (MDE) analysis for this specification only. This focus is unlikely to alter the substantive interpretation, as the cumulative and non-cumulative specifications yielded substantively similar results across the main and robustness analyses. In the following, we first describe the sensitivity-analysis framework and then present the resulting contamination estimates and adjusted MDEs.

Sensitivity-Analysis Framework. Based on the precision of the main-analysis estimates, we calculate conventional MDEs at 80% power from the cluster-robust standard errors as approximately $2.80 \times SE$ (two-sided $\alpha = .05$; Bloom, 1995). Such an MDE is the smallest true effect size that the analysis has an 80% chance of detecting as statistically significant at the 5% level (Bloom, 1995, p. 547). These MDEs are expressed in standard-deviation units as in the main text and range from 0.08 to 0.09 SD in lower secondary education (SC3) and from 0.13 to 0.22 SD in elementary school (SC2) (see Table E.1). However, these MDEs are expressed on the attenuated observed-effect scale and do not yet account for the potential control-group contamination. To adjust these MDEs for this attenuation, let λ denote the share of control observations that are in fact treated.

Under the assumptions of parallel trends and effect homogeneity, the estimated effect can be approximated as $(1 - \lambda)$ times the true effect, because the share λ of contaminated control observations absorbs part of the treatment effect. Accordingly,

$$MDE_{true} = \frac{MDE_{observed}}{1 - \lambda} . \quad (3)$$

For illustration, if 25% of control observations are in fact treated ($\lambda = .25$), an observed effect of 0.15 SD would correspond to a true effect of $0.15/(1 - .25) = 0.20$ SD.

We construct λ from two components. The first is the number m of unobserved students in classroom c and grade g , m_{cg} , calculated as the difference between the teacher-reported and observed class size. For classrooms without teacher-reported class-size information (approximately 20% in elementary school and 40% in lower secondary education), we impute the grade-specific mean teacher-reported class size. Among classrooms with valid reports, teacher-reported class size averages 21.9 students in elementary school and 25.6 students in lower secondary education across grade levels. This imputation assumes that the grade-specific means observed among reporting classrooms provide a reasonable approximation for classrooms with missing reports. Where the observed number of students exceeds the reported or imputed class size, we set the number of unobserved students to zero. The second component of λ is the annual probability r that any given unobserved student is retained. For this purpose, we use official retention rates from the school years corresponding to the respective NEPS cohorts (see Appendix B, Table B.1). For elementary school, we set $r = 1.2\%$, corresponding to the official Grade 3 retention rate. Grade 3 is the only grade for which an official retention rate is available and in which retention can affect outcomes observed within the SC2 analytic sample; we therefore apply this rate to all relevant prior grades

in SC2. For lower secondary education, we consider two scenarios, $r = 1.6\%$ and $r = 3.0\%$, corresponding to the lowest and highest official retention rates observed across Grades 5–8, in which retention can affect outcomes observed within the SC3 analytic sample.

Assuming a common retention probability r and independent retention events among the m_{cg} unobserved students, the probability that at least one unobserved student in a control classroom is retained in a given grade is

$$1 - (1 - r)^{m_{cg}} . \quad (4)$$

For illustration, if a classroom has 10 unobserved students ($m = 10$) and the annual retention probability is 2% ($r = .02$), the probability that at least one of these students is retained is $1 - (1 - .02)^{10} \approx .18$, or about 18%.

Because treatment status is absorbing, the probability of control-group contamination accumulates over prior grades in which an unobserved retention event could have occurred. For a control observation in classroom c and grade g , we therefore use the numbers of unobserved students in that classroom across all relevant prior grades observed within the respective panel. We denote the resulting contamination probability for classroom c in grade g by p_{cg} :

$$p_{cg} = 1 - (1 - r)^{\sum_{s \in K_c(g)} m_{cs}} \quad (5)$$

where s indexes prior grade levels and $K_c(g)$ denotes the set of prior grades observed for classroom c within the respective panel in which an unobserved retention event could have occurred, with the relevant prior grades drawn from Grades 1–3 in SC2 and Grades 5–8 in SC3. Retention in Grade 4 or Grade 9 would affect classmates only in the following grade, which lies outside the

respective analytic window. For illustration, if a Grade 4 control classroom has 8, 10, and 7 unobserved students in Grades 1, 2, and 3, respectively ($m_{c1} = 8$, $m_{c2} = 10$, and $m_{c3} = 7$), then $\sum_{s \in K_c(4)} m_{cs} = 25$. With an annual retention probability of 2% ($r = 0.02$), the implied contamination probability is $p_{c4} = 1 - (1 - 0.02)^{25} \approx 0.40$, or about 40%. In other words, this observation, although classified as a control, has an approximately 40% probability of being truly treated.

Finally, we approximate λ separately for SC2 and SC3 by averaging p_{cg} across all control observations within each educational stage. Substituting the estimated λ into equation (3) yields the contamination-adjusted MDE. Because these calculations rely on simplifying assumptions, these adjusted MDEs should be interpreted as sensitivity benchmarks rather than as exact bias corrections.

Results. Turning to the results, applying this procedure yields $\lambda = .172$ (17.2%) in SC2. In SC3, $\lambda = .161$ (16.1%) under the lower retention-rate scenario ($r = .016$) and $\lambda = .252$ (25.2%) under the higher scenario ($r = .030$). Put differently, about 17% of control observations in SC2 and 16–25% in SC3 are expected to be truly treated because an unobserved classmate was retained. Table E.1 translates these contamination estimates into contamination-adjusted MDEs. In SC2, the observed-scale MDEs range from 0.13 to 0.22 SD; accounting for $\lambda = .172$ increases them to 0.16–0.26 SD. In SC3, the observed-scale MDEs range from 0.08 to 0.09 SD. They increase to 0.10–0.11 SD under the lower retention-rate scenario and to 0.11–0.12 SD under the higher scenario. Relative to the observed-scale MDEs, the adjustment increases the detectable-effect thresholds by approximately 21% in SC2, 19% in the lower SC3 scenario, and 34% in the higher SC3 scenario.

These results indicate that incomplete classroom observation reduces the sensitivity of the analyses but does not fundamentally alter their interpretation. Under the assumptions of the sensitivity analysis, true effects equal in magnitude to the adjusted MDEs would have an 80% chance of producing an estimate that is statistically significant at the 5% level, while larger effects would be detected with greater power. Taken together with the overall pattern of the main analysis, these results suggest that the true effects are either null, as indicated by most of the main estimates, or, if present, likely too small to be detected reliably with our data. For the most precisely estimated outcomes, the adjusted MDEs are about 0.10 SD in SC3 and 0.16 SD in SC2, although thresholds are higher for some outcomes, reaching up to 0.12 and 0.26 SD, respectively. The implications of these findings for policy evaluation are discussed in the Discussion and Conclusion section.

Table E.1: Observed and Contamination-Adjusted Minimum Detectable Effects

	GPA	Math Grade	German Grade	Math Comp.	Vocab./ Reading Comp.	Science Comp.
<i>SC2 (Elementary School)</i>						
Observed MDE	.14	.18	.16	.22	.13	.18
Adjusted MDE ($r = 1.2\%$)	.17	.22	.20	.26	.16	.22
<i>SC3 (Lower Secondary Level)</i>						
Observed MDE	.08	.09	.09	.08	.09	–
Adjusted MDE ($r = 1.6\%$)	.10	.10	.11	.10	.11	–
Adjusted MDE ($r = 3.0\%$)	.11	.12	.12	.11	.12	–

Note: All MDEs refer to the non-cumulative specification, 80% power, and a two-sided $\alpha = .05$, and are expressed in standard-deviation units. Adjusted MDEs scale the observed MDEs upward to account for potential attenuation from unobserved classmate retention, based on the assumed annual retention rate (r) and the estimated number of unobserved classmates. Science competence was not assessed in SC3. All values are rounded to two decimal places.
Source: NEPS data; authors' calculations.

Appendix F: Robustness to Alternative Survey Weights

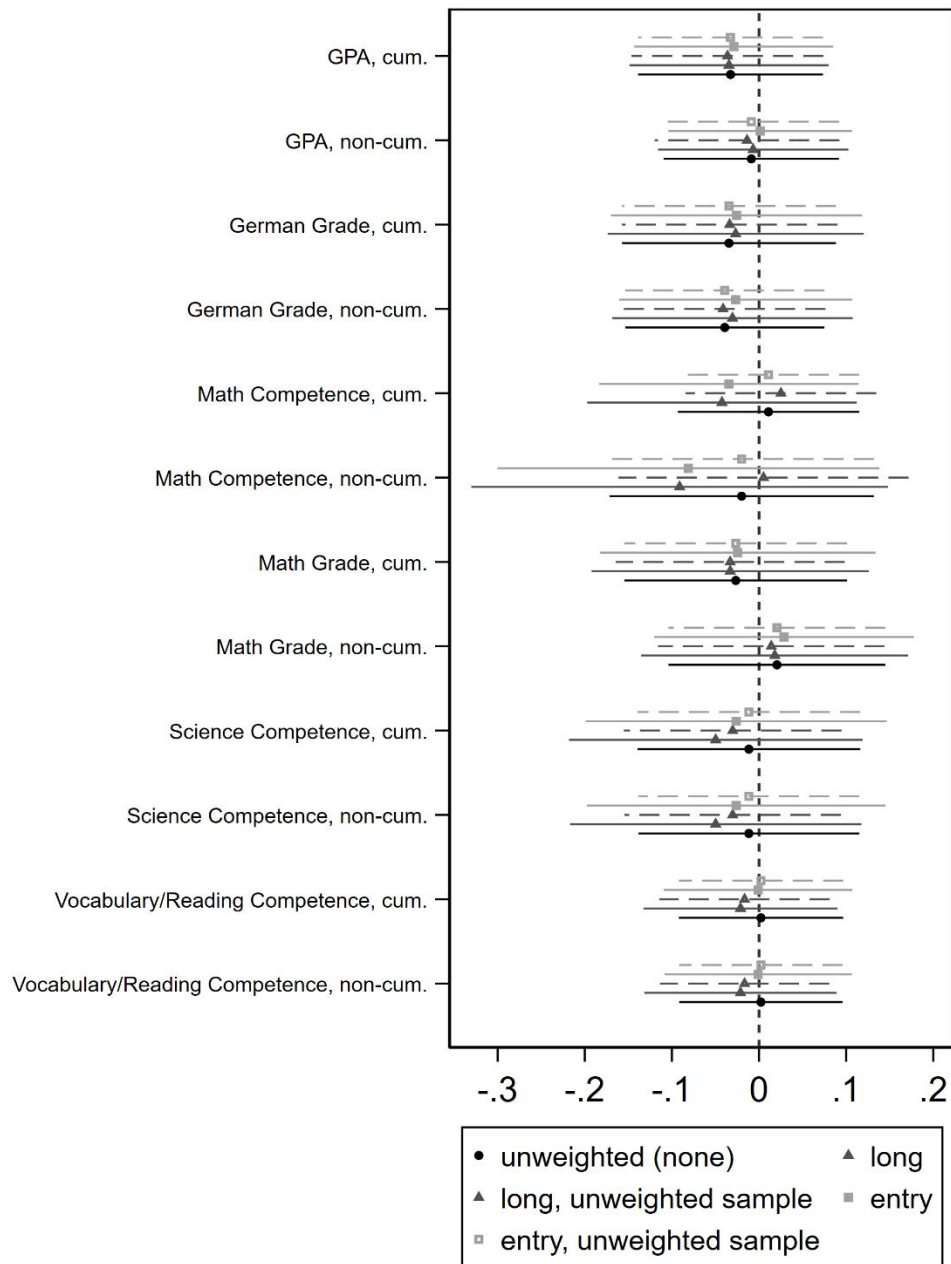


Figure F.1: Replication of Figure 2 Using Alternative Survey Weights (SC2)

Note: The figure replicates Figure 2 for SC2 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

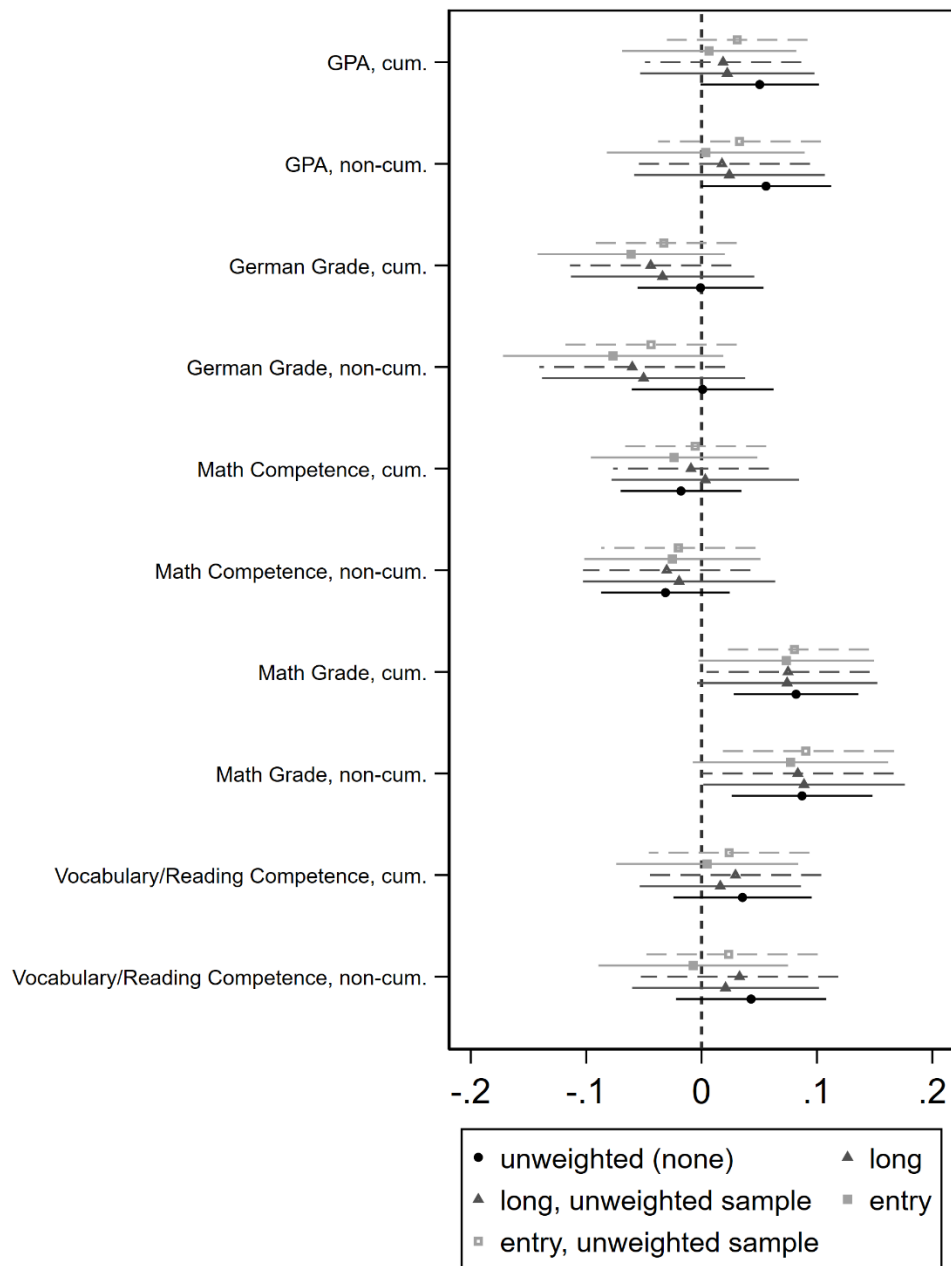


Figure F.2: Replication of Figure 2 Using Alternative Survey Weights (SC3)

Note: The figure replicates Figure 2 for SC3 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

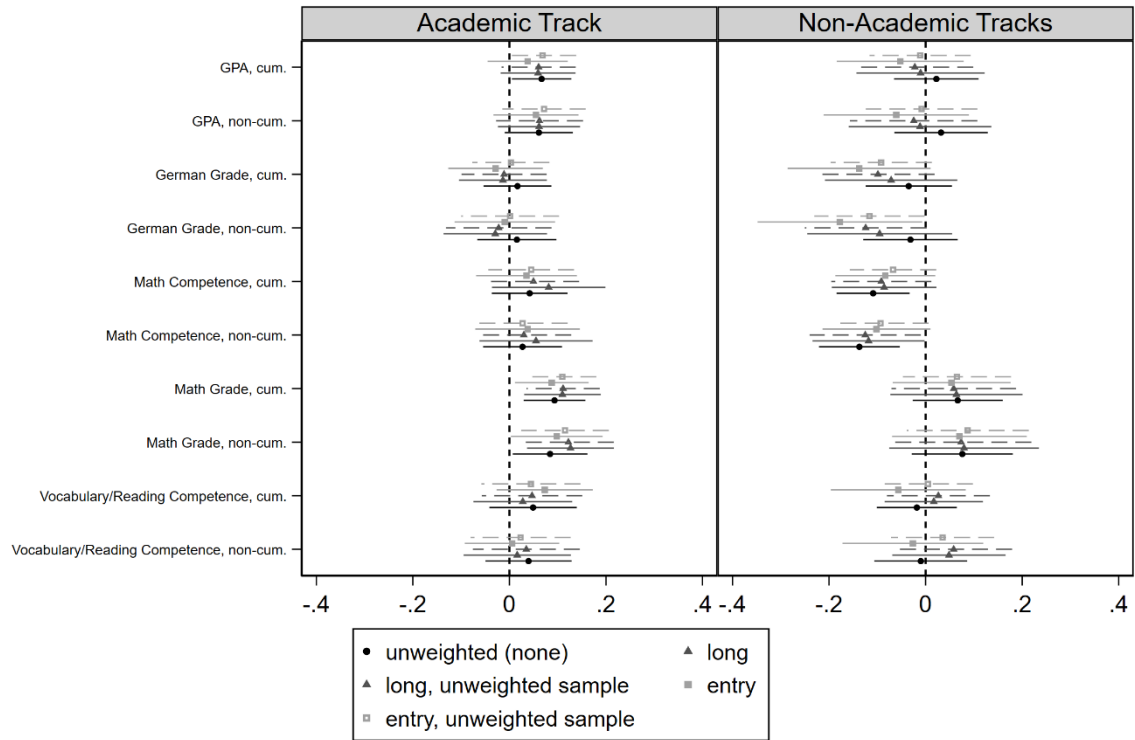


Figure F.3: Replication of Figure D.1 Using Alternative Survey Weights

Note: The figure replicates Figure D.1 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

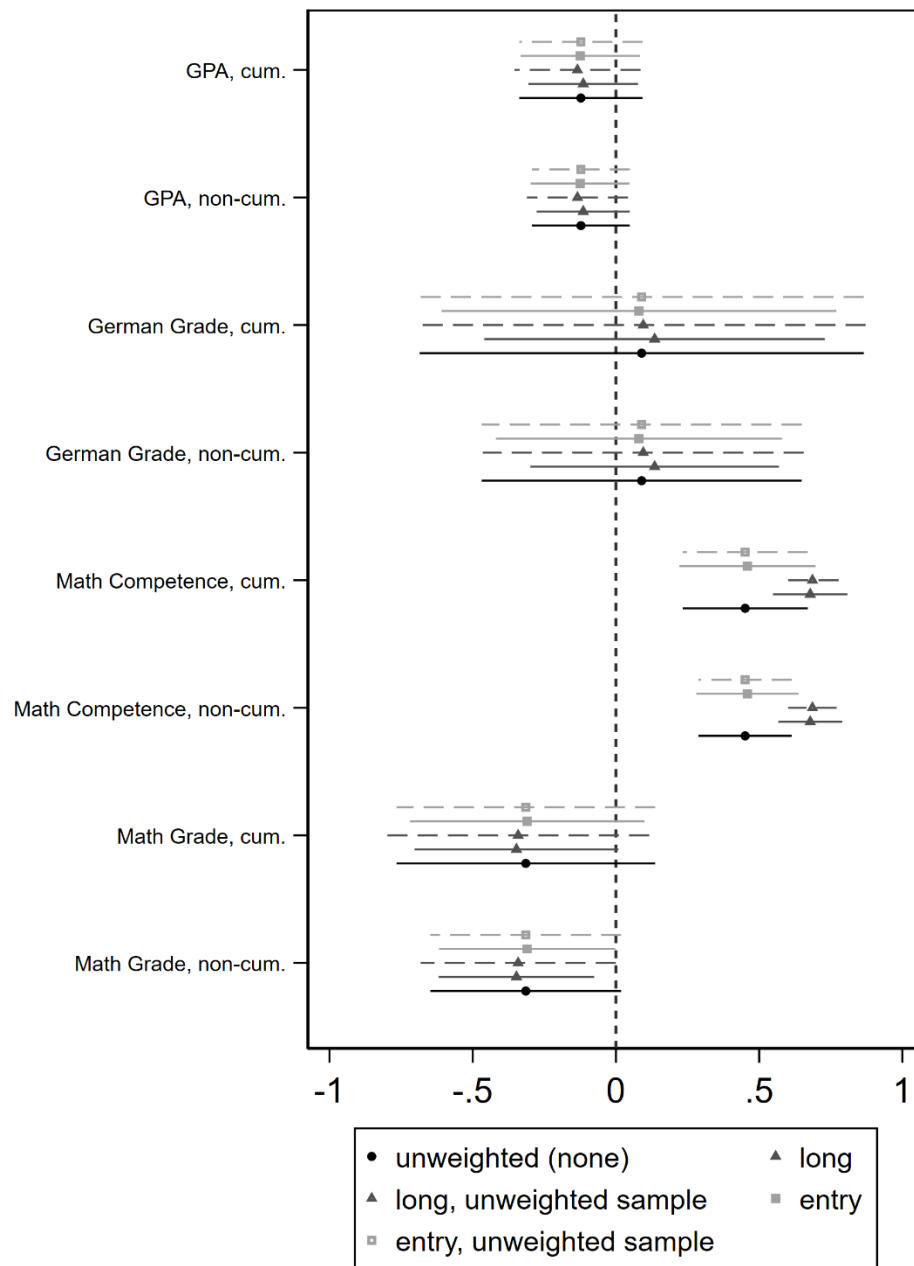


Figure F.4: Replication of Figure E.1 Using Alternative Survey Weights (SC2)

Note: The figure replicates Figure E.1 for SC2 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

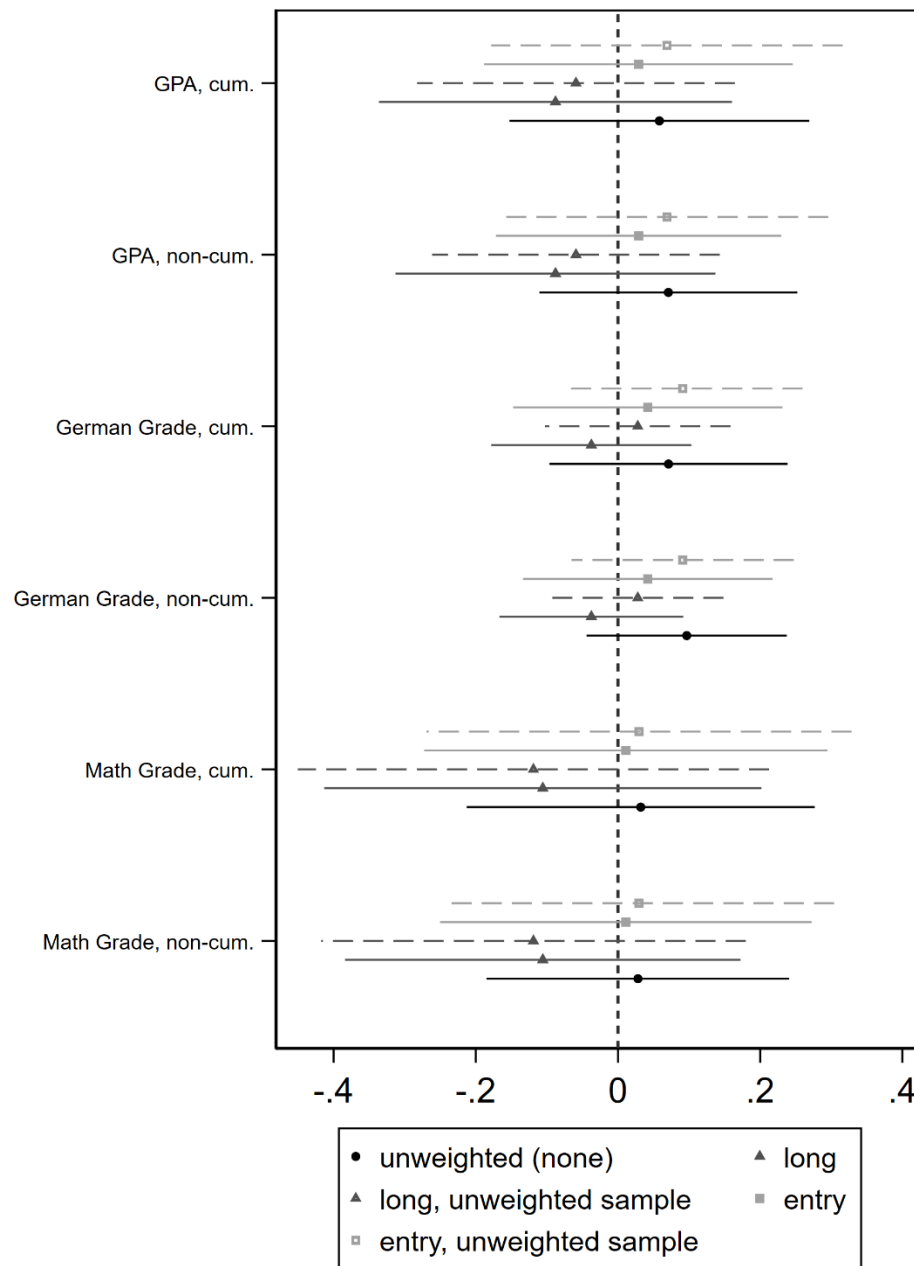


Figure F.5: Replication of Figure E.1 Using Alternative Survey Weights (SC3)

Note: The figure replicates Figure E.1 for SC3 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

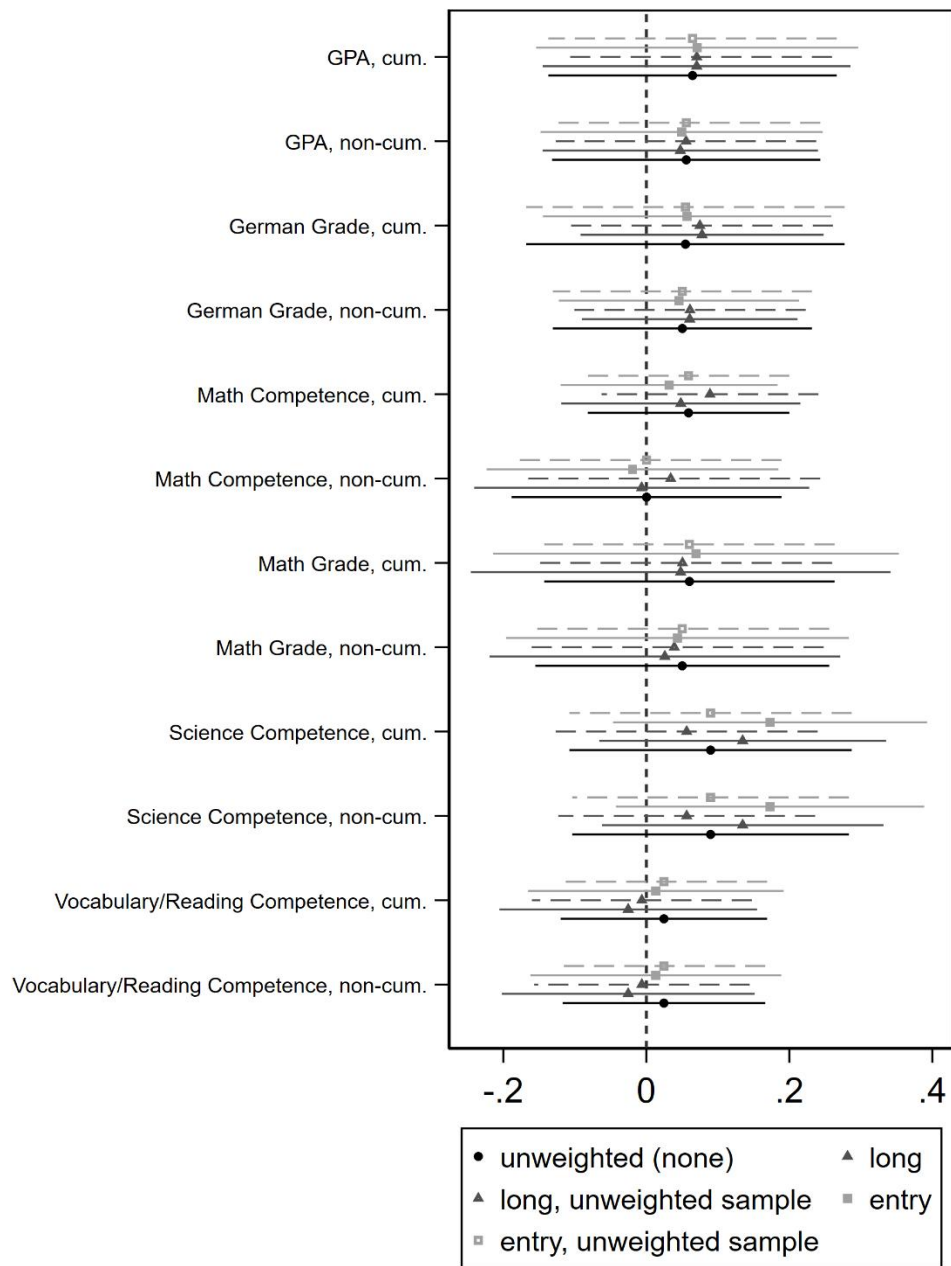


Figure F.6: Replication of Figure E.2 Using Alternative Survey Weights

Note: The figure replicates Figure E.2 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

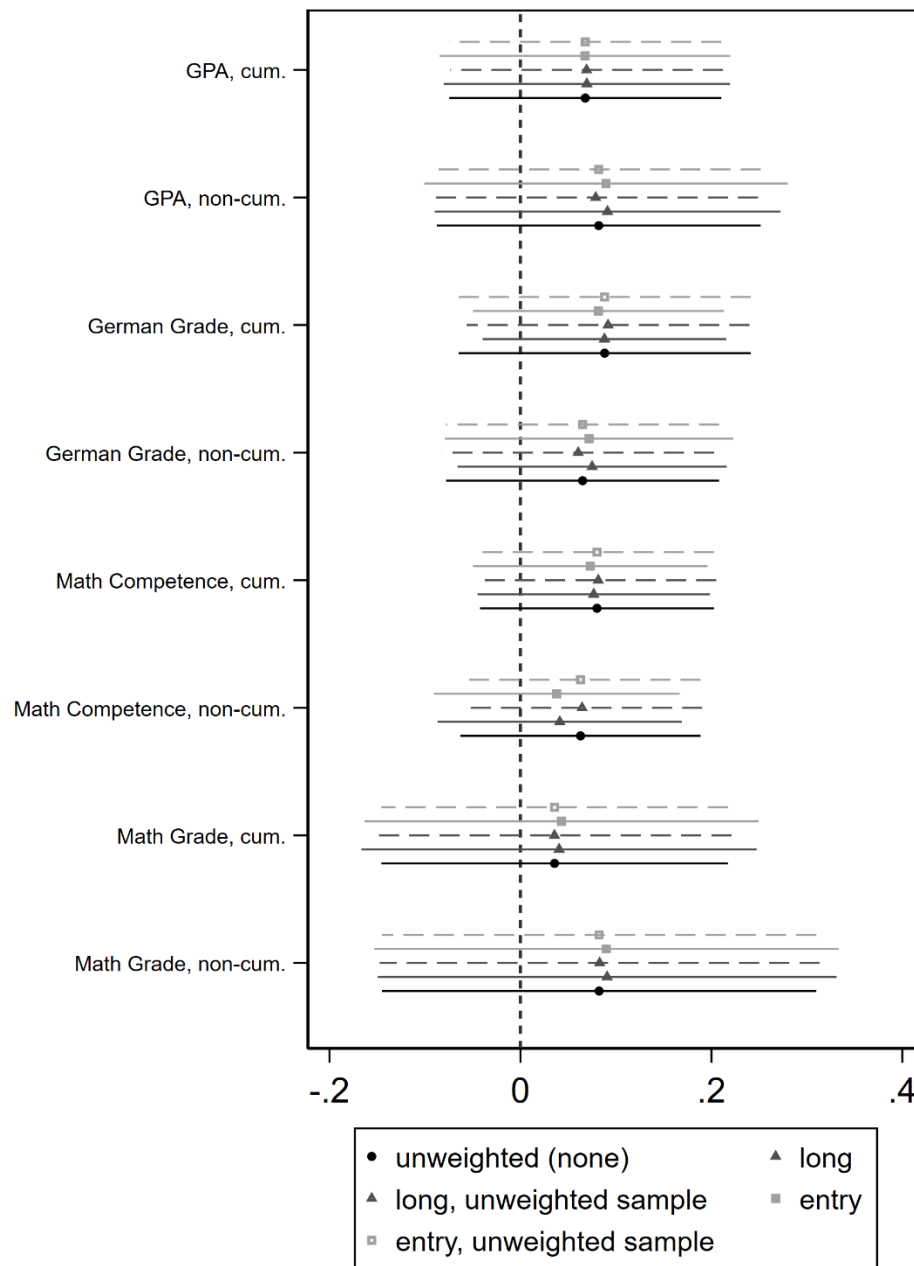


Figure F.7: Replication of Figure E.3, Panel (a), Using Alternative Survey Weights (SC2)

Note: The figure replicates Figure E.3, Panel (a), for SC2 using alternative survey-weighting specifications. All models use the reduced sample with observed class-size information but do not control for class size. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

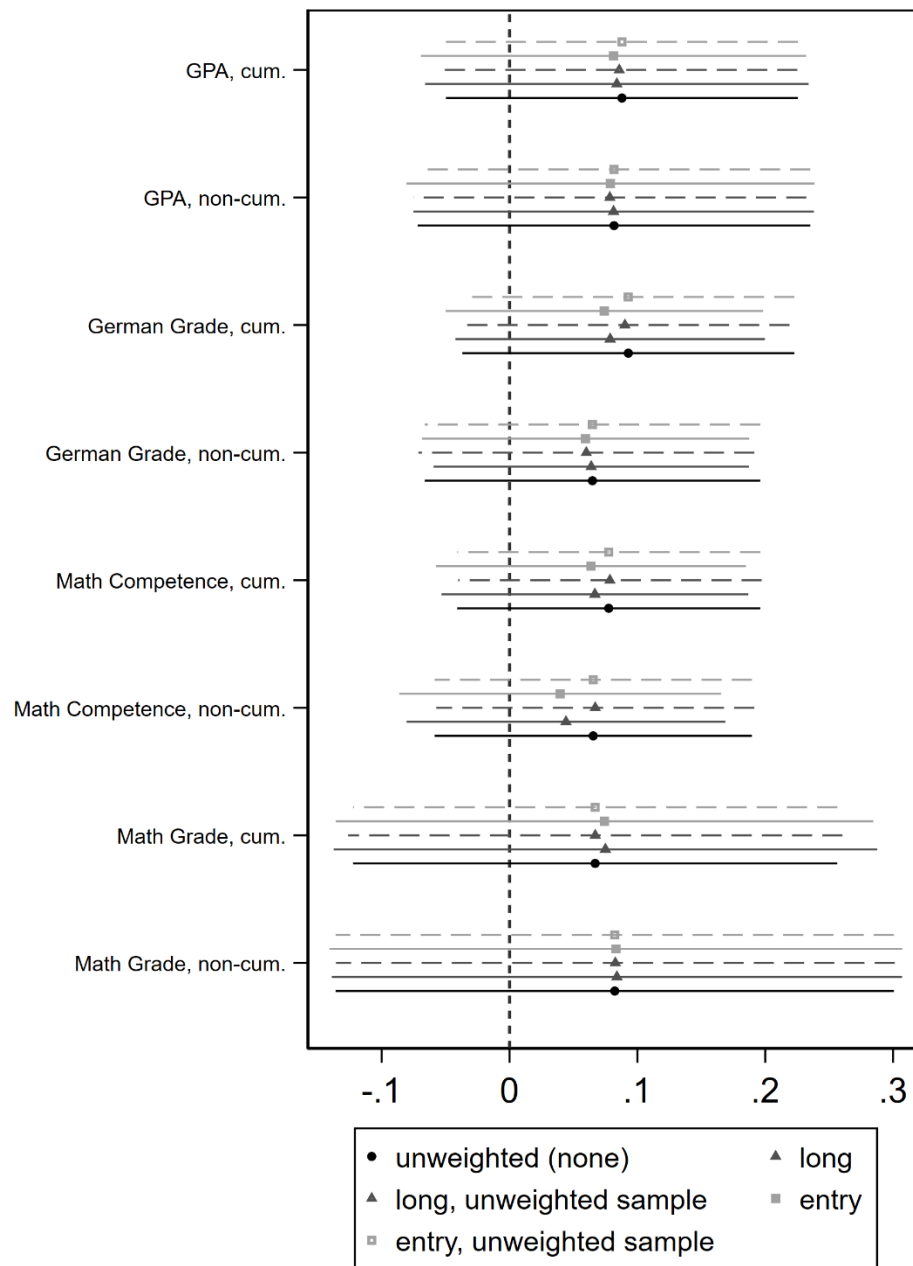


Figure F.8: Replication of Figure E.3, Panel (b), Using Alternative Survey Weights (SC2)

Note: The figure replicates Figure E.3, Panel (b), for SC2 using alternative survey-weighting specifications. All models use the reduced sample and additionally control for class size. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

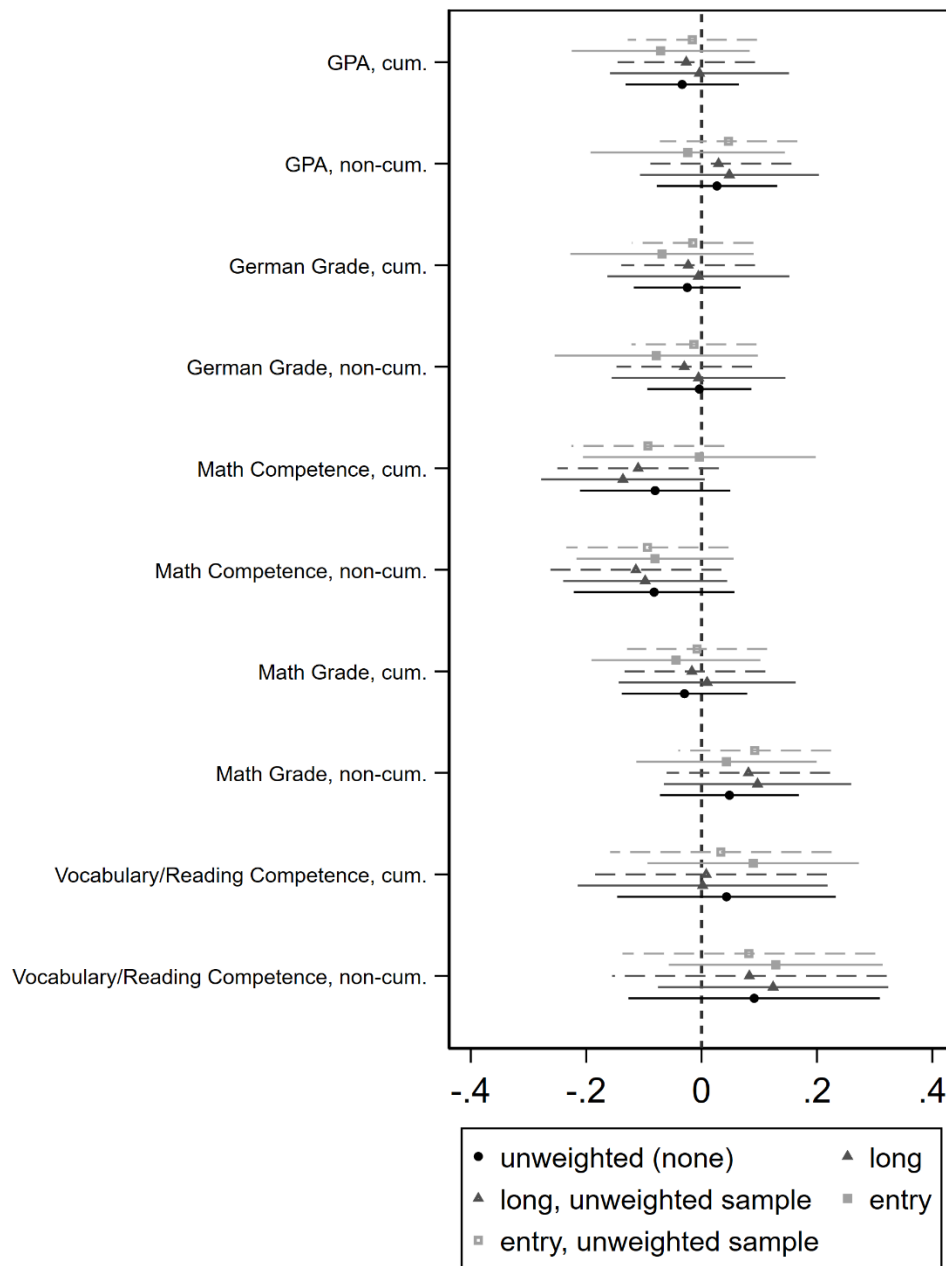


Figure F.9: Replication of Figure E.3, Panel (a), Using Alternative Survey Weights (SC3)

Note: The figure replicates Figure E.3, Panel (a), for SC3 using alternative survey-weighting specifications. All models use the reduced sample with observed class-size information but do not control for class size. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

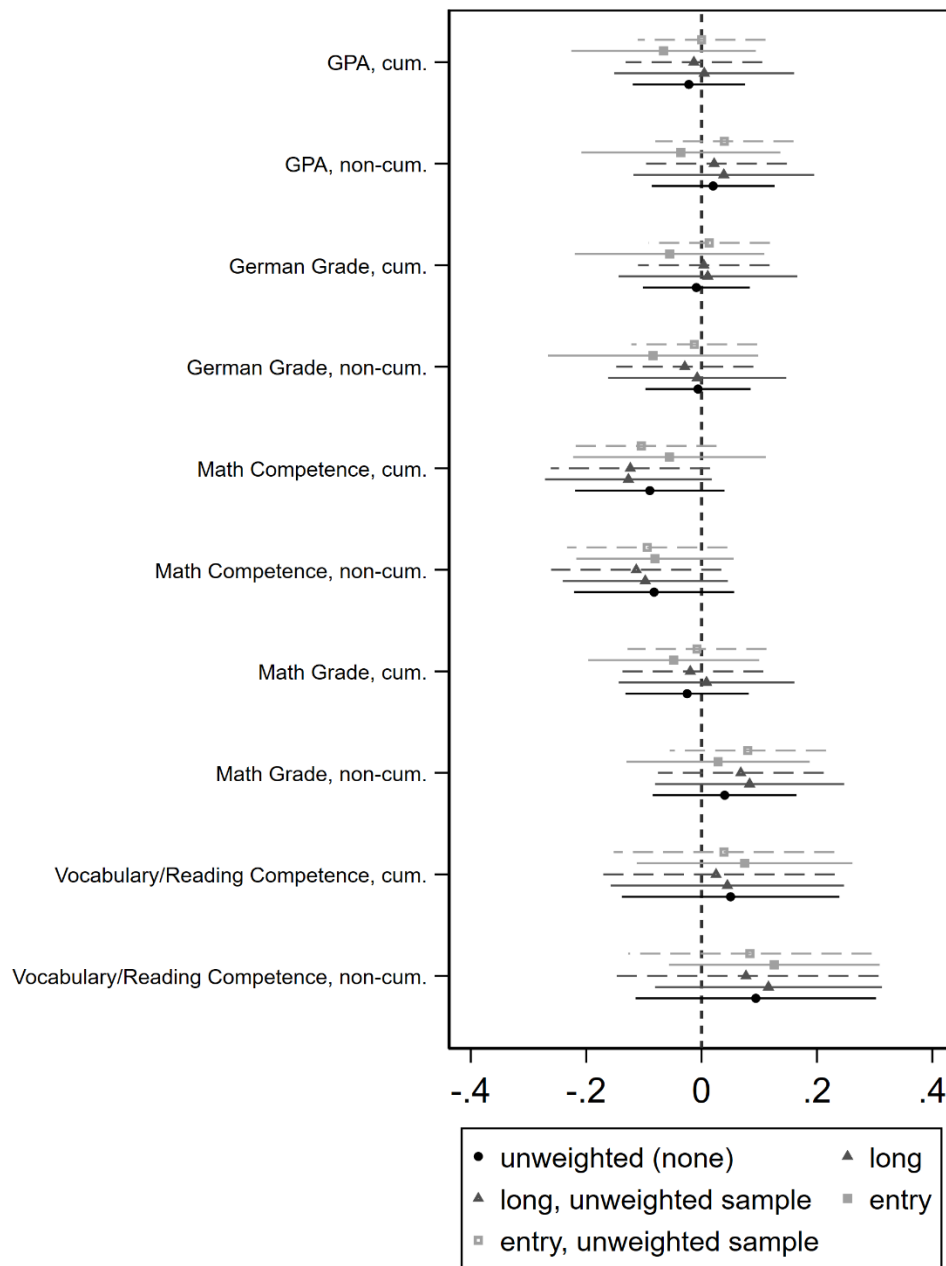


Figure F.10: Replication of Figure E.3, Panel (b), Using Alternative Survey Weights (SC3)

Note: The figure replicates Figure E.3, Panel (b), for SC3 using alternative survey-weighting specifications. All models use the reduced sample and additionally control for class size. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

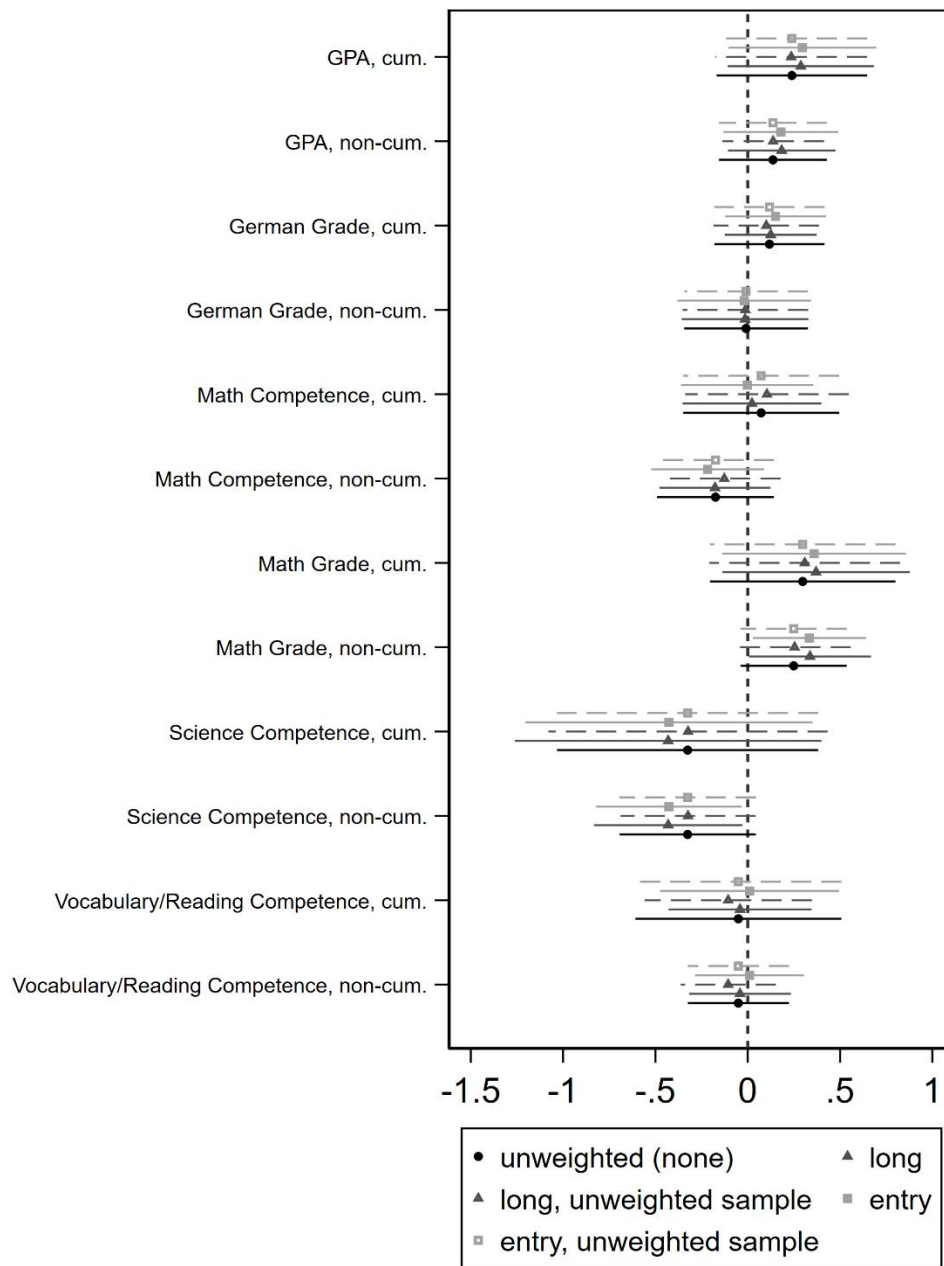


Figure F.11: Replication of Figure E.4 Using Alternative Survey Weights (SC2)

Note: The figure replicates Figure E.4 for SC2 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

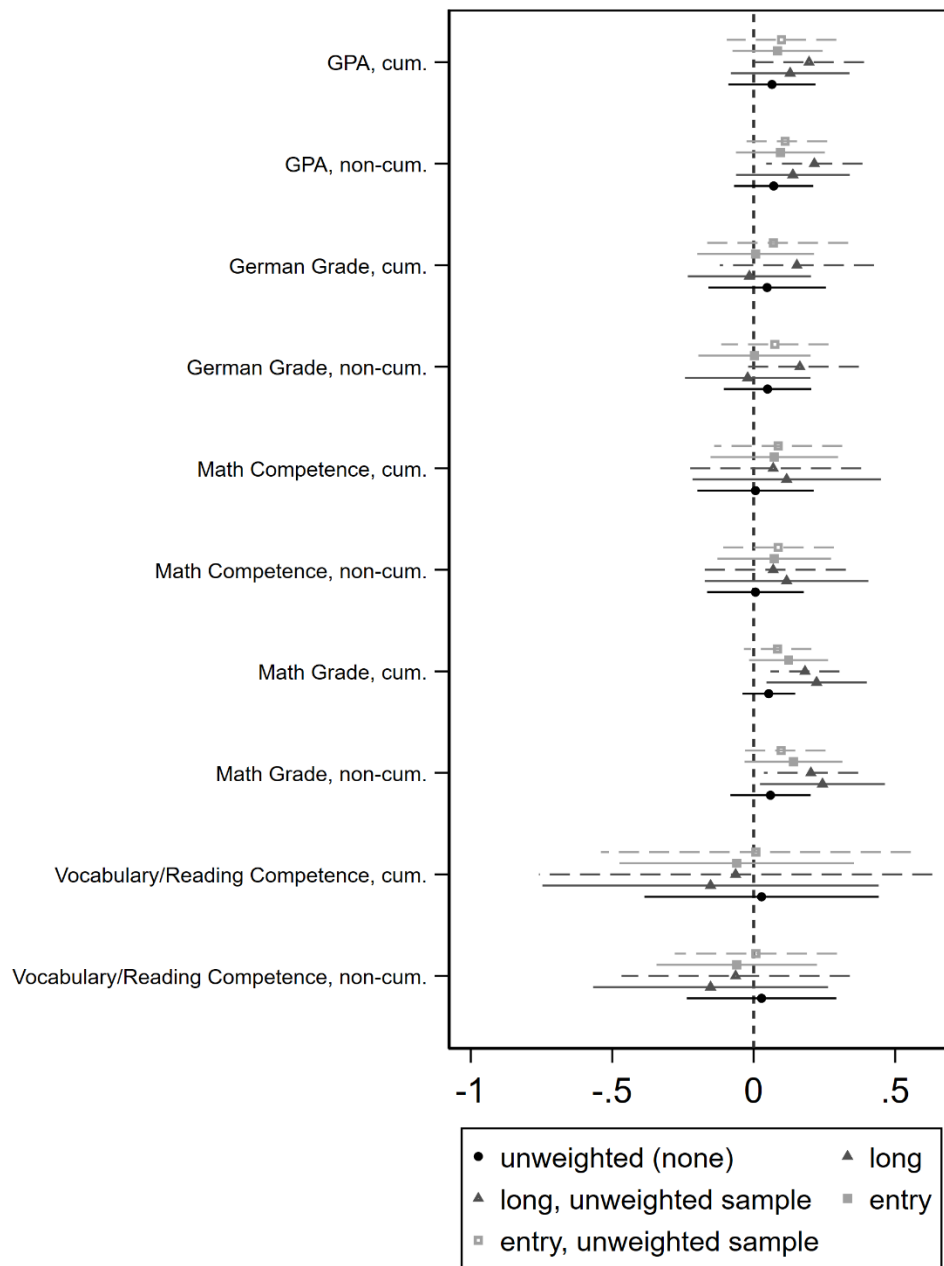


Figure F.12: Replication of Figure E.4 Using Alternative Survey Weights (SC3)

Note: The figure replicates Figure E.4 for SC3 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

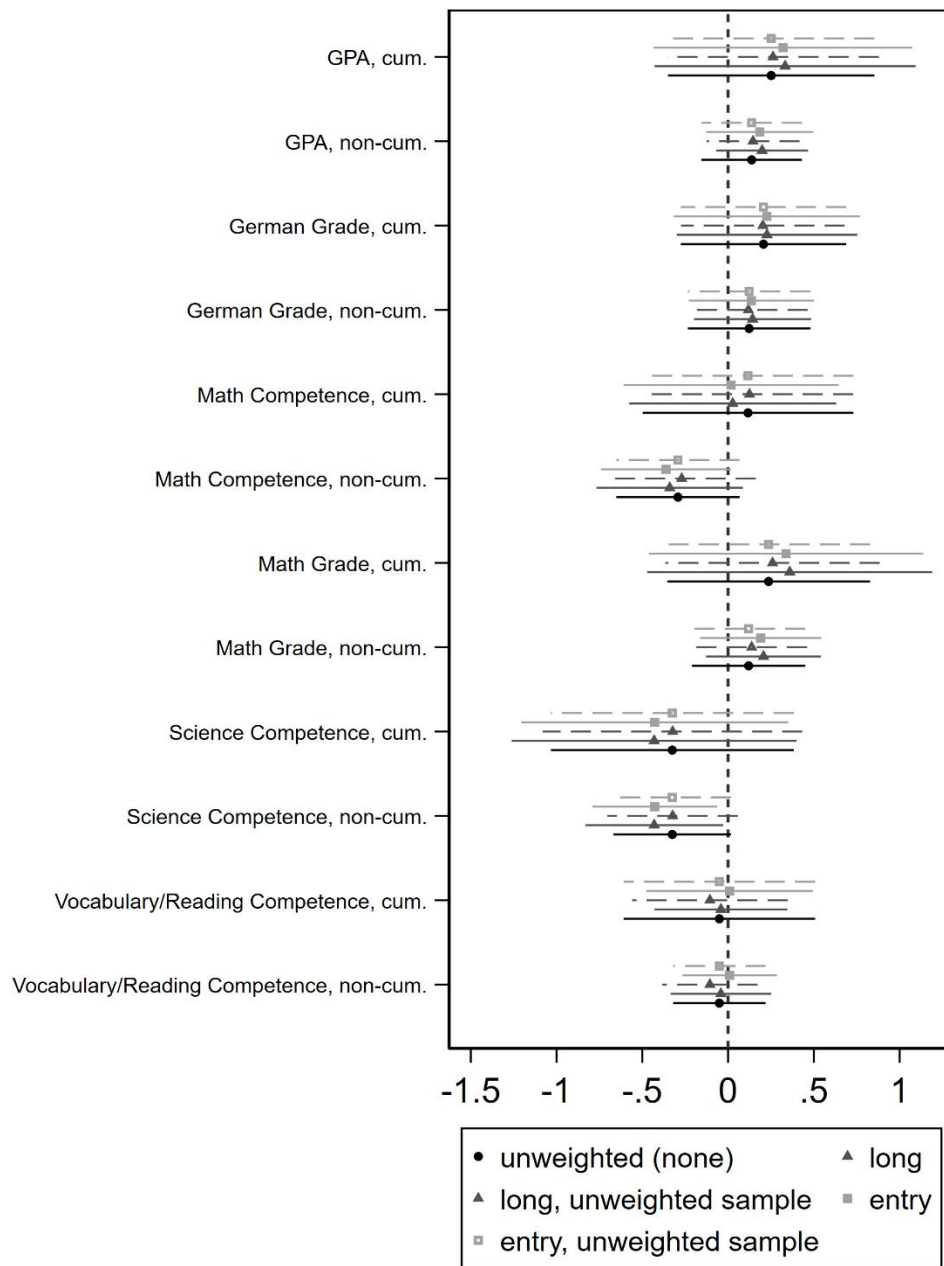


Figure F.13: Replication of Figure E.5 Using Alternative Survey Weights (SC2)

Note: The figure replicates Figure E.5 for SC2 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

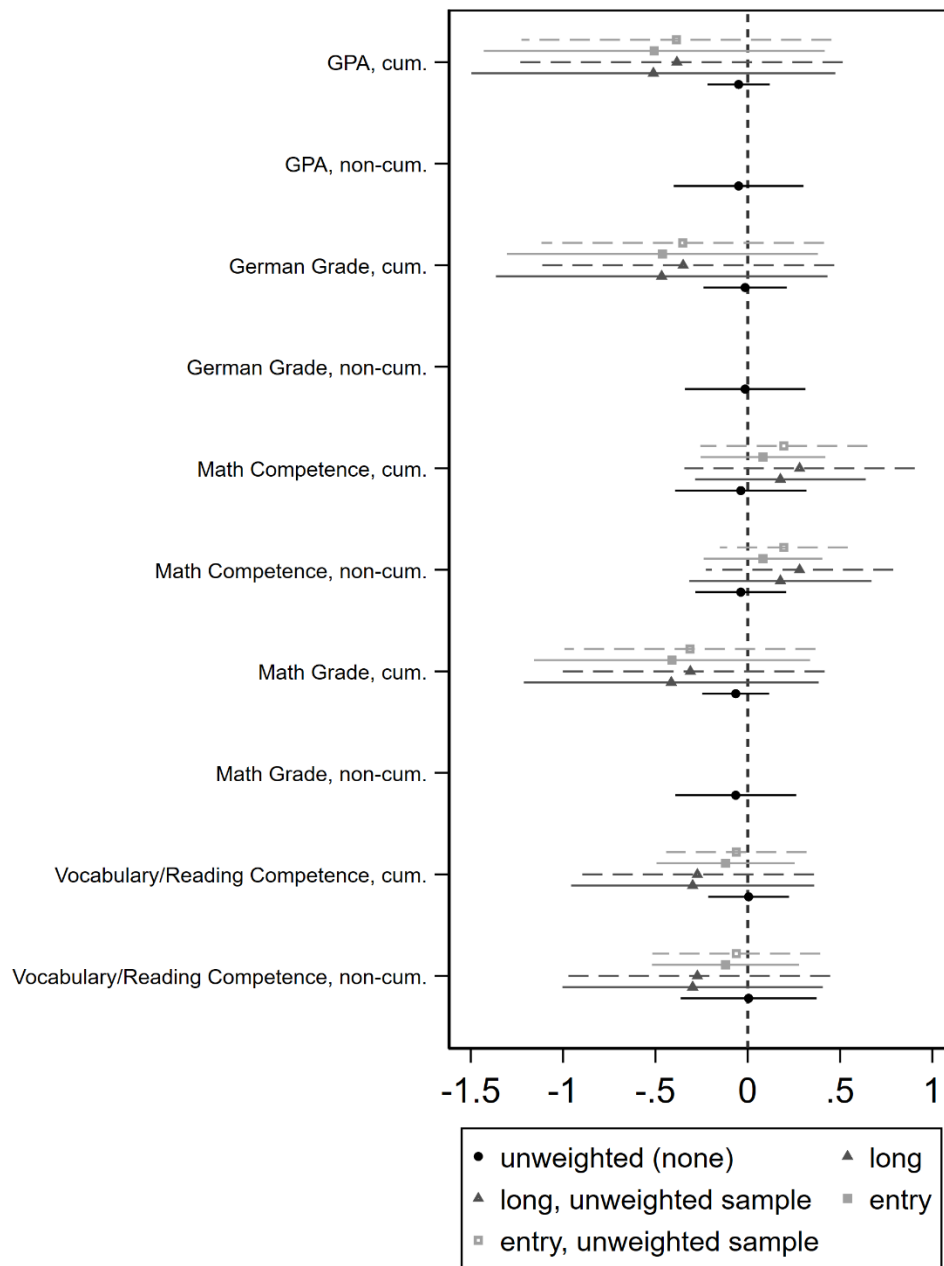


Figure F.14: Replication of Figure E.5 Using Alternative Survey Weights (SC3)

Note: The figure replicates Figure E.5 for SC3 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

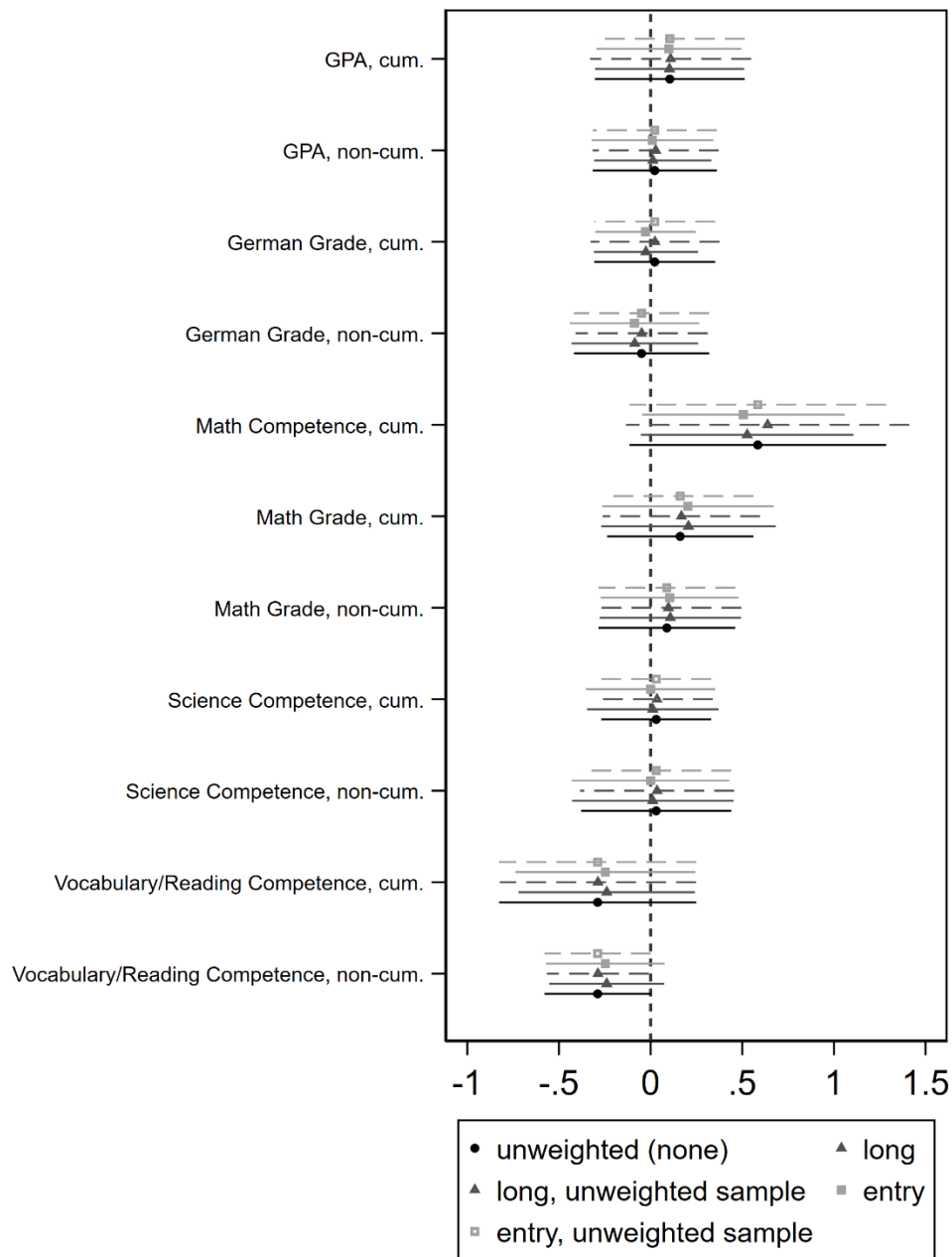


Figure F.15: Replication of Figure E.6 Using Alternative Survey Weights (SC2)

Note: The figure replicates Figure E.6 for SC2 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.

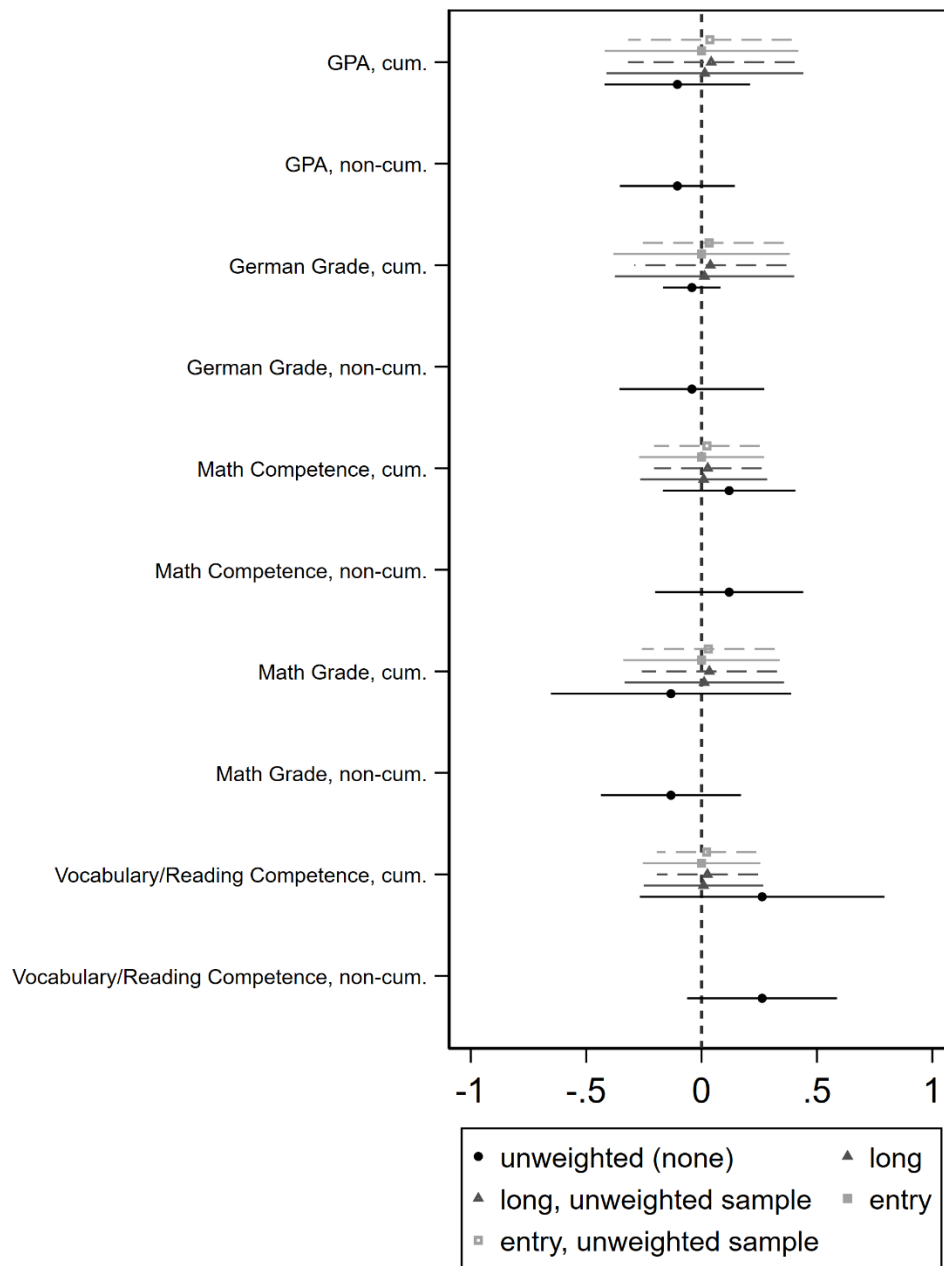


Figure F.16: Replication of Figure E.6 Using Alternative Survey Weights (SC3)

Note: The figure replicates Figure E.6 for SC3 using alternative survey-weighting specifications. Estimates are shown for the original unweighted sample, using entry and longitudinal survey weights, and unweighted on the corresponding entry-weight and longitudinal-weight samples. The latter two estimates isolate changes due to sample composition from those due to weighting. Horizontal lines indicate 95% confidence intervals; all outcomes are expressed in standard-deviation units.

Source: NEPS data; authors' illustration.