The Economics of Adaptive Capacity: Why Positive Average Effects Can Mask Severe Harm

Oleksandr Movshuk

University of Toyama

Email: movshuk@eco.u-toyama.ac.jp

January 19, 2026

Abstract

We model adaptation to adversity as allocation of scarce adaptive capacity under constraints. Mild adversity can trigger compensatory responses by students, families, and institutions that partially offset, but do not eliminate, harm already inflicted, generating a positive average effect under low observed intensity. As exposure severity rises, adaptive capacity becomes exhausted, producing monotonic deterioration and large losses for vulnerable minorities. Using Japanese PISA data and machine learning, we document a three-regime dose–response: positive average effects at low intensity, sign reversal at higher thresholds, and substantial losses in the upper tail. Psychological resilience and family resources account for a sizable share of treatment effect heterogeneity. Machine learning uncovers concentrated harm among small, otherwise hidden subpopulations (1–2%) experiencing losses of 40–60 points that conventional stratification fails to detect.

JEL Codes: C14, C21, I21, J13

1 Introduction

Students reporting bullying victimization score 11–18 points higher on standardized tests than non-bullied peers in Japanese PISA data. We stress at the outset that this positive average effect does *not* indicate that bullying is beneficial. Rather, it reflects compensatory responses by students, families, and institutions to harm already inflicted. As exposure severity increases, estimated

effects deteriorate and reverse sign: students in the upper tail of the bullying intensity distribution experience substantial losses, exceeding 40 points at the 99th percentile.

We interpret this empirical pattern through an economic framework of adaptation under finite capacity constraints. Mild adversity may trigger compensatory responses by students, families, and institutions that offset direct harm, while severe exposure exhausts adaptive capacity and generates large losses for vulnerable minorities. Importantly, our empirical results are observational. Throughout the paper, we treat estimated effects as reduced-form patterns consistent with this adaptive-capacity mechanism under a selection-on-observables assumption, rather than as *precisely identified causal effects* or evidence that adversity itself is beneficial.

Our contributions cover theory, empirics, and methodology. **Theoretically**, we formalize adaptation as scarce resource allocation with threshold effects, proving two key results. First, linear regression coefficients can be negative even when average treatment effects are positive (Proposition 1), reconciling our findings with prior meta-analyses. Second, under right-skewed intensity, cutoff-indexed effects deteriorate monotonically (Proposition 2), generating testable predictions about severity gradients.

Empirically, we test the theoretical predictions using Japanese PISA data and machine learning methods for heterogeneous treatment effects. The estimated severity gradient exhibits monotonic deterioration at high cutoffs in both cohorts. In 2018, where resilience is observed, variable importance analysis indicates that student resilience explains 16.1% of treatment effect heterogeneity, the largest share among all predictors. In 2015, where resilience is unavailable, age leads (17.8%), likely proxying developmental differences in coping capacity. Across both cohorts, family socioeconomic status ranks among the top predictors and accounts for roughly one-third (35–38%) of the decomposition-based variation in estimated effects, consistent with the resource mobilization interpretation.

Methodologically, we show that machine learning overcomes detection failures when vulnerability arises through high-dimensional interactions. With p binary covariates, exhaustive interaction testing requires $2^p - 1$ specifications and suffers severe power loss under multiple testing correction. By optimizing over the partition space, machine learning uncovers minorities (1–2%) experiencing losses of 40–60 points that conventional stratification misses, including young immigrants in 2015 and males in atypical schools in 2018, whose losses are comparable to those at the top 1% of bullying intensity.

Relation to literature. Our framework builds on the continuous treatment methods of Hirano and Imbens (2004) and the marginal treatment effect literature developed by Heckman and Vytlacil (2005), which analyze how treatment effects vary with intensity and unobserved resistance to treatment. We depart from these approaches by varying the treatment definition itself through thresholding realized intensity, rather than modeling a single continuous dose–response

function. This perspective makes explicit how aggregation across heterogeneous intensity regimes can produce positive average effects at low thresholds while concealing severe harm concentrated in the upper tail of the exposure distribution. Recent work combines machine learning with semi-parametric methods to estimate heterogeneous treatment effects (Wager and Athey, 2018; Athey et al., 2019). We adapt these tools for diagnostic purposes, identifying realized harm rather than optimizing treatment allocation. Several economics studies document treatment effects that vary sharply across intensity and baseline characteristics (Glewwe et al., 2009; Aizer and Doyle Jr, 2015; Pencavel, 2015), consistent with capacity constraints. Within the bullying literature, Eriksen et al. (2014) estimate long-run earnings effects using Danish administrative data, while Brown and Taylor (2011) examine neighborhood correlates; neither examines the nonlinear dose–response structure central to our framework.

The remainder proceeds as follows. Section 2 develops the theoretical framework. Section 3 describes data. Section 4 presents identification and estimation. Section 5 documents empirical results. Section 6 discusses limitations. Section 7 concludes.

2 The Economics of Adaptive Capacity: A Unified Framework

2.1 The Central Insight: Adaptation as Production Under Scarcity

The adversity paradox, where exposure is associated with higher average achievement yet severe harm for a minority, reveals a fundamental property of human adaptation: *adversity is both a destructive force and a trigger that mobilizes compensatory resources, with the balance determined by the scarcity and depletion of adaptive capacity*.

We formalize this within an economic framework in which students, families, and institutions allocate scarce adaptive resources under imperfect information about individual adaptive capacity and unobserved vulnerability thresholds. Adversity serves two simultaneous functions. First, it imposes direct damage through cognitive and emotional costs that reduce achievement via standard channels. Second, it triggers capacity mobilization by deploying previously dormant adaptive resources such as parental time, peer support, institutional intervention, and psychological resilience.

The adversity paradox emerges from the nonlinear interaction between these two functions under heterogeneous and depletable adaptive capacity. At low intensity, resource mobilization may partially offset direct costs, producing positive average associations in observational data. At high intensity, capacity constraints bind, and direct damage overwhelms adaptation as resources exhaust. Positive associations do not imply that adversity is beneficial; they reflect post-hoc responses to harm that has already occurred.

2.2 The Three-Regime Structure of Adaptive Response

Consider a student endowed with adaptive capacity $K_i > 0$, a multidimensional resource that includes psychological capital (resilience, self-efficacy, growth mindset), familial resources (parental time, financial capacity for tutoring or counseling), and institutional support (teacher attention, peer networks, school counseling services).

The dormancy regime ($d = 0$) is a conceptual baseline corresponding to students who report no bullying. All empirical estimands are defined relative to this no-adversity group. The mobilization regime begins at low but strictly positive intensity levels, below individual adaptive thresholds.

Adversity of intensity d generates both direct harm and triggers compensatory responses. We model the student's achievement production function as:

$$Y_i(d) = Y_i(0) - \underbrace{\lambda \cdot d}_{\text{direct damage}} + \underbrace{g_i(d, K_i)}_{\text{adaptive response}}, \tag{1}$$

where $Y_i(0)$ denotes baseline achievement in the absence of adversity (dormancy regime), $\lambda > 0$ captures inherent harm from adversity, and $g_i(d, K_i)$ represents the compensatory response that depends on both intensity d and the student's adaptive capacity K_i. The critical innovation is modeling g_i as having three distinct regimes that reflect the economics of scarce resource allocation under capacity constraints.

With no adversity in dormancy regime, adaptive resources remain unused: $g_i(0, K_i) = 0$. Families allocate resources to baseline levels. Students lack triggers to mobilize effort or seek support. Mild adversity ($0 < d < d_i^*$) activates latent capacity. Parents observe bullying, recognize a need for intervention, and reallocate resources. Students experience 'stress inoculation', moderate challenges that build coping skills (Meichenbaum, 1985). Peers and teachers mobilize support networks. Because capacity is abundant relative to demand in this regime, marginal returns are increasing:

$$g_i(d, K_i) = \beta_{1i} \cdot d \quad \text{for } d < d_i^*, \quad \beta_{1i} > \lambda, \tag{2}$$

where $\beta_{1i} > \lambda$ ensures the measured compensatory effect exceeds measured direct damage on the observed outcome, yielding $Y_i(d) > Y_i(0)$. This positive association on test scores does not imply net benefit; unmeasured costs (psychological distress, foregone leisure, parental stress) may exceed measured academic gains.

Beyond the threshold d_i^*, adaptive demand exceeds capacity. Resource exhaustion operates because parental time and attention are finite. A family investing 8 hours daily supporting a bullied child cannot double that indefinitely, and psychological reserves deplete faster than they replenish.

Diminishing returns appear as the marginal product of compensatory investment declines. Coordination failure emerges as severe adversity fragments support networks. In this regime, marginal costs of adaptation rise sharply while marginal benefits fall:

$$g_i(d, K_i) = \beta_{1i} \cdot d - \beta_{2i} \cdot (d - d_i^*)^2 \quad \text{for } d \geq d_i^*, \quad \beta_{2i} > 0. \tag{3}$$

The quadratic penalty captures accelerating harm as capacity is progressively exceeded. Achievement collapses as the system transitions from productive adaptation to crisis management. Combining these regimes, the treatment effect $\tau_i(d) \equiv Y_i(d) - Y_i(0)$ can be written as

$$\tau_i(d) = \tilde{\beta}_{1i}\, d - \beta_{2i}(d - d_i^*)^2 \mathbb{1}\{d \geq d_i^*\}, \tag{4}$$

where $\tilde{\beta}_{1i} \equiv \beta_{1i} - \lambda > 0$ denotes the net mobilization slope. For notational simplicity, we define the *net mobilization slope* $\beta_{1i} \equiv \tilde{\beta}_{1i} - \lambda =$ (compensatory response slope) − (direct damage rate) > 0, so that

$$\tau_i(d) = \beta_{1i} d - \beta_{2i}(d - d_i^*)^2 \mathbb{1}\{d \geq d_i^*\}.$$

Throughout the remainder of the paper, β_{1i} refers to this net mobilization slope unless otherwise noted.

Remark 1 (Notation Convention). Throughout this paper, $\tau_i(d)$ denotes the individual-level dose–response function mapping intensity to treatment effect for student i. When we aggregate, $\tau(x; c) \equiv \mathbb{E}[\tau_i(d_i) \mid X_i = x, d_i \geq c]$ denotes the conditional average treatment effect (CATE) among those with covariates x exceeding cutoff c, and $\theta(c) \equiv \mathbb{E}[\tau_i(d_i) \mid d_i \geq c]$ denotes the population average among all those above cutoff c. The cutoff-indexed estimand $\theta(c)$ is our primary target, tracing the dose–response as we vary the intensity threshold.

Individual adaptive thresholds d_i^* are unobserved and not identified in the data. They serve as a conceptual device to describe heterogeneous responses to intensity, not as estimable parameters. Our empirical analysis therefore focuses on population-level patterns in cutoff-indexed average effects. References to "regime transitions" denote changes in average responses across increasingly restrictive intensity subsets, rather than observed transitions at the individual level.

2.3 Visual Illustration: The Three-Regime Structure

Figure 1 visualizes the core mechanism. The figure displays dose–response curves for three students with different adaptive capacities: high (K_H, blue), medium (K_M, orange), and low (K_L, red). Each curve follows the functional form $\tau_i(d) = \beta_{1i} d - \beta_{2i}(d - d_i^*)^2 \mathbb{1}\{d \geq d_i^*\}$, exhibiting dormancy ($d = 0$), mobilization ($0 < d < d_i^*$, upward sloping), and depletion ($d \geq d_i^*$) regimes.

The vertical line at $d = 6$ illustrates why conventional analysis fails to detect depletion-regime

harm: at identical bullying intensity d, the high-capacity student shows a positive association ($\tau_H(6) = +2.3$), the medium-capacity student shows a modest positive association ($\tau_M(6) = +1.0$), while the low-capacity student suffers substantial harm ($\tau_L(6) = -0.9$).

2.4 Heterogeneous Capacity and the Distribution of Vulnerability

The threshold d_i^* is student-specific, determined by interactions across multiple dimensions: individual resilience (students with high self-efficacy and growth mindset have larger β_{1i} and smaller β_{2i}, yielding higher d_i^*), familial resources (high-ESCS families have greater capacity to mobilize compensatory investments), institutional context (adequate counseling and low student-teacher ratios provide external buffers that raise d_i^*), and baseline stress load (students already facing other stressors have pre-depleted capacity, lowering d_i^*).

This heterogeneity generates a latent distribution of vulnerability. At any intensity level d, the population partitions into mobilizers ($i : d < d_i^*$) whose compensatory responses offset observed harm, and the depleted ($i : d \geq d_i^*$) suffering capacity exhaustion. Students in the mobilization regime are not "benefiting from bullying"; rather, they are coping with harm through resource mobilization that may or may not fully compensate for unobserved costs. Vulnerability is conditional on realized intensity. A student with low capacity appears in the compensating group if they experience mild exposure but in the depleted group if they experience severe exposure. Standard stratification by observable characteristics fails to detect vulnerability because it averages over the intensity distribution, missing the critical interaction between d and $d_i^*(X_i)$.

2.5 Why Average Effects Reverse Sign: The Aggregation Problem

When a continuous treatment intensity is dichotomized at a cutoff, the resulting estimand averages treatment effects across distinct response regimes. Individuals above the cutoff may lie either below or above their adaptive threshold d_i^*, so the average effect mixes mobilization and depletion responses. As the cutoff increases, this composition shifts toward the upper tail where depletion dominates, generating sign reversals in average effects. Formally, define cutoff-indexed estimand:

$$\theta(c) = \mathbb{E}[Y_i(d_i) - Y_i(0) \mid d_i \geq c], \tag{5}$$

which averages the treatment effect at realized intensity d_i among those exceeding cutoff c.

If the intensity distribution $F(d)$ is right-skewed, most exposed students experience low-to-moderate intensity where $d < d_i^*$. An inclusive definition ($c \approx 0$) yields:

$$\theta(0) = \mathbb{E}[\beta_{1i} d_i - \beta_{2i}(d_i - d_i^*)^2 \mathbb{1}\{d_i \geq d_i^*\}] > 0, \tag{6}$$

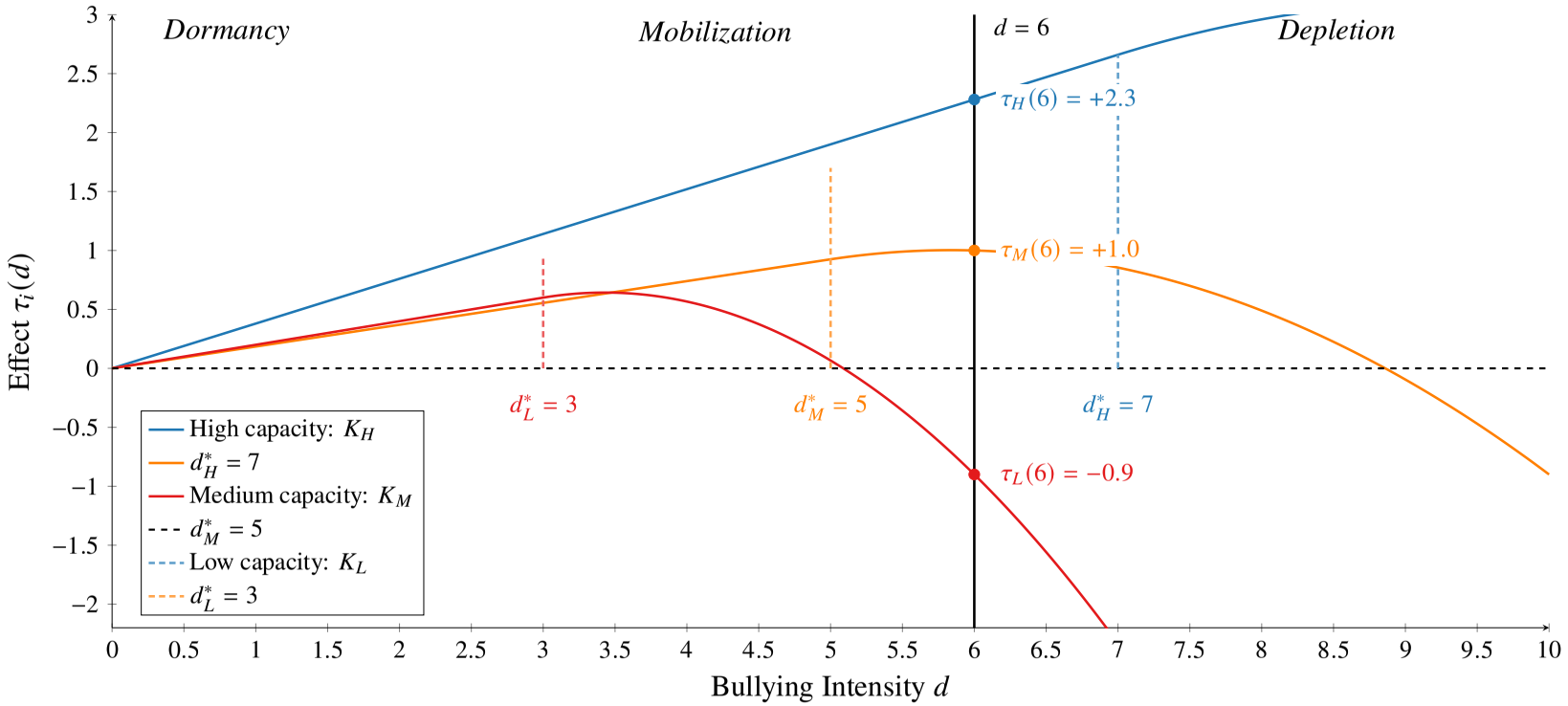


Figure 1: Three-Regime Dose–Response Structure with Heterogeneous Adaptive Capacity

Notes: The figure illustrates stylized individual dose–response functions under heterogeneous adaptive capacity. Parameter values are chosen for expositional clarity rather than estimated from data. For each individual i, achievement initially rises with exposure as compensatory responses are mobilized, may continue to increase briefly beyond the mobilization threshold d_i^* until the turning point d_i^{**}, and then declines once the quadratic penalty dominates. The turning point $d_i^{**} \equiv d_i^* + \beta_{1i}/(2\beta_{2i})$ marks the onset of strict deterioration. Monotonic decline applies only for exposure levels $d \geq d_i^{**}$ and, at the population level, for cutoffs $c \geq \bar{d}^{**}$. The "Mobilization" regime label refers to the intensity range where compensatory responses offset measured harm, producing positive associations on the observed outcome. This does not imply that adversity is beneficial; students in this regime are coping with harm, not benefiting from it. The figure is intended to convey the qualitative three-regime structure rather than quantitative magnitudes.

because the distribution concentrates in the mobilization regime. As the cutoff becomes more restrictive, the estimand increasingly weights the upper tail where depletion dominates. Consequently, $\theta(c)$ declines monotonically and eventually reverses sign. At the 99th percentile, nearly all students satisfy $d_i \gg d_i^*$, yielding

$$\theta(c_{99}) = \mathbb{E}[\beta_{1i} d_i - \beta_{2i}(d_i - d_i^*)^2 \mid d_i \geq c_{99}] < 0. \tag{7}$$

2.6 Formal Characterization: Two Key Results

We now formalize when linear regression yields negative coefficients despite positive average effects, and when cutoff-indexed effects decline monotonically.

Proposition 1 (Reconciling Positive Average Effects with Negative OLS Slopes)**.** *Suppose individual dose–response functions take the form*

$$\tau_i(d) = \beta_{1i} d - \beta_{2i}(d - d_i^*)^2 \mathbb{1}\{d \geq d_i^*\},$$

where $\beta_{1i} > 0$ denotes the net mobilization slope as defined in Section 2 (compensatory response minus direct damage), treatment intensity $d_i \sim F$ has support $[0, \bar{d}]$, and $(\beta_{1i}, \beta_{2i}, d_i^*)$ are mean-independent of d_i. Let $Y_i = Y_i(0) + \tau_i(d_i)$ with $\mathbb{E}[Y_i(0) \mid d_i] = \mathbb{E}[Y_i(0)]$. Then the population OLS coefficient from regressing Y_i on d_i satisfies

$$\beta_{\text{OLS}} = \underbrace{\mathbb{E}[\beta_{1i}]}_{\substack{\textit{average net} \\ \textit{mobilization slope}}} - \underbrace{\mathbb{E}[\beta_{2i}] \cdot \Lambda}_{\substack{\textit{depletion} \\ \textit{penalty}}}, \tag{8}$$

where

$$\Lambda \equiv \frac{\mathbb{E}\big[(d_i - d_i^*)^2 (d_i - \mathbb{E}[d_i]) \mathbb{1}\{d_i \geq d_i^*\}\big]}{Var(d_i)} > 0$$

whenever $\Pr(d_i > d_i^* + \epsilon) > 0$ *for some* $\epsilon > 0$ *and* d_i *is non-degenerate.*

Consequently, even when the average net marginal effect $\mathbb{E}[\beta_{1i}]$ is positive (i.e., compensatory mobilization exceeds direct damage on average), sufficient mass in the high-intensity depletion regime generates $\beta_{\text{OLS}} < 0$.

Proof. See Appendix A.1.

OLS estimates a best linear approximation to a nonlinear dose–response function. Proposition 1 shows that the negative OLS coefficient arises not merely because some slopes are negative, but because curvature is concentrated in the high-intensity tail. The term Λ captures how the quadratic depletion penalty covaries with intensity: it is large when substantial mass lies in the region $d_i \geq d_i^*$

where the penalty activates. Even if most students remain in the mobilization regime, a small mass of high-intensity observations can dominate the aggregate covariance and overturn the positive average marginal effect. But $\theta(0)$ estimates the average level of the measured treatment effect among exposed students, which can be positive if most remain in the mobilization regime. A positive $\theta(0)$ indicates successful compensation on the measured outcome, not that exposure is beneficial overall. This reconciles our findings with meta-analyses reporting negative correlations (Nakamoto and Schwartz, 2010; Kowalski et al., 2014).

Proposition 2 (Severity Gradient and Rate of Deterioration)**.** *Suppose (i) individual dose–response functions take the form* $\tau_i(d) = \beta_{1i}d - \beta_{2i}(d - d_i^*)^2\mathbb{1}\{d \geq d_i^*\}$ *with* $\beta_{2i} > 0$*, (ii) the intensity distribution* $F(d)$ *is continuously differentiable with density* $f(d) > 0$ *on* $(0, \bar{d})$*, and (iii) the population upper envelope of turning points* $\bar{d}^{**} \equiv \sup_i\{d_i^* + \beta_{1i}/(2\beta_{2i})\}$ *satisfies* $\bar{d}^{**} < \bar{d}$*. Define the cutoff-indexed estimand* $\theta(c) = \mathbb{E}[\tau_i(d_i) \mid d_i \geq c]$ *and the boundary effect* $\bar{\tau}(c) = \mathbb{E}[\tau_i(c)]$*. Then, for all* $c \geq \bar{d}^{**}$*,*

$$\frac{d\theta(c)}{dc} = h(c)\big[\theta(c) - \bar{\tau}(c)\big] < 0, \tag{9}$$

where $h(c) = f(c)/[1 - F(c)]$ *is the hazard rate of the intensity distribution.*

Proof. See Appendix A.2.

Interpretation. Once the cutoff c exceeds the highest individual turning point $\bar{d}^{**}$, all remaining observations lie in the strictly declining portion of their dose–response curves. Raising the cutoff therefore removes relatively better-off cases near the boundary and leaves increasingly severe exposures in the conditioning set. As a result, the average effect among those above the cutoff must decline with c.

2.7 The Detection Problem: Why Vulnerability Hides in High Dimensions

Traditional heterogeneity analysis relies on prespecified interactions between treatment and observed covariates. This approach is ill-suited because correct specification requires more than identifying which covariates interact and how they interact. Under our framework, interactions are intensity-dependent: the same covariate configuration can generate positive or negative effects depending on whether realized exposure lies below or above an individual's adaptive threshold. As a result, researchers must also correctly specify the intensity range at which an interaction becomes active. The detection problem is therefore three-dimensional, requiring simultaneous identification of relevant covariates, their interaction structure, and the intensity levels at which these interactions operate. For conventional methods, this is difficult very difficult task.

Generalized random forests (Wager and Athey, 2018; Athey et al., 2019) solve the detection problem by optimizing over the partition space rather than testing predetermined hypotheses. At

each node, the forest searches all possible splits of all covariates to maximize treatment effect heterogeneity, automatically discovering nonlinear thresholds and high-dimensional interactions without prespecifying functional form. Policy trees (Athey and Wager, 2021) translate the forest's partition into interpretable rules.

3 Data and Descriptive Statistics

3.1 Data Sources and Sample Construction

We analyze data from the Programme for International Student Assessment (PISA), focusing on Japan in 2015 (n = 6,647 students across 198 schools) and 2018 (n = 6,109 students across 183 schools), the first two PISA waves measuring student-level bullying victimization (OECD, 2016, 2019).

Japan provides a useful setting. The PISA Japan samples are large with high response rates. Japan's relatively homogeneous education system, with standardized curricula and low between-school resource variation, helps isolate within-student differences. In our PISA data, only about 1% of students are foreign-born, creating small but identifiable immigrant groups and allowing a test of whether machine learning can detect vulnerable minorities missed by conventional stratification.

Table 1 summarizes key characteristics. Both cohorts exhibit similar demographics, including near-equal gender balance, a mean age of 15.78 years, and low immigration rates (approximately 1%). Under our inclusive definition, bullying prevalence is 46.4% in 2015 and 36.8% in 2018, implying that the majority of students are classified in the dormancy regime, where exposure levels are mild relative to students' adaptive capacity.

3.2 Variable Construction

3.2.1 Outcome: Academic Achievement

We construct a composite achievement measure from PISA's reading, mathematics, and science scores using principal components analysis. The first component explains approximately 89% of variance in both cohorts, capturing general academic proficiency rather than domain-specific skills.

3.2.2 Treatment: Bullying Victimization

PISA measures bullying using six items capturing victimization frequency over the past 12 months on a four-point scale. Following PISA methodology, we aggregate these items into a continuous bullying intensity index using item response theory (IRT) (see OECD, 2017, 2020, for details). Our baseline treatment defines bullying as reporting any victimization, that is, responding other

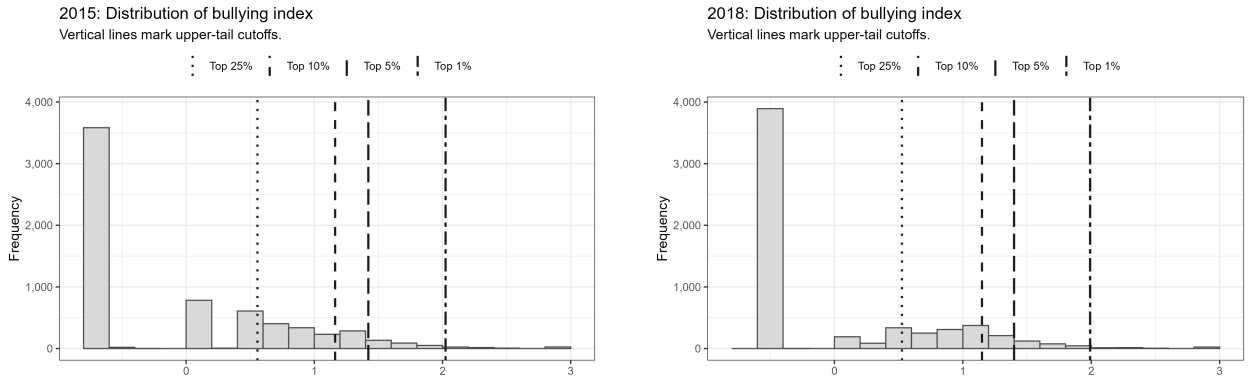
Table 1: Sample Characteristics

Characteristic	2015	2018
Students (n)	6,647	6,109
Schools (k)	198	183
Test score, mean (SD)	529.1 (91.6)	520.0 (91.9)
Bullied (%)	46.4	36.8
Male (%)	50.2	48.9
Mean age (years)	15.78	15.78
Born abroad (%)	0.9	1.1
ESCS index, mean	−0.18	−0.11
Number of covariates	13	16

Notes: PISA sampling weights applied. Test score is first principal component of reading, math, and science. ESCS constructed using Japan-specific factor loadings. Sample sizes reflect students with complete data on all covariates; 40 students in 2015 and 159 students in 2018 are excluded due to missing bullying index values.

than "never" to at least one item. This intentionally inclusive definition captures a broad range of exposure, including mild and infrequent episodes.

Figure 2 illustrates distributional features central to our framework. In both cohorts, the bullying index exhibits substantial mass at the lower bound and pronounced right skewness (0.95 in 2015 and 1.21 in 2018), with considerable upper-tail variation. These features imply that most exposed students remain in the mobilization regime with $d < d_i^*$, while a small fraction transition into depletion. Detailed distributional diagnostics reported in Table A3 in Appendix D.2 confirm these patterns and satisfy the conditions required for Proposition 2.



(a) 2015 (b) 2018

Figure 2: Distribution of Bullying Index with Upper-Tail Thresholds

Notes: Vertical lines indicate 75th, 90th, 95th, and 99th percentiles. Both distributions are strongly right-skewed with pronounced mass at the lower bound.

3.2.3 Covariates: Measuring Adaptive Capacity

Both cohorts include measures of psychological capital (student resilience, parental emotional support), familial resources (ESCS: Economic, Social, and Cultural Status index), and institutional support (teacher support, staff shortage, student–teacher ratio). Staff shortage is a school-level index measuring principals' perceptions of inadequate or poorly qualified teaching staff.

A key limitation is that resilience, capturing self-efficacy and growth mindset, is observed only in 2018. This measure directly captures the psychological component of adaptive capacity K_i underlying individual thresholds d_i^*.

Immigration status is measured by student's country of birth. In Japan's context, the "born abroad" category is heterogeneous: it includes both foreign nationals (recent immigrants, children of migrant workers) and Japanese citizens born overseas (e.g., children of businesspeople or diplomats stationed abroad, known as *kikokushijo*). These subgroups likely differ substantially in language proficiency, cultural integration, and adaptive capacity. PISA data do not permit disaggregation. The 2015 finding that young foreign-born students experience depletion (Section 5.7) should therefore be interpreted cautiously, as effects may be driven by a specific subgroup within this heterogeneous category.

4 Empirical Framework

4.1 Identification

Let Y_i denote student i's test score, $d_i \geq 0$ denote continuous bullying intensity, and $X_i \in \mathbb{R}^p$ denote pre-treatment covariates. For each cutoff c, we define a binary exposure indicator $W_i(c) = \mathbb{1}\{d_i \geq c\}$, where $W_i(c) = 1$ corresponds to being in the upper tail of the realized intensity distribution.

To keep the estimand aligned with our empirical implementation, we define cutoff-specific binary potential outcomes $Y_i(w; c)$, $w \in \{0, 1\}$, where $Y_i(1; c)$ is the potential outcome under $W_i(c) = 1$ and $Y_i(0; c)$ is the potential outcome under $W_i(c) = 0$. Our primary estimand is the average effect among those classified as exposed under cutoff c:

$$\theta(c) \equiv \mathbb{E}[\, Y_i(1; c) - Y_i(0; c) \mid W_i(c) = 1 \,]\,. \tag{10}$$

We also consider the conditional analogue

$$\tau(x; c) \equiv \mathbb{E}[\, Y_i(1; c) - Y_i(0; c) \mid X_i = x,\, W_i(c) = 1 \,]\,, \tag{11}$$

which characterizes how treatment effects vary across observable characteristics among students

exposed at cutoff c. This estimand conditions on both covariates X_i and exposure status $W_i(c) = 1$, differing from the standard CATE definition that conditions only on X_i.

We adopt this definition because our diagnostic objective is to identify heterogeneity *within the exposed population*, where regime transitions from mobilization to depletion occur. Conditioning on exposure status ensures that $\tau(x; c)$ characterizes effect variation among students actually experiencing adversity at or above threshold c, rather than averaging over unexposed students for whom the treatment effect is definitionally zero.

This conditional estimand serves two purposes. First, it aligns with our conceptual framework in which adaptive capacity and vulnerability depend on student, family, and institutional characteristics captured in X_i. Second, it corresponds to the object estimated by machine learning methods, which recover heterogeneity in effects as a function of covariates rather than only population averages. The population estimand $\theta(c)$ is then obtained by averaging these conditional effects over the covariate distribution of the exposed group, so that $\theta(c) = \mathbb{E}[\tau(X_i; c) \mid W_i(c) = 1]$.

Our objective is diagnostic: to distinguish realized harm (the depletion regime) from compensatory responses that offset direct damage (the mobilization regime). Identification relies on three standard assumptions:

Assumption 1 (Selection on Observables): For each cutoff c, conditional on observed pre-treatment covariates X_i, assignment to the cutoff-defined exposure $W_i(c) = \mathbb{1}\{d_i \geq c\}$ is as good as random. Formally,

$$\{Y_i(1; c), Y_i(0; c)\} \perp\!\!\!\perp W_i(c) \mid X_i.$$

This assumption allows differences in outcomes between exposed and unexposed students at cutoff c to be interpreted as reduced-form effects consistent with causal impacts under the maintained identification assumptions. We evaluate this assumption through sensitivity analysis (Section 5.9, Appendix E.1).

Assumption 2 (Overlap): $0 < \mathbb{P}[W_i(c) = 1 \mid X_i = x] < 1$ for all x in the support of X_i.

We assess overlap by examining propensity score distributions (Figure 3 for 2018; Figure A1 for 2015). Both distributions show substantial density throughout their support. Fewer than 1% of observations fall below 0.1 or above 0.9, and no observations fall outside $[0.05, 0.95]$. We do not implement propensity score trimming, but results are robust to excluding observations with extreme propensity scores (available upon request).

Assumption 3 (SUTVA, cutoff-specific): For each cutoff c, student i's observed outcome equals the potential outcome under their realized cutoff-specific exposure status:

$$Y_i = Y_i\big(W_i(c); c\big).$$

This is an approximation given bullying's social nature. Spillovers operating through classroom

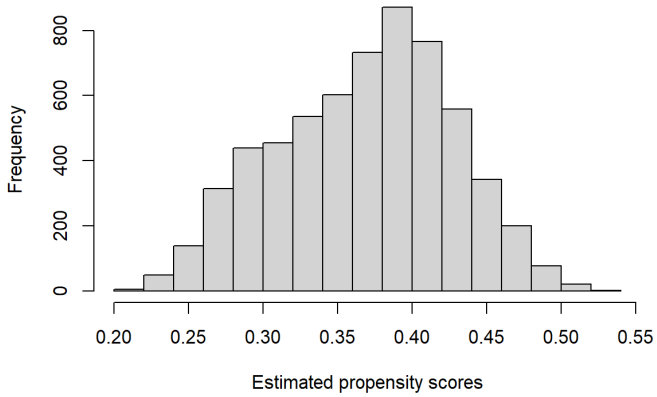


Figure 3: Distribution of Estimated Propensity Scores (2018)

Notes: Propensity scores from regression forest. Distribution satisfies overlap with density away from boundaries.

climate or peer dynamics could bias estimates in either direction. Our estimates capture reduced-form effects combining direct impacts and any spillovers.

Interpretive Caution. Our identification strategy relies on selection on observables. While we include rich controls spanning demographics, family background, psychological resources, and school characteristics, unobserved confounders may remain. We therefore interpret estimates as conditional associations consistent with causal effects under maintained assumptions, rather than definitively established causal impacts. The convergent evidence from multiple diagnostics (Section 5.4) strengthens but cannot prove the causal interpretation.

4.2 Estimation

Our estimation strategy evaluates two testable predictions of the adaptive-capacity framework. The first is the *adversity paradox*: under an inclusive definition of bullying intensity, average effects are non-negative because most exposed students remain in the mobilization regime. The second is the existence of *regime transitions*: as exposure intensity increases, average effects deteriorate and eventually become negative as students transition into the depletion regime. Proposition 2 formalizes this second prediction by implying that the cutoff-indexed estimand $\theta(c)$ should be non-increasing for sufficiently large c.

To test these predictions, we implement a three-stage double machine learning procedure. First, we estimate propensity scores $\hat{e}(x)$ using a regression forest. Second, we estimate baseline outcomes $\hat{m}(x)$ using a separate regression forest. Third, we grow a causal forest on residualized

outcomes and treatments. This orthogonalization removes direct confounding, allowing the causal forest to recover heterogeneity in treatment effects associated with differences in adaptive capacity K_i and realized intensity d_i. Details are provided in Appendix C.1.

Our primary estimand is the average treatment effect among the exposed at cutoff c, estimated as $\hat{\tau}_{\text{ATT}} = \sum_i W_i(c)\hat{\tau}(X_i; c)/\sum_i W_i(c)$. Standard errors account for clustering at the school level and incorporate PISA sampling weights. To trace regime transitions, we repeat the estimation across a sequence of nested treatment definitions defined by progressively higher intensity thresholds: any victimization, top 25%, top 10%, top 5%, and top 1% of the bullying intensity distribution.

4.3 Addressing Reverse Causality

A fundamental concern is reverse causality: high-achieving students may attract bullying through peer jealousy. We address this through four diagnostics.

First, we exploit the severity gradient. Under pure reverse causality, positive associations should persist across intensity levels. Instead, we observe sign reversal, inconsistent with a uniform selection mechanism. Second, we test whether treatment effects vary with predicted baseline achievement. Third, we examine whether predicted achievement directly predicts bullying exposure. Fourth, we implement bounds following Oster (2019) to assess how much selection on unobservables would be required to eliminate estimated effects.

4.4 Detecting Hidden Vulnerabilities Through Policy Trees

We construct policy trees (Athey and Wager, 2021) to translate heterogeneous effects into interpretable targeting rules identifying students in the depletion regime. Policy trees recursively partition the covariate space to maximize treatment effect differences across groups. We constrain depth to two levels (four terminal nodes) for interpretability.

5 Results

5.1 Regime I: Average Effects Under Low Observed Intensity

Table 2 establishes the adversity paradox: students reporting any bullying victimization score 17.5 points higher (2015) and 11.3 points higher (2018) than non-bullied peers, with confidence intervals excluding zero. Under an inclusive definition aggregating mild to severe experiences, victimization associates with *higher* achievement.

This positive average reflects that most exposed students experience low-to-moderate intensity with $d_i < d_i^*$ (Figure 2), where compensatory responses offset direct damage on the measured

outcome. These responses are triggered by harm already inflicted and should not be interpreted as evidence that exposure itself is beneficial. The positive association indicates successful coping, not that bullying improves outcomes. Students would likely perform at least as well, and suffer less unmeasured harm, in the absence of victimization.

5.2 Regime Transitions: Testing Proposition 2

As cutoffs increase beyond students' adaptive thresholds d_i^*, we progressively select into the depletion regime, where capacity exhaustion dominates. Table 2 tests this prediction by estimating $\theta(c)$ across five nested cutoffs.

Results broadly confirm the three-regime structure, with one apparent local irregularity. In 2018, the point estimate for the top 25% cutoff slightly exceeds that for any victimization (11.3 $\rightarrow$ 12.9). This difference is small relative to sampling uncertainty: the implied difference of 1.6 points has a standard error exceeding 4 points and is therefore statistically indistinguishable from zero ($z \approx 0.32$). Consequently, the ordering of these two point estimates is uninformative about the underlying population dose–response.

Beyond the top decile, effects deteriorate monotonically in both cohorts. Point estimates fall from modest losses to large negative values, and this ordering is robust to sampling uncertainty. The sign reversal around the top decile marks the onset of the depletion regime, where average effects become negative and continue to decline with increasing severity.

We define the *depletion-vulnerable population* as students above the 90th percentile of bullying intensity, where average treatment effects turn negative. This threshold-based definition is empirically motivated: Table 2 shows that the sign reversal from positive to negative average effects occurs between the top 25% (+7.5 to +12.9 points) and top 10% (−9.8 to −6.5 points) cutoffs in both cohorts. The 90th percentile thus marks the approximate onset of the depletion regime at the population level, though individual students may transition at different intensity levels depending on their adaptive capacity d_i^*.

Within this group, losses escalate sharply across severity bands: students in the top decile incur modest declines of 7–10 points (*early depletion*), those in the top 5% lose roughly 20 points (*substantial depletion*), and those in the top 1% suffer losses exceeding 40 points (*severe depletion*).

5.3 Selection Diagnostics

Positive average effects could partly reflect selection if higher-achieving students are more likely to recognize or disclose mild victimization. Appendix D.3 reports diagnostics.

Figure A2 shows bullying incidence rises modestly from 43.5% (lowest quintile) to 50.2% (highest) in 2015, a 6.7pp gradient. In 2018, the gradient is much smaller (1.2pp). Figure A3

Table 2: Empirical Evidence for Three-Regime Structure: Monotonic Deterioration in Treatment Effects Across Bullying Intensity Definitions

	2015 Cohort			2018 Cohort			Δ from Baseline
Definition	**ATT**	**N**	**Share**	**ATT**	**N**	**Share**	**(2015 / 2018)**
Any victimization	+17.5*** [12.8, 22.2]	3,084	0.464	+11.3*** [5.6, 17.0]	2,248	0.368	—
()							
Top 25% (quartile)	+7.5** [2.1, 13.0]	1,702	0.256	+12.9*** [6.8, 18.9]	1,625	0.266	−10.0 / +1.6
(Late mobilization)							
Top 10% (decile)	−9.8** [-18.2, -1.4]	711	0.107	−6.5 [-15.1, 2.1]	819	0.134	−27.3 / −17.8
(Early depletion)							
Top 5%	−18.6*** [-29.8, -7.5]	379	0.057	−20.0*** [-31.7, -8.4]	446	0.073	−36.1 / −31.3
(Substantial depletion)							
Top 1%	−58.8*** [-82.1, -35.6]	120	0.018	−45.7*** [-62.3, -29.1]	226	0.037	−76.3 / −57.0
(Severe depletion)							

Notes: ATT estimates from causal forests with cluster-robust standard errors at school level incorporating PISA sampling weights; 95% confidence intervals in brackets. The final column reports the change from baseline (Any victimization) separately for each cohort (2015 / 2018). The "Top 1%" category includes all students at or above the 99th percentile of the bullying index. Due to the discrete nature of the IRT-scaled index with mass points, this threshold captures 1.8% of students in 2015 and 3.7% in 2018. We retain this labeling convention for comparability with the continuous-treatment literature. Proposition 2 predicts monotonic deterioration only for $c \geq \bar{d}^{**}$. The "Mobilization regime" label refers to intensity levels where compensatory responses offset measured academic harm, producing positive associations; this does not imply that exposure is beneficial overall. The "Depletion regime" labels refer to intensity levels where harm exceeds compensatory capacity.

documents that bullied students have higher ESCS on average.

Yet this selection pattern cannot explain our estimated effects. If positive results were driven by selection alone, controlling for observables should attenuate estimates. Instead, covariate adjustment increases estimates (from +19.54 to +21.16 in 2015; Table A6), and propensity distributions exhibit strong overlap (Figures 3 and A1). Most importantly, selection cannot account for monotonic deterioration from positive to severely negative effects along the intensity gradient.

5.4 Testing Reverse Causality

Four diagnostic tests assess whether reverse causality, in which high achievers attract bullying through peer jealousy, explains positive average effects.

The severity gradient provides the strongest evidence against reverse causality. Under jealousy-driven targeting, positive associations should persist across intensity levels. Instead, effects reverse sign from +17.5 to −58.8 points, exhibiting the regime transition pattern predicted by the adaptive capacity framework but inconsistent with a uniform selection mechanism.

Treatment effects are uniform across predicted achievement quintiles, with no gradient ($p > 0.8$, Table A4). Most decisively, predicted achievement has a significantly negative coefficient when predicting bullying exposure: $\beta = -0.430$ (SE = 0.047, $p < 0.001$) in 2015 and $\beta = -0.083$ (SE = 0.033, $p = 0.011$) in 2018. High achievers are less likely to be bullied; this pattern is inconsistent with jealousy-driven targeting.

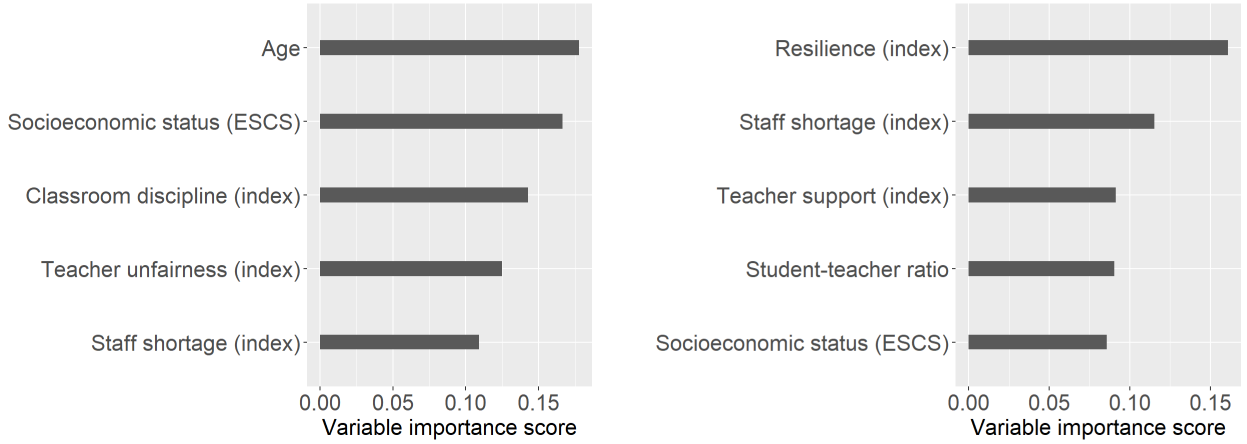
Oster bounds (Table A6) indicate that unobservables would need to select in the opposite direction from observables in 2015, or be nearly nine times as important in 2018, to eliminate the estimated effects. We report these bounds as heuristic sensitivity diagnostics because our primary estimands are nonlinear and obtained via machine learning, so the assumptions underlying Oster (2019) do not strictly apply. These results support the interpretation that positive average effects reflect compensatory mobilization rather than reverse causality.

5.5 Evidence for Adaptive Capacity as a Determinant of Regime Transitions

To identify characteristics shaping individual adaptive thresholds, we examine variable importance from the causal forest (Figure 4). Across cohorts, measures of psychological, familial, and institutional resources rank highest. Student resilience is the most important predictor in 2018 (16.1%). In 2015, where resilience is unavailable, age leads (17.8%), likely proxying developmental differences in coping capacity. Socioeconomic status and school staff shortage consistently appear among the top five predictors in both cohorts, mapping directly to the adaptive capacity framework: ESCS reflects household capacity for compensatory investment, and staff shortages proxy institutional

constraints on external support. The prominence of age in 2015 may reflect that younger students have less developed adaptive resources, consistent with the policy tree finding that younger foreign-born students experience the most severe depletion.

Variable importance captures predictive relevance for treatment effect heterogeneity rather than causal effects. Even so, the stable prominence of capacity-related variables is consistent with the interpretation that regime transitions are driven by differences in adaptive capacity, as implied by our conceptual framework.



(a) 2015: Age leads (17.8%)

(b) 2018: Resilience leads (16.1%)

Figure 4: Variable Importance for Treatment Effect Heterogeneity

Notes: Variables ranked by contribution to treatment effect heterogeneity. Prominence of resilience in 2018 (16.1%) is consistent with adaptive capacity framework, though variable importance reflects predictive relevance rather than causal importance.

5.6 Heterogeneity Tests and Conventional Stratification

Our empirical strategy relies on causal forests, which are designed to estimate heterogeneous rather than average treatment effects. This focus is important because the adaptive-capacity framework predicts pronounced nonlinearity in responses to bullying intensity, with large negative effects concentrated in the tails of the exposure distribution. If treatment effects were in fact homogeneous, machine learning methods would add little beyond conventional estimators, and the resulting patterns would simply reflect average effects that can be recovered by standard methods. Testing for heterogeneity is therefore a prerequisite for interpreting forest-based estimates as evidence of differential responses rather than estimation noise.

Calibration tests provide varying evidence for treatment effect heterogeneity, with marginal significance in 2015 ($p = 0.095$) and strong rejection in 2018 ($p = 0.002$, Table A7). These results are consistent with treatment effect heterogeneity, with evidence that varies in strength across

cohorts. This pattern suggests that heterogeneity arises from complex interactions across multiple dimensions of student and school characteristics, which conventional stratification approaches may detect only partially and with limited power.

Consistent with this interpretation, traditional stratification by gender, socioeconomic status, age, and immigration status yields predominantly positive average effects across all categories (Appendix F.1). Only immigrants exhibit negative point estimates (−63.2 in 2015 and −6.0 in 2018), but neither estimate is statistically significant, illustrating the limited ability of coarse subgroup analysis to isolate vulnerable populations.

These findings illustrate a fundamental detection problem. Stratification by observable characteristics X_i averages over the full intensity distribution, obscuring the interaction between individual adaptive thresholds $d_i^*(X_i)$ and realized exposure intensity d_i. As a result, conventional approaches fail to detect heterogeneity that emerges primarily in the upper tail of the intensity distribution.

5.7 Detecting the Depletion Regime Through Policy Trees

Policy trees are used solely as diagnostic tools to show how high-dimensional heterogeneity can hide severe harm in small subpopulations. They are not intended to produce targeting rules or policy recommendations. Because the identified groups are small and differ across cohorts, the results should be interpreted as descriptive evidence of latent vulnerability rather than as stable criteria for intervention.

5.7.1 2015: Younger Students Born Abroad Experience Depletion

Figure 5 displays the 2015 tree. The primary split at age 16 separates younger students (+12.40 points) from older students (+28.08 points). Among younger students, native-born show positive associations consistent with successful mobilization (+13.11 points) while students born abroad ($n = 45$, 0.7%) experience severe depletion (−58.28 points), nearly identical to the 58.8-point loss at top 1% intensity (reported in Table 2).

The mechanism likely reflects low adaptive capacity d_i^*. However, as noted in Section 3, the "born abroad" category is heterogeneous, including both recent immigrants (who may face language barriers and cultural adaptation stress) and Japanese citizens born overseas (*kikokushijo*, who typically have stronger institutional support upon return). If the effect is driven primarily by the former subgroup, language barriers may limit families' ability to mobilize compensatory resources, while baseline stress from cultural adaptation reduces cognitive and emotional reserves. This interpretation should be treated as tentative given the heterogeneity of the category.

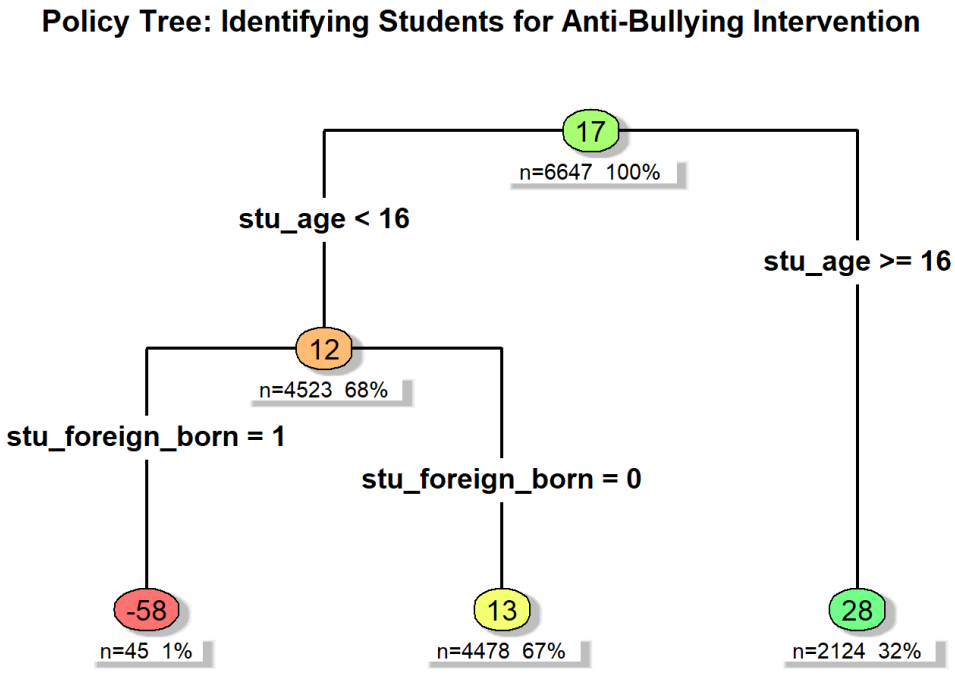


Figure 5: 2015 Policy Tree: Young Immigrants Transition to Depletion Regime

Notes: Depth-2 tree on Augmented Inverse Probability Weighted (AIPW) scores. Younger immigrants ($n = 45$, 0.7%) suffer severe capacity exhaustion, losing 58 points, a magnitude comparable to that of the most severely exposed students.

5.7.2 2018: Males in Atypical Institutional Settings Experience Depletion

Figure 6 displays the 2018 tree. Among females, older students (n = 997, 16%) experience early depletion: −10.95 points. Among males, those in conventional schools (student-teacher ratio > 2) show positive associations consistent with successful mobilization (+21.08 points) while males in atypical settings (ratio < 2, n = 91, 1.5%) experience substantial-to-severe depletion: −40.62 points.

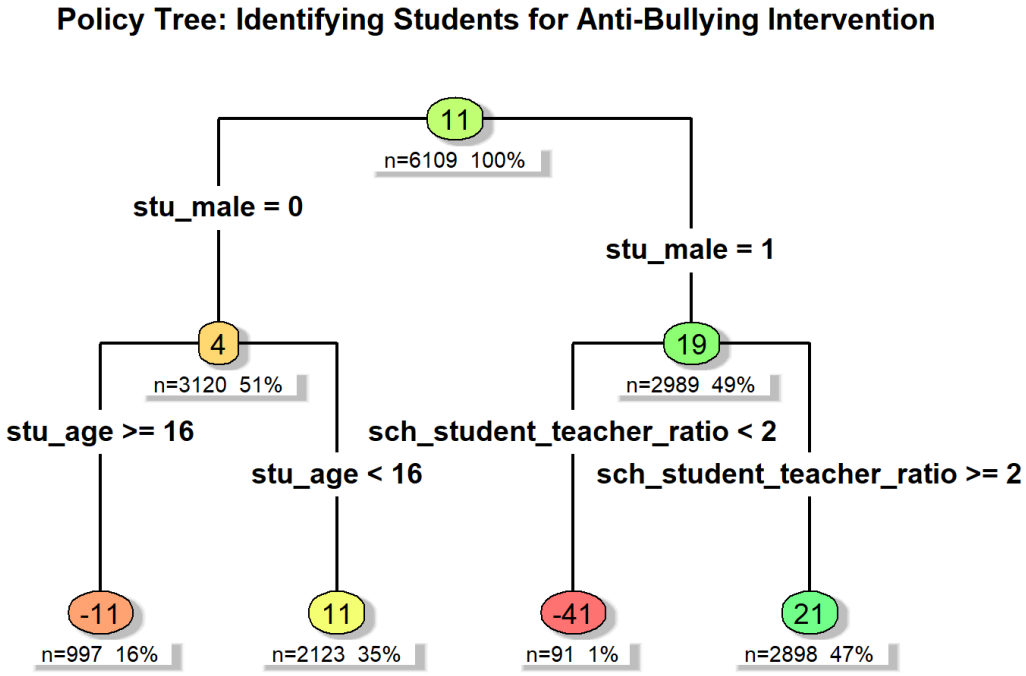


Figure 6: 2018 Policy Tree: Males in Atypical Schools Transition to Depletion Regime

Notes: Depth-2 tree. Males in schools with ratios < 2 (n = 91, 1.5%) experience substantial-to-severe capacity exhaustion, losing 41 points.

Very low student–teacher ratios (< 2) likely reflect special education environments, alternative schools, or small specialized programs serving at-risk populations[1]. Students in these settings often enter with reduced baseline adaptive capacity, lowering d_i^*. As a result, even moderate bullying exposure can quickly exceed their remaining adaptive reserves.

[1] Although PISA provides only anonymized public-use microdata, the school-level information indicates that student–teacher ratios below 2 are rare, accounting for approximately 3% of schools in the 2018 wave. Individual schools cannot be identified, but the distributional evidence suggests these settings are highly atypical.

5.7.3 Synthesis

Table 3 compares policy tree results with severity-gradient findings. Within the roughly 10% of students exhibiting negative average effects, policy trees isolate small minorities (approximately 1–2%) experiencing the most severe losses, comparable in magnitude to those observed at the top 1% of bullying intensity. These results illustrate that machine learning methods can, in principle, uncover extreme vulnerability concentrated in the upper tail of the intensity distribution that conventional stratification fails to detect.

At the same time, the policy tree results are intended as diagnostic illustrations rather than as a basis for targeting. While effect magnitudes align closely with the depletion-regime predictions (approximately −40 to −60 points), the specific observable characteristics identified should not be interpreted as stable markers of vulnerability without further validation.

Table 3: Severe Capacity Exhaustion Within Depletion-Regime Populations

Cohort	Subgroup	N	%	Effect	Top 1%
2015	Younger born abroad	45	0.7%	−58.28	−58.8
2018	Males, ratio < 2	91	1.5%	−40.62	−45.7
	Older females	997	16.3%	−10.95	

Notes: Vulnerable subgroups from policy trees compared to top 1% of intensity. The identified vulnerable subgroups differ entirely across cohorts (young immigrants in 2015 versus males in atypical schools in 2018), raising concerns about overfitting or genuine context dependence. With only 45–91 observations in vulnerable subgroups and only two cross-sections available, we cannot distinguish between these explanations.

5.8 Decomposition Analysis: Role of Family Resources

Decomposition analysis indicates that ESCS covaries with approximately one-third of total treatment effect heterogeneity: 37.7% in 2015 and 35.4% in 2018 (Appendix F.2, Figure A6). This is a descriptive decomposition measuring statistical association, not causal mediation. The pattern is consistent with the resource mobilization mechanism but could also reflect confounding: high-ESCS families may differ on unobserved dimensions (e.g., parental involvement, neighborhood quality) that independently affect both bullying exposure and academic achievement. The sequential ignorability assumptions required for causal mediation interpretation (Imai et al., 2010) are implausible here: ESCS likely correlates with unobserved determinants of both bullying exposure (e.g., school quality, neighborhood characteristics) and academic achievement (e.g., parental involvement, cognitive stimulation at home). We interpret these results as suggestive correlational evidence that family resources covary with treatment effect heterogeneity, not as estimates of causal mediation pathways.

5.9 Sensitivity Analysis

Sequentially removing key covariates produces changes of less than 2% of baseline estimates (Appendix E.1, Figure A4). The severity gradient does not appear to be driven by specific controls. Although unobserved confounding cannot be ruled out, explaining the sharp sign reversal at high intensity would require unobservables strongly correlated with bullying severity and acting opposite to observed covariates. Accordingly, we interpret the results as reduced-form evidence consistent with the adaptive-capacity mechanism under maintained assumptions.

5.10 Reconciling with Prior Literature

Our positive average effects appear to contradict meta-analyses reporting negative correlations between bullying and achievement (Nakamoto and Schwartz, 2010; Kowalski et al., 2014). Proposition 1 resolves this apparent difference in findings.

The divergence arises because the estimands answer different questions. The cutoff-indexed estimand $\theta(c)$ averages treatment effect *levels* among those exposed above threshold c. OLS estimates a best linear approximation to a nonlinear dose–response, dominated by marginal effects where curvature is active. When depletion is threshold-activated, negative curvature concentrates in the upper tail, generating excess covariance that can overturn positive average effects.

Figure 7 illustrates with a numerical example (complete derivations are in Appendix A.3). Consider $\tau(d) = 10d - 4(d-5)^2\mathbb{1}\{d \geq 5\}$ with $d \sim \text{Uniform}[0, 12]$. Despite severe harm at high intensity ($\tau(12) = -76$), the average treatment effect is positive ($\theta(0) \approx 21.9$) because most mass lies in the mobilization regime. Yet OLS yields a negative slope ($\beta_{\text{OLS}} \approx -3.50$) because it captures average marginal effects dominated by the depletion regime's negative curvature.

Both characterizations correctly summarize different aspects of the same underlying three-regime structure. Meta-analyses asking whether more exposure hurts recover negative gradients because they estimate slopes in the depletion regime. Our analysis finds positive average associations because most students remain in the mobilization regime where compensatory responses offset harm on measured outcomes. The apparent contradiction dissolves once we recognize that average *levels* and average *slopes* are distinct estimands that can have opposite signs under nonlinear dose–response with threshold-activated depletion.

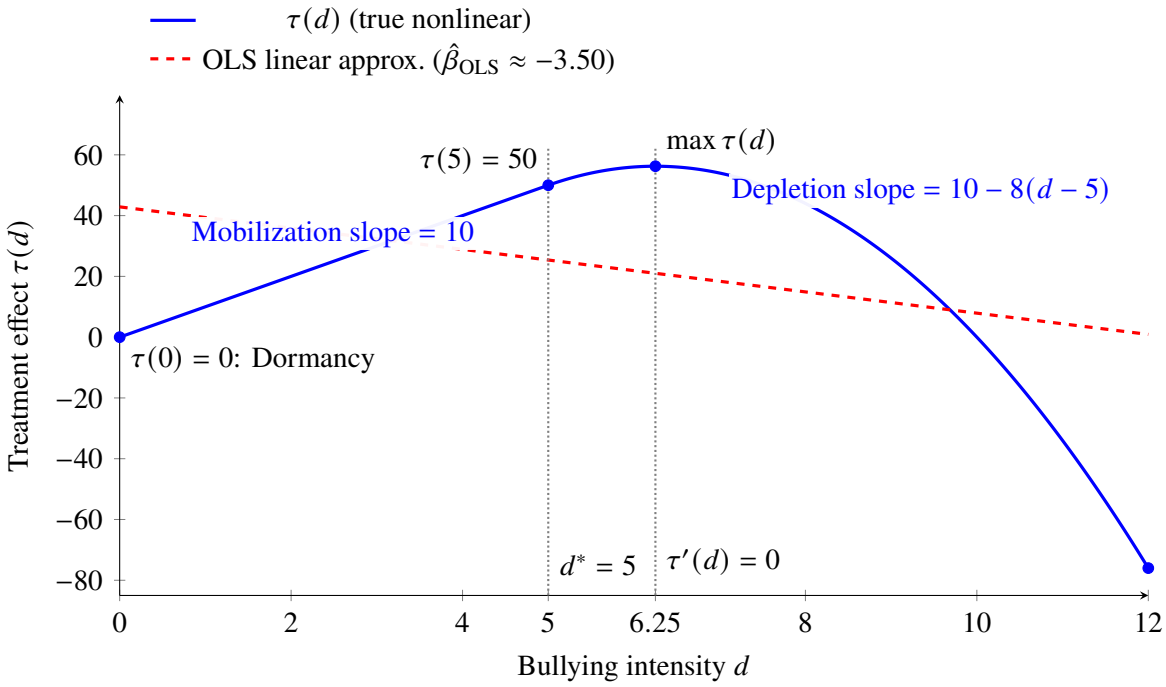


Figure 7: Reconciling Positive Average Effects with Negative OLS Slopes

Notes: Numerical illustration with $\tau(d) = 10d - 4(d-5)^2\mathbb{1}\{d \geq 5\}$ under $d \sim \text{Uniform}[0, 12]$. The average treatment effect $\theta(0) \approx 21.9 > 0$ because most mass lies in the mobilization regime where effects are positive. Yet OLS yields $\beta_{\text{OLS}} \approx -3.50 < 0$ because it estimates the average slope, which is dominated by negative curvature in the depletion regime. Both characterizations are correct: they measure different aspects of the same nonlinear dose–response structure.

6 Limitations and Future Research

6.1 Temporal Instability of Depletion-Vulnerable Subgroups

The identities of populations most vulnerable to capacity exhaustion differ across cohorts, with younger foreign-born students identified in 2015 and males in atypical schools in 2018. This temporal instability admits two interpretations.

One possibility is genuine context dependence. The mapping from observable characteristics to adaptive capacity, $X_i \to d_i^*(X_i)$, may shift with cohort composition and social conditions. Under this interpretation, the three-regime dose–response structure remains stable even as the observable correlates of vulnerability change.

An alternative explanation is statistical overfitting. Policy trees search a high-dimensional covariate space and may identify patterns that fail to replicate, particularly when vulnerable groups are small. With only two cross-sections, out-of-sample validation is not feasible.

Several features favor the context-dependence interpretation. Effect magnitudes for identified subgroups closely match those observed at the top 1% of intensity, and the cutoff at which effects turn negative (around the 90th percentile) is stable across cohorts. Nonetheless, these findings argue against fixed targeting rules. Effective policy design requires replication across additional cohorts to identify populations with low adaptive capacity d_i^*.

6.2 Generalizability

All findings are based on Japanese PISA cohorts. Japan's education system has distinctive features that may shape adaptive capacity. While the three-regime framework should apply broadly to settings with heterogeneous adaptive capacity, continuous exposure intensity, and nonlinear dose–response relationships, the specific groups with low d_i^* and the thresholds at which regime transitions occur are likely to differ across countries.

Applying the framework to PISA data from other countries, or to TIMSS data (Mullis et al., 2020), would help assess the external validity of the three-regime framework.

6.3 Cross-Sectional Design

We cannot determine whether the positive effects observed in the mobilization regime persist over time or instead reflect short-run compensatory effort that later gives way to depletion. The data capture test performance but not longer-run outcomes such as educational attainment, mental health, or labor market success. If successful adaptation in the mobilization regime builds durable resilience, long-run effects may exceed short-run effects. Conversely, if compensatory effort

imposes hidden costs (parental stress, foregone leisure), the apparent positive associations observed in the mobilization regime may erode or reverse over time.

6.4 Small Sample Sizes at Distributional Extremes

Our strongest findings, losses of 59 points and 46 points at the top 1% bullying intensity, rest on 120 observations in 2015 and 226 observations in 2018 (Table 2). While confidence intervals appropriately reflect this uncertainty, point estimates in small samples can be sensitive to outliers.

6.5 Measurement Limitations

Self-reported bullying victimization may be affected by social desirability bias and recall error. Students may underreport severe victimization and overreport mild experiences, and if reporting propensity correlates with achievement or adaptive capacity, estimated effects may reflect both true dose–response relationships and measurement artifacts.

The bullying scale exhibits good internal consistency, with Cronbach's alpha of 0.800 in 2015 and 0.819 in 2018, exceeding conventional benchmarks for acceptable reliability ($\alpha \geq 0.7$). This suggests that random measurement error is unlikely to drive the results. However, internal consistency does not address systematic misreporting, particularly severity-dependent recall within the 12-month window.

The consistency of regime patterns across two cohorts provides some reassurance, but validation against administrative or external records would further strengthen confidence about bullying indicator.

7 Conclusion

This paper documents an *adversity paradox* in observational data: students reporting bullying victimization score higher on average than non-bullied peers, while those experiencing severe exposure suffer large losses. Across the intensity distribution, estimated effects shift from positive to sharply negative, producing a 76-point swing in the 2015 data and a 57-point swing in the 2018 data. Minor local non-monotonicities at low cutoffs are quantitatively negligible relative to estimation uncertainty and do not alter the conclusion that average effects decline sharply once exposure exceeds students' adaptive capacity.

Positive average effects on test scores do not imply that bullying is beneficial. Students experiencing bullying may suffer psychological distress, anxiety, diminished well-being, and other costs that our academic outcome measure does not capture. The positive associations we document

reflect families and institutions successfully mobilizing resources to protect academic performance despite harm inflicted, not evidence that adversity improves outcomes. A student who copes successfully with bullying is not better off than a student who was never bullied; they have simply managed to avoid the worst academic consequences of an experience that likely imposed substantial unmeasured costs.

We interpret this pattern through an economic theory of adaptation under finite capacity constraints, stressing that the empirical results provide reduced-form evidence consistent with the proposed mechanism rather than establishing definitive causal effects.

The paper makes three contributions. First, we provide a unified theoretical framework in which adversity simultaneously inflicts direct damage and triggers compensatory resource mobilization. When adaptive capacity is abundant, compensatory mobilization offsets measurable harm; when capacity is exhausted, harm accelerates and becomes visible in outcomes. This framework reconciles our positive average findings with prior meta-analyses reporting negative associations, showing that average *levels* and average *slopes* are distinct estimands that can have opposite signs under nonlinear dose–response.

Second, we demonstrate that machine learning methods can detect vulnerable populations that conventional stratification misses. Causal forests and policy trees identify small minorities (0.7–1.5%) experiencing losses of 41–58 points, comparable to the most severely exposed students, despite positive average effects across all conventional demographic groups. This illustrates how optimizing over the partition space uncovers harm hidden by high-dimensional interactions.

Third, we provide evidence that psychological resilience, family socioeconomic resources, and institutional capacity influence individual adaptive thresholds, namely bullying intensities at which students transition from offsetting harm to experiencing losses. These findings suggest that vulnerability is jointly determined by exposure intensity and available coping resources, not by exposure alone.

Our findings carry an important interpretive implication. The positive average effects reflect post hoc adaptation, families and institutions responding to harm already inflicted, rather than evidence that adversity itself is beneficial. Because individual thresholds are unobservable, we cannot predict ex ante who will cope successfully and who will be overwhelmed. This unpredictability, combined with the severity of harm to those who cannot adapt, reinforces the conclusion that **preventing bullying across all levels of severity remains the appropriate policy response**.

More broadly, the mechanism we document extends beyond bullying to any setting in which individuals, families, or institutions face adversity under finite capacity constraints. The central lesson is that favorable average effects can conceal severe harm to vulnerable minorities. Effective policy should therefore look beyond averages, identify hidden vulnerabilities, and reduce exposure to adversity that disproportionately burdens those least able to cope.

Online Appendix

Contents

1 Introduction **1**

2 The Economics of Adaptive Capacity: A Unified Framework **3**

2.1 The Central Insight: Adaptation as Production Under Scarcity 3

2.2 The Three-Regime Structure of Adaptive Response 4

2.3 Visual Illustration: The Three-Regime Structure 5

2.4 Heterogeneous Capacity and the Distribution of Vulnerability 6

2.5 Why Average Effects Reverse Sign: The Aggregation Problem 6

2.6 Formal Characterization: Two Key Results 8

2.7 The Detection Problem: Why Vulnerability Hides in High Dimensions 9

3 Data and Descriptive Statistics **10**

3.1 Data Sources and Sample Construction 10

3.2 Variable Construction 10

3.2.1 Outcome: Academic Achievement 10

3.2.2 Treatment: Bullying Victimization 10

3.2.3 Covariates: Measuring Adaptive Capacity 12

4 Empirical Framework **12**

4.1 Identification 12

4.2 Estimation 14

4.3 Addressing Reverse Causality 15

4.4 Detecting Hidden Vulnerabilities Through Policy Trees 15

5 Results **15**

5.1 Regime I: Average Effects Under Low Observed Intensity 15

5.2 Regime Transitions: Testing Proposition 2 16

5.3 Selection Diagnostics 16

5.4 Testing Reverse Causality 18

5.5 Evidence for Adaptive Capacity as a Determinant of Regime Transitions 18

5.6 Heterogeneity Tests and Conventional Stratification 19

5.7 Detecting the Depletion Regime Through Policy Trees 20

5.7.1 2015: Younger Students Born Abroad Experience Depletion 20

5.7.2 2018: Males in Atypical Institutional Settings Experience Depletion 22
5.7.3 Synthesis 23
5.8 Decomposition Analysis: Role of Family Resources 23
5.9 Sensitivity Analysis 24
5.10 Reconciling with Prior Literature 24

6 Limitations and Future Research **26**
6.1 Temporal Instability of Depletion-Vulnerable Subgroups 26
6.2 Generalizability 26
6.3 Cross-Sectional Design 26
6.4 Small Sample Sizes at Distributional Extremes 27
6.5 Measurement Limitations 27

7 Conclusion **27**

A Theoretical Foundations **32**
A.1 Proof of Proposition 1 (Aggregation Bias) 32
A.2 Proof of Proposition 2 (Monotonic Deterioration) 32
A.3 Numerical Example: Reconciling Positive Levels with Negative Slopes 34

B Data Construction **36**
B.1 PISA Sampling and Weighting 36
B.2 Outcome Variable Construction 36
B.3 Bullying Index Construction 36
B.4 ESCS Construction 36
B.5 Covariate Availability Across PISA Waves 36
B.6 Full Descriptive Statistics 36

C Estimation Methodology **39**
C.1 Generalized Random Forests Procedure 39
C.2 Policy Tree Optimization 39

D Identification Diagnostics **40**
D.1 Propensity Score Overlap 40
D.2 Distributional Properties of Bullying Index 40
D.3 Selection Bias Investigation 40
D.4 Reverse Causality Tests 43
D.4.1 Treatment Effects by Predicted Achievement Quintile 43

D.4.2 Predictors of Bullying Exposure 43
D.4.3 Oster Bounds for Omitted Variable Bias 44

E Robustness and Sensitivity **45**
E.1 Sensitivity to Omitted Confounders 45
E.2 Calibration Tests for Treatment Effect Heterogeneity 45

F Heterogeneity and Mechanisms **47**
F.1 Conventional Subgroup Analysis 47
F.2 Decomposition Analysis: Role of Family Resources 47

A Theoretical Foundations

We present the formal proofs underlying Propositions 1 and 2, followed by a numerical example illustrating how positive average treatment effects can coexist with negative OLS slopes.

A.1 Proof of Proposition 1 (Aggregation Bias)

Let $Y_i = Y_i(0) + \tau_i(d_i)$ and assume $\mathbb{E}[Y_i(0) \mid d_i] = \mathbb{E}[Y_i(0)]$. The population OLS slope is

$$\beta_{\text{OLS}} = \frac{\text{Cov}(Y_i, d_i)}{\text{Var}(d_i)}.$$

Since $\text{Cov}(Y_i(0), d_i) = 0$,

$$\text{Cov}(Y_i, d_i) = \text{Cov}\big(\tau_i(d_i), d_i\big) = \text{Cov}\big(\tilde{\beta}_{1i} d_i - \beta_{2i} H_i, d_i\big),$$

where $H_i \equiv (d_i - d_i^*)^2 \mathbb{1}\{d_i \geq d_i^*\}$.

By mean-independence:

$$\text{Cov}(\tilde{\beta}_{1i} d_i, d_i) = \mathbb{E}[\tilde{\beta}_{1i}]\text{Var}(d_i),$$
$$\text{Cov}(\beta_{2i} H_i, d_i) = \mathbb{E}[\beta_{2i}]\text{Cov}(H_i, d_i).$$

Define

$$\Lambda \equiv \frac{\mathbb{E}\big[(d_i - d_i^*)^2 (d_i - \mathbb{E}[d_i]) \mathbb{1}\{d_i \geq d_i^*\}\big]}{\text{Var}(d_i)} > 0.$$

Then

$$\beta_{\text{OLS}} = \mathbb{E}[\tilde{\beta}_{1i}] - \mathbb{E}[\beta_{2i}] \cdot \Lambda.$$

□

A.2 Proof of Proposition 2 (Monotonic Deterioration)

Define $S(c) = 1 - F(c)$ and

$$\Theta(c) \equiv \mathbb{E}[\tau_i(d_i) \mathbb{1}\{d_i \geq c\}] = \int_c^{\bar{d}} \bar{\tau}(d) f(d)\, dd,$$

where $\bar{\tau}(d) \equiv \mathbb{E}[\tau_i(d)]$.

The cutoff-indexed estimand is $\theta(c) = \Theta(c)/S(c)$.

By the quotient rule:

$$\frac{d\theta}{dc} = \frac{S(c) \cdot \Theta'(c) - \Theta(c) \cdot S'(c)}{S(c)^2}.$$

Computing derivatives:

$$\Theta'(c) = -\bar{\tau}(c) f(c),$$
$$S'(c) = -f(c).$$

Substituting:

$$\frac{d\theta}{dc} = \frac{f(c)}{S(c)} \cdot \left[\theta(c) - \bar{\tau}(c)\right] = h(c) \cdot \left[\theta(c) - \bar{\tau}(c)\right].$$

It remains to establish that $\bar{\tau}(c) > \theta(c)$ for sufficiently large c. Under the maintained model, for $d \geq d_i^*$:

$$\tau_i(d) = \beta_{1i} d - \beta_{2i}(d - d_i^*)^2,$$

with derivative

$$\tau_i'(d) = \beta_{1i} - 2\beta_{2i}(d - d_i^*).$$

This derivative equals zero at the turning point $d_i^{**} \equiv d_i^* + \frac{\beta_{1i}}{2\beta_{2i}}$, is positive for $d < d_i^{**}$, and negative for $d > d_i^{**}$. Thus $\tau_i(d)$ is *not* monotonically decreasing throughout the depletion regime; it first increases after crossing d_i^*, reaches a maximum at d_i^{**}, then decreases.

Define the population upper envelope of turning points:

$$\bar{d}^{**} \equiv \sup_i \left\{ d_i^* + \frac{\beta_{1i}}{2\beta_{2i}} \right\}.$$

For $c \geq \bar{d}^{**}$, all individuals with $d_i \geq c$ satisfy $d_i > d_i^{**}$, placing them in the strictly decreasing portion of their individual dose–response curves. In this region, students at exactly $d_i = c$ have the *lowest* intensity among those satisfying $d_i \geq c$, and therefore the *highest* treatment effects within this group (since τ_i is strictly decreasing for $d > d_i^{**}$).

Consequently, the average effect at the boundary exceeds the conditional average:

$$\bar{\tau}(c) = \mathbb{E}[\tau_i(c)] > \mathbb{E}[\tau_i(d_i) \mid d_i \geq c] = \theta(c).$$

Since $h(c) > 0$ and $\theta(c) - \bar{\tau}(c) < 0$, we conclude $d\theta/dc < 0$ for all $c \geq \bar{d}^{**}$.

Rate of Deterioration. For right-skewed distributions characteristic of bullying intensity, the hazard rate $h(c) = f(c)/[1-F(c)]$ increases in the upper tail. Simultaneously, the gap $[\bar{\tau}(c)-\theta(c)]$ widens because the quadratic depletion penalty $\beta_{2i}(d-d_i^*)^2$ accelerates for higher-intensity students.

Together, these mechanisms imply that $|d\theta/dc|$ increases as $c \to \bar{d}$, generating accelerating deterioration in the upper tail.

Empirical Interpretation. Our finding of monotonic deterioration for $c \geq Q_{90}$ is consistent with $\bar{d}^{**} \leq Q_{90}$, that is, by the 90th percentile of intensity, all students have passed their individual turning points and entered the strictly decreasing regime. This condition cannot be directly verified since individual parameters $(\beta_{1i}, \beta_{2i}, d_i^*)$ are unobserved, but it is plausible given that the turning point d_i^{**} depends on the ratio $\beta_{1i}/(2\beta_{2i})$, which should be bounded for students with finite adaptive capacity. □

A.3 Numerical Example: Reconciling Positive Levels with Negative Slopes

This example clarifies how positive average treatment effects can coexist with negative marginal slopes under nonlinear dose–response functions.

Consider a homogeneous population with dose–response function

$$\tau(d) = 10d - 4(d-5)^2 \mathbb{1}\{d \geq 5\},$$

where intensity follows $d_i \sim \text{Uniform}[0, 12]$. This specification exhibits three regimes. In *dormancy* ($d = 0$), we have $\tau(0) = 0$. In the *mobilization regime* ($0 < d < 5$), the treatment effect is $\tau(d) = 10d > 0$ with constant slope +10. In the *depletion regime* ($d \geq 5$), the treatment effect becomes $\tau(d) = 10d - 4(d-5)^2$, with slope

$$\tau'(d) = 10 - 8(d-5),$$

which equals zero at $d = 6.25$ and turns negative thereafter.

Despite substantial harm at high intensity (e.g., $\tau(12) = 120 - 196 = -76$), the **average measured treatment effect remains positive**:

$$\theta(0) = \mathbb{E}[\tau(d_i)] = \mathbb{E}[10d_i] - 4 \cdot \mathbb{E}[(d_i - 5)^2 \mid d_i \geq 5] \cdot \Pr(d_i \geq 5) = 60 - 4 \cdot \frac{49}{3} \cdot \frac{7}{12} = \frac{197}{9} \approx 21.9 > 0.$$

Yet OLS estimates an average *slope*. Since the upper tail features large negative values of $\tau(d)$, treatment effects and intensity become negatively correlated. The OLS coefficient is:

$$\beta_{\text{OLS}} = \frac{\text{Cov}(\tau(d_i), d_i)}{\text{Var}(d_i)} = 10 - 4 \cdot \frac{\text{Cov}\big((d_i - 5)^2 \mathbb{1}\{d_i \geq 5\}, d_i\big)}{\text{Var}(d_i)} = 10 - 4 \cdot \frac{5831}{1728} = -\frac{1511}{432} \approx -3.50 < 0.$$

Figure 7 visualizes this reconciliation. The blue solid curve shows the true nonlinear dose–

response with positive mobilization slope (+10) transitioning to negative depletion slope at $d^* = 5$, reaching maximum effect at $d = 6.25$ before collapsing. The dashed line shows the OLS linear approximation with negative slope (≈ -3.50), which captures the dominant pattern in the depletion regime where curvature is active.

The hazard rate interpretation from Proposition 2 explains why deterioration accelerates. For Uniform$[0, 12]$, the hazard rate is $h(c) = 1/(12-c)$, increasing from $1/12$ at $c = 0$ to ∞ as $c \to 12$. Since $d\theta/dc = h(c)\,[\theta(c) - \bar{\tau}(c)]$ and $\bar{\tau}(c) > \theta(c)$ in the depletion regime, a rising hazard rate accelerates deterioration in the upper tail, consistent with the observed 27-point decline to the top decile and an additional 49-point drop at the top percentile.

B Data Construction

B.1 PISA Sampling and Weighting

PISA employs a two-stage stratified sampling design. Schools are sampled with probability proportional to size within strata defined by region and school type. Within sampled schools, 35 students are randomly selected. We apply final student weights provided by PISA throughout our analysis.

B.2 Outcome Variable Construction

We construct a composite achievement measure from PISA's reading, mathematics, and science domain scores using principal components analysis. For both cohorts, we standardize each domain score, extract the first principal component, and rescale to the original score metric.

The first component explains 89.6% of total variance in 2015 and 89.4% in 2018, reflecting strong cross-domain correlation characteristic of general academic ability.

B.3 Bullying Index Construction

PISA measures bullying victimization using six questions capturing frequency of distinct behaviors over 12 months. Each item is reported on a four-point ordinal scale. We aggregate using item response theory (IRT) following PISA methodology.

B.4 ESCS Construction

We construct the Economic, Social, and Cultural Status (ESCS) index from its underlying components using Japan-specific factor loadings: parental occupation, parental education, and household possessions. ESCS captures familial resources for mobilization.

B.5 Covariate Availability Across PISA Waves

Table A1 documents which covariates are available in each PISA wave.

B.6 Full Descriptive Statistics

Table A2 presents detailed descriptive statistics for both analytical samples.

Table A1: Covariate Availability Across PISA Waves

Variable	**2015**	**2018**
Demographics		
Age	✓	✓
Gender	✓	✓
Student born abroad	✓	✓
Parent born abroad	✓	✓
Socioeconomic Status		
ESCS composite index	✓	✓
Psychological Capital		
Resilience	—	✓
Parental emotional support	✓	✓
Teacher support	✓	✓
School Characteristics		
Private vs. public	✓	✓
Student–teacher ratio	✓	✓
Staff shortage	✓	✓

Notes: Variables categorized by role in adaptive capacity framework. Resilience, a primary determinant of individual adaptive thresholds d_i^*, is available only in the 2018 wave.

Table A2: Descriptive Statistics: PISA 2015 and 2018 Japan Analytical Samples

	2015 (n = 6,647)		**2018** (n = 6,109)	
Variable	Mean	SD	Mean	SD
Outcome				
Test score (composite)	529.1	91.6	520.0	91.9
Treatment				
Bullied (any victimization)	0.464	—	0.368	—
Demographics				
Age (years)	15.78	0.29	15.78	0.29
Male	0.502	0.500	0.489	0.500
Born abroad	0.009	0.095	0.011	0.105
Familial Resources				
ESCS index	−0.18	0.70	−0.11	0.73
Psychological Capital				
Resilience	—	—	−0.005	0.89
Parental support	−0.007	0.91	−0.066	0.88
Teacher support	0.001	0.92	−0.009	0.91
Institutional Capacity				
Private school	0.272	—	0.307	—
Student–teacher ratio	11.43	4.17	11.90	4.29
Staff shortage	−0.068	0.95	0.071	0.89

Notes: Variables categorized by role in adaptive capacity framework. Psychological and school climate variables standardized via IRT. Sample sizes reflect students with complete data on all covariates; 40 students in 2015 and 159 students in 2018 are excluded due to missing bullying index values.

C Estimation Methodology

C.1 Generalized Random Forests Procedure

We implement a three-stage double machine learning procedure:

1. **Stage 1:** Estimate $\hat{e}(x) = \hat{\mathbb{E}}[W_i \mid X_i = x]$ using a regression forest.
2. **Stage 2:** Estimate $\hat{m}(x) = \hat{\mathbb{E}}[Y_i \mid X_i = x]$ using a separate regression forest.
3. **Stage 3:** Grow a causal forest on residualized outcomes $\tilde{Y}_i = Y_i - \hat{m}(X_i)$ and residualized treatments $\tilde{W}_i = W_i - \hat{e}(X_i)$.

The causal forest estimates conditional treatment effects $\tau(x; c)$ by recursively partitioning the covariate space to maximize treatment effect heterogeneity. Each tree employs honest estimation, using separate subsamples for determining splits and for estimating treatment effects within leaves (Athey and Imbens, 2016). This sample splitting reduces overfitting and ensures that estimated heterogeneity reflects genuine variation in treatment effects rather than noise induced by adaptive partitioning.

C.2 Policy Tree Optimization

Policy trees (Athey and Wager, 2021) translate estimated treatment effect heterogeneity into simple, interpretable decision rules. Splits are chosen to separate observations with systematically different treatment effects rather than to improve outcome prediction. At each node, we select the split that maximizes

$$\Delta = \frac{\left(\sum_{i \in L} \hat{\Gamma}_i\right)^2}{|L|} + \frac{\left(\sum_{i \in R} \hat{\Gamma}_i\right)^2}{|R|}, \tag{12}$$

where L and R denote the left and right child nodes and $\hat{\Gamma}_i$ are doubly robust scores for the individual treatment effect. The objective Δ rewards splits that create child nodes with large and opposing average treatment effects while penalizing small sample sizes.

We constrain tree depth to two levels (four terminal nodes) to maintain interpretability.

D Identification Diagnostics

We assess the plausibility of identification assumptions by examining propensity score overlap, distributional properties required by Proposition 2, potential selection bias, and reverse causality.

D.1 Propensity Score Overlap

We assess the overlap assumption by examining the distribution of estimated propensity scores. Figure A1 shows the distribution for the 2015 cohort; the corresponding distribution for 2018 appears in Figure 3 in the main text. Both distributions show substantial density throughout their support. Fewer than 1% of observations fall below 0.1 or above 0.9, and none fall outside $[0.05, 0.95]$, indicating that the overlap assumption is satisfied.

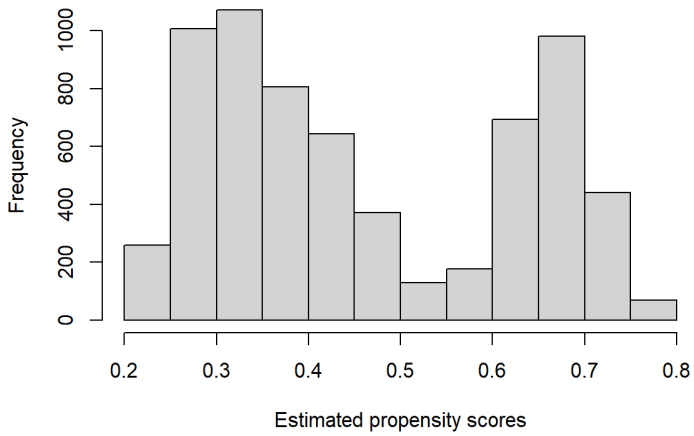


Figure A1: Distribution of Estimated Propensity Scores (2015)

Notes: Propensity scores from regression forest. Distribution satisfies overlap with density away from boundaries.

D.2 Distributional Properties of Bullying Index

Proposition 2 requires that the intensity distribution be right-skewed with positive density throughout its support. Table A3 presents distributional diagnostics verifying these conditions.

D.3 Selection Bias Investigation

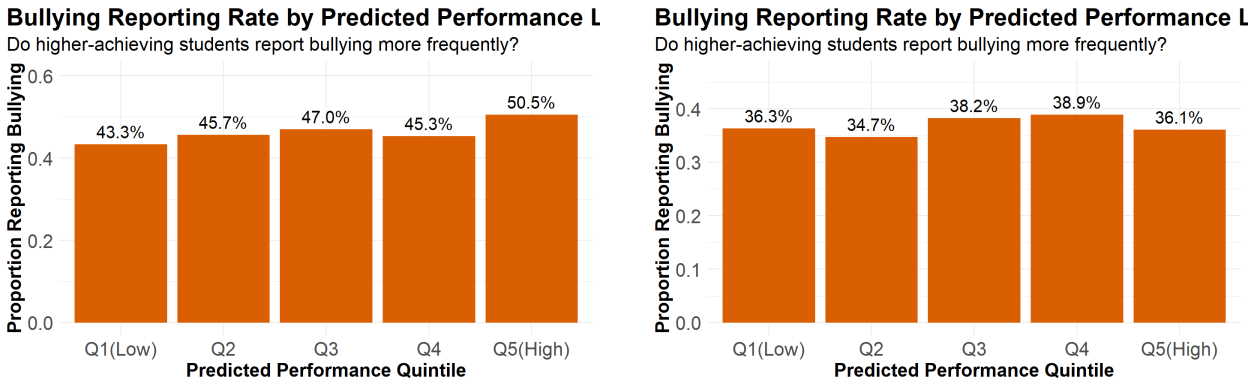
Positive average effects could partly reflect selection if higher-achieving students are more likely to recognize or disclose mild victimization. Figure A2 shows bullying incidence by quintile of

Table A3: Bullying Index Distribution Diagnostics

Statistic	**2015**	**2018**
Sample and location		
N (finite values)	6,607	5,950
Mean	−0.009	−0.041
Median	−0.633	−0.541
Dispersion		
Standard deviation	0.783	0.753
IQR	1.189	1.067
Range	3.483	3.387
Shape		
Skewness	0.947	1.212
Excess kurtosis	−0.085	0.357
Upper-tail thresholds		
75th percentile	0.557	0.526
90th percentile	1.162	1.149
95th percentile	1.421	1.399
99th percentile	2.024	1.991

Notes: Diagnostics for standardized bullying index. Strong right skewness (0.95 in 2015, 1.21 in 2018) with pronounced mass at lower bound implies most students experience low-moderate intensity (mobilization regime) while approximately 10% experience severe victimization (depletion regime).

predicted test scores. In 2015, incidence increases from 43.5% (Q1) to 50.2% (Q5), a 6.7 percentage point gradient. In 2018, the gradient is much smaller (1.2 percentage points). This modest positive gradient suggests higher-achieving students may be slightly more likely to report victimization, but the magnitude is too small to explain the large positive treatment effects we estimate.



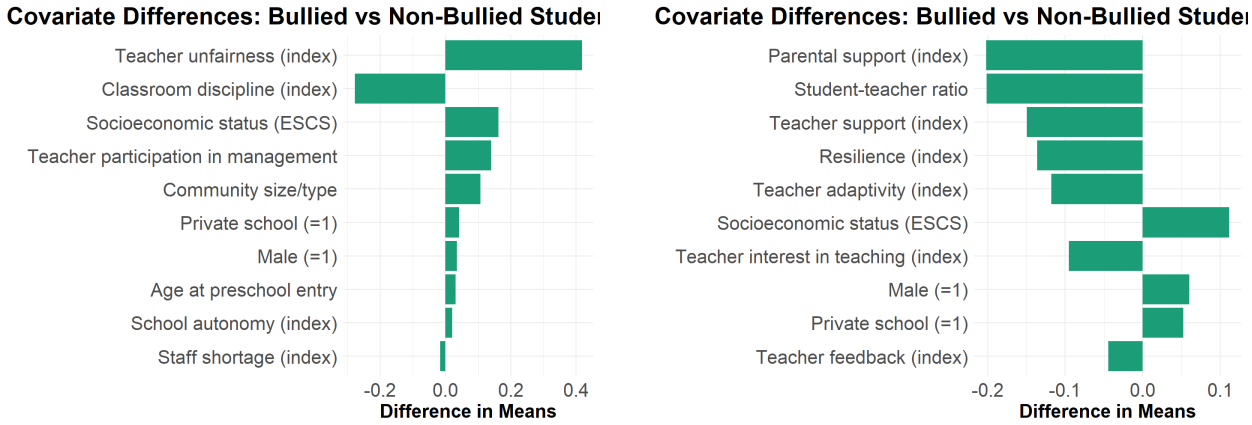
(a) 2015: 6.7pp gradient

(b) 2018: 1.2pp gradient

Figure A2: Bullying Reporting Rates by Predicted Achievement Quintile

Notes: Bullying incidence by quintile of predicted test scores. Positive gradient suggests higher-achieving students may be more likely to report mild victimization, but this selection works against the mobilization mechanism.

Figure A3 shows differences in mean covariate values (bullied minus non-bullied). Bullied students have slightly higher ESCS on average in both cohorts, consistent with the selection pattern above.



(a) 2015 cohort

(b) 2018 cohort

Figure A3: Covariate Differences Between Bullied and Non-Bullied Students

Notes: Differences in mean covariate values (bullied minus non-bullied). Bullied students report higher ESCS in both cohorts.

Yet this selection pattern cannot explain our estimated effects. If positive results were driven

by selection alone, controlling for observables should attenuate estimates. Instead, covariate adjustment increases estimates (from +19.54 to +21.16 in 2015; Table A6). Most importantly, selection cannot account for monotonic deterioration from positive to severely negative effects along the intensity gradient.

D.4 Reverse Causality Tests

A fundamental concern is reverse causality: high-achieving students may attract bullying through peer jealousy. We address this through three diagnostic tests.

D.4.1 Treatment Effects by Predicted Achievement Quintile

If reverse causality drives results, treatment effects should be largest among students with the highest predicted baseline achievement. Table A4 presents treatment effects estimated within quintiles of predicted achievement. Effects are approximately uniform across quintiles, with a slightly negative gradient (−1.35 in 2015, −0.99 in 2018). This pattern is opposite to what reverse causality would predict.

Table A4: Treatment Effects by Predicted Achievement Quintile

	2015 Cohort			2018 Cohort		
Quintile	**N**	**ATT**	**SE**	**N**	**ATT**	**SE**
Q1 (Lowest)	1,330	+19.7	5.16	1,222	+9.9	6.67
Q2	1,329	+19.2	5.07	1,222	+9.6	5.85
Q3	1,329	+10.8	5.17	1,221	+18.0	5.51
Q4	1,329	+17.9	4.97	1,222	+9.9	5.95
Q5 (Highest)	1,330	+18.4	4.98	1,222	+8.9	5.42
Gradient (Q5−Q1)		−1.35	7.17		−0.99	8.59
p-value		0.850			0.909	

Notes: Treatment effects estimated within quintiles of predicted achievement. Gradient is slightly negative in both years, opposite to reverse causality prediction.

D.4.2 Predictors of Bullying Exposure

Under reverse causality, predicted achievement should positively predict bullying exposure. Table A5 presents logistic regression results predicting bullying status. Predicted achievement has a significantly *negative* coefficient in both cohorts: $\beta = -0.430$ (SE = 0.047, $p < 0.001$) in 2015 and $\beta = -0.083$ (SE = 0.033, $p = 0.011$) in 2018. High achievers are *less* likely to be bullied, contradicting the jealousy-driven targeting hypothesis.

Table A5: Predictors of Bullying Exposure

	2015		2018	
Predictor	Coef.	SE	Coef.	SE
Predicted achievement (std.)	-0.430^{***}	0.047	-0.083^{*}	0.033
ESCS (std.)	0.598^{***}	0.047	0.205^{***}	0.033
Age	0.117	0.086	0.020	0.093
Male	0.121^{*}	0.050	0.246^{***}	0.053

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Notes: Logistic regression predicting bullying exposure. Negative coefficient on predicted achievement contradicts jealousy-driven targeting.

D.4.3 Oster Bounds for Omitted Variable Bias

Table A6 presents bounds following Oster (2019) to assess how much selection on unobservables would be required to eliminate estimated effects.

Table A6: Oster Bounds: Robustness to Omitted Variable Bias

	2015		2018	
Specification	Coef.	R^2	Coef.	R^2
Bivariate	+19.54	0.011	+13.21	0.005
With controls	+21.16	0.218	+12.77	0.229
Coefficient change	+1.62		−0.44	
δ for $\beta^* = 0$	−4.13		+8.86	

Notes: δ indicates ratio of selection on unobservables to selection on observables required to produce $\beta^* = 0$. In 2015, $\delta < 0$ indicates adding controls increases the coefficient, implying unobservables would need to select in the opposite direction from observables. In 2018, $\delta = 8.86$ indicates unobservables would need to be nearly 9× as important as observables to eliminate the estimated effect.

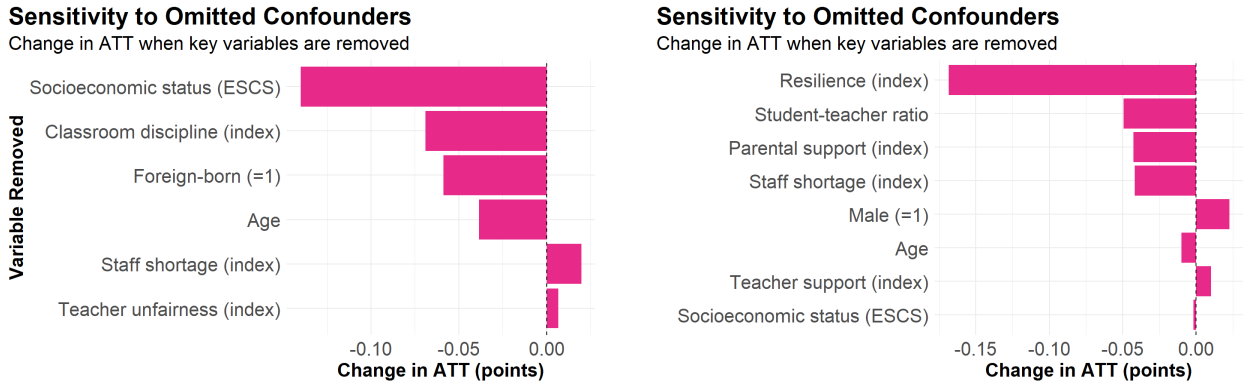
We report these bounds as heuristic sensitivity diagnostics because our primary estimands are nonlinear and obtained via machine learning, so the assumptions underlying Oster (2019) do not strictly apply. These results support the interpretation that positive average effects reflect compensatory mobilization rather than reverse causality.

E Robustness and Sensitivity

We assess sensitivity to specification choices and test whether the causal forest detects genuine heterogeneity.

E.1 Sensitivity to Omitted Confounders

Figure A4 shows changes in the estimated ATT when key covariates are sequentially excluded. All changes represent less than 2% of baseline estimates. Removing resilience (2018), the primary determinant of d_i^* in our framework, changes ATT by only −0.16 points.



(a) 2015: Maximum change −0.12 points

(b) 2018: Maximum change −0.16 points

Figure A4: Sensitivity to Omitted Confounders

Notes: Each bar reports change in ATT when indicated covariate is excluded. Results are highly stable across specifications.

The severity gradient does not appear to be driven by specific controls. Although unobserved confounding cannot be ruled out, explaining the sharp sign reversal at high intensity would require unobservables strongly correlated with bullying severity and acting opposite to observed covariates.

E.2 Calibration Tests for Treatment Effect Heterogeneity

Our empirical strategy relies on causal forests to detect treatment effect heterogeneity. If treatment effects were homogeneous, machine learning methods would add little beyond conventional estimators. We test for heterogeneity using the calibration procedure proposed by Chernozhukov et al. (2020).

Table A7 presents calibration test results. The "Mean forest prediction" coefficient tests whether the average forest prediction matches the true average effect; values near 1 indicate good calibration. The "Differential prediction" coefficient tests whether the forest detects genuine heterogeneity; significant values reject the null of homogeneous effects.

Table A7: Calibration Tests for Treatment Effect Heterogeneity

	2015			2018		
Coefficient	Est.	SE	p	Est.	SE	p
Mean forest prediction	1.016	0.135	< 0.001	1.044	0.258	< 0.001
Differential prediction	0.489	0.374	0.095	0.955	0.325	0.002

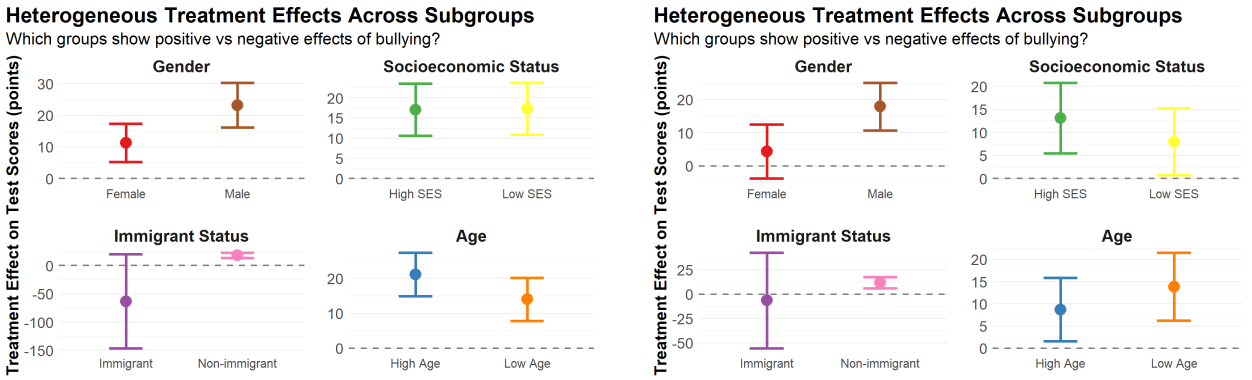
Notes: Calibration test following Chernozhukov et al. (2020). Heterogeneity is marginally significant in 2015 ($p = 0.095$) and strongly significant in 2018 ($p = 0.002$).

Results provide varying evidence for heterogeneity: marginal significance in 2015 ($p = 0.095$) and strong rejection of homogeneity in 2018 ($p = 0.002$). These results support the use of causal forests to characterize treatment effect variation, particularly in 2018 where resilience is observed.

F Heterogeneity and Mechanisms

F.1 Conventional Subgroup Analysis

Figure A5 reports treatment effects estimated within predetermined demographic strata: gender, socioeconomic status (above/below median ESCS), age (above/below median), and immigration status.



(a) 2015 subgroup effects

(b) 2018 subgroup effects

Figure A5: Traditional Subgroup Analysis

Notes: Effects estimated within predetermined strata. All subgroups show positive effects except immigrants (negative but statistically insignificant in both cohorts).

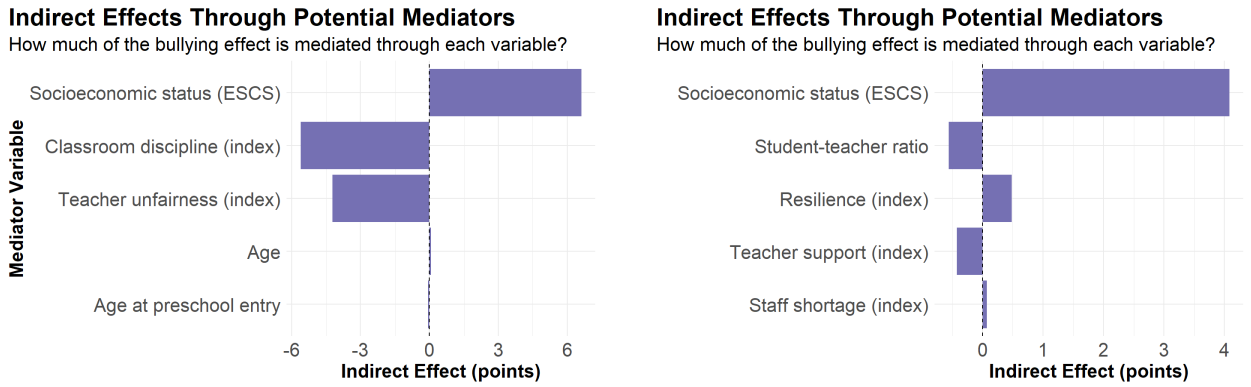
All subgroups show positive average effects except immigrants, who exhibit negative point estimates (-63.2 in 2015 and -6.0 in 2018). Neither immigrant estimate is statistically significant due to small sample sizes ($n \approx 60$ in 2015 and $n \approx 70$ in 2018).

The pattern reveals a fundamental detection problem: stratification by observable characteristics averages over the full intensity distribution, obscuring the interaction between individual adaptive thresholds $d_i^*(X_i)$ and realized exposure intensity d_i. Conventional approaches therefore fail to detect heterogeneity that emerges primarily in the upper tail of the intensity distribution.

F.2 Decomposition Analysis: Role of Family Resources

To investigate mechanisms underlying treatment effect heterogeneity, we decompose the variation in estimated individual treatment effects by covariate. Figure A6 shows the share of total effect variation statistically associated with each covariate.

ESCS covaries with approximately one-third of total treatment effect heterogeneity: 37.7% in 2015 and 35.4% in 2018. This pattern is consistent with the resource mobilization mechanism in our theoretical framework: high-ESCS families have greater capacity to mobilize compensatory resources (tutoring, counseling, parental time) that offset harm on measured academic outcomes.



(a) 2015 decomposition

(b) 2018 decomposition

Figure A6: Decomposition of Treatment Effect Heterogeneity by Covariate

Notes: Share of treatment effect variation statistically associated with each covariate. ESCS covaries with approximately one-third of total heterogeneity in both cohorts (37.7% in 2015, 35.4% in 2018). This decomposition is descriptive and does not establish causal mediation pathways.

This is a descriptive decomposition measuring statistical association, not causal mediation. The sequential ignorability assumptions required for causal mediation interpretation (Imai et al., 2010) are implausible here: ESCS likely correlates with unobserved determinants of both bullying exposure (e.g., school quality, neighborhood characteristics) and academic achievement (e.g., parental involvement, cognitive stimulation at home). We interpret these results as suggestive correlational evidence that family resources covary with treatment effect heterogeneity, not as estimates of causal mediation pathways.

References

Aizer, A. and Doyle Jr, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, 130(2):759–803.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Brown, S. and Taylor, K. (2011). Bullies in the hood: The relationship between bullying, school performance, and area income. *European Journal of Political Economy*, 27(4):662–673.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2020). Generic machine learning inference on heterogeneous treatment effects in randomized experiments. *Proceedings of the National Academy of Sciences*, 117(48):30039–30045.

Eriksen, T. L. M., Nielsen, H. S., and Simonsen, M. (2014). The effects of bullying on academic performance and adult earnings. *Labour Economics*, 31:213–228.

Glewwe, P., Kremer, M., and Moulin, S. (2009). Many children left behind? textbooks and test scores in kenya. *American Economic Journal: Applied Economics*, 1(1):112–135.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 73–84. Wiley.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., and Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, 140(4):1073–1137.

Meichenbaum, D. (1985). *Stress Inoculation Training*. Pergamon Press, New York.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., and Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Nakamoto, J. and Schwartz, D. (2010). Is peer victimization associated with academic achievement? a meta-analytic review. *Social Development*, 19(2):221–242.

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD Publishing, Paris.

OECD (2017). *PISA 2015 Technical Report*. OECD Publishing, Paris.

OECD (2019). *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD Publishing, Paris.

OECD (2020). *PISA 2018 Technical Report*. OECD Publishing, Paris.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Pencavel, J. (2015). The productivity of working hours. *The Economic Journal*, 125(589):2052–2076.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.