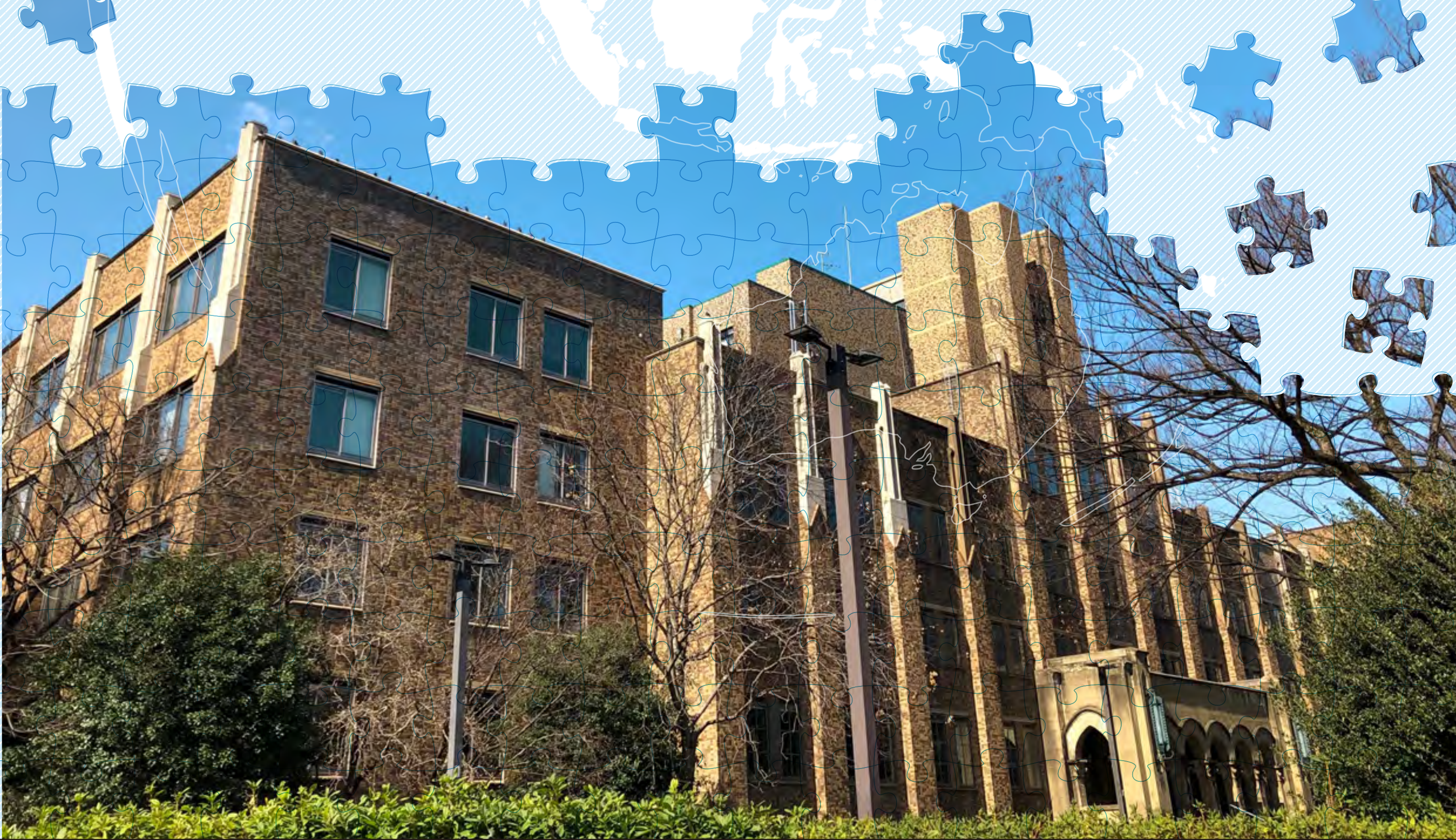
Center for Social Research and Data Archives, Institute of Social Science, The University of Tokyo

SUSTAINABLE DEVELOPMENT GOALS

CSRDA supports the Sustainable Development Goals

CSRDA Discussion Paper

Another Look at Primary and Secondary Effects: Causal Decomposition of Inequalities in Educational Opportunity



No. 139

Date July. 2026

Name Sho Fujihara

SDGs

4 QUALITY EDUCATION

10 REDUCED INEQUALITIES

Another Look at Primary and Secondary Effects: Causal Decomposition of Inequalities in Educational Opportunity

1 Introduction

The disparities in educational attainment by socioeconomic background are known as inequalities of educational opportunity (IEO) (Boudon 1974), and there is an extensive body of research on their extent and trends (Shavit and Blossfeld 1993; Breen et al. 2009). Among several factors, academic performance is considered a critical one in generating IEO, and studies have sought to decompose IEO into primary and secondary effects of stratification, both theoretically and empirically (Boudon 1974; Breen and Goldthorpe 1997; Erikson et al. 2005).

On primary effects, Boudon (1974) writes:

> The lower the social status, the poorer the cultural background — hence the lower the school achievement, and so on. These are what we have called the primary effects of stratification ... The primary effects of stratification may be represented in a Cartesian space ... Lower-class children tend to be located in one "corner" of the Cartesian space. Conversely, upper-class children are more likely to be located at the opposite "corner" (Boudon 1974, 29).

And on secondary effects:

> [E]ven if two youngsters are located at the same point of the Cartesian space described earlier, the probabilities that they will choose a [a: more likely to lead to a higher social status, author's note] rather than b are likely to differ. Inasmuch as the youngsters are by assumption located at the same point of the Cartesian space, they do not differ with respect to the primary effects of stratification, but they do differ with respect to its secondary effects (Boudon 1974, 30).

Although Boudon referred to the primary effect as cultural inequality, the relationship between class and academic performance can also be produced by other factors such as nutrition, health, and discrimination (Breen and Goldthorpe 1997). Sociologists have mainly focused on the secondary effects, namely inequalities in rational decision-making derived from cost-benefit calculations that persist after the influence of performance is accounted for (Boudon 1974; Erikson and Jonsson 1996; Breen and Goldthorpe 1997; Jonsson and Erikson 2000). Boudon himself later characterized this choice mechanism as cognitive rather than narrowly utilitarian, deriving educational decisions from a system of contextualized arguments rather than from cost-benefit calculations alone (Boudon 1998). More recent empirical work directly measures class differences in subjective probabilities of educational success and returns (Abbiati et al. 2018; Barone et al. 2018).

An influential body of work on this theme begins with Erikson et al. (2005) and its subsequent extensions (Jackson et al. 2007; Jackson 2013a). The policy implications differ depending on which effect dominates. When primary effects do, interventions targeting performance gaps are most effective. When secondary effects dominate, policy should address the structural constraints and incentive structures that shape educational decisions (Jackson 2013b). Deriving such implications, however, requires causal rather than associational identification, since intervening on factors that are only associated with attainment cannot reduce inequalities if those associations do not reflect causal effects. As several authors have noted, identification of the relevant causal effects poses serious challenges (Erikson et al. 2005; Jackson et al. 2007; Jackson 2013b; Morgan 2012; Morgan et al. 2013).

Studies that have applied a causal inference framework to examine primary and secondary effects include those by Morgan (2012) and Morgan et al. (2013). Morgan (2012) thoroughly articulates the methodological challenges of causal identification, assesses conditioning strategies for meeting them, and concludes that available data support only an associational decomposition.

Building on this line of work, the present study addresses two interrelated research questions, formulated to be directly translatable into identifiable statistical quantities:

1. By how much would educational disparities between social classes be reduced if academic performance distributions were equalized across classes, and how much disparity would remain?

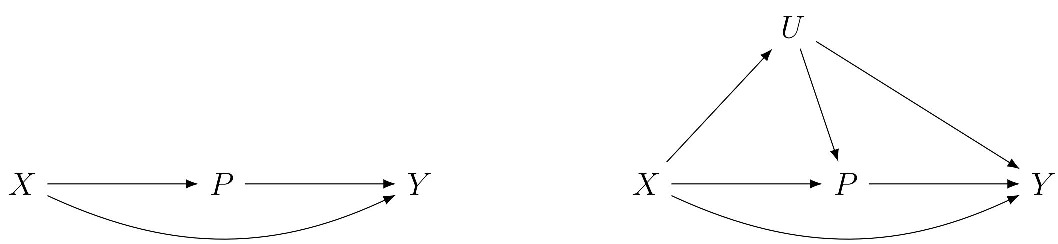
2. How do these magnitudes compare to those produced by conventional decomposition methods that condition on performance?

This study follows the estimand-first workflow advocated by Lundberg (2022), proceeding from a theoretical concept to a research question, to a precise statistical estimand, and only then to an estimator. The specific decomposition technique adopted is the causal decomposition analysis of Park et al. (2022, 2023a,b), which targets the change in an observed disparity under a counterfactual intervention on a mediator while treating the social category itself as a grouping variable. Section 2 reviews previous research and its methodological challenges. Section 3 proceeds through five explicit steps: from Boudon's theoretical concepts of primary and secondary effects, to the data-generating process represented by a directed acyclic graph, to the formal definition of the quantities to be estimated, to the assumptions required to express these quantities in terms of observable data, to the procedure for producing numerical estimates. Section 4 introduces the data and variables. Section 5 applies both conventional and the proposed methods to Japanese data. Section 6 discusses limitations, and Section 7 concludes.

2 Previous Research

The conventional approach to decomposing IEO into primary and secondary effects, developed by Erikson et al. (2005) and extended by Jackson et al. (2007); Jackson (2013a), fits class-specific normal distributions for academic performance and class-specific logistic regressions for the transition, then computes counterfactual transition probabilities by exchanging the performance distribution or the choice pattern between classes. The contributions of primary and secondary effects to IEO are read off these counterfactuals, typically as differences in log-odds. The analysis can be implemented in the *DECIDE* R package (Kartsonaki 2022), and methodological details and standard errors are discussed in Kartsonaki et al. (2013). An alternative decomposition for non-linear models, the KHB method of Karlson and Holm (2011), addresses the scale identification problem that complicates direct/indirect effect comparisons in logit and probit specifications, and Karlson (2013) extends it with a parametric summary measure that aggregates across multiple class contrasts under a proportionality assumption. These methods improve best practice within the conventional framework, but they still condition on academic performance and report effects on the log-odds scale. As Morgan (2012) argues, such estimates conflate choice-based secondary effects with confounding and average over directional asymmetries. Adjusting for confounders in these models also reshapes the disparity being decomposed rather than only reallocating it. These limitations motivate the causal decomposition approach pursued here, which adjusts confounders for identification while leaving the target disparity intact. Applications have examined cross-time and cross-national patterns (Jackson and Jonsson 2013). Erikson and Rudolphi (2010) apply this framework to six Swedish cohorts born between 1948 and 1982 and find that the decline in IEO at the transition to upper-secondary schooling reflects corresponding declines in both primary and secondary effects, with the relative magnitude of secondary effects depending on whether ability is measured by grades or by cognitive tests. Schindler and Lörz (2012) use KHB to show that growing class disparities in German tertiary enrollment between 1976 and 2005 are driven primarily by secondary effects.

The standard graphical representation of the underlying model is shown in Figure 1a (Erikson et al. 2005), with class (X), performance (P), and attainment (Y). The path $X \rightarrow P \rightarrow Y$ corresponds to the primary effect, and the residual $X \rightarrow Y$ corresponds to the secondary effect (Morgan et al. 2013). Erikson et al. (2005) emphasises that a more realistic model also includes an unobserved component U (Figure 1b), such as an anticipatory decision, that affects both P and Y. As Morgan (2012) shows, conditioning on P when P is a collider between X and U induces a spurious association between X and U (Elwert and Winship 2014), biasing both PE and SE estimates. Morgan (2012) therefore recommends abandoning causal interpretation in favour of a simpler associational analysis. The present article takes a different path: rather than abandoning causal interpretation, it defines a precise estimand within the causal decomposition analysis framework of Park et al. (2022, 2023a,b) and addresses Morgan's three specific concerns about the conventional method (weak causal warrant, log-odds averaging, and inattention to heterogeneity). Online Appendix G provides a fuller discussion of Morgan's critique and how the present approach responds to each point.



(a) Simple model **(b) More realistic model**

Figure 1: Basic directed acyclic graphs for how social category X are associated with educational attainment Y

Note: X: social category; P: performance; Y: educational attainment; U: confounders.
Alt text: Two side-by-side directed acyclic graphs. Panel (a), the simple model, has three nodes—social class X, academic performance P, and educational attainment Y—with arrows from X to P, from P to Y, and a curved arrow from X directly to Y. Panel (b), the more realistic model, adds a node U above P, with arrows from X to U, from U to P, and from U to Y.

3 From Theory to Estimand

3.1 Theoretical Framework and Research Questions

Boudon (1974) distinguished primary effects, the channel through which social background influences attainment through academic performance, from secondary effects, the remaining influence after performance differences are accounted for, traditionally attributed to differential educational choices based on perceived costs and benefits. Boudon's theory admits several empirical translations, each defined by the counterfactual scenario it adopts. The one pursued here asks: if academic performance distributions were equalized across social classes, how much of the observed educational disparity would be reduced, and how much would remain? The reduction in the disparity corresponds to what Boudon would call primary effects, and the remaining disparity corresponds to secondary effects. This formulation treats class as a grouping variable that defines the disparities of interest (Lundberg 2022; Jackson and VanderWeele 2018) rather than as a manipulable cause, with the intervention on academic performance, not on class. It thereby resolves the identification problem that Morgan (2012) raises against the causal effect of class itself while preserving the spirit of Boudon's distinction.

This operationalization is one possible reading of Boudon rather than a faithful translation. Boudon's original secondary effects refer specifically to differential choice behavior at given levels of achievement and cultural capital (Boudon 1974, 30), while the estimand defined here captures every pathway from class to attainment that does not pass through measured performance, including measurement error and non-choice channels. Boudon also appears to have conceived of the performance differences themselves as observed group differences, so the counterfactual equalization of performance distributions across classes, on which the present estimand rests, is our formalization rather than an idea found in his account. Online Appendix A discusses the alternative causal mediation framework and the reasons it is not adopted here. The decomposition proposed in this article is therefore best understood as another look at Boudon's distinction: a disparity-decomposition counterpart that prioritizes identification at the cost of leaving the choice-mechanism interpretation, which calls for explicit modeling of class-differentiated beliefs (Breen and Goldthorpe 1997), to future work.

3.2 Data-Generating Process

The data-generation process explaining disparities in educational attainment due to socioeconomic background is illustrated using a directed acyclic graph (DAG) in Figure 2, based on the DAGs used in Park et al. (2022, 2023a,b). Initially, as in previous studies (Erikson et al. 2005), we consider nodes for X (social class), P (performance), and Y (educational attainment), with causal pathways $X \to Y$ and $X \to P \to Y$. Additionally, factors behind class, performance, and educational attainment, such as race and neighborhood disadvantage, are represented as L_1, C_1, and C_2, and are assumed to be related through a historical context denoted as H (VanderWeele and Robinson 2014), rather than through direct causal connections, as noted in Morgan et al. (2013). Age, period, cohort, and the societal context in which one was raised are included in C_1. Race and ethnicity-related factors are in C_2, and socio-economic characteristics related to class are in L_1. The C_1 and C_2 are enclosed in squares as baseline covariates, which are considered for adjustment when examining disparities in Y across the classes defined by X. The choice of which baseline covariates to adjust depends on the type of disparities being investigated. This choice is made at the outset of the analysis. It fixes the target disparity before any decomposition, and everything that follows only allocates that fixed quantity between primary and secondary effects.

In addition, X is related to Y through pathways such as $X \leftarrow H \to L_1 \to \cdots \to Y$. L_2 (previously denoted as U) is influenced by X and L_1 and in turn affects P and Y. Such exposure-induced confounders raise identification difficulties for causal mediation but are accommodated by causal decomposition analysis. Online Appendix H provides a technical note with examples relevant to the present setting.

The pathway $X \to L_2 \to P \to Y$ represents a primary effect because it clarifies how X influences P and subsequently connects to Y. Conversely, the path $X \to L_2 \to Y$ does not pass through P and thus pertains to a secondary effect. What matters is not which L_2 variables count as primary or secondary effects, but how they link X to Y. The distinction between primary and secondary effects depends on whether the pathway passes through P (primary effect) or not (secondary effect). Whether a given L_2 variable (such as cram school attendance or study time) should be treated as part of the performance-mediated mechanism or as a confounder of the P–Y relationship is a theoretical choice. Section 5.2 examines both interpretations by comparing model specifications that differ in the role assigned to these variables.

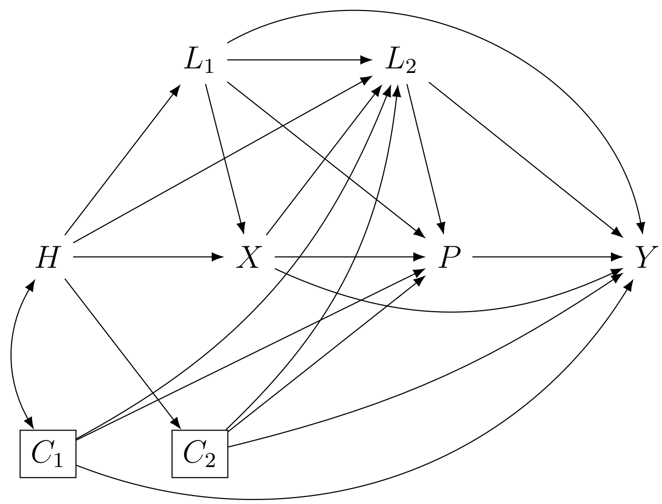


Figure 2: A more realistic directed acyclic graph for how social category X are associated with educational attainment Y

Note: X: social category; P: performance; Y: educational attainment; C_1: confounders of P and Y positioned before X; C_2: confounders of P and Y positioned after X; H: the history that produced the associations among X, C_1, C_2, L_1, and L_2.
Alt text: A directed acyclic graph showing the assumed data-generating process. A horizontal row contains H (history), social class X, academic performance P, and educational attainment Y; L_1 and L_2 sit above the row; the boxed observed confounders C_1 and C_2 sit below. H is a common origin with arrows to X, L_1, L_2, and the confounders. X has arrows to L_2, P, and Y. Each of L_1, L_2, C_1, and C_2 has arrows to P and Y, and P has an arrow to Y.

Based on Figure 2, we examine which pathways can be considered as IEO, primary effects, and secondary effects, respectively. As already noted, this can be viewed as a problem of decomposing the total effect into direct and indirect effects in mediation analysis (Morgan et al. 2013).

3.3 Estimand

The primary target estimand is IEO itself, the observed disparity in educational attainment between social classes. The secondary target is how IEO changes when academic performance distributions are equalized across classes, with the reduction defining the primary effect and the remaining disparity defining the secondary effect. Following Erikson et al. (2005) and Morgan et al. (2013), three classes are considered: service class (S), intermediate (I), and working (W). Y is a binary indicator of university enrollment. Baseline covariates $\boldsymbol{C}$ are considered for adjustment.

Given baseline covariates $\boldsymbol{C} = \boldsymbol{c}$, the disparity between S and W is defined as follows:

$$IEO_{SW|\boldsymbol{c}} = \mathbb{E}[Y_i|X_i = S, \boldsymbol{c}] - \mathbb{E}[Y_i|X_i = W, \boldsymbol{c}]. \tag{1}$$

This quantity is an observed associational difference rather than a causal effect, as it may include both causal and non-causal pathways linking social class to educational attainment as shown in Figure 2.

Let $Y_i(\mathcal{P}_{p|X=S,\boldsymbol{c}})$ refer to the potential outcome Y_i for individual i if their performance level P were set to a random draw from the distribution of the performance level ($\mathcal{P}$) for individuals where $X = S$ and $\boldsymbol{C} = \boldsymbol{c}$. Therefore, $\mathbb{E}[Y_i(\mathcal{P}_{p|X=S,\boldsymbol{c}})|X_i = W, \boldsymbol{c}]$ is the average counterfactual outcome for group W, if their performance level P was set to a random draw from the distribution for group S with the same baseline covariates levels $\boldsymbol{c}$. Then the primary and secondary effects when equalizing to the performance distribution to that of group S can be defined as follows:

$$PE_{\bar{S}W|\boldsymbol{c}} = \mathbb{E}[Y_i(\mathcal{P}_{p|X=S,\boldsymbol{c}})|X_i = W, \boldsymbol{c}] - \mathbb{E}[Y_i|X_i = W, \boldsymbol{c}], \text{and} \tag{2}$$

$$SE_{\bar{S}W|\boldsymbol{c}} = \mathbb{E}[Y_i|X_i = S, \boldsymbol{c}] - \mathbb{E}[Y_i(\mathcal{P}_{p|X=S,\boldsymbol{c}})|X_i = W, \boldsymbol{c}]. \tag{3}$$

The bar over S symbolizes that we are taking the performance distribution of class S and applying it to another class, rather than considering the actual class S itself. $PE_{\bar{S}W|\boldsymbol{c}}$ measures the difference between the counterfactual outcome if group W had the performance level of group S and the actual outcome for group W. If $PE_{\bar{S}W|\boldsymbol{c}}$ is not zero, this reflects the part of the gap due to performance differences between S and W. $SE_{\bar{S}W|\boldsymbol{c}}$ measures the difference between the actual outcome for group S and the counterfactual outcome for group W if they had the same performance level P as group S. It quantifies the change in outcomes when performance levels between S and W are equalized. If $SE_{\bar{S}W|\boldsymbol{c}}$ is not zero, this reflects the part of the gap arising from factors other than performance. The sum of the primary and secondary effects equals the disparity, IEO observed between students from service and working classes.

$$IEO_{SW|\boldsymbol{c}} = PE_{\bar{S}W|\boldsymbol{c}} + SE_{\bar{S}W|\boldsymbol{c}}. \tag{4}$$

Similarly, let $\mathbb{E}[Y_i(\mathcal{P}_{p|X=W,\boldsymbol{c}})|X_i = S, \boldsymbol{c}]$ be the average counterfactual outcome for group S if their performance level P was set to a random draw from the distribution for the group W with the baseline covariates $\boldsymbol{c}$. Then the primary effect and secondary effects can be defined as follows:

$$PE_{S\bar{W}|\boldsymbol{c}} = \mathbb{E}[Y_i|X_i = S, \boldsymbol{c}] - \mathbb{E}[Y_i(\mathcal{P}_{p|X=W,\boldsymbol{c}})|X_i = S, \boldsymbol{c}], \text{and} \tag{5}$$

$$SE_{S\bar{W}|\boldsymbol{c}} = \mathbb{E}[Y_i(\mathcal{P}_{p|X=W,\boldsymbol{c}})|X_i = S, \boldsymbol{c}] - \mathbb{E}[Y_i|X_i = W, \boldsymbol{c}]. \tag{6}$$

Therefore, the sum of the primary and secondary effects equals the IEO.

$$IEO_{SW|\boldsymbol{C}} = PE_{S\bar{W}|\boldsymbol{C}} + SE_{S\bar{W}|\boldsymbol{C}}. \tag{7}$$

In this way, within the potential-outcome framework, we can decompose the IEO into primary and secondary effects as one possible decomposition.

3.4 Identification

The counterfactual quantities in PE and SE are not directly observed. Identification under causal decomposition analysis requires three standard assumptions: conditional independence between potential outcomes and performance given class and the covariate set, positivity (the performance distributions overlap sufficiently across classes), and consistency (Jackson and VanderWeele 2018; Park et al. 2023a,b). Under these assumptions, the counterfactual expectations are non-parametrically identified by the g-formula of Robins (1986). The explicit identification formulas and the technical statement of the three assumptions are given in Online Appendix A. The most consequential assumption is the first: no unobserved confounders of the performance-attainment relationship. This assumption is examined in the sensitivity analysis in Section 5.3. The positivity condition is examined empirically in Online Appendix B, where the kernel densities of academic performance overlap substantially across classes, with at least 90.8% of each class lying within the shared support.

A crucial feature distinguishes this framework from the conventional two-stage model. In the conventional approach, the only way to adjust for a confounder is to enter it into the performance and transition models, which simultaneously changes the disparity being decomposed. Conditioning on a pre-class confounder L_1 (such as parental education or neighborhood) shifts the target from the observed disparity $\mathbb{E}[Y_i|X_i = S] - \mathbb{E}[Y_i|X_i = W]$ to a within-L_1 adjusted disparity that may be only a fraction of the gap of substantive interest, while conditioning on an exposure-induced L_2 additionally removes the part of the class effect that operates through L_2. The disparity decomposition decouples these two roles: the target estimand remains the observed disparity, with primary and secondary effects summing to IEO by construction, whereas L_1 and L_2 enter only the identification of the performance–attainment effect, held at each group's own covariate distribution in the g-formula. Confounders therefore correct the allocation between primary and secondary effects without altering the target disparity being decomposed.

3.5 Estimation

Among the available estimation methods (Lundberg 2022; Park et al. 2022, 2023a), we use the imputation-based approach because academic performance is a continuous variable. This approach builds on the g-formula of Robins (1986). The mathematical formulation and technical details are given in Park et al. (2023a). Here we describe the procedure in natural language.

The procedure works through three steps. First, we establish the relationship between social class and academic performance, adjusting for baseline characteristics. This allows us to predict what each student's academic performance would be under a counterfactual scenario, such as what performance levels working-class students would achieve if they followed the same performance distribution as service-class students. Second, we use these counterfactual performance values to predict educational outcomes. We take each working-class student and ask: given their actual family resources, neighborhood characteristics, and other non-performance factors (L_1 and L_2), what would their probability of university enrollment be if they had the academic performance typical of service-class students? This substitution isolates the role of performance differences while holding the measured non-performance factors (L_1 and L_2) constant. Third, we aggregate these individual predictions to obtain group-level estimates. By averaging the counterfactual outcomes across all working-class students, we obtain the expected university enrollment rate for working-class students under service-class performance patterns. Similarly, we can reverse the analysis to predict what would happen to service-class students if they had working-class performance patterns. The comparison between actual outcomes and these counterfactual predictions yields the primary and secondary effects.

For the causal decomposition, we used the *causal.decomp* package in R (Park et al. 2023a,b). Confidence intervals for the decomposition components are obtained from the nonparametric bootstrap implemented in the package (1,000 samples) within each imputed dataset and combined across the 20 imputations by Rubin's rules (Online Appendix C).

4 Data and Variables

The Japanese setting is informative for the primary/secondary effects question because progression to upper-secondary and higher education remains predominantly examination-based, even as recommendation- and portfolio-based entries have expanded in recent years (Yamamoto and Brinton 2010). Formal selection occurs early, at age 15, on a relatively transparent meritocratic criterion, which makes academic performance an especially salient channel through which class background operates (Kariya and Rosenbaum 1987). At the same time, families' class-differentiated investments in shadow education (cram schools, private tutoring, and selective private or national junior high schools) are widespread and well measured (Yamamoto and Brinton 2010; Entrich 2018), allowing these pre-transition resources to enter the analysis as observable confounders rather than as a purely residual category. Direct decomposition evidence for Japan is nonetheless limited. Using PISA 2003 data, Shirakawa (2017) reports that secondary effects account for about 60–70% of the IEO in educational expectations, an estimate based on the conventional method and on expected rather than actual attainment. The present analysis instead examines actual university enrollment observed through 2023, with adjustment for performance–attainment confounders. The case is therefore not presented as exceptional but as one in which the institutional organization of selection makes the distinction between performance-mediated and non-performance-mediated class effects empirically tractable.

The data used in this study were the Panel Survey of Junior High School Students and Mothers. This survey involved pairs of students, who were in their final year of junior high school (ages 14 to 15, born from April 2000 to March 2001) and their mothers (mean age = 45.2, SD = 3.94). The sample was drawn from the existing online research panel maintained by Ipsos Japan. The Ipsos panel for Japan is constructed primarily through random sampling from the Basic Resident Register (the official residential registry maintained by Japanese municipalities), with a smaller component recruited through referrals from already enrolled panelists to compensate for the increasing restrictions on Basic Resident Register access under Japan's personal-information protection regime. Households with a child in the third year of junior high school and an available mother were invited to participate. This recruitment design is a common constraint of contemporary Japanese sample surveys and limits the strict representativeness of the analytic sample.

The initial phase of the survey, a mail survey, commenced in October 2015. Out of 4,117 invited child-parent pairs, 1,854 pairs provided valid responses from October 2015 to January 2016, yielding a valid response rate of 45.0%. This survey captured data prior to the children's entry into high school, including parental socioeconomic status and background. A follow-up mail survey in December 2017, when the children were in their second year of high school, gathered information about their high school experiences. From December 2019, the survey moved online (LimeSurvey). The dataset includes details such as socioeconomic background (based on the mothers' responses), self-reported junior high school grades, and their educational attainments up to 2023.

The outcome variable, Y, is a binary variable indicating whether an individual enrolled in a university by 2023. The social category X is a three-category version of the Erikson-Goldthorpe-Portocarero class scheme Erikson and Goldthorpe (1992), consisting of S: service class (I+II), I: intermediate class (III and IVab), and W: working class (V+VI, IVc+VIIab).

Academic performance P is a composite score derived from responses regarding grades in multiple subjects asked in the third year of junior high school. The question posed was, "Approximately where do your current grades fall within your grade level?" and it covered the subjects "Overall," "Japanese," "Mathematics," "Science," "Social Studies," and "English." The response options were "Top," "Above average," "Average," "Below average," and "Bottom." Scores were assigned with "Bottom" receiving 1 point and "Top" receiving 5 points. A principal component analysis was applied to these scores. The first principal component had an eigenvalue of 4.42 and the second an eigenvalue of 0.49, with the first component explaining 73.7% of the total variance. The scores on the first principal component were transformed to have a mean of 0 and a standard deviation of 1, and these transformed scores are used as the variable for academic performance. Although prior research has established reasonable accuracy of self-reported grades (Nakazawa 2022), self-assessment bias cannot be ruled out.

The variable L_1, which is associated with social class but not influenced by it, encompasses several important factors. These include the father's and mother's years of education, the neighborhood advantage index, and the distance to the nearest university. It also considers the number of siblings, the individual's birth order, whether any of the grandparents attended college (coded as 1 for yes), the mother's age, the birth month, and the population of the municipality in which the student resides. These components capture various aspects of family background and environmental conditions.

The variable L_2, which is influenced by social class and affects both academic performance and university enrollment, comprises three substantive groups. The first group is economic resources: logged income, logged savings, and the student's and mother's subjective perceptions of wealth. The second group is family investments in academic performance: attendance at cram school, sessions with a private tutor, weekday and weekend study time at non-school educational facilities, weekday and weekend study time at home, and the type of junior high school attended (selective private or national vs. public). The third group is educational aspirations: the child's and mother's university aspirations in ninth grade. The four CD models in Section 5.2 incrementally add these groups as confounders and thereby span the range of defensible choices about which L_2 pathways belong to the performance-mediated mechanism and which do not. Educational aspirations are also likely to be influenced by academic performance, so model specifications that include and exclude them are reported separately.

Since this survey targets students born in the same year, there is no need to adjust for age or year of birth. Gender was included as a baseline covariate C. Consequently, IEO is estimated with adjustment for gender.

University enrollment by 2023 is observed for 1,585 of the 1,854 initial pairs (85.5%). The outcome is unobserved for the other 269 pairs because of panel attrition or item non-response. The analysis retains all 1,854 pairs. Missing values, including unobserved outcomes, were multiply imputed with the Amelia II package in R (Honaker et al. 2011), generating 20 imputed datasets. The descriptive statistics are displayed in Table 1.

Table 1: Descriptive statistics

Node	Variable	*Mean*	*SD*
Y	University enrollment (university = 1)	0.604	
$X = S$	Service (S)	0.389	
$X = I$	Intermediate (I)	0.312	
$X = W$	Working (W)	0.299	
P	Academic performance	0.000	1.000
C_1	Gender (women = 1)	0.508	
L_{11}	Father's years of education	14.065	2.110
L_{12}	Mother's years of education	13.551	1.461
L_{13}	Neighborhood advantage index	0.562	0.255
L_{14}	Distance to the nearest university (logged)	8.228	1.110
L_{15}	Number of siblings	1.424	0.835
L_{16}	Order of birth	1.838	0.846
L_{17}	Grandparents' education (any college = 1)	0.239	
L_{18}	Mother's age/10	4.520	0.394
L_{19}	Birth month (April = 0)	5.330	3.396
L_{110}	Population (logged)	11.784	1.010
L_{21}	Logged income	6.377	0.590
L_{22}	Logged savings	5.450	1.676
L_{23}	Student's subjective wealth	2.335	0.950
L_{24}	Mother's subjective wealth	1.890	0.812
L_{25}	Attendance at cram school	0.661	
L_{26}	Sessions with a private tutor	0.048	
L_{27}	Study time at non-school educational facilities on weekdays	0.900	0.779
L_{28}	Study time at home on weekdays	0.777	0.571
L_{29}	Study time at non-school educational facilities on weekends	0.848	0.945
L_{210}	Study time at home on weekends	1.053	0.764
L_{211}	Sector of junior high school	0.083	
L_{212}	Child's university aspiration in ninth grade	0.638	
L_{213}	Mother's university aspiration for the child in ninth grade	0.675	

Note: Values shown are pooled estimates from 20 multiply imputed datasets using Rubin's rules. Missing data were imputed using the Amelia II package. Y is the outcome variable indicating university enrollment. X represents social class. P is standardized academic performance (mean = 0, SD = 1). C_1 is the baseline covariate, gender (female = 1). $L_{11} - L_{213}$ are confounders. The sample size reflects the complete cases after imputation ($n = 1854$).

5 Results

5.1 Conventional Method

Figure 3 summarizes the three descriptive relationships that motivate the decomposition: (a) the class gradient in university enrollment (Y by X), (b) the class gradient in academic performance (P by X), and (c) the performance–enrollment relationship within each class (Y by P and X). The enrollment rate is 0.72 for the service class, 0.63 for the intermediate class, and 0.43 for the working class. These rates define the disparities that the decompositions below take as their target. Adjusting for gender, IEO is 0.285 (95% confidence interval [0.230, 0.341]) for the service–working comparison, 0.092 [0.032, 0.152] for the service–intermediate comparison, and 0.193 [0.123, 0.264] for the intermediate–working comparison (Online Appendix C). The methods that follow differ only in how they allocate these fixed totals between PE and SE. Performance differs across classes by about 0.4 standard deviations between the service and working classes and about 0.2 standard deviations between adjacent classes (service–intermediate and intermediate–working). The performance–enrollment slopes are broadly comparable across classes, but the class-specific intercepts differ substantially: at any given performance level, service-class students are markedly more likely to enroll than working-class students.

We then apply the conventional decomposition methods of Erikson et al. (2005) and Morgan (2012). For Erikson et al.'s method we use the *DECIDE* package in R (Kartsonaki 2022). For Morgan's probability-based version of the standard model we use a logistic regression of university enrollment on class, performance, and their interaction. The estimated parameters and the underlying logistic models are reported in Tables 2 and 3. Numerical details are discussed in Online Appendix F.

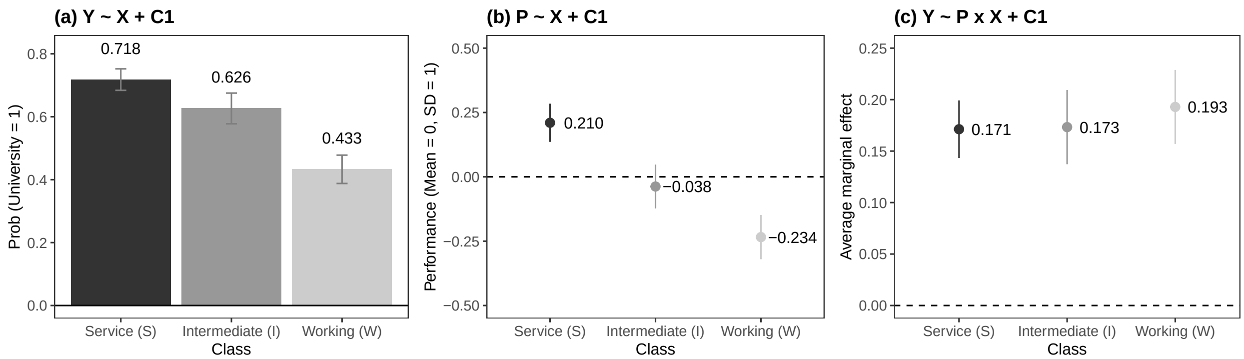


Figure 3: Descriptive Patterns of Educational Attainment and Performance by Social Class

Note: Panel A shows the probability of university enrollment by social class with 95% confidence intervals. Panel B displays the mean academic performance (standardized with mean = 0, SD = 1) by social class with 95% confidence intervals. Panel C presents the average marginal effects of academic performance on university enrollment for each social class from a logistic regression model with interaction terms. All models control for gender (C_1). Error bars represent 95% confidence intervals.
Alt text: A three-panel figure, with 95% confidence intervals shown throughout. Panel A is a bar chart of the probability of university enrollment by social class (Service, Intermediate, Working). Panel B plots the mean standardized academic performance (mean 0, standard deviation 1) by class. Panel C plots the average marginal effect of academic performance on university enrollment by class, from a logistic model with class-by-performance interactions. Values are highest for the Service class and lowest for the Working class in Panels A and B.

Table 2: Estimated parameters.

Class (X)	n	$\hat{\Pr}(Y=1\|X)$	$\hat{\mu}_X$	$\hat{\sigma}_X$	$\hat{\alpha}_X$	$\hat{\beta}_X$	$-\hat{\alpha}_X/\hat{\beta}_X$
Service (S)	722	0.718	0.210	0.967	0.890	1.039	−0.856
Intermediate (I)	573	0.626	−0.038	0.993	0.625	0.862	−0.725
Working (W)	559	0.433	−0.234	0.994	−0.109	0.923	0.119
Full Sample	1854	0.604	0.000	1.000	0.673	0.989	−0.681

Note: Parameters are pooled estimates from 20 multiply imputed datasets using Rubin's rules. n is the sample size per class. $\hat{\Pr}(Y=1|X)$ is the proportion making the transition to university for each class X, $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation of academic performance within each class. $\hat{\alpha}$ and $\hat{\beta}$ are the estimated logistic regression coefficients where the transition probability is modeled as $\Pr(Y=1|P) = 1/(1+\exp(-(\alpha_X + \beta_X P)))$ for each class X. The ratio $-\hat{\alpha}_X/\hat{\beta}_X$ represents the performance threshold at which the transition probability equals 0.5. The full sample row shows the overall parameters when class distinctions are ignored.

Figure 4 provides a graphical representation of the primary and secondary effects, following Erikson et al. (2005), using the estimated parameters from Table 2, assuming that academic performance follows a normal distribution within each class. The distance between the class-specific performance distributions represents the primary effect, while the distance between the class-specific transition curves represents the secondary effect.

Table 3 reports three logistic regression models predicting university enrollment from class, performance, and their interaction. Table 4 reports the observed and counterfactual transition probabilities from Erikson et al.'s and Morgan's methods. Following Morgan (2012), the decomposition is presented in probability differences rather than log-odds ratios.

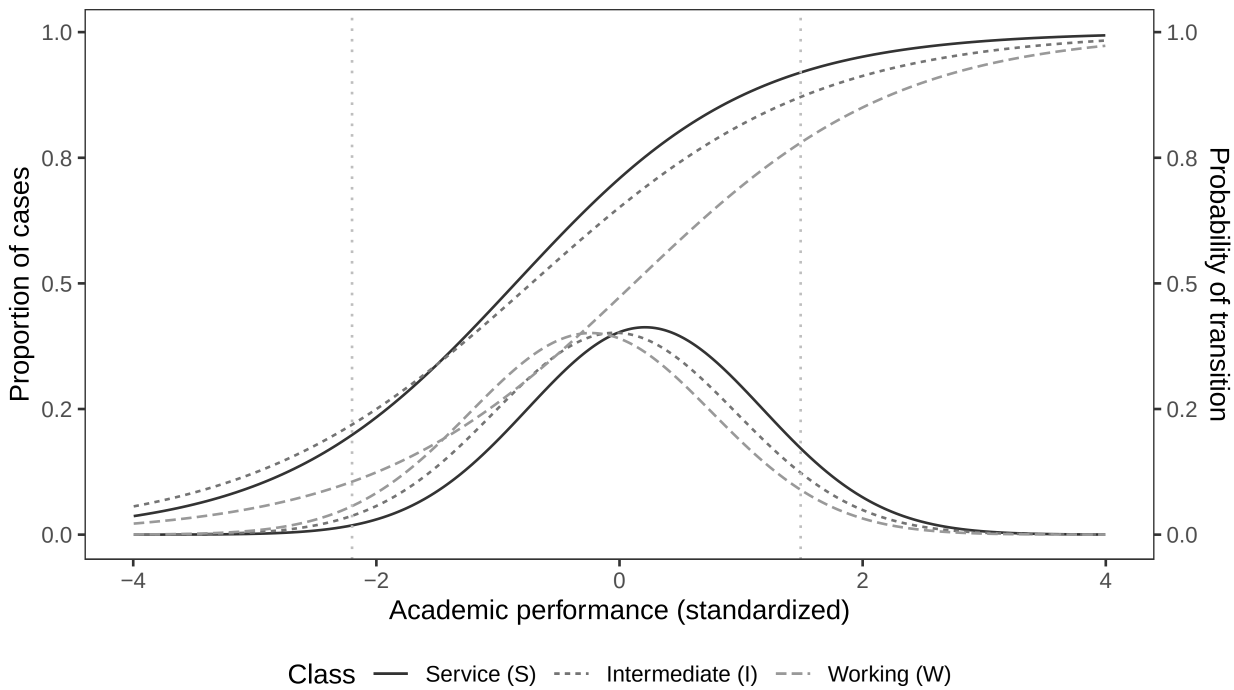


Figure 4: Performance distributions and transition probabilities by class

Note: Theoretical transition probabilities (rising curves, right axis) and performance distributions (bell curves, left axis) calculated from the estimated parameters ($\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\alpha}_X$, $\hat{\beta}_X$) for each class in Table 2. Solid line: Service class (S); Dashed line: Intermediate class (I); Dotted line: Working class (W).
Alt text: A combined density-and-line plot. For each social class it shows a bell-shaped academic-performance distribution (left axis) and an upward-sloping curve giving the probability of university enrollment as a function of standardized performance (right axis). The Service class is drawn with a solid line, the Intermediate class with a dashed line, and the Working class with a dotted line. The curves are computed from the estimated class-specific parameters in Table 2, with the Service class enrolling at lower performance levels than the Working class.

Table 3: Logistic regression models for university enrollment

	Model 1	Model 2	Model 3
Constant	1.104	1.082	1.086
	(0.102)	(0.113)	(0.114)
Intermediate ($X = I$)	−0.423	−0.250	−0.272
	(0.138)	(0.147)	(0.148)
Working ($X = W$)	−1.215	−1.004	−1.011
	(0.126)	(0.138)	(0.140)
Gender (women = 1)	−0.319	−0.365	−0.362
	(0.101)	(0.111)	(0.111)
Performance		0.948	1.035
		(0.069)	(0.117)
× Intermediate ($X = I$)			−0.167
			(0.169)
× Working ($X = W$)			−0.097
			(0.168)
N	1,854	1,854	1,854

Note: Pooled estimates from 20 multiply imputed datasets using Rubin's rules. Standard errors in parentheses.

Table 4: Estimated probabilities of university enrollment

Class	S	I	W
(a) Probabilities from Table 2			
Service (S)	0.718	0.669	0.518
Intermediate (I)	0.670	0.626	0.469
Working (W)	0.631	0.590	0.431
(b) Probabilities from Model 3 in Table 3			
Service (S)	0.718	0.671	0.523
Intermediate (I)	0.670	0.627	0.472
Working (W)	0.632	0.591	0.433

Note: Rows represent the performance distribution of the specified class; columns represent the educational choice pattern of the specified class. Diagonal entries show observed probabilities. Panel (a) uses estimated parameters from Table 2; panel (b) uses predicted probabilities from Model 3 in Table 3. Values are pooled estimates from 20 multiply imputed datasets.

The Erikson et al. and Morgan methods yield almost identical results. The decomposition is summarised in Figure 5. Full numerical values are in Online Appendix C. Conventional methods estimate a primary-effect share of about 30% for the S–W gap, about 18–20% for I–W, and about 48–52% for S–I. Across all comparisons, secondary effects account for roughly half or more of the disparity.

The two values of PE within each pairwise comparison, obtained by equalizing to one or the other class's performance distribution, are not identical. Following Morgan (2012), these asymmetries are not estimation noise but reflect a substantive property of the decomposition: the estimates average individual-level effects over the performance distribution of the class from which counterfactual movement originates, and because primary effects make these distributions differ across classes, direction-specific estimates differ in expectation. Online Appendix D shows that this directional gap widens when the underlying disparity is small. Collapsing them into a single log-odds-based measure, as in earlier work, obscures this heterogeneity.

5.2 Causal Decomposition Method

Table 4 also summarises the decomposition results using causal decomposition analysis with varying sets of control variables. Unlike the conventional methods, which exchange the class-specific performance distribution and the class-specific choice pattern between groups, the causal decomposition holds L_1 and L_2 fixed and equalizes only the performance distribution across classes (see Section 3). We compare these results with the traditional methods to assess both compatibility and the impact of controlling for confounders. Figure 5 depicts the contributions of PE and SE to IEO of all models including Erikson et al.'s and Morgan's methods.

The four causal decomposition models differ in how they specify L_2, the set of variables that are influenced by social class and that may affect both academic performance and educational attainment. CD Model 1 conditions only on gender. CD Model 2 adds L_1 pre-class confounders together with the economic-resource components of L_2 (income, savings, and subjective wealth) but excludes family investments in performance such as cram school attendance, private tutoring, study time, and private or national school sector. CD Model 3 adds these investment variables as confounders, and CD Model 4 further adds educational aspirations as proxies for an anticipatory decision. The four models correspond to four interpretive stances on how much of the class effect on attainment should be attributed to the performance-mediated pathway. Private and national school sector is treated alongside cram school and study time as a performance-related investment, on the view that the choice to send a child to a selective private or national junior high school is, like cram school attendance, a goal-directed family investment intended to raise academic performance, even though private-school environments may also affect attainment through non-performance channels such as peer networks and exam-oriented curricula.

CD Model 1, which mirrors the conventional methods of Erikson et al. and Morgan in conditioning only on gender, yields similar results to both. The total inequality measures are virtually identical (0.285 for the S–W comparison), and the proportions attributed to primary effects are close across the three approaches for the service-working comparison, at approximately 29–32%, with wider direction-specific variation for service-intermediate (38–59%) and intermediate-working (15–20%) comparisons. These results are compatible with the conventional methods when minimal controls are used.

CD Model 2 retains family investments in performance (cram school, private tutoring, study time, and private/national school sector) as part of the primary-effect pathway rather than as confounders to be adjusted away. Under this specification, the PE share for the S–W comparison is approximately 24%, with corresponding values for the S–I comparison (34–40%) and I–W comparison (10–13%). These values are modestly lower than CD Model 1 because pre-class confounders (L_1) and economic-resource components of L_2 (income, savings, subjective wealth) have already been adjusted away.

CD Model 3 additionally treats cram school, private tutoring, study time, and school sector as confounders of the P–Y relationship. The PE share for the S–W comparison is approximately 23%, with corresponding values of 34–38% for S–I and 10–12% for I–W. The gap between CD Model 2 and CD Model 3 is substantively small: treating cram school, private tutoring, and study time as confounders rather than as mechanisms produces almost no change in the estimated PE/SE allocation. Conditional on the pre-class confounders and economic resources already adjusted in CD Model 2, additionally treating the family investments in performance (cram school, private tutoring, study time, and school sector) as confounders adds little further P–Y confounding, so the PE/SE allocation does not depend strongly on whether these investment variables are treated as mechanism or as confounders.

CD Model 4 further includes educational aspirations. Here primary effects decrease further: for S–W comparisons to 12–14% (from 23% in CD Model 3), for S–I comparisons to 19–22% (from 34–38%), and for I–W comparisons to 4–5% (from 10–12%). Morgan (2012) identifies two permissible interpretations of the role of educational aspirations that cannot be resolved empirically. Under the first interpretation, aspirations indicate an anticipatory decision that constitutes the choice-based secondary effect itself (Jonsson and Erikson 2000; Erikson et al. 2005), so conditioning on them purges a portion of the genuine secondary effect and biases the PE estimate downward. CD Model 4 then gives a lower bound for PE. Under the second interpretation, aspirations are characteristics shaped by earlier socialization that mediate, rather than constitute, the secondary effect (Morgan 2012), in which case CD Model 4 represents the most comprehensive adjustment for confounding. Because educational aspirations measured concurrently with academic performance may also be partly shaped by performance itself, overcontrol bias remains a concern under either reading.

Across the four specifications, the models span the plausible range of the PE share for the S–W gap. CD Model 1, which mirrors conventional decomposition methods, marks the upper end at approximately 30%. CD Models 2 and 3, which control for L_1 and the economic-resource components of L_2 and differ only in whether family investments in performance are also treated as confounders, give nearly identical PE shares around 23–25%. The insensitivity of the central estimate to the mechanism-versus-confounder interpretation of cram school and study time is substantive: once economic resources and school sector are adjusted for, classifying family investments as part of the primary mechanism (CD

Model 2) or as confounders (CD Model 3) makes little practical difference to the decomposition. CD Model 4, which additionally treats educational aspirations as confounders, marks the lower end at 12–14%. Simulation-based 95% confidence intervals (Online Appendix C) show that the S–W primary effect remains positive under every specification. Equalizing to the service-class performance distribution, PE is 0.081 [0.049, 0.114] under CD Model 1 and 0.034 [0.009, 0.058] under CD Model 4. For the smaller I–W gap, by contrast, the intervals include zero once confounders are adjusted. Across this range, secondary effects account for between roughly 68% and 88% of the S–W gap. Even under the specification most favorable to the performance-mediated pathway, most of the class gap in attainment runs through channels other than measured performance.

The numerical values underlying Figure 5 are reported in Online Appendix C.

5.3 Robustness

We conducted three robustness checks. First, following the method of Park et al. (2023b), we examined the sensitivity of CD Model 3 estimates to an unmeasured P–Y confounder under varying assumed strengths. The PE estimates remain robust at moderate levels of unmeasured confounding: an unmeasured variable would need to explain a larger share of variance in both P and Y than any single observed covariate (parental education, family income, or cram school attendance) to eliminate the estimated primary effect. Second, we re-estimated the four models with the outcome redefined as entry into any form of higher education (specialized training schools, junior colleges, technical colleges, 4-year universities, and graduate schools). The S–W disparity in higher education entry is smaller (IEO $=$ 0.147 versus 0.285 for the main 4-year-university outcome), but across the four CD models the PE share remains in a comparable range (about 5–37%, versus 12–32% for the main outcome), preserving the qualitative finding that secondary effects account for the majority of the disparity. Detailed results for both checks are reported in Online Appendix E and Online Appendix D respectively. Third, the reduction in the primary-effect share when confounders are adjusted is robust to the outcome-model specification. Replacing

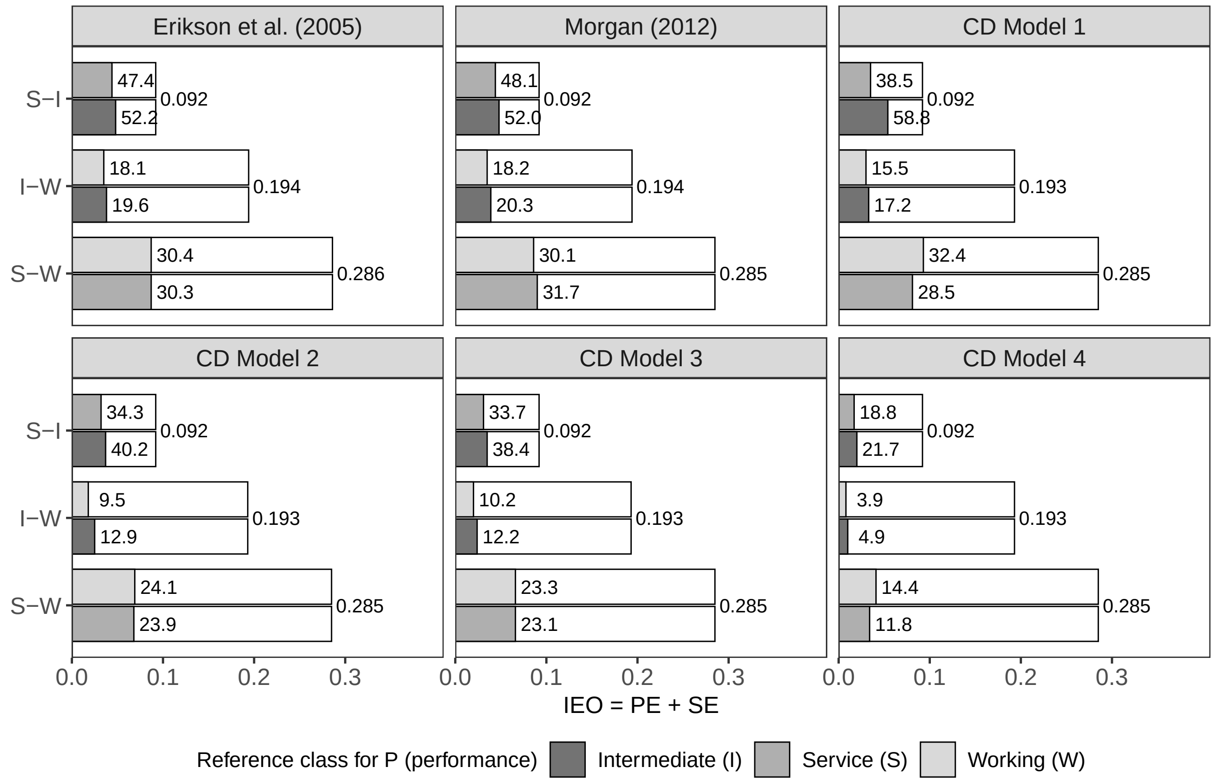


Figure 5: Decomposition of Inequalities in Educational Opportunity (IEO) into Primary Effects (PE) and Secondary Effects (SE): Results from Traditional Methods (Erikson and Morgan) and Causal Decomposition Models with Different Sets of Controls

Note: Bars represent total inequality (IEO) between social classes. The axis label indicates the comparison between two classes: 'S-W' indicates the comparison between Service and Working classes; 'S-I' indicates the comparison between Service and Intermediate classes; and 'I-W' indicates the comparison between Intermediate and Working classes. Gray portions indicate primary effects (PE), representing the contribution of performance differences to educational inequality. Numbers next to the gray portions show PE percentages for each comparison pattern. The shade of gray indicates the reference class whose performance distribution is applied (e.g., S indicates equalizing to the Service class performance distribution, W to the Working class distribution). CD Model 1 includes only gender (C_1) as a control. CD Model 2 adds L_1 (factors associated with but not influenced by social class, such as parental education and neighborhood characteristics) and the economic-resource components of L_2 (income, savings, and subjective wealth), treating family investments in performance as part of the primary-effect mechanism. CD Model 3 additionally treats these investments (cram school, private tutoring, study time, and private/national school sector) as confounders of the P–Y relationship. CD Model 4 further adds the child's and mother's educational aspirations. The decrease in primary effects from CD Model 1 to CD Model 4 shows how controlling for confounders reduces the estimated contribution of academic performance to educational inequality.
Alt text: A grid of horizontal bar charts, one panel per method (Erikson, Morgan, and the four Causal Decomposition models). Within each panel, bars give the total inequality (IEO) for the class comparisons S–W, I–W, and S–I. The gray portion of each bar is the primary effect (the performance-mediated share) and the remaining portion is the secondary effect; labels next to the gray portions give the primary-effect percentage. The primary-effect share decreases systematically from the Erikson and Morgan methods through the CD models, while the total length of each bar (total IEO) stays roughly constant.

the logistic outcome regression with a range of statistical learners (penalized regression, splines, generalized additive models, tree ensembles, gradient boosting, Bayesian additive regression trees, a neural network, and a Super Learner), the primary-effect share for the S–W gap still falls when confounders are added, by about 4 to 7 percentage points for nine of the ten learners, with only gradient boosting (already low without adjustment) essentially unchanged. The same reduction holds for the S–I and I–W comparisons (Online Appendix I). The adjusted level is less precisely identified across learners (roughly 23 to 26% for the S–W gap), but its reduction relative to the gender-only specification is consistent.

6 Limitations

This analysis has several limitations. We highlight four that we consider most consequential for interpreting the results. First, the operationalization of the intervention is conceptually idealized. Performance is itself produced by social-class-linked inputs (study time, cram school attendance, family resources), so an intervention that altered P alone, leaving these inputs untouched, is not coherent as a real-world policy. The decomposition reported here should be read as a benchmark indicating the upper limit of disparity reduction achievable through performance-targeted interventions, not as an estimate of the effect of any specific policy.

Second, the measurement of academic performance is imperfect. Self-reported grades may carry measurement error and class-specific reference-point biases. Although prior research finds the self-reports broadly accurate (Nakazawa 2022), the present scale cannot rule out systematic differences in how students from different social backgrounds evaluate their own performance.

Third, the mechanisms behind the secondary effect remain unspecified. The causal decomposition identifies the magnitude of SE but not what drives it. The reported decomposition also reports population-level averages and does not capture the unit-level heterogeneity emphasized by Morgan (2012), who shows that policy interventions shifting the enrollment threshold or shifting non-performance characteristics will have differential effects across classes that aggregate decompositions cannot reveal. The aspirations specification (CD Models 3 vs. 4) further illustrates the difficulty of separating choice from confounding without independent measurement of beliefs.

Finally, as already noted in Section 3.1, the present operationalization is one reading of Boudon's framework rather than its definitive translation. The estimand defined here captures every non-performance-mediated pathway, which is broader than choice alone and includes measurement error and structural mechanisms. Alternative operationalizations, including direct modeling of choice mechanisms, joint analysis of multiple measured mediators (Park et al. 2024), and a belief-equalization estimand parallel to the performance-equalization estimand proposed here, are promising directions for future research (Abbiati et al. 2018; Barone et al. 2018).

7 Conclusion

This study applies a causal decomposition approach, distinct from both conventional associational decomposition and causal mediation analysis, to Boudon's primary and secondary effects. The framework defines IEO as an observed disparity and treats academic performance as the target of a counterfactual equalizing intervention, which preserves the theoretical distinction between performance-mediated and non-performance-mediated pathways while circumventing the identification problem that the causal effect of social class itself raises. The approach addresses the methodological challenges articulated by Morgan (2012), and the same framework can be applied to disparities defined by race, ethnicity, gender, or childhood disability, and their intersections.

Primary effects are smaller than conventional estimates suggest. Once P–Y confounders are accounted for, the share of IEO attributable to performance differences narrows from about 30% to between 12 and 24%. We can draw two implications from these results.

First, conventional decompositions risk overstating the role of academic performance in generating IEO and, by extension, the expected returns to performance-targeted policies. Second, and as the counterpart of the first, the share attributed to secondary effects rises from about 70% under conventional methods to between 76 and 88% once confounders are adjusted. Even under the idealized intervention of equalizing performance distributions across classes, then, the majority of the observed gap would remain. This enlarged share should not be interpreted as a direct estimate of choice-based mechanisms. It combines choice, measurement error, and structural channels not captured by P. What positive interventions would close the remaining gap is a question the present analysis cannot answer. Answering it will require richer measurement of class-specific beliefs about costs, success probabilities, returns, and downward mobility risks (Abbiati et al. 2018; Barone et al. 2018), together with the framework developed here applied to belief-equalization as well as performance-equalization estimands.

References

Abbiati, Giovanni, Gianluca Argentin, Carlo Barone, and Antonio Schizzerotto. 2018. "Information Barriers and Social Stratification in Higher Education: Evidence from a Field Experiment." *The British Journal of Sociology* 69:1248–1270.

Barone, Carlo, Moris Triventi, and Giulia Assirelli. 2018. "Explaining Social Inequalities in Access to University: A Test of Rational Choice Mechanisms in Italy." *European Sociological Review* 34:554–569.

Boudon, Raymond. 1974. *Education, Opportunity, and Social Inequality*. Wiley.

Boudon, Raymond. 1998. "Limitations of Rational Choice Theory." *American Journal of Sociology* 104:817–828.

Breen, Richard and John H. Goldthorpe. 1997. "Explaining Educational Differentials: Towards a Formal Rational Action Theory." *Rationality and Society* 9:275–305.

Breen, Richard, Ruud Luijkx, Walter Müller, and Reinhard Pollak. 2009. "Nonpersistent Inequality in Educational Attainment: Evidence from Eight European Countries." *American Journal of Sociology* 114:1475–1521.

Elwert, Felix and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40:31–53.

Entrich, Steve R. 2018. *Shadow Education and Social Inequalities in Japan*. Cham: Springer International Publishing.

Erikson, Robert and John H. Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford: Clarendon.

Erikson, Robert, John H. Goldthorpe, Michelle Jackson, Meir Yaish, and David R. Cox. 2005. "On Class Differentials in Educational Attainment." *Proceedings of the National Academy of Sciences* 102:9730–9733.

Erikson, Robert and Jan O. Jonsson. 1996. "Explaining Class Inequality in Education: The Swedish Test Case." In *Can Education Be Equalized? The Swedish Case in Comparative Perspective*, edited by Robert Erikson and Jan O. Jonsson, pp. 1–63. Boulder: Westview Press.

Erikson, Robert and F. Rudolphi. 2010. "Change in Social Selection to Upper Secondary School–Primary and Secondary Effects in Sweden." *European Sociological Review* 26:291–305.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "AMELIA II: A Program for Missing Data." *Journal of statistical software* 45:1–47.

Jackson, John W. and Tyler J. VanderWeele. 2018. "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities:." *Epidemiology* 29:825–835.

Jackson, Michelle (ed.). 2013a. *Determined to Succeed?: Performance versus Choice in Educational Attainment*. Stanford, California: Stanford University Press.

Jackson, Michelle. 2013b. "How Is Enequality of Educational Opportunity Generated? The Case for Primary and Secondary Effects." In *Determined to Succeed?: Performance versus Choice in Educational Attainment*, edited by Michelle Jackson, pp. 1–33. Stanford, California: Stanford University Press.

Jackson, Michelle, Robert Erikson, John H. Goldthorpe, and Meir Yaish. 2007. "Primary and Secondary Effects in Class Differentials in Educational Attainment: The Transition to A-Level Courses in England and Wales." *Acta Sociologica* 50:211–229.

Jackson, Michelle and Jan O. Jonsson. 2013. "Why Does Inequality of Educational Opportunity Vary across Countries? Primary and Secondary Effects in Comparative Context." In *Determined to Succeed?: Performance versus Choice in Educational Attainment*, edited by Michelle Jackson, pp. 306–337. Stanford, California: Stanford University Press.

Jonsson, Jan O. and Robert Erikson. 2000. “Understanding Educational Inequality: The Swedish Experience.” *L’Année sociologique* 50:345–382.

Kariya, Takehiko and James E. Rosenbaum. 1987. “Self-Selection in Japanese Junior High Schools: A Longitudinal Study of Students’ Educational Plans.” *Sociology of Education* 60:168.

Karlson, Kristian Bernt. 2013. “Summarizing Primary and Secondary Effects.” *Research in Social Stratification and Mobility* 33:72–82.

Karlson, Kristian Bernt and Anders Holm. 2011. “Decomposing Primary and Secondary Effects: A New Decomposition Method.” *Research in Social Stratification and Mobility* 29:221–237.

Kartsonaki, Christina. 2022. “DEComposition of Indirect and Direct Effects (Package ‘DECIDE’).”

Kartsonaki, Christina, Michelle Jackson, and David R. Cox. 2013. “How Is Enequality of Educational Opportunity Generated? The Case for Primary and Secondary Effects.” In *Determined to Succeed?: Performance versus Choice in Educational Attainment*, edited by Michelle Jackson, pp. 34–55. Stanford, California: Stanford University Press.

Lundberg, Ian. 2022. “The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories.” *Sociological Methods & Research* p. 00491241211055769.

Morgan, Stephen L. 2012. “Models of College Entry in the United States and the Challenges of Estimating Primary and Secondary Effects.” *Sociological Methods & Research* 41:17–56.

Morgan, Stephen L., Michael W. Spiller, and Jennifer J. Todd. 2013. “Class Origins, High School Graduation, and College Entry in the United States.” In *Determined to Succeed?:*

Performance versus Choice in Educational Attainment, edited by Michelle Jackson, pp. 279–305. Stanford, California: Stanford University Press.

Nakazawa, Wataru. 2022. “Self-Reported Academic Grades in Social Surveys: Implications and Reliability during Third Year of Junior High School.” *Shakaigaku Hyoron (Japanese Sociological Review)* 72:487–503.

Park, Soojin, Suyeon Kang, and Chioun Lee. 2023a. “Choosing an Optimal Method for Causal Decomposition Analysis with Continuous Outcomes: A Review and Simulation Study.” *Sociological Methodology* p. 00811750231183711.

Park, Soojin, Suyeon Kang, Chioun Lee, and Shujie Ma. 2023b. “Sensitivity Analysis for Causal Decomposition Analysis: Assessing Robustness toward Omitted Variable Bias.” *Journal of Causal Inference* 11.

Park, Soojin, Namhwa Lee, and Rafael Quintana. 2024. “Causal Decomposition Analysis With Time-Varying Mediators: Designing Individualized Interventions to Reduce Social Disparities.” *Sociological Methods & Research* p. 00491241241264562.

Park, Soojin, Xu Qin, and Chioun Lee. 2022. “Estimation and Sensitivity Analysis for Causal Decomposition in Health Disparity Research.” *Sociological Methods & Research* p. 00491241211067516.

Robins, James. 1986. “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect.” *Mathematical Modelling* 7:1393–1512.

Schindler, Steffen and Markus Lörz. 2012. “Mechanisms of Social Inequality Development: Primary and Secondary Effects in the Transition to Tertiary Education Between 1976 and 2005.” *European Sociological Review* 28:647–660.

Shavit, Yossi and Hans-Peter Blossfeld (eds.). 1993. *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. Boulder, Colorado: Westview Press.

Shirakawa, Toshiyuki. 2017. "Socioeconomic Differences In Educational Expectation: Primary and Secondary Effects in International Context." *Kansai Sociological Review* 16:72–84.

VanderWeele, Tyler J. and Whitney R. Robinson. 2014. "On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables:." *Epidemiology* 25:473–484.

Yamamoto, Yoko and Mary C. Brinton. 2010. "Cultural Capital in East Asian Educational Systems: The Case of Japan." *Sociology of Education* 83:67–83.

Online Appendix to "Another Look at Primary and Secondary Effects: Causal Decomposition of Inequalities in Educational Opportunity"

This document contains supplementary material for the article "Another Look at Primary and Secondary Effects: Causal Decomposition of Inequalities in Educational Opportunity." References to "Figure 1," "Figure 2," and other numbered objects in the main text refer to figures and tables in the published article.

A. Causal Mediation Estimands and Why They Are Not Adopted

Causal mediation analysis offers a natural alternative to the disparity decomposition framework adopted in the main text. It decomposes the causal effect of social class on educational attainment into direct and indirect components (Hong 2015; VanderWeele 2015; Brand et al. 2019; Klein and Kühhirt 2021; Shi et al. 2021; Wodtke and Zhou 2020; Zhou and Wodtke 2019). This section provides the formal definitions of these estimands, states their identification conditions, and explains why they are not adopted here.

Let $Y_i(x)$ denote the potential outcome for individual i when class is set to x, where $x \in \{S, I, W\}$, and let $Y_i(x,p)$ denote the potential outcome when class is set to x and performance is set to p. The average treatment effect (ATE) of class contrasting S and W is

$$ATE_{SW} = \mathbb{E}[Y_i(S)] - \mathbb{E}[Y_i(W)]. \tag{A1}$$

The average Controlled Direct Effect (CDE) contrasting S and W when performance is fixed at p is

$$CDE_{SW}(p) = \mathbb{E}[Y_i(S,p)] - \mathbb{E}[Y_i(W,p)]. \tag{A2}$$

Identification of CDE requires (1) no unobserved confounders between X and Y, and (2) no confounders between P and Y. Fixing P at a single value p implies an intervention in which all individuals are forced to that performance value, which is neither theoretically supported nor practically feasible.

Let $P_i(x)$ denote the performance score for individual i when class is set to x. The Natural Direct Effect (NDE) and Natural Indirect Effect (NIE) are

$$NDE_{SW} = \mathbb{E}[Y_i(S, P_i(S))] - \mathbb{E}[Y_i(W, P_i(S))], \tag{A3}$$

$$NIE_{SW} = \mathbb{E}[Y_i(W, P_i(S))] - \mathbb{E}[Y_i(W)], \tag{A4}$$

$$ATE_{SW} = NDE_{SW} + NIE_{SW}. \tag{A5}$$

The composition assumption is $Y_i(S) = Y_i(S, P_i(S))$ and $Y_i(W) = Y_i(W, P_i(W))$. Identification requires, in addition to (1) and (2), (3) no unobserved confounders between X and P, and (4) no unobserved confounders influenced by X between P and Y (VanderWeele 2015). If one targets the randomized analogues of the natural effects, assumption (4) is replaced by a randomized-intervention condition (VanderWeele et al. 2014; Wodtke and Zhou 2020), but the requirement to identify the causal effect of X on Y remains.

These estimands are not adopted in the main analysis for two related reasons. First, the structural problem identified by Morgan (2012) concerns assumption (1): backdoor paths from X to Y through race, neighborhood, and institutional features make the causal effect of X on Y implausible to identify in observational data. Second, condition (4) is violated when X affects L_2 as it does in Figure 2 of the main text. The estimands above are also conceptually demanding because they require the analyst to specify a counterfactual intervention on social class itself. The disparity decomposition framework adopted in the main text avoids both difficulties by treating X as a grouping variable that defines the disparity rather than as a manipulable cause, and by defining the intervention on P alone. Boudon noted that "IEO rates are subject to variations as a function of national context, point in time, and school level" (Boudon 1974, 41), suggesting that he conceived of IEO as observed differences in attainment by social background rather than as the causal effect of background on attainment, which is consistent with the present framework. The same reading applies to performance: Boudon appears to have treated class differences in performance as observed quantities, and the counterfactual equalization of performance distributions is a formalization introduced here rather than an idea found in his account.

Formal Definition of the Disparity Decomposition Estimands. For class S versus W, the disparity is

$$IEO_{SW|\boldsymbol{c}} = \mathbb{E}[Y_i|X_i = S, \boldsymbol{c}] - \mathbb{E}[Y_i|X_i = W, \boldsymbol{c}]. \tag{A6}$$

Let $Y_i(\mathcal{P}_{p|X=x,\boldsymbol{c}})$ denote the potential outcome when individual i's performance is set to a random draw from the conditional distribution $\Pr(P|X = x, \boldsymbol{c})$. The primary and secondary effects equalizing to class S are

$$PE_{\bar{S}W|\boldsymbol{c}} = \mathbb{E}[Y_i(\mathcal{P}_{p|X=S,\boldsymbol{c}})|X_i = W, \boldsymbol{c}] - \mathbb{E}[Y_i|X_i = W, \boldsymbol{c}], \tag{A7}$$

$$SE_{\bar{S}W|\boldsymbol{c}} = \mathbb{E}[Y_i|X_i = S, \boldsymbol{c}] - \mathbb{E}[Y_i(\mathcal{P}_{p|X=S,\boldsymbol{c}})|X_i = W, \boldsymbol{c}], \tag{A8}$$

with $IEO_{SW|\boldsymbol{c}} = PE_{\bar{S}W|\boldsymbol{c}} + SE_{\bar{S}W|\boldsymbol{c}}$. The corresponding quantities equalizing to class W are defined symmetrically.

Identification Conditions. The following assumptions are sufficient to identify PE and SE from observable data (Jackson and VanderWeele 2018; Park et al. 2023a,b):

C1: Conditional Independence $Y_i(p) \perp\!\!\!\perp P_i|X_i = x, \boldsymbol{L}_i = \boldsymbol{l}, \boldsymbol{C}_i = \boldsymbol{c}$ for all $x \in \{S, I, W\}$.

C2: Positivity $f(p \mid X = x, \boldsymbol{L} = \boldsymbol{l}, \boldsymbol{C} = \boldsymbol{c}) > 0$ for all $x \in \{S, I, W\}$ and all p in the support of the reference-class performance distribution, where f denotes the conditional density of P. Because P is continuous, this is a common-support condition on overlapping densities rather than a positive point mass. It is examined empirically in Appendix B.

C3: Consistency If $P_i = p$, then $Y_i = Y_i(p)$.

Condition C1 requires no unobserved P–Y confounders. Under C1–C3, the counterfactual expectations are identified non-parametrically by the g-formula of Robins (1986):

$$\mathbb{E}[Y_i(\mathcal{P}_{p|X=S,\boldsymbol{c}})|X_i = W, \boldsymbol{c}] = \sum_{p,\boldsymbol{l}} \mathbb{E}[Y_i|X_i = W, p, \boldsymbol{l}, \boldsymbol{c}] \Pr(\boldsymbol{l}|X_i = W, \boldsymbol{c}) \Pr(p|X_i = S, \boldsymbol{c}), \quad \text{(A9)}$$

and symmetrically for $\mathbb{E}[Y_i(\mathcal{P}_{p|X=W,\boldsymbol{c}})|X_i = S, \boldsymbol{c}]$. Following Park et al. (2023a,b), the summation is written for notational simplicity. With continuous P and mixed $\boldsymbol{L}$, it should be understood as integration with respect to the relevant conditional distributions. The imputation-based estimator described in Section 3.6 of the main text operates on the empirical distribution.

B. Common Support of Academic Performance

The identification of primary and secondary effects requires the positivity condition C2 (Section 3.4 of the main text): all combinations of class and performance levels relevant to the counterfactual integrand must have non-zero conditional probability. This section reports an empirical diagnostic of the overlap of academic performance distributions across the three classes.

Table A1 reports descriptive statistics of the standardized performance variable by class. The empirical minimum and maximum coincide at -2.2 and 1.49 for all three classes, indicating that the Likert-scale-based composite is bounded in the same range across classes. Inter-class differences appear in the central tendency: medians are 0.29, -0.05, and -0.23 for the service, intermediate, and working classes respectively, corresponding to roughly half a standard deviation between the top and bottom classes.

Table A1: Empirical support of academic performance by class

Class	n	Min	5th pct	25th pct	Median	75th pct	95th pct	Max
Service (S)	655	−2.20	−1.60	−0.36	0.29	1.03	1.49	1.49
Intermediate (I)	496	−2.20	−1.81	−0.64	−0.05	0.70	1.49	1.49
Working (W)	487	−2.20	−2.05	−0.81	−0.23	0.55	1.49	1.49

The shared support, defined as the intersection of the [5th, 95th] percentile ranges across the three classes, is approximately $[-1.60, 1.49]$. Within this shared range, 95.0%, 93.5%, and 90.8% of the service, intermediate, and working classes respectively are observed. Figure A1 displays the kernel density estimates of academic performance by class, both overall and stratified by gender. The three distributions are unimodal and broadly overlap across the entire support, consistent with the positivity condition C2.

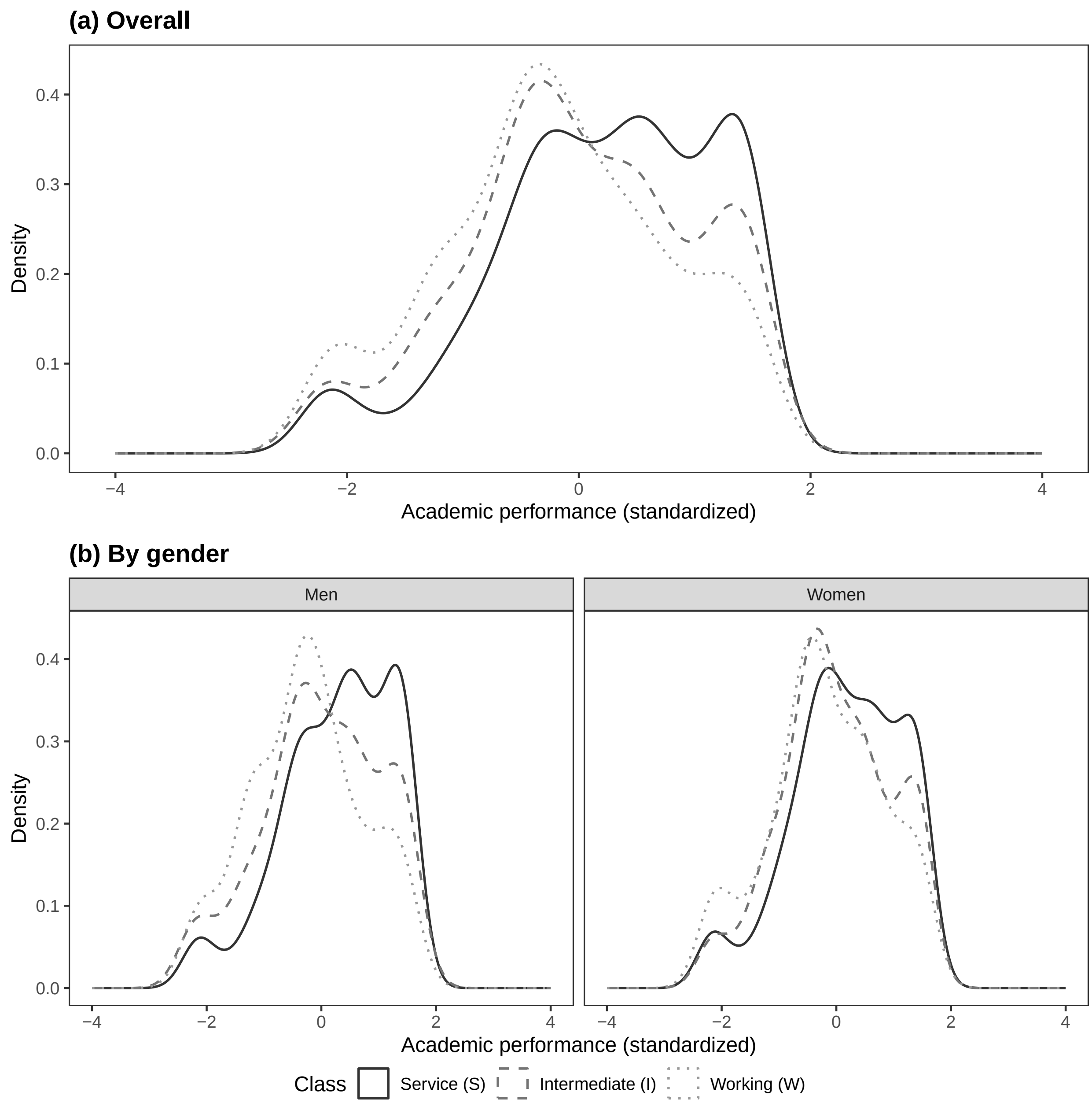


Figure A1: Empirical distributions of academic performance by class

Note: Kernel density estimates of standardized academic performance by class. Solid line: Service class (S); dashed line: Intermediate class (I); dotted line: Working class (W).

Alt text: A two-panel kernel-density figure of standardized academic performance. Panel (a) shows the overall distributions for the Service (solid line), Intermediate (dashed line), and Working (dotted line) classes. Panel (b) repeats the plot stratified by gender. All distributions are unimodal and overlap substantially across the full range, with central locations shifting rightward for higher classes.

C. Full Decomposition Results

This section reports the numerical values underlying Figure 5 of the main text. Table A2 reports the full decomposition for the four CD models alongside Erikson et al.'s and Morgan's methods, with the main outcome (Y = entry into a 4-year university or higher).

Table A2: Decomposition of IEO into Primary and Secondary Effects

Comparison	Performance	IEO	PE	SE	PE (%)	SE (%)
(a) Erikson et al. (2005)						
$S-I$	S	0.092	0.044	0.048	47.396	52.604
$S-I$	I	0.092	0.048	0.044	52.153	47.847
$S-W$	S	0.286	0.087	0.200	30.252	69.748
$S-W$	W	0.286	0.087	0.199	30.392	69.608
$I-W$	I	0.194	0.038	0.156	19.598	80.402
$I-W$	W	0.194	0.035	0.159	18.127	81.873
(b) Morgan (2012)						
$S-I$	S	0.092	0.044	0.048	48.082	51.918
$S-I$	I	0.092	0.048	0.044	52.015	47.985
$S-W$	S	0.285	0.090	0.195	31.709	68.291
$S-W$	W	0.285	0.086	0.199	30.075	69.925
$I-W$	I	0.194	0.039	0.154	20.313	79.687
$I-W$	W	0.194	0.035	0.158	18.187	81.813
(c1) CD Model 1						
$S-I$	S	0.092	0.035	0.057	38.451	61.549
$S-I$	I	0.092	0.054	0.038	58.760	41.240
$S-W$	S	0.285	0.081	0.204	28.545	71.455
$S-W$	W	0.285	0.093	0.193	32.441	67.559
$I-W$	I	0.193	0.033	0.160	17.224	82.776
$I-W$	W	0.193	0.030	0.163	15.477	84.523
(c2) CD Model 2						
$S-I$	S	0.092	0.032	0.061	34.315	65.685
$S-I$	I	0.092	0.037	0.055	40.203	59.797
$S-W$	S	0.285	0.068	0.217	23.878	76.122
$S-W$	W	0.285	0.069	0.217	24.071	75.929
$I-W$	I	0.193	0.025	0.168	12.933	87.067
$I-W$	W	0.193	0.018	0.175	9.544	90.456
(c3) CD Model 3						
$S-I$	S	0.092	0.031	0.061	33.659	66.341
$S-I$	I	0.092	0.035	0.057	38.441	61.559
$S-W$	S	0.285	0.066	0.219	23.113	76.887
$S-W$	W	0.285	0.066	0.219	23.260	76.740
$I-W$	I	0.193	0.024	0.170	12.249	87.751
$I-W$	W	0.193	0.020	0.174	10.201	89.799
(c4) CD Model 4						
$S-I$	S	0.092	0.017	0.075	18.811	81.189
$S-I$	I	0.092	0.020	0.072	21.731	78.269
$S-W$	S	0.285	0.034	0.252	11.814	88.186
$S-W$	W	0.285	0.041	0.244	14.390	85.610
$I-W$	I	0.193	0.010	0.184	4.927	95.073
$I-W$	W	0.193	0.008	0.186	3.901	96.099

Note: IEO = PE + SE. Performance column indicates which class's performance distribution is applied. PE (%) and SE (%) show proportions of IEO. (a) Erikson et al.'s method uses parameters from Table 2. (b) Morgan's method uses Model 3 in Table 3. (c1)–(c4) are causal decomposition models: CD Model 1 controls for gender only; CD Model 2 adds L_1 and economic-resource components of L_2 (income, savings, subjective wealth), treating family investments in performance (cram school, private tutoring, study time, private/national school sector) as part of the primary-effect mechanism; CD Model 3 additionally treats those investment variables as confounders of the P–Y relationship; CD Model 4 further adds educational aspirations as anticipatory-decision proxies. Values are pooled from 20 multiply imputed datasets.

Table A3 adds 95% confidence intervals for the causal decomposition models. Within each imputed dataset, the sampling variance of each component is obtained from the non-parametric bootstrap implemented in the *causal.decomp* package (1,000 bootstrap samples). The within-imputation variance is then combined with the between-imputation variance across the 20 imputations using Rubin's rules with Barnard–Rubin degrees of freedom. The intervals for the primary effect exclude zero under all four specifications for the S–W and S–I comparisons. For the I–W comparison, whose primary effect is smallest, the intervals include zero once performance–attainment confounders are adjusted (CD Models 2–4). Under the adjusted specifications, the allocation of the I–W gap to the performance-mediated pathway is therefore not statistically distinguishable from zero.

Table A3: Decomposition of IEO with 95% confidence intervals (causal decomposition models)

Comparison	Performance	*IEO*	*PE*	*SE*
(c1) CD Model 1				
$S - I$	S	0.092 [0.032, 0.152]	0.035 [0.009, 0.062]	0.057 [0.003, 0.111]
$S - I$	I	0.092 [0.032, 0.152]	0.054 [0.026, 0.083]	0.038 [-0.020, 0.096]
$S - W$	S	0.285 [0.230, 0.341]	0.081 [0.049, 0.114]	0.204 [0.147, 0.261]
$S - W$	W	0.285 [0.230, 0.340]	0.093 [0.062, 0.124]	0.193 [0.135, 0.251]
$I - W$	I	0.193 [0.123, 0.264]	0.033 [0.003, 0.064]	0.160 [0.091, 0.229]
$I - W$	W	0.193 [0.122, 0.265]	0.030 [0.001, 0.059]	0.163 [0.094, 0.233]
(c2) CD Model 2				
$S - I$	S	0.092 [0.033, 0.152]	0.032 [0.010, 0.054]	0.061 [0.005, 0.116]
$S - I$	I	0.092 [0.032, 0.152]	0.037 [0.013, 0.061]	0.055 [-0.003, 0.114]
$S - W$	S	0.285 [0.230, 0.341]	0.068 [0.037, 0.099]	0.217 [0.159, 0.275]
$S - W$	W	0.285 [0.231, 0.340]	0.069 [0.040, 0.097]	0.217 [0.159, 0.274]
$I - W$	I	0.193 [0.122, 0.265]	0.025 [-0.003, 0.053]	0.168 [0.099, 0.238]
$I - W$	W	0.193 [0.122, 0.265]	0.018 [-0.005, 0.042]	0.175 [0.104, 0.246]
(c3) CD Model 3				
$S - I$	S	0.092 [0.032, 0.152]	0.031 [0.009, 0.053]	0.061 [0.004, 0.118]
$S - I$	I	0.092 [0.032, 0.152]	0.035 [0.011, 0.060]	0.057 [-0.002, 0.115]
$S - W$	S	0.285 [0.230, 0.341]	0.066 [0.036, 0.096]	0.219 [0.161, 0.278]
$S - W$	W	0.285 [0.230, 0.341]	0.066 [0.037, 0.095]	0.219 [0.160, 0.278]
$I - W$	I	0.193 [0.123, 0.264]	0.024 [-0.004, 0.051]	0.170 [0.100, 0.240]
$I - W$	W	0.193 [0.122, 0.264]	0.020 [-0.004, 0.044]	0.174 [0.103, 0.244]
(c4) CD Model 4				
$S - I$	S	0.092 [0.032, 0.152]	0.017 [0.001, 0.034]	0.075 [0.015, 0.134]
$S - I$	I	0.092 [0.032, 0.152]	0.020 [0.001, 0.039]	0.072 [0.013, 0.132]
$S - W$	S	0.285 [0.230, 0.341]	0.034 [0.009, 0.058]	0.252 [0.192, 0.311]
$S - W$	W	0.285 [0.230, 0.341]	0.041 [0.016, 0.066]	0.244 [0.185, 0.304]
$I - W$	I	0.193 [0.122, 0.265]	0.010 [-0.009, 0.028]	0.184 [0.112, 0.255]
$I - W$	W	0.193 [0.122, 0.265]	0.008 [-0.007, 0.022]	0.186 [0.114, 0.258]

Note: IEO = PE + SE. The Performance column indicates which class's performance distribution is applied. Point estimates are pooled across 20 multiply imputed datasets and correspond to those in Table A2. The 95% confidence intervals in brackets combine the nonparametric-bootstrap sampling variance within each imputed dataset (1,000 bootstrap samples, as implemented in the *causal.decomp* package) with the between-imputation variance, using Rubin's rules with Barnard–Rubin degrees of freedom. Model specifications are as in Table A2.

D. Robustness: Entry into Higher Education

This section reports a robustness analysis in which the outcome is redefined as entry into any form of higher education, including specialized training schools (*senshu gakkō*), junior colleges (*tanki daigaku*), technical colleges (*kōsen*), 4-year universities, and graduate schools. The motivation is that Boudon's secondary effects concern educational continuation decisions broadly, not specifically the choice of 4-year university. Redefining Y in this broader way provides a robustness check on the main results. Table A4 reports the full decomposition with this alternative outcome, and Figure A2 presents it in the same format as Figure 5 of the main text, so the robustness of the pattern can be assessed by direct comparison.

Table A4: Decomposition of IEO into Primary and Secondary Effects (higher education as outcome)

Comparison	Performance	*IEO*	*PE*	*SE*	*PE* (%)	*SE* (%)
(a) Erikson et al. (2005)						
$S-I$	S	0.059	0.029	0.031	48.286	51.714
$S-I$	I	0.059	0.027	0.032	45.083	54.917
$S-W$	S	0.156	0.057	0.099	36.387	63.613
$S-W$	W	0.156	0.047	0.109	29.936	70.064
$I-W$	I	0.097	0.024	0.073	25.084	74.916
$I-W$	W	0.097	0.022	0.075	22.264	77.736
(b) Morgan (2012)						
$S-I$	S	0.055	0.027	0.028	49.854	50.146
$S-I$	I	0.055	0.026	0.029	46.630	53.370
$S-W$	S	0.147	0.056	0.091	38.055	61.945
$S-W$	W	0.147	0.045	0.102	30.717	69.283
$I-W$	I	0.092	0.024	0.068	26.042	73.958
$I-W$	W	0.092	0.021	0.072	22.561	77.439
(c1) CD Model 1						
$S-I$	S	0.055	0.025	0.030	46.043	53.957
$S-I$	I	0.055	0.026	0.029	47.263	52.737
$S-W$	S	0.147	0.054	0.093	37.009	62.991
$S-W$	W	0.147	0.046	0.101	31.368	68.632
$I-W$	I	0.092	0.025	0.068	26.750	73.250
$I-W$	W	0.092	0.015	0.078	15.748	84.252
(c2) CD Model 2						
$S-I$	S	0.055	0.026	0.029	47.191	52.809
$S-I$	I	0.055	0.011	0.044	20.558	79.442
$S-W$	S	0.147	0.051	0.096	34.494	65.506
$S-W$	W	0.147	0.026	0.121	17.864	82.136
$I-W$	I	0.092	0.026	0.067	27.891	72.109
$I-W$	W	0.092	0.000	0.092	0.393	99.607
(c3) CD Model 3						
$S-I$	S	0.055	0.025	0.030	45.292	54.708
$S-I$	I	0.055	0.007	0.048	12.591	87.409
$S-W$	S	0.147	0.049	0.098	33.162	66.838
$S-W$	W	0.147	0.020	0.127	13.798	86.202
$I-W$	I	0.092	0.026	0.067	27.757	72.243
$I-W$	W	0.092	0.001	0.094	−1.365	101.365
(c4) CD Model 4						
$S-I$	S	0.055	0.019	0.036	33.850	66.150
$S-I$	I	0.055	0.002	0.056	−2.993	102.993
$S-W$	S	0.147	0.036	0.111	24.404	75.596
$S-W$	W	0.147	0.007	0.140	4.516	95.484
$I-W$	I	0.092	0.022	0.071	23.538	76.462
$I-W$	W	0.092	0.005	0.097	−5.559	105.559

Note: Y = entry into any form of higher education (vocational school, junior college, technical college, university, or graduate school). IEO = PE + SE. Performance column indicates which class's performance distribution is applied. PE (%) and SE (%) show proportions of IEO. Model specifications correspond to those in Table A2. Values are pooled from 20 multiply imputed datasets.

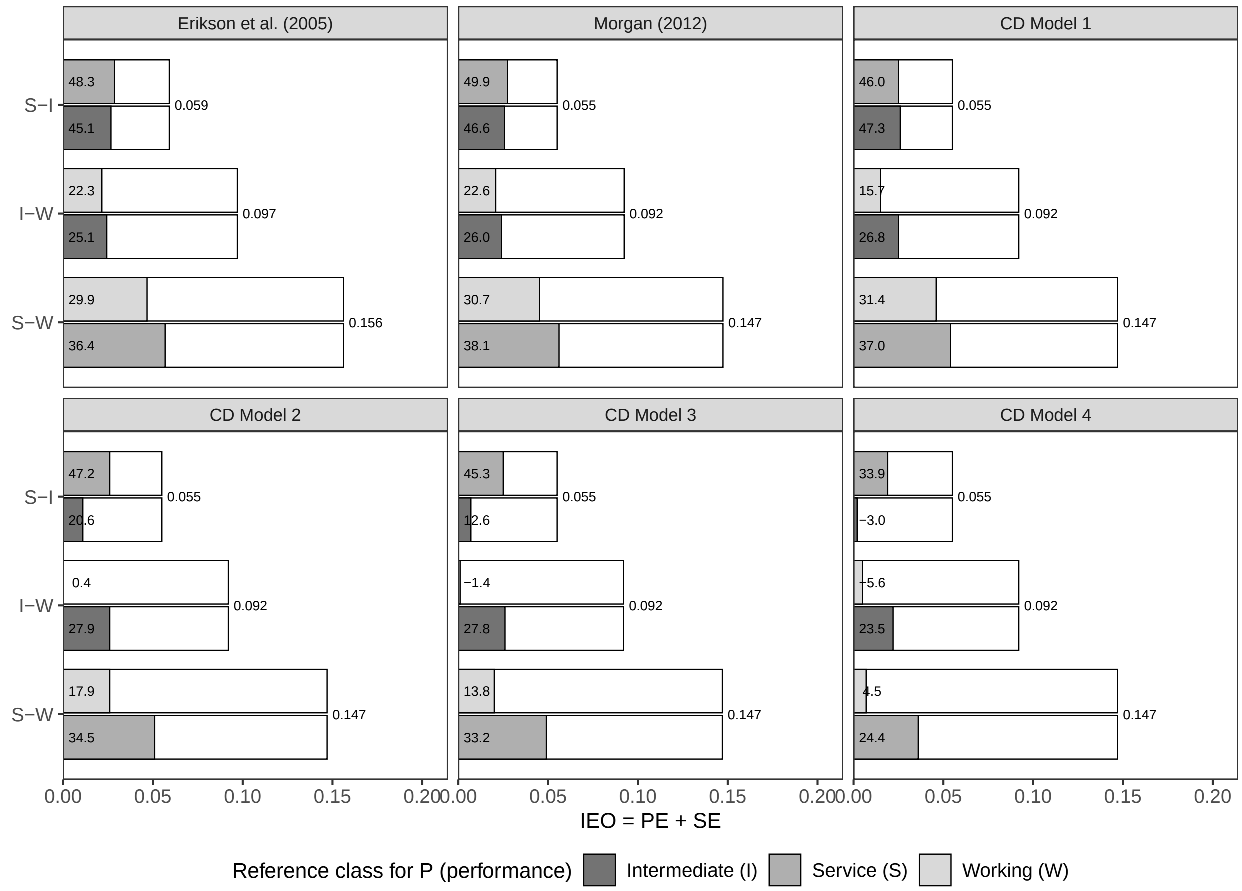


Figure A2: Decomposition of Inequalities in Educational Opportunity (IEO) into Primary Effects (PE) and Secondary Effects (SE) for Entry into Higher Education: Results from Traditional Methods (Erikson and Morgan) and Causal Decomposition Models with Different Sets of Controls

Note: This figure parallels Figure 5 of the main text, with the outcome redefined as entry into any form of higher education (vocational school, junior college, technical college, 4-year university, or graduate school). Bars represent total inequality (IEO) between social classes; the label on each group of bars indicates the comparison ('S-W' = Service vs. Working, 'I-W' = Intermediate vs. Working, 'S-I' = Service vs. Intermediate). Gray portions indicate primary effects (PE), the contribution of performance differences; numbers on the gray portions give the PE percentage, where S, I, and W denote the class whose performance distribution is applied. Specifications match those in Table A2 and Table A4. As with the main 4-year-university outcome, the primary-effect share is similar for the conventional methods and CD Model 1 and declines through CD Model 4 while the total length of each bar (total IEO) stays constant, confirming that the qualitative finding is not specific to the 4-year-university definition.
Alt text: A three-by-two grid of horizontal bar charts, one panel per method (Erikson, Morgan, and CD Model 1 through CD Model 4). Within each panel, bars give the total inequality (IEO) for the comparisons S–W, I–W, and S–I, with the gray portion the primary-effect (performance-mediated) share and the remainder the secondary effect. The primary-effect share is similar for the conventional methods and CD Model 1 and decreases through CD Model 4, while the total bar length stays roughly constant.

A further feature of these results is that the primary-effect share is more sensitive to the reference direction (the class whose performance distribution is imposed) than for the main outcome, particularly under the most heavily adjusted models. This is the substantive direction-dependence discussed in the main text following Morgan (2012), amplified here by the smaller disparities of this broader outcome: the S–I disparity is only 0.055, so the share is a ratio with a small denominator that becomes unstable, and occasionally slightly negative, once the performance-mediated component approaches zero. It is therefore not a sign of estimation error, and the qualitative pattern, a primary-effect share that declines as performance–attainment confounders are added, is unchanged.

E. Sensitivity Analysis for Unmeasured Performance–Attainment Confounding

The validity of the disparity decomposition rests on Condition C1, namely that there are no unmeasured confounders between academic performance (P) and educational attainment (Y) conditional on the observed covariates. To assess robustness to potential violations of C1, I conducted a sensitivity analysis following Park et al. (2023b), which extends the framework of Park et al. (2022).

The analysis examines how an unmeasured confounder U associated with both P and Y would affect the estimates. The strength of U is characterized by two partial R^2 quantities: $R^2_{P,U}$, the share of variance in P explained by U after conditioning on observed covariates, and $R^2_{Y,U}$, the corresponding share of variance in Y. The "critical value" is the combination of these parameters at which the primary effect becomes zero.

Figure A3 displays the sensitivity surfaces for the working-class versus service-class comparison under CD Model 3. Panel A shows how the primary effect (disparity reduction) changes under different confounding scenarios. Panel B shows the corresponding changes in the secondary effect (disparity remaining). The contour lines indicate effect sizes at different combinations of $R^2_{P,U}$ and $R^2_{Y,U}$.

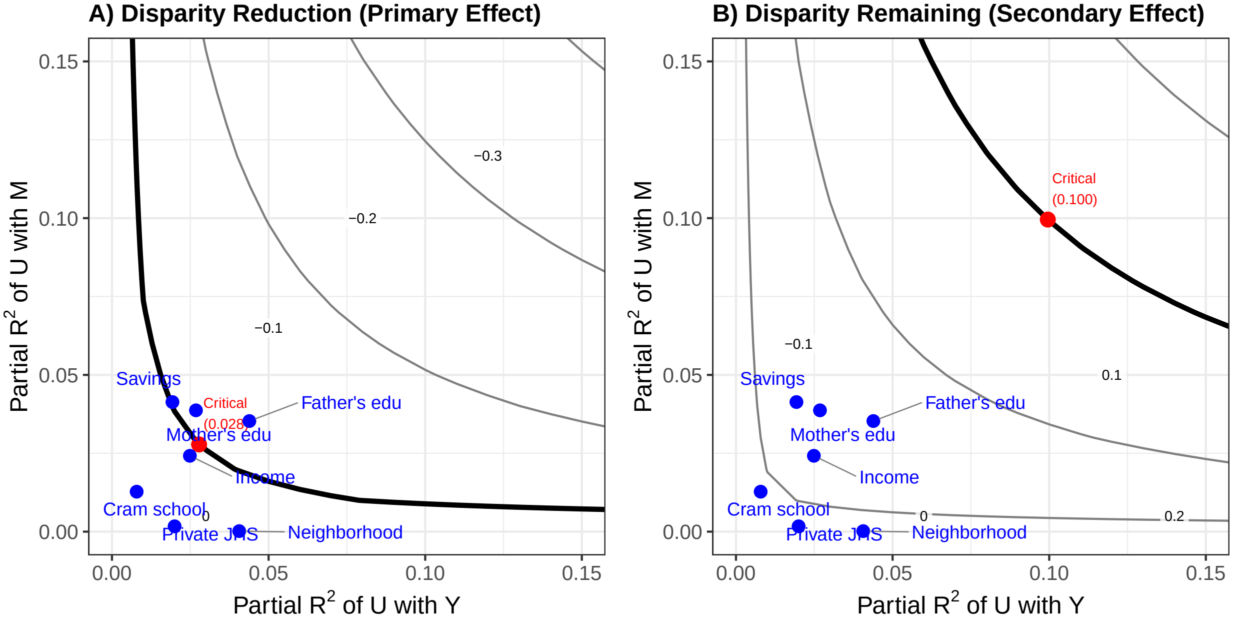


Figure A3: Sensitivity Analysis: Robustness of Primary and Secondary Effects to Unmeasured Confounding (Working Class vs Service Class Comparison)

Note: Panel A shows how the primary effect (disparity reduction) changes under different levels of unmeasured confounding between academic performance (P) and university enrollment (Y). Panel B shows the corresponding changes in the secondary effect (disparity remaining). The x-axis represents the partial R^2 of an unmeasured confounder U with university enrollment, and the y-axis represents the partial R^2 of U with academic performance. Contour lines indicate effect sizes at different confounding levels, with the bold line indicating zero effect. The red point marks the critical R^2 values (approximately 0.028, 0.028), where the primary effect would equal zero. Blue points represent observed covariates as benchmarks: father's education, mother's education, neighborhood characteristics, family income, savings, cram school attendance, and private junior high school attendance. All observed covariates fall below the critical threshold, suggesting that an unmeasured confounder would need to be moderately stronger than any single observed covariate to eliminate the primary effect. Based on the method proposed by Park et al. (2023a).

Alt text: A two-panel contour plot. Both panels share the same axes: the horizontal axis is the partial R^2 of an unmeasured confounder U with university enrollment Y, and the vertical axis is the partial R^2 of U with academic performance P. Panel A shows contours of the primary effect and Panel B the contours of the secondary effect under varying confounding strengths, with a bold contour at zero effect. A red point marks the critical values (approximately 0.028, 0.028) at which the primary effect reaches zero, and blue points mark observed covariates (parental education, neighborhood characteristics, family income, savings, cram school attendance, and private junior high attendance), all of which fall below the critical threshold.

For the W versus S comparison under CD Model 3, the critical R^2 values are approximately $(0.028, 0.028)$. An unmeasured confounder would therefore need to explain about 2.8% of the variance in both P and Y after controlling for the observed covariates to fully eliminate the estimated primary effect. As a benchmark, the partial R^2 values for prominent observed covariates are: father's education $(0.012, 0.008)$, mother's education $(0.008, 0.005)$, family income $(0.015, 0.011)$, and cram school attendance $(0.021, 0.018)$. These observed confounders all fall below the critical threshold, suggesting that an unmeasured confounder would need to be stronger than any single observed covariate to reduce the estimated PE to zero.

A structural property of this decomposition is worth noting. Because IEO is an observed quantity that requires no causal assumptions, unmeasured confounding between P and Y does not bias the total disparity but reallocates it between PE and SE: any upward bias in PE produces an equal downward bias in SE. The causal identification challenge therefore lies entirely in the allocation of IEO between the two components, not in the overall magnitude of inequality. Two sources of bias also operate in opposite directions on PE estimates. Omitting P–Y confounders inflates PE upward when the confounding is positive, as is plausible here, while measurement error in P attenuates the mediated pathway and deflates PE downward. In the absence of measurement error, PE estimates from conventional methods may therefore be read as an upper bound for the contribution of academic performance, and accounting for such confounding reduces the estimated share of PE.

F. Numerical Details for the Conventional Method

This section provides numerical details for the conventional decomposition methods discussed in Section 5.1 of the main text. Performance is highest in the service class ($\hat{\mu}_S = 0.210$), followed by the intermediate class ($\hat{\mu}_I = -0.038$) and the working class ($\hat{\mu}_W = -0.234$), with little difference in standard deviations across classes ($\hat{\sigma}_X \in [0.967, 0.994]$).

The estimated logistic regression slopes $\hat{\beta}_X$ correspond to implied average marginal effects of 0.179, 0.175, and 0.192 for the service, intermediate, and working classes respectively (computed from the class-specific parameters in Table 2 under the within-class normality assumption), indicating a broadly comparable performance–enrollment relationship across classes. The ratio $-\hat{\alpha}_X/\hat{\beta}_X$ gives the performance score at which the predicted enrollment probability equals 0.5: -0.856 for the service class, -0.725 for the intermediate class, and 0.119 for the working class, indicating that service-class students enroll even at low performance whereas working-class students enroll only at high performance.

Table A5 reports observed and counterfactual transition probabilities. The diagonal entries correspond to the observed enrollment probabilities (0.718 for S, 0.626 for I, 0.431 for W in panel (a)). Off-diagonal entries give what-if probabilities: the entry in row X and column X' gives the enrollment probability when the performance distribution corresponds to class X but the choice pattern corresponds to class X'. For example, equalizing the performance distribution of W to that of S while keeping W's choice pattern gives a counterfactual probability of 0.518, so the primary effect is $0.518 - 0.431 = 0.087$ and the secondary effect is $0.718 - 0.518 = 0.200$, with the IEO decomposed as $(0.718 - 0.431) = (0.518 - 0.431) + (0.718 - 0.518) = PE + SE$.

Table A5: Estimated probabilities of university enrollment

Class	S	I	W
(a) Probabilities from Table 2			
Service (S)	0.718	0.669	0.518
Intermediate (I)	0.670	0.626	0.469
Working (W)	0.631	0.590	0.431
(b) Probabilities from Model 3 in Table 3			
Service (S)	0.718	0.671	0.523
Intermediate (I)	0.670	0.627	0.472
Working (W)	0.632	0.591	0.433

Note: Rows represent the performance distribution of the specified class; columns represent the educational choice pattern of the specified class. Diagonal entries show observed probabilities. Panel (a) uses estimated parameters from Table 2; panel (b) uses predicted probabilities from Model 3 in Table 3. Values are pooled estimates from 20 multiply imputed datasets.

For the largest disparity (S–W, IEO = 0.286), PE equals 0.087 both when aligning to S and when aligning to W, with PE shares of approximately 30% in either direction. For I–W (IEO = 0.194), PE values are 0.038 and 0.035 (shares of 20% and 18%). For S–I (IEO = 0.092), PE values are 0.044 and 0.048 (shares of 47% and 52%). Across all pairs, the PE share ranges from about 18% to 52%, and the SE share from about 48% up to 82% of the disparity. Shirakawa (2017) reports broadly similar shares (60–70% for the secondary effect) using PISA 2003 data for Japan.

G. Morgan's Three Challenges and How the Present Approach Responds

Morgan (2012) identifies three specific challenges to the standard primary/secondary effects decomposition. This section restates each challenge and the response of the disparity decomposition framework adopted in this article.

First challenge: Weak causal warrant. Morgan argues that the net coefficient of class in a logit model controlling for performance lacks a strong causal warrant for interpretation as a "choice-based" secondary effect, because backdoor paths from class to attainment through race, neighborhood characteristics, and institutional features cannot be purged with conventional conditioning. The conventional method therefore conflates choice-based secondary effects with backdoor confounding. The present approach addresses this by redefining IEO as a *disparity* (an observed associational difference) rather than as a causal effect of class. The intervention to be evaluated is on academic performance, not on class. The identification problem for the causal effect of class itself is therefore sidestepped, while causal identification is retained for the counterfactual intervention on P.

Second challenge: Reliance on log-odds-based measures. Morgan criticises the convention of summarising PE and SE through log-odds-based relative-importance measures that average over the direction of counterfactual movement. Such measures obscure substantive heterogeneity because the same nominal contrast produces different decompositions depending on the direction of equalisation. The present approach uses probability-based effects throughout, consistent with Morgan's recommendation.

Third challenge: Inattention to unit-level heterogeneity. Morgan emphasises that PE and SE estimates are class-specific averages weighted by the performance distribution of the class from which counterfactual movement originates, and that primary effects themselves generate class differences in these distributions. Direction-specific estimates therefore differ in expectation, and averaging across directions discards substantive information about the unit-level heterogeneity of class effects. The present approach reports direction-specific estimates ($PE_{\bar{S}W}$ and $PE_{S\bar{W}}$) for every pairwise comparison and interprets the asymmetries between them as substantive properties of the decomposition rather than as estimation noise.

H. Exposure-Induced Confounders in the Present Setting

The vector L_2 (previously denoted as U in Erikson et al. (2005)'s notation) includes variables that are influenced by social class and that in turn affect both academic performance and educational attainment. Examples in the present data include private middle school attendance, cram school participation, and long hours of extra-curricular study. Such "exposure-induced" confounders raise serious identification difficulties for causal mediation analysis, because conditioning on the mediator P when an unmeasured exposure-induced confounder is present generates collider bias, while not conditioning leaves the mediator-outcome confounding unresolved (Wodtke and Zhou 2020). Within the disparity decomposition framework adopted here, by contrast, L_2 enters the identification formulas as part of the joint distribution that is held fixed in the counterfactual integrand. Its presence is accommodated without requiring identification of the causal effect of class on attainment. The substantive question of whether to treat a given L_2 variable as a confounder of the P–Y relationship or as part of the performance-mediated mechanism is a theoretical choice, and the four CD models in the main text encompass both possibilities for the central candidate variables (cram school, private tutoring, study time, and school sector).

I. Robustness to Outcome-Model Specification

The central empirical claim of this article is that conventional methods overstate the performance-mediated channel: adjusting for performance–attainment confounders reduces the estimated primary-effect share relative to the conventional, gender-only specification. To assess whether this reduction depends on the parametric logistic outcome model, we re-estimated the decomposition for all three class comparisons (each equalizing the lower class to the higher class's performance distribution), replacing the logistic outcome regression with a range of statistical learners, and we compared the primary-effect share without confounder adjustment (CD Model 1, gender only) to that under the full adjustment set (CD Model 3). The mediator model and counterfactual intervention are held fixed. Only the outcome regression varies.

Table A6 reports the results. Across the three comparisons and ten learners, the primary-effect share falls when confounders are adjusted in 28 of the 30 cases. The only exceptions are gradient boosting for the S–W and I–W gaps, whose unadjusted estimates are already among the lowest and therefore leave little to reduce. For the headline S–W gap, the better-behaved learners give a reduction of about 4 to 5 percentage points and an adjusted level near 23 to 26% (the high-variance single regression tree gives inflated levels but still shows the reduction). The I–W gap behaves similarly, with reductions of about 3 percentage points. The S–I gap has a small observed disparity (IEO = 0.092), so its share is a ratio with a small denominator and is correspondingly sensitive. Here the smooth, well-specified learners (logistic regression, regression splines, the generalized additive model, Bayesian additive regression trees, and the neural network) agree on a reduction of 6 to 8 percentage points, whereas the tree-based and penalized learners produce unstable levels. Random forests in particular cannot extrapolate the performance effect and drive the adjusted share toward zero. The direction of the reduction is nonetheless consistent across comparisons and learners. We therefore emphasize the direction and approximate magnitude of the reduction rather than the precise level, and defer a fuller treatment of efficient and doubly robust estimation, cross-fitting, and the functional-form sensitivity of the adjusted level to separate methodological work.

Table A6: Primary-effect share (%) under alternative outcome models, by class comparison

Outcome regression	CD Model 1 (gender only)	CD Model 3 (full adjustment)	Reduction (pp)
(a) Service vs. working (S–W gap, performance equalized to S; IEO = 0.285)			
Logistic (baseline)	27.6	23.0	−4.6
Penalized logistic (lasso)	30.3	25.9	−4.4
Natural splines in P	28.0	23.5	−4.5
Generalized additive model	28.2	23.3	−4.9
Regression tree	47.5	40.9	−6.6
Random forest	29.7	25.7	−4.0
Gradient boosting	24.4	25.8	+1.3
Bayesian additive regression trees	27.1	23.5	−3.6
Neural network	28.0	22.8	−5.3
Super Learner (ensemble)	30.3	26.0	−4.2
(b) Service vs. intermediate (S–I gap, performance equalized to S; IEO = 0.092)			
Logistic (baseline)	41.3	34.9	−6.4
Penalized logistic (lasso)	49.7	28.8	−20.8
Natural splines in P	38.7	33.0	−5.7
Generalized additive model	44.8	38.3	−6.5
Regression tree	39.2	11.4	−27.8
Random forest	25.5	0.0	−25.4
Gradient boosting	49.7	35.7	−14.0
Bayesian additive regression trees	38.2	32.1	−6.1
Neural network	42.8	34.4	−8.4
Super Learner (ensemble)	45.4	27.9	−17.5
(c) Intermediate vs. working (I–W gap, performance equalized to I; IEO = 0.193)			
Logistic (baseline)	15.5	12.1	−3.4
Penalized logistic (lasso)	19.6	16.6	−3.0
Natural splines in P	15.6	12.4	−3.1
Generalized additive model	15.7	12.2	−3.4
Regression tree	46.6	40.3	−6.4
Random forest	29.4	22.4	−7.0
Gradient boosting	11.0	15.2	+4.2
Bayesian additive regression trees	16.5	14.8	−1.6
Neural network	15.9	12.3	−3.7
Super Learner (ensemble)	19.9	17.0	−2.8

Note: Primary-effect share (the performance-mediated component as a percentage of the observed disparity, IEO) for each class comparison, equalizing the lower class to the higher class's performance distribution. CD Model 1 conditions on gender only; CD Model 3 adds the full confounder set (L_1–L_{10} and LL_1–LL_{11}). Estimates use the imputation-based g-formula with the outcome regression replaced by each learner, pooled across 20 multiply imputed datasets (Amelia; analytic $n = 1{,}854$), with stochastic learners using a single draw per imputation. Because the S–I gap has a small disparity (IEO = 0.092), its share is a ratio with a small denominator and is correspondingly sensitive. The high-variance tree-based learners (single regression tree and random forest) extrapolate the performance effect poorly and give unstable adjusted levels.

References

Boudon, Raymond. 1974. *Education, Opportunity, and Social Inequality*. Wiley.

Brand, Jennie E., Ravaris Moore, Xi Song, and Yu Xie. 2019. "Why Does Parental Divorce Lower Children's Educational Attainment? A Causal Mediation Analysis." *Sociological Science* 6:264–292.

Erikson, Robert, John H. Goldthorpe, Michelle Jackson, Meir Yaish, and David R. Cox. 2005. "On Class Differentials in Educational Attainment." *Proceedings of the National Academy of Sciences* 102:9730–9733.

Hong, Guanglei. 2015. *Causality in a Social World: Moderation, Mediation and Spill-over*. Chichester, UK: John Wiley & Sons, Ltd.

Jackson, John W. and Tyler J. VanderWeele. 2018. "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities:." *Epidemiology* 29:825–835.

Klein, Markus and Michael Kühhirt. 2021. "Direct and Indirect Effects of Grandparent Education on Grandchildren's Cognitive Development: The Role of Parental Cognitive Ability." *Sociological Science* 8:265–284.

Morgan, Stephen L. 2012. "Models of College Entry in the United States and the Challenges of Estimating Primary and Secondary Effects." *Sociological Methods & Research* 41:17–56.

Park, Soojin, Suyeon Kang, and Chioun Lee. 2023a. "Choosing an Optimal Method for Causal Decomposition Analysis with Continuous Outcomes: A Review and Simulation Study." *Sociological Methodology* p. 00811750231183711.

Park, Soojin, Suyeon Kang, Chioun Lee, and Shujie Ma. 2023b. "Sensitivity Analysis for Causal Decomposition Analysis: Assessing Robustness toward Omitted Variable Bias." *Journal of Causal Inference* 11.

Park, Soojin, Xu Qin, and Chioun Lee. 2022. "Estimation and Sensitivity Analysis for Causal Decomposition in Health Disparity Research." *Sociological Methods & Research* p. 00491241211067516.

Robins, James. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7:1393–1512.

Shi, Baoyi, Christine Choirat, Brent A. Coull, Tyler J. VanderWeele, and Linda Valeri. 2021. "CMAverse: A Suite of Functions for Reproducible Causal Mediation Analyses." *Epidemiology* 32:e20.

Shirakawa, Toshiyuki. 2017. "Socioeconomic Differences In Educational Expectation: Primary and Secondary Effects in International Context." *Kansai Sociological Review* 16:72–84.

VanderWeele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.

VanderWeele, Tyler J., Stijn Vansteelandt, and James M. Robins. 2014. "Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder:." *Epidemiology* 25:300–306.

Wodtke, Geoffrey T. and Xiang Zhou. 2020. "Effect Decomposition in the Presence of Treatment-induced Confounding: A Regression-with-residuals Approach." *Epidemiology* 31:369–375.

Zhou, Xiang and Geoffrey T. Wodtke. 2019. "A Regression-with-Residuals Method for Estimating Controlled Direct Effects." *Political Analysis* 27:360–369.