

Causal Effect of Multicultural Background on Academic Achievement: A Double Machine Learning Approach*

Won Gyoung Kim[†] Pilsun Choi[‡]

Abstract Despite sustained government efforts to support multicultural students in Korea, little empirical evidence exists on whether a multicultural background actually causes lower academic achievement. This paper applies the Double/Debiased Machine Learning (DML) framework of Chernozhukov *et al.* (2018) to three panel cohorts from the Korean Children and Youth Panel Survey 2010. Multiple machine learning methods—including Lasso, Ridge, Random Forest, Boosted Trees, and Neural Networks—flexibly control for socioeconomic and educational covariates, while Neyman-orthogonal score functions and cross-fitting ensure robustness to regularization bias and functional form misspecification. Our principal finding is that, after controlling for observed socioeconomic and educational covariates, there is no statistically robust evidence that multicultural background lowers academic achievement independently of these factors. For the Grade 1 and Grade 4 cohorts, the raw achievement gap is no longer statistically distinguishable from zero once covariates are accounted for. For the Grade 7 cohort, some specifications yield marginally significant negative effects, but these are not robust across learners or ensemble methods. These results suggest that the observed gap is driven by socioeconomic factors rather than multicultural status itself, calling for a reassessment of policies that target academic support exclusively at multicultural students.

Keywords Double Machine Learning, causal Inference, multicultural students, academic achievement, partially linear regression.

JEL Classification I21, I24, C23, C55.

*This paper was written as part of Konkuk University's research support program for its faculty on sabbatical leave in 2026.

[†]College of Education, Texas A&M International University, 5201 University Boulevard, Laredo, Texas, USA 78041. E-mail: won.kim@tamiu.edu.

[‡]Corresponding author, Department of International Trade, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea 05029. E-mail: pchoi@konkuk.ac.kr.

1. INTRODUCTION

The rapid growth of multicultural families in Korea has prompted the government to implement a wide range of support policies for students from multicultural backgrounds. According to the 2025 Basic Education Statistics (Ministry of Education and KEDI, 2025), the number of multicultural students (from international marriage and foreign worker families) enrolled in Korean primary and secondary schools reached approximately 202,000, accounting for 4.0% of all students—a roughly 4.3-fold increase from 47,000 when the survey began in 2012. In response, the government has consistently prioritized academic support for multicultural students across all four rounds of the Basic Plan for Multicultural Family Support Policy, a legally mandated five-year plan established under the Multicultural Families Support Act.

The premise underlying these policies—that multicultural background, in itself, depresses academic capacity—is evident in the plans themselves. One of the objectives of the First Basic Plan for Multicultural Family Support Policy (2010–2012) was to “strengthen support for the children of multicultural families and cultivate global talent,” with detailed tasks including the strengthening of tailored educational support and the expansion of infrastructure for children struggling to adjust to school (Ministry of Gender Equality and Family, 2010). The current Fourth Basic Plan for Multicultural Family Policy (2023–2027) likewise sets “guaranteeing an equal starting line for multicultural children and adolescents” as one of its objectives, with detailed tasks such as raising the learning capacity of school-age multicultural children (Ministry of Gender Equality and Family, 2023). That learning support for multicultural children and adolescents appears, without exception, as a component of the support measures for multicultural families reflects a perception that multicultural background itself works negatively on academic capacity—that is, a perception that multicultural background exerts a direct negative effect even after controlling for factors such as parental income and Korean language proficiency.

However, this policy orientation has largely been based on presumptive rather than empirical reasoning. Kim and Choi (2025) point out that few studies have rigorously examined whether students from multicultural backgrounds actually underperform academically relative to their non-multicultural peers once relevant confounding factors are controlled for. Much of the existing evidence cited in policy documents is indirect. For instance, the Fourth Basic Plan (Ministry of Gender Equality and Family, 2023) highlights a gap in tertiary education enrollment rates between the general population (71.5%) and multicultural family children (40.5%), but lower enrollment does not necessarily

reflect lower academic achievement. The academic literature on multicultural students' achievement has also been limited. Most empirical studies, such as Park (2017), Yun and Oh (2020), and Yeon and Choi (2020), have used the Multicultural Adolescents Panel Study (MAPS) and analyzed only multicultural students without a non-multicultural comparison group.

A notable exception is Kim and Choi (2025), who used the Korean Children and Youth Panel Survey (KCYPs) 2010 data and estimated a random-effects panel regression model with interaction terms. Their results showed no statistically significant or robust evidence that multicultural background lowers academic performance after controlling for household income, parental education, private tutoring, school engagement, and other covariates. Furthermore, they found no evidence that the relationship between these explanatory variables and academic achievement differs by multicultural status.

While Kim and Choi (2025) represent a pioneering contribution, their analysis relies on a conventional panel regression framework. In observational studies of this kind, where treatment assignment—here, being from a multicultural background—is not randomized, the validity of causal claims depends critically on the correct specification of the functional relationship between confounders and outcomes. Traditional parametric approaches impose strong assumptions on this functional form, and misspecification can lead to biased estimates of the treatment effect. This limitation motivates the present study.

This paper applies the Double/Debiased Machine Learning (DML) framework proposed by Chernozhukov *et al.* (2018) to estimate the causal effect of multicultural background on academic achievement. The DML framework addresses two fundamental challenges that arise when machine learning (ML) methods are used in causal inference. First, ML estimators of nuisance functions (e.g., the conditional expectation of the outcome or the propensity score) are necessarily regularized, inducing a bias that contaminates naïve plug-in estimators of the treatment effect parameter. Second, ML methods tend to overfit the training data, creating spurious correlations between estimation errors and structural disturbances. Chernozhukov *et al.* (2018) show that these problems can be overcome by combining two ingredients: (1) Neyman-orthogonal score functions that are locally insensitive to perturbations in nuisance parameter estimates, and (2) cross-fitting, an efficient form of sample splitting that eliminates overfitting bias. The resulting DML estimator is $\sqrt{N}$ -consistent, approximately unbiased, and normally distributed, enabling valid confidence intervals and hypothesis tests regardless of the specific ML method used to estimate the nuisance functions.

We provide a methodologically rigorous causal analysis of a policy-relevant question that has been addressed only by conventional methods. By leveraging the flexibility of ML methods to approximate potentially complex confounding relationships while retaining the inferential guarantees of the DML framework, we obtain estimates that are robust to functional form misspecification. We replicate the analysis across all three panel cohorts of the KCYPS 2010—the Grade 1, Grade 4, and Grade 7 cohorts—to assess the robustness of the findings. If multicultural background truly exerts a causal effect on academic achievement, this should be detectable consistently across cohorts that span different developmental stages.

2. METHODOLOGY

2.1. PARTIALLY LINEAR REGRESSION MODEL

To estimate the effect of multicultural background on academic achievement, we adopt the partially linear regression (PLR) model, which serves as the lead example in Chernozhukov *et al.* (2018). The PLR model dates back to Robinson (1988), who established root- N -consistent semiparametric estimation of the linear coefficient while leaving the nuisance function g_0 nonparametric, using kernel-based partialling-out. We adopt the DML estimator rather than Robinson’s original kernel-based procedure for two related reasons. First, our enriched control set Z_{it} is high-dimensional, and kernel or series estimators of the nuisance functions suffer from the curse of dimensionality in this regime, whereas modern ML learners combined with Neyman-orthogonal scores and cross-fitting accommodate high-dimensional controls under weak rate conditions. Second, DML does not require the researcher to commit to a single functional form or smoothing approach for g_0 and m_0 ; the orthogonalization makes the estimator of θ_0 first-order insensitive to regularization bias in these nuisance estimates. This is precisely why a machine-learning approach is valuable from an econometric standpoint, and it constitutes part of the methodological contribution of the present study. The PLR model is specified as:

$$\begin{aligned} Y_{it} &= D_i \theta_0 + g_0(X_{it}) + U_{it}, & E[U_{it} | X_{it}, D_i] &= 0, \\ D_i &= m_0(X_{it}) + V_{it}, & E[V_{it} | X_{it}] &= 0, \end{aligned}$$

where Y_{it} is the academic achievement score of student i at time t , D_i is the binary multicultural background indicator (the treatment variable), X_{it} is a vector of control variables, and U_{it} and V_{it} are disturbance terms. The parameter of

interest, θ_0 , measures the average causal effect of multicultural background on academic achievement, conditional on the controls. The function $g_0(\cdot)$ captures the potentially complex, nonlinear relationship between the controls and the outcome, while $m_0(\cdot)$ is the conditional expectation of the treatment given controls, which for a binary treatment coincides with the propensity score. We use the time-invariant treatment D_i (rather than a time-varying D_{it}) because multicultural background is fixed for each student across all waves; this is the empirically correct specification for our setting. The control variables X_{it} are nonetheless time-varying, so $m_0(X_{it}) = E[D_i | X_{it}]$ is a well-defined conditional expectation that may differ across waves through the time-varying controls, even though its argument predicts a time-invariant outcome. Crucially, neither g_0 nor m_0 needs to be parametrically specified; both are treated as high-dimensional nuisance functions to be estimated by ML methods.

The PLR model is appropriate for our setting because the treatment variable D_i enters linearly, which allows us to interpret θ_0 directly as the treatment effect parameter, while the confounding relationship is left unrestricted. This is the key advantage over the conventional parametric panel regression used in Kim and Choi (2025), where the function $g_0(\cdot)$ is implicitly assumed to be linear in X_{it} .

2.2. THE DML ESTIMATION PROCEDURE

A naïve approach to estimating θ_0 would be to use an ML estimator of g_0 and plug it directly into the estimating equation. However, as Chernozhukov *et al.* (2018) demonstrate, this leads to an estimator that is not $\sqrt{N}$ -consistent due to regularization bias. The bias arises because regularized ML estimators—such as Lasso, random forests, or neural networks—necessarily trade off bias and variance, and the resulting bias in the nuisance function estimate propagates into the estimate of θ_0 .

The DML procedure overcomes this problem through two key ingredients. The first is the use of a Neyman-orthogonal score function. For the PLR model, the orthogonal score is:

$$\psi(W; \theta, \eta) = \{Y - D\theta - g(X)\} \{D - m(X)\}, \quad \eta = (g, m), \quad (1)$$

where $W = (Y, D, X)$ denotes the vector of observable variables entering the score, and $\eta_0 = (g_0, m_0)$ is the true nuisance parameter. The Neyman orthogonality property means that the Gateaux derivative of $E[\psi(W; \theta_0, \eta_0)]$ with respect to η vanishes at η_0 . Intuitively, the moment condition is locally insensitive to perturbations in the nuisance parameter estimates, so that biased ML estimates

of g_0 and m_0 do not strongly contaminate the estimate of θ_0 . In the PLR model, this orthogonalization is achieved by “partialling out” the confounding effect of X from the treatment variable D , forming the residual $V = D - m_0(X)$. The regularization bias, which depends on the product of estimation errors in g_0 and m_0 , can vanish even when both nuisance functions are estimated at relatively slow rates, provided the product of their convergence rates is $o(N^{-1/2})$.

The second ingredient is cross-fitting, a form of sample splitting that eliminates the bias induced by overfitting. Without sample splitting, estimation errors and structural disturbances are correlated because the same data are used to estimate both the nuisance functions and the parameter of interest. Cross-fitting addresses this by partitioning the sample into K folds. For each fold k , the nuisance functions are estimated using all observations except those in fold k , and the resulting estimates are used to form the score function for observations in fold k . The final estimator aggregates across all folds, thereby using the full sample for estimation of θ_0 without sacrificing efficiency.

Formally, we implement the DML2 algorithm of Chernozhukov *et al.* (2018, Definition 3.2). Let $(I_k)_{k=1}^K$ be a K -fold random partition of the observation indices. For each fold k , we estimate nuisance functions using data in I_k^c to obtain estimators $\hat{g}_{0,k}$ and $\hat{m}_{0,k}$. The DML2 estimator $\tilde{\theta}_0$ solves:

$$\frac{1}{K} \sum_{k=1}^K E_{n,k} [\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})] = 0.$$

Under the regularity conditions stated in Chernozhukov *et al.* (2018, Theorem 4.1), the DML estimator satisfies:

$$\sigma^{-1} \sqrt{N} (\tilde{\theta}_0 - \theta_0) \xrightarrow{d} N(0, 1),$$

where $\sigma^2 = (E[V^2])^{-1} E[V^2 U^2] (E[V^2])^{-1}$. The key rate condition requires that the product of the root-mean-squared estimation errors for g_0 and m_0 converges to zero faster than $N^{-1/2}$. This is a substantially weaker requirement than the $o(N^{-1/4})$ rate for each nuisance function separately, and it allows one function to be estimated at a slow rate provided the other is estimated sufficiently well.

Following the recommendation of Chernozhukov *et al.* (2018, Remark 3.1), we set $K = 5$ for our main results. Chernozhukov *et al.* (2018) find moderate values of K (such as 4 or 5) to perform better than $K = 2$ in simulations and applications, since larger K leaves more observations in each auxiliary fold for estimating the high-dimensional nuisance functions; with our cohort sample sizes (roughly 8,700 to 11,700 person-year observations), $K = 5$ leaves about 80% of

the data for nuisance estimation in each fold while retaining a sizeable evaluation fold. To account for the randomness introduced by the particular sample partition, we repeat the entire DML estimation $S = 20$ times, where S denotes the number of repetitions with independently drawn random sample partitions, and report the median estimate and the median-based variance estimator (Chernozhukov *et al.*, 2018, Definition 3.3), which is more robust to outliers than the mean-based alternative. The choice of $S = 20$ is large enough to stabilize the across-split variation in the reported estimates while remaining computationally feasible given the repeated cross-fitting across multiple learners and cohorts.

2.3. MACHINE LEARNING METHODS FOR NUISANCE FUNCTION ESTIMATION

A strength of the DML framework is that the theoretical guarantees hold for a broad array of ML methods used to estimate the nuisance functions g_0 and m_0 , as long as they satisfy the required rate conditions. Following the empirical illustrations in Chernozhukov *et al.* (2018, Section 6), we employ several ML methods and report results across all of them to demonstrate robustness.

Lasso is an ℓ_1 -penalized least squares estimator that performs simultaneous estimation and variable selection; we augment the raw covariate set with all second-order terms (squares and pairwise interactions) to allow for nonlinear effects within a sparse linear framework. Ridge is an ℓ_2 -penalized least squares estimator that shrinks coefficients toward zero without setting any to exactly zero, and is effective when predictors are highly correlated. Elastic Net is a convex combination of the Lasso (ℓ_1) and Ridge (ℓ_2) penalties that inherits the variable selection property of the Lasso while maintaining the grouping effect of Ridge. For all three penalized methods, the penalty parameter (and, for Elastic Net, the mixing ratio) is chosen by cross-validation.

Random Forest is an ensemble of decision trees, each fitted on a bootstrap sample with a random subset of features considered at each split. Random forests are well suited to capturing nonlinear relationships and interactions among covariates without explicit specification. We use 1,000 trees with default tuning parameters. Gradient Boosting is an iterative ensemble method that sequentially fits weak learners (typically shallow trees) to the residuals of the previous stage, with regularization parameters chosen by cross-validation. Gradient boosting is known for its strong predictive performance across a variety of data-generating processes.

Neural Network with Early Stopping is a feedforward neural network trained with backpropagation, where training is halted when performance on a validation

set ceases to improve, thereby preventing overfitting.

For each ML method, the nuisance functions $g_0(X) = E[Y | X]$ and $m_0(X) = E[D | X]$ are estimated separately. In the PLR model with the partialling-out score (1), m_0 is the conditional mean of the treatment given controls, which for a binary treatment is the propensity score. Since our treatment variable (multicultural background) is binary, the ML estimator of m_0 is obtained either from a classification model or from a regression model; in either case, predicted values are used as the estimated propensity scores. We emphasize that $m_0(Z_{it}) = E[D_i | Z_{it}]$ is a predictive conditional expectation used to orthogonalize the treatment with respect to the controls, not a structural or behavioral model of how multicultural background is determined. It serves only to construct the residualized treatment $V_{it} = D_i - m_0(Z_{it})$, capturing statistical association in whatever direction it runs; consequently, the fact that multicultural background is fixed well before many of the controls are realized does not affect the interpretation of this object, which is a projection rather than a causal selection equation.

2.4. ADAPTATION TO PANEL DATA

The KCYPS 2010 data have a panel structure with repeated observations on the same individuals over time. The DML framework as presented in Chernozhukov *et al.* (2018) assumes independent and identically distributed observations. A standard approach for handling panel data in conventional econometrics is to include individual fixed effects and time fixed effects. In a traditional linear fixed-effects model, individual fixed effects absorb all time-invariant regressors via the within-transformation, making it mechanically impossible to estimate the coefficient on a time-invariant treatment variable. This is the reason Kim and Choi (2025) adopted a random-effects model in their conventional panel regression. In the DML framework, the issue manifests differently but leads to the same conclusion. Recall that the DML partialling-out procedure constructs the orthogonalized regressor $V_{it} = D_i - m_0(Z_{it})$, where $m_0(Z_{it}) = E[D_i | Z_{it}]$ is the conditional expectation of the treatment given the controls. Suppose the control set Z_{it} contained a full set of individual indicators (i.e., one dummy per student). Because the treatment D_i is itself time-invariant and takes a single value within each student, these indicators would predict D_i without error, so that $m_0(Z_{it}) = D_i$ and hence $V_{it} = 0$ for every observation. With no residual variation in the orthogonalized regressor, the score (1) is identically zero and θ_0 is unidentified. Thus, individual indicators must be excluded from the control set in either the conventional or the DML framework when the treatment is time-invariant.

To flexibly account for latent individual-specific heterogeneity and common

time factors without relying on strong functional form assumptions, we follow the conditioning strategy discussed in Chernozhukov *et al.* (2024, Chapter 9). Specifically, we augment the raw control variables X_{it} with three additional sets of constructed variables: (i) within-individual averages $\bar{X}_i$, computed as the time-series mean of each control variable for each student, which serve as proxies for unobserved individual-level characteristics; (ii) within-time-period averages $\bar{X}_t$, computed as the cross-sectional mean of each control variable at each survey wave, which capture common time-specific shocks affecting all students; and (iii) initial conditions $X_{i,0}$ and $Y_{i,0}$, the baseline values of the controls and the outcome at the first available survey wave, which capture the student's starting position. Conditioning on the within-individual averages $\bar{X}_i$ is a nonparametric generalization of the correlated-random-effects device of Mundlak (1978) and Chamberlain (1982), in which time averages of the covariates proxy for unobserved individual effects that are correlated with the observed regressors. It is therefore important to be explicit about what this strategy can and cannot absorb: it captures the component of individual heterogeneity that is correlated with the time-varying observables (through their within-individual means), but any individual-specific factor that is orthogonal to the included controls remains unaddressed. In this sense the approach is weaker than a full fixed-effects specification, which would absorb all time-invariant heterogeneity but, as noted above, would also eliminate the between-individual variation needed to identify the effect of a time-invariant treatment. The conditioning strategy thus occupies a middle ground, trading the full heterogeneity control of fixed effects for the ability to identify the treatment effect. Although the within-individual averages and initial conditions are themselves time-invariant, their inclusion does not reintroduce the identification problem discussed above: unlike a full set of individual indicators, these summary variables do not uniquely identify each student and therefore do not perfectly predict the treatment, so that the orthogonalized regressor $V_{it} = D_i - m_0(Z_{it})$ retains residual variation. What must be excluded is the set of individual indicators, not time-invariant controls in general. We also include the time index t to allow for deterministic time trends. Formally, with T_i denoting the number of waves observed for student i , these constructed variables are defined as

$$\bar{X}_i = \frac{1}{T_i} \sum_{s=1}^{T_i} X_{is}, \quad \bar{X}_t = \frac{1}{N_t} \sum_{j=1}^{N_t} X_{jt}, \quad (X_{i,0}, Y_{i,0}) = (X_{i1}, Y_{i1}),$$

where the index s runs over the waves at which student i is observed, re-indexed from 1 to T_i in chronological order (so that $s = 1$ corresponds to the first wave

in which the student appears, regardless of its original wave number), j indexes students within the cross-sectional average, and N_t is the number of students observed in wave t . The full set of control variables entering the nuisance functions is thus

$$Z_{it} = (X_{it}, \bar{X}_i, \bar{X}_t, X_{i,0}, Y_{i,0}, t). \quad (2)$$

This strategy leverages the assumption that flexibly conditioning on these observables—including individual-specific and time-specific summary variables—is sufficient to account for latent sources of heterogeneity that might otherwise confound the estimated effect of multicultural background on academic achievement.

Under this conditioning strategy, the nuisance functions in the PLR model become $g_0(Z_{it}) = E[Y_{it} | Z_{it}]$ and $m_0(Z_{it}) = E[D_i | Z_{it}]$, both of which are estimated by the ML methods described in Section 2.3 using the enriched control set Z_{it} . The ML learners can then exploit the rich structure of these augmented controls—including interactions among time-varying, individual-mean, and period-mean variables—to flexibly approximate the confounding relationship without requiring the researcher to pre-specify a particular functional form.

In the DML implementation, the cross-fitting is performed at the individual level: all observations from the same student are assigned to the same fold. This ensures that the nuisance function estimates for a given fold are constructed from data that are independent of the observations used to estimate the treatment effect in that fold, thereby preserving the within-individual correlation structure and preventing information leakage. For simplicity in this analysis, and following the approach taken in Chernozhukov *et al.* (2024, Chapter 9), we assume that once all sources of dependence are accounted for by the enriched set of observed controls, the disturbance U_{it} may be treated as independent across individuals i and over time t . The inclusion of individual-level means, time-period means, and initial conditions in the control set makes this assumption more plausible than it would be with the raw controls alone, as these constructed variables absorb much of the individual-specific and time-specific variation that would otherwise induce dependence in the disturbances.

We now collect the conditions under which the DML estimator of θ_0 is $\sqrt{N}$ -consistent and asymptotically normal in our panel setting. These conditions adapt the regularity conditions of Chernozhukov *et al.* (2018, Assumption 4.1) for the PLR model to the enriched control set Z_{it} defined in (2).

Assumption 1 (Conditional Exogeneity). Conditional on the enriched control

set Z_{it} , the treatment is mean-independent of the structural disturbance:

$$E[U_{it} | Z_{it}, D_i] = E[U_{it} | Z_{it}], \quad (3)$$

where $U_{it} = Y_{it} - D_i\theta_0 - g_0(Z_{it})$. Equivalently, once Z_{it} is conditioned on, D_i is as good as randomly assigned with respect to academic achievement.

Assumption 2 (Partially Linear Structure and Neyman Orthogonality). The data satisfy the PLR model $Y_{it} = D_i\theta_0 + g_0(Z_{it}) + U_{it}$ and $D_i = m_0(Z_{it}) + V_{it}$ with $E[V_{it} | Z_{it}] = 0$, and the score (1) satisfies the Neyman orthogonality condition $\partial_\eta E[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0$.

Assumption 3 (Quality of Nuisance Estimation). The ML estimators $\hat{g}_0$ and $\hat{m}_0$ are consistent in $L^2(P)$, and the product of their root-mean-squared convergence rates satisfies

$$\|\hat{g}_0 - g_0\|_{P,2} \times \|\hat{m}_0 - m_0\|_{P,2} = o(N^{-1/2}).$$

This product condition is weaker than requiring each nuisance function to converge at the $o(N^{-1/4})$ rate, and it permits one function to be estimated slowly provided the other is estimated sufficiently well.

Assumption 4 (Cross-fitting and Disturbance Independence). Estimation uses K -fold cross-fitting in which all observations from the same student are assigned to the same fold. Conditional on the enriched control set Z_{it} , the disturbances U_{it} are independent across students i and over time t .

Assumption 1 is the unconfoundedness condition that identifies θ_0 ; equation (3) makes precise the informal statement that, conditional on Z_{it} , treatment is as good as randomly assigned. Assumption 2 is satisfied by construction of the orthogonal score in (1). Assumption 3 is the only non-primitive condition; the required rates are attainable for the ML methods used here under standard structured assumptions on the nuisance functions, and we therefore do not restate the method-specific primitive conditions (see Chernozhukov *et al.*, 2018, Remark 4.1). Assumption 4 warrants further comment. The independence of U_{it} across i and t is a strong simplifying assumption in a panel context; we adopt it following Chernozhukov *et al.* (2024, Chapter 9) on the premise that the enriched controls Z_{it} , by absorbing individual-specific and time-specific variation through $\bar{X}_i$, $\bar{X}_t$, and the initial conditions, account for the dependence that would otherwise remain in the raw disturbances. Cross-fitting at the student level further preserves the within-student correlation structure during nuisance estimation and prevents information leakage across folds. We acknowledge that residual dependence may remain. While we do not pursue a separate sensitivity analysis for

this assumption, the cross-fitting design and the enriched controls are intended to mitigate its impact, and the omitted-variable-bias analysis in Section 2.5 addresses the related and more consequential concern of unobserved confounding in the exogeneity condition.

Assumption 1 is the same unconfoundedness assumption that underlies the analysis of Kim and Choi (2025). The advantage of the DML approach is that the confounding adjustment is performed nonparametrically via ML, rather than through a pre-specified linear function, thereby reducing the risk of bias due to functional form misspecification. The use of Neyman-orthogonal scores and cross-fitting provides inferential guarantees (valid confidence intervals and hypothesis tests) that are robust to the use of complex, high-dimensional ML estimators for the nuisance functions. By “robust” we mean, in a precise econometric sense, that the first-order effect of the nuisance estimation error on the moment condition vanishes (Neyman orthogonality), so that the $\sqrt{N}$ -asymptotic normality and the validity of the resulting confidence intervals for θ_0 are preserved even when g_0 and m_0 are estimated at slower-than-parametric rates. By replicating the analysis across all three KCYPS 2010 cohorts using multiple ML methods, we assess the robustness of the estimated treatment effect across different developmental stages and estimation strategies.

2.5. SENSITIVITY TO UNOBSERVED CONFOUNDING

Assumption 1 (conditional exogeneity) cannot be tested directly, and the KCYPS 2010 does not contain certain potentially relevant covariates, most notably the student’s and parents’ Korean language proficiency, the family’s immigration motive (labor, marriage, or study), and the degree of social-network integration. If such variables affect both multicultural classification and academic achievement, the estimated effect is biased. We therefore assess the robustness of our estimates to unobserved confounding using the omitted-variable-bias framework of Chernozhukov *et al.* (2022), which extends classical sensitivity analysis to the DML setting and does not require the omitted variables to be observed.

For the partially linear model, the bias induced by an unobserved confounder admits the sharp characterization $|\theta_s - \theta| \leq B = SC_Y C_D$, where θ_s is the estimate from the observed specification and θ is the bias-corrected target. The scale term S is identified from the observed data; in the PLR case it equals $\sqrt{E(Y - g_s)^2 / E[V_s^2]}$, the ratio of the outcome-residual standard deviation to the treatment-residual standard deviation, both of which are by-products of the cross-fitting procedure. The two sensitivity parameters C_Y and C_D are not identified: C_Y^2 is the proportion of residual variation in the outcome that the

confounder would explain, and C_D^2 is the corresponding proportion for the treatment. Bounding the bias therefore reduces to postulating plausible values for C_Y^2 and C_D^2 .

We summarize the analysis through the robustness value, defined as the smallest level of confounding strength (with C_Y^2 and C_D^2 set equal) at which the confidence bound for θ first includes zero (Chernozhukov *et al.*, 2022, Definition 2). A large robustness value indicates that only strong unobserved confounding could overturn the result; a small value indicates the opposite. Because Chernozhukov *et al.* (2022) define this quantity geometrically rather than in closed form, we obtain it by numerically solving for the confounding level at which the bias-adjusted confidence bound reaches zero. For the confidence bounds we apply the point-estimate standard error symmetrically to both ends of the identified interval; this omits the asymmetric variance correction in the influence function of Chernozhukov *et al.* (2022, Theorem 4) and is conservative, in that it yields a larger robustness value than the exact procedure.

To anchor the sensitivity parameters in substantive terms rather than arbitrary values, we benchmark them against the explanatory power of the observed covariates, asking how strong the unobserved confounding would have to be relative to the strongest observed controls. We also note that, because multicultural status is a rare binary treatment (approximately one percent of the sample), the treatment-residual variance $E[V_s^2]$ is small, which mechanically inflates the scale term S and compresses the robustness value; we therefore interpret the robustness value relative to the benchmarked confounding strength rather than against an absolute threshold. The results of this analysis are reported in Section 3.3.

3. EMPIRICAL RESULTS

3.1. DATA

We use the KCYPS 2010, a nationally representative longitudinal survey conducted by the National Youth Policy Institute. The survey tracks three cohorts of students—those in Grade 1, Grade 4, and Grade 7 as of 2010—over seven waves from 2010 to 2016. The initial sample sizes are 2,342, 2,378, and 2,351 for the three cohorts, respectively. Since the academic achievement measure (a self-assessed score on a five-point scale for each of five subjects: Korean, Mathematics, English, Science, and Social Studies) was not collected in the first wave, we use data from Waves 2 through 7 for the Grade 1 and Grade 4 cohorts, and Waves 2 through 6 for the Grade 7 cohort (as Wave 7 corresponds to the first year of university for this group). The outcome variable is the average of the five

subject scores, recoded so that higher values indicate better performance.

We acknowledge that this self-assessed measure is an imperfect proxy for actual academic achievement, and we consider two distinct sources of measurement error. The first is classical measurement error, in which the self-assessment reflects true achievement plus idiosyncratic noise. Because this error is in the outcome variable, it leaves the treatment-effect estimate unbiased and inflates only its standard error; averaging across five subjects mitigates such noise relative to a single-subject self-report. The second is the reference-group effect: students evaluate themselves relative to their peers, so if multicultural students attend schools with lower-achieving peer compositions, their self-assessments may be biased upward relative to their absolute achievement. Such a bias would compress the multicultural–non-multicultural gap in self-assessed scores below the true gap in achievement, and could therefore mask a negative effect. We note, however, that a strong masking effect of this kind would be difficult to reconcile with the positive and near-zero point estimates obtained for the Grade 1 and Grade 4 cohorts, as the estimates in Section 3.2 below will show; the reference-group concern is most pertinent to the residual negative estimates in the Grade 7 cohort, where we accordingly interpret the magnitude with caution. The KCYPS 2010 does not contain objective test scores, so we cannot directly verify these results against an external benchmark. With objective scores we would expect classical noise to be reduced—increasing statistical power—and the reference-group bias to be removed; assessing whether our conclusions are reproduced under such a measure is an important direction for future work.

The treatment variable is a binary indicator for multicultural background, defined as membership in an international marriage family, a foreign worker family, a North Korean defector (*saeteomin*) family, or a Korean-Chinese (*joseon-jok*) family. This variable is time-invariant for each student. The proportion of multicultural students is approximately 1% across all three cohorts, which is consistent with the actual population share during the survey period.

The set of control variables follows Kim and Choi (2025) and includes: gender, annual household income, father’s and mother’s education, guardian’s life satisfaction, self-reported health, hours of private tutoring, after-school program participation, pre-primary private education experience (Grade 1 cohort only), parental care, parental abuse, school learning engagement, peer relations, and teacher relations. The last five variables are constructed as the arithmetic mean of multiple survey items measured on four-point Likert scales.

The data and variables used in this study are essentially identical to those in Kim and Choi (2025). The estimation samples contain 11,701 person-year

observations over six waves for the Grade 1 cohort, 10,659 person-year observations over six waves for the Grade 4 cohort, and 8,676 person-year observations over five waves for the Grade 7 cohort.

3.2. ESTIMATION RESULTS

Table 1 reports, in addition to the OLS specifications and the individual ML learners introduced in Section 2.3, two aggregation schemes. Best selects for each nuisance function the ML learner attaining the lowest cross-fitted residual root mean squared error, and re-runs the DML procedure with that pair. Least Squares (LS) Model Average instead regresses the outcome and the treatment on the cross-fitted predictions of all nine specifications (the three OLS specifications and the six ML learners) by least squares, and estimates θ_0 from the residuals of these two regressions.

As a benchmark, we first examine the OLS No Controls specification, which uses only a constant to predict both the outcome and the treatment. The point estimate is approximately equal to the simple difference in mean academic achievement between the two groups. The estimates are -0.130 (Grade 1), -0.277 (Grade 4), and -0.345 (Grade 7), all statistically significant at the 5% level or better. These raw gaps suggest that, without controlling for any confounders, multicultural students score on average 0.13 to 0.35 points lower (on a 1–5 scale) than their non-multicultural peers, with the gap widening in older cohorts.

However, this raw gap cannot be interpreted as a causal effect. Multicultural and non-multicultural students differ not only in their cultural background but also in household income, parental education, access to private tutoring, and numerous other characteristics that independently affect academic achievement. It is precisely to filter out these confounding factors and isolate the pure effect of multicultural background that we proceed from OLS with basic controls to OLS with all controls and, ultimately, to the DML estimators that flexibly control for a rich set of covariates.

Once controls are introduced, the picture changes substantially. For the Grade 1 cohort, most controlled specifications yield point estimates that are positive or very close to zero, and none is statistically significant at conventional levels. The OLS Basic Controls estimate is 0.074 (SE = 0.050), and the DML estimates using various ML learners range from -0.029 (Neural Network) to 0.085 (Best), all statistically insignificant. The LS Model Average yields a positive estimate of 0.206 that is marginally significant at the 10% level, but this is likely an artifact of the ensemble procedure rather than a genuine effect. This result indicates that the raw negative gap observed in the uncontrolled speci-

Model	Cohort		
	Grade 1	Grade 4	Grade 7
OLS No Controls	−0.1296** (0.0583)	−0.2772*** (0.0926)	−0.3450*** (0.0858)
OLS Basic Controls	0.0744 (0.0498)	−0.0459 (0.0715)	−0.1417* (0.0831)
OLS All Controls	0.0453 (0.0481)	−0.0511 (0.0581)	−0.1687*** (0.0615)
LassoCV	0.0435 (0.0480)	−0.0502 (0.0579)	−0.1489** (0.0609)
RidgeCV	0.0430 (0.0476)	−0.0479 (0.0580)	−0.1509** (0.0608)
ENetCV	0.0444 (0.0482)	−0.0505 (0.0580)	−0.1492** (0.0609)
Random Forest	0.0723 (0.0659)	−0.0525 (0.0787)	−0.1475** (0.0658)
Boosted Trees	0.0816 (0.1983)	−0.0936 (0.1858)	−0.1037 (0.1820)
NN (Early Stopping)	−0.0294 (0.0538)	−0.0594 (0.0627)	−0.1658*** (0.0575)
Best	0.0847 (0.1987)	−0.0797 (0.1640)	−0.1339 (0.1610)
Least Squares Model Average	0.2055* (0.1085)	−0.1635 (0.1467)	0.0140 (0.0959)

Table 1: DML ESTIMATES OF THE EFFECT OF MULTICULTURAL BACKGROUND ON ACADEMIC ACHIEVEMENT. Point estimates use the DML2 algorithm with $K = 5$ folds. Standard errors are in parentheses and are heteroskedasticity-robust (HC1); except for Least Squares Model Average, which reports the standard error from a single combined regression, they are aggregated across the $S = 20$ partitions by the median-based variance estimator. ***, $p < 0.01$; **, $p < 0.05$; *, $p < 0.1$.

cation is entirely attributable to differences in observed covariates between the two groups: once household income, parental education, and other covariates are accounted for, multicultural background has no detectable direct effect on academic achievement in the youngest cohort.

For the Grade 4 cohort, the controlled estimates are uniformly negative but small in magnitude, ranging from -0.046 (OLS Basic) to -0.164 (LS Model Average). Crucially, none of these estimates is statistically significant at the 5%

level, and most are insignificant even at the 10% level. The Boosted Trees and Best estimators produce somewhat larger negative point estimates but with correspondingly large standard errors, indicating considerable estimation uncertainty. The Grade 4 results thus echo the Grade 1 findings: after controlling for observable confounders, there is no robust evidence that multicultural background depresses academic achievement.

The Grade 7 cohort presents the most nuanced picture. The controlled estimates are negative across all individual learners, and several achieve statistical significance. The OLS All Controls estimate is -0.169 ($SE = 0.061$, significant at 1%), and the regularized linear methods (Lasso, Ridge, Elastic Net) produce estimates in the range of -0.149 to -0.151 , significant at 5%. The Neural Network with early stopping yields -0.166 (significant at 1%), and the Random Forest estimate is -0.148 (significant at 5%). However, the Boosted Trees and Best estimators, which have substantially larger standard errors, are statistically insignificant, and the LS Model Average is 0.014 and far from significant. Thus, even in the cohort where the evidence for a negative effect is strongest, the magnitude and significance of the estimate vary across learners, and the ensemble methods fail to find a statistically significant effect. It is also useful to gauge the substantive magnitude of these estimates rather than their statistical significance alone. Expressed in standard-deviation units of the outcome (whose standard deviations are 0.73, 0.87, and 0.81 for the three cohorts), the magnitudes are negligible in the younger cohorts: for the Grade 1 cohort the better-performing learners straddle zero in sign and none exceeds about 0.10 standard deviations in absolute value, and for the Grade 4 cohort they fall in a narrow band of roughly 0.06 to 0.07 standard deviations. The Grade 7 estimates from the better-performing learners correspond to roughly 0.18 to 0.21 standard deviations, which is small by conventional benchmarks—near the lower threshold for what is usually regarded as a small effect—and amounts to about 0.15 points, or under four percent, on the five-point scale. Hence, even where a statistically significant negative estimate appears, its substantive magnitude is modest.

Our results are broadly consistent with those of Kim and Choi (2025), who used a conventional random-effects panel regression with the same data. They reported that the multicultural dummy coefficient was negative but only marginally significant for the Grade 1 cohort (at 5%) and Grade 4 cohort (at 10%), and paradoxically positive and highly significant for the Grade 7 cohort. They concluded that no statistically robust evidence exists for a negative causal effect of multicultural background on academic achievement. Our DML analysis largely corroborates this conclusion with one notable difference: whereas

the Grade 7 result of Kim and Choi (2025) was confounded by the presence of interaction terms in their specification (making direct interpretation of the multicultural dummy difficult), our PLR model provides a cleaner estimate of the average treatment effect. In the Grade 7 cohort, some of our DML specifications do find a statistically significant negative effect, but the lack of robustness across learners—particularly the insignificance of the ensemble estimate—cautions against strong conclusions.

This example also illustrates the practical importance of considering several learners rather than relying on a single pre-specified choice. As Chernozhukov *et al.* (2024, Chapter 9) emphasize in a different empirical context, it is generally unknown *ex ante* which learner will perform best, and there can be substantial differences between the better-performing learners and the others both in terms of predictive accuracy and the resulting DML estimates and standard errors. In our application, the Boosted Trees and Best estimators exhibit markedly larger standard errors than the regularized linear methods and Random Forest, suggesting instability in the nuisance function estimates. The large standard errors of the ensemble estimators (Best and LS Model Average) appear to stem from this instability rather than from a genuine absence of an effect: the ensemble combines learner predictions, so the poor and noisy performance of the Boosted Trees learner, together with near-collinearity among the learner predictions, inflates the variance of the combined estimator. We therefore place greater interpretive weight on the better-performing individual learners, whose estimates are mutually consistent, than on the ensemble estimators in this particular application. The consistency of results across the better-performing learners (Lasso, Ridge, Elastic Net, Random Forest, and Neural Network) is reassuring and suggests that the conclusions are not driven by the idiosyncrasies of any single ML method. We note, however, that for the Grade 7 cohort these better-performing learners do yield negative and significant estimates, and we do not dismiss them: as discussed in Section 3.3, the most plausible unobserved confounders would attenuate rather than strengthen these estimates, so we interpret them as an upper bound on the magnitude of any true negative effect rather than as evidence against it. The somewhat stronger negative estimates in Grade 7 relative to the younger cohorts may also reflect grade-specific factors not fully captured by our controls, such as the intensity and level of prior private tutoring, which becomes increasingly consequential at higher grades; we return to this as a limitation.

3.3. SENSITIVITY ANALYSIS RESULTS

We now apply the omitted-variable-bias analysis described in Section 2.5 to assess the robustness of our findings to potential unobserved confounders, most notably the student's and parents' Korean language proficiency and social-network integration.

We begin with the direction of the potential bias, which is central to interpreting the analysis. The omitted variables of concern are, on average, lower for multicultural students: children in international-marriage and foreign-worker families tend to have lower Korean language proficiency and weaker social-network integration, and these in turn are associated with lower academic achievement. An omitted variable that is negatively associated with the treatment and positively associated with the outcome biases the estimated treatment effect downward. Omitting such variables therefore makes the estimated effect of multicultural background more negative than the true effect, not less. Controlling for Korean proficiency or social-network integration, had they been available, would thus be expected to attenuate the negative estimates toward zero, reinforcing rather than overturning our conclusion of no adverse effect. This is most relevant for the Grade 7 cohort, where the only residual negative estimates appear; the direction of the likely bias implies that these estimates represent, if anything, an upper bound on the magnitude of any true negative effect.

The robustness value provides a complementary, quantitative summary, but it must be read in light of two features of our setting. First, the robustness value measures only the strength of confounding required to overturn a result and is, by construction, agnostic to its direction: it presumes the worst-case adversarial confounder. Given the directional argument above, this worst case is implausible here, since the most threatening omitted variables in fact work in our favor. Second, because multicultural status is a rare binary treatment (approximately one percent of the sample), the treatment-residual variance is small and the scale term S is correspondingly large (between 5.6 and 6.2 across cohorts). This mechanically compresses the robustness value, so its absolute level is not directly comparable to the conventional thresholds used in settings with balanced treatments. For this reason, where a residual estimate remains to be defended—the Grade 7 cohort—we report the robustness value alongside a benchmark against the observed covariates rather than against an absolute standard.

We apply the sensitivity analysis to the RidgeCV specification, which is representative of the better-performing learners (Lasso, Ridge, Elastic Net, Random Forest, and Neural Network) whose point estimates cluster tightly together; for the Grade 7 cohort, for instance, these learners all yield estimates between -0.148 and -0.166 . Because the scale term S depends on the residual vari-

ances rather than on the point estimate, the robustness value is virtually unchanged across these learners, so the choice of RidgeCV is not consequential. For the Grade 1 and Grade 4 cohorts, the 95% confidence bound already includes zero before any confounding adjustment, so no positive degree of confounding is needed to overturn significance; this simply restates the absence of a detectable effect in these cohorts, and we therefore do not report a robustness value for them. For the Grade 7 cohort, the confidence-bound robustness value is 0.9% and the point-estimate robustness value is 2.6%. This small robustness value reflects the compression induced by the rare treatment rather than genuine fragility; interpreted together with the directional argument, it does not indicate that a plausible confounder would strengthen a negative effect.

To place the robustness value on an interpretable scale, we benchmark it against the explanatory power of the observed controls. For each observed covariate we compute its partial R^2 with the treatment—that is, the share of the treatment-residual variation it explains given all other controls—which is measured on the same scale as the treatment-side sensitivity parameter C_D^2 underlying the robustness value. Across all three cohorts, no single observed covariate reaches a partial R^2 above 0.46%: the strongest control explains 0.45%, 0.43%, and 0.23% of the treatment-residual variation in the Grade 1, Grade 4, and Grade 7 cohorts, respectively, and the vast majority of controls fall well below 0.1%. These magnitudes are small precisely because the treatment is rare, which limits how much any covariate can explain about it.

The comparison is informative for the Grade 7 cohort, where the confidence-bound robustness value is 0.9%. The strongest observed control there contributes only 0.23%, roughly one-quarter of the robustness value, and the point-estimate robustness value of 2.6% exceeds it by more than tenfold. In other words, for unobserved confounding to overturn the result, a single omitted variable would have to be several times more strongly associated with multicultural status than any of the rich set of socioeconomic and educational controls we already include—including household income, parental education, and private tutoring. We regard this as implausible, particularly given that the most frequently cited omitted variables (Korean language proficiency, social-network integration) would, as argued above, move the estimate toward zero rather than away from it. We focus the benchmark on the treatment side because the rare treatment makes C_D the binding constraint; the same logic applies with even more room on the outcome side.

We caution that this comparison is made one covariate at a time. If several omitted dimensions acted jointly and independently, their combined contribution

could exceed that of any single observed control. Even so, overturning the result would require the joint unobserved contribution to surpass the robustness value while acting in the adverse direction, against the directional argument above; we view this combination as unlikely.

4. DISCUSSION AND POLICY IMPLICATIONS

Our DML analysis yields a clear finding: once observed socioeconomic and educational covariates are flexibly controlled for, there is no robust evidence that multicultural background has a direct effect on academic achievement—that is, an effect operating independently of these observed channels. For the two younger cohorts (Grade 1 and Grade 4), the estimated treatment effects are small, often positive in sign, and statistically insignificant across all ML learners. For the oldest cohort (Grade 7), some individual learners yield negative and statistically significant estimates, but the results are not robust: the Best and LS Model Average estimators—which aggregate across learners—are both statistically insignificant. Taken together, these findings suggest that the raw achievement gap between multicultural and non-multicultural students is largely—and in some cohorts, entirely—accounted for by differences in socioeconomic and educational background variables rather than by multicultural status per se.

We note an important point of interpretation. Several of the controls—most clearly time-varying household income and parental care during the student’s schooling—may lie on the causal pathway from multicultural background to achievement rather than being common causes that precede it. Multicultural background is fixed at birth, whereas these variables are realized over the student’s life, so they may function as mediators rather than confounders. If so, our estimate is closer to a controlled direct effect—the effect of multicultural background net of these observed channels—than to a total effect, and our finding of no robust effect should be read as the absence of a direct effect, while leaving open the possibility of an indirect effect operating through the mediating variables.¹ We emphasize that this interpretation does not weaken our policy conclusion; if anything, it sharpens it. The relevant question for policy is whether multicultural status per se, as a demographic category, warrants targeted academic support, or whether the appropriate targets are the underlying factors. A finding that the achievement gap operates through income, parental resources, and language environment—rather than through membership in the

¹We thank an anonymous referee for raising this point, which led us to reframe our estimand as a controlled direct effect.

category itself—directly supports reorienting policy from the category to those factors. We also note that variables determined before the student's birth, such as parental education, are more naturally interpreted as background confounders than as mediators, so the control set mixes both types; we therefore do not attempt a full mediation decomposition and instead interpret our estimate as the effect net of the observed socioeconomic channels.

These results carry important policy implications. The Korean government's multicultural family support policies have consistently included academic support programs targeted specifically at multicultural students, premised on the assumption that these students are academically disadvantaged. Our findings challenge this premise. If the observed achievement gap is driven primarily by socioeconomic factors (lower household income, lower parental education) that happen to be correlated with multicultural background, then policies targeting academic support exclusively at multicultural students are misspecified: they address the symptom rather than the cause.

A more effective approach would be to design universal support programs that address the underlying structural disadvantages—such as low-income tutoring subsidies, educational resource programs for underserved areas, and language support for students with limited Korean proficiency—without conditioning eligibility on multicultural status. This would not only avoid the stigmatizing effect of singling out multicultural students as academically deficient but would also benefit non-multicultural students from similarly disadvantaged backgrounds who are currently excluded from multicultural-specific programs. As Kim and Choi (2025) note, the U.S. approach under the Every Student Succeeds Act (2015) offers a useful model: rather than targeting a single demographic category, it guarantees equal educational opportunity for all educationally disadvantaged groups, including those from diverse linguistic and cultural backgrounds.

That said, our findings do not imply that multicultural students face no challenges. Language barriers, cultural adaptation difficulties, and identity-related stresses are real and well-documented. The point is that these challenges do not appear to translate into lower academic achievement once socioeconomic background is adequately controlled for. Targeted interventions for specific difficulties (e.g., Korean language instruction for recently arrived students, culturally responsive pedagogy) may still be warranted, but they should be designed based on the specific need rather than on the blanket assumption that multicultural background implies academic deficiency.

5. CONCLUSION

This paper applied the Double/Debiased Machine Learning framework of Chernozhukov *et al.* (2018) to estimate the causal effect of multicultural background on academic achievement using three panel cohorts from the KCYPS 2010. By employing Neyman-orthogonal score functions and cross-fitting with multiple ML learners—including Lasso, Ridge, Elastic Net, Random Forest, Boosted Trees, and Neural Networks—we obtained estimates that are robust to functional form misspecification and regularization bias.

Our principal finding is that, after flexibly controlling for socioeconomic and educational covariates, there is no statistically robust evidence that multicultural background lowers academic achievement independently of those observed factors. The raw achievement gap visible in unconditional comparisons is substantially attenuated—and, in the younger cohorts, no longer statistically distinguishable from zero—once confounders are accounted for. This result is consistent with the findings of Kim and Choi (2025) using conventional panel regression methods, and it holds across multiple ML estimation strategies.

These findings call for a critical reassessment of the policy paradigm that frames multicultural students as academically disadvantaged and designs targeted learning-support programs on that basis. Rather than policies anchored to the demographic category of “multicultural background,” what is needed are universal, factor-based interventions that address specific structural disadvantages—low income, limited Korean proficiency, insufficient educational infrastructure—regardless of a student’s cultural background. Such an approach would improve both the efficiency and the equity of educational support policies in Korea’s increasingly diverse society.

REFERENCES

- Chamberlain, G. (1982). “Multivariate regression models for panel data,” *Journal of Econometrics* 18, 5–46.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal* 21, C1–C68.
- Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A. and Syrgkanis, V. (2022). “Long story short: Omitted variable bias in causal machine learning,” NBER Working Paper No. 30302, National Bureau of Economic Research.

- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M. and Syrgkanis, V. (2024). *Applied causal inference powered by ML and AI*, online textbook.
- Kim, W. G. and Choi, P. S. (2025). “An empirical analysis of the impact of multicultural background on academic achievement,” *Journal of Social Science* 32, 182–202.
- Ministry of Education and Korean Educational Development Institute (2025). *2025 basic education statistics*, published August 28, 2025.
- Ministry of Gender Equality and Family (2010). *First basic plan for multicultural family support policy (2010–2012)*.
- Ministry of Gender Equality and Family (2023). *Fourth basic plan for multicultural family policy (2023–2027)*.
- Mundlak, Y. (1978). “On the pooling of time series and cross section data,” *Econometrica* 46, 69–85.
- Park, M. H. (2017). “Structural relationships of factors affecting academic achievement of multicultural students,” *Multicultural Education Studies* 10, 27–55.
- Robinson, P. M. (1988). “Root- N -consistent semiparametric regression,” *Econometrica* 56, 931–954.
- Yeon, E. M. and Choi, H. S. (2020). “Estimating changes in academic achievement of multicultural adolescents and exploring predictive factors using piecewise growth modeling,” *Journal of the Korea Academia-Industrial Cooperation Society* 21, 370–379.
- Yun, Y. R. and Oh, B. H. (2020). “The effect of multicultural mentoring program participation on multicultural students’ academic achievement,” *Korean Journal of Elementary Education* 31, 1–16.