

Identification and estimation of semiparametric multilayered sample selection tools

Dongwoo Kim

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP10/26

IDENTIFICATION AND ESTIMATION OF SEMIPARAMETRIC MULTILAYERED SAMPLE SELECTION MODELS

DONGWOO KIM

Simon Fraser University and Korea University

Many selection problems are multilayered: agents first decide whether to participate and then sort among ordered or unordered categories. This paper shows that the sorting layer changes the geometry of identification. Unlike binary selection, in which selection bias can be summarized by a scalar control function, ordered and multinomial sorting generally produce multi-index control functions whose dimension determines the continuous covariate variation needed for identification. I establish matched non-identification and point-identification results for both architectures, showing how nonlinearity in the selection structure can substitute for excluded variables. I also show how additional structural restrictions reduce the control-function dimension and make estimation practical. I propose \sqrt{n} -consistent two-step sieve plug-in estimators and apply the framework to gender wage gaps among Korean college graduates. Accounting for sorting reshapes the entry-level gap along the firm-size margin, where the corrected female coefficient turns positive for large-firm employment.

1. INTRODUCTION

Heckman (1974, 1979) pioneered the econometric analysis of selection bias by modeling the joint determination of participation and outcomes. Heckman's insight has become one of the most influential ideas in microeconometrics. Yet Heckman's framework treats selection as a binary event: an individual either participates or not. In many empirical settings, the selection structure is considerably richer. For instance, conditional on entering the labor force, workers sort into specific occupations, firms, or industries. The outcome of interest is shaped by *two layers* of selection: participation and sorting among alternatives. Ignoring

Dongwoo Kim: dongwook@sfu.ca

The author gratefully acknowledges support from the Social Sciences and Humanities Research Council of Canada under the Insight Grant (435-2024-0322).

the sorting layer conflates within-occupation wage effects with between-occupation composition effects, potentially distorting policy implications.

This paper develops semiparametric models for *multilayered selection* that achieve point identification by leveraging nonlinearity in the selection mechanism rather than exclusion restrictions. I consider two distinct selection architectures. *Vertical sorting* arises when categories can be meaningfully ordered (for example, by job quality, firm size, or amenity provision) and can be modeled through ordered threshold-crossing processes. *Horizontal sorting* arises when categories are unordered (such as STEM vs. non-STEM jobs) and requires a multinomial choice framework. For each architecture, I characterize the resulting selection bias, establish conditions under which the outcome equation parameters are point identified, and propose computationally tractable sieve-based estimators.

The central identification challenge in multilayered selection is that the selection bias function generally depends on multiple selection indices: the threshold functions delineating categories or the utility indices governing multinomial choice. This substantially complicates identification relative to the binary selection case, where the bias can be a function of scalar selection probability. I establish identification results for both selection architectures. First, when the ordered selection process is governed by a single index, the selection bias reduces to a function of that index, so that a single continuously distributed covariate together with nonlinearity in the selection index suffices for point identification without an exclusion restriction. When the ordered selection process is fully nonparametric, the selection bias becomes a function of *two* indices simultaneously. I show that at least three continuous covariates are required to identify the outcome parameters, and that the requirement is binding by exhibiting an explicit non-identification result when fewer continuous covariates are available. This sharp increase in the identification requirement is a consequence of the richer index structure and is, to my knowledge, new in the literature.

Second, for horizontal sorting modeled as multinomial choice with $K + 1$ categories (including an outside option like unemployment), the bias correction function generally depends on K indices. I show that additional structural restrictions on the preference heterogeneity reduce the dimensionality of the selection bias. Under a multinomial logit se-

lection, the bias collapses to a single-index control function. A distinctive feature of the multinomial logit specification is that the nonlinearity condition for identification is automatically satisfied even with a linear utility specification. This contrasts with the ordered case, where the control function is approximately linear under Gaussian errors, making identification fragile without exclusion restrictions. Under a weaker exchangeability condition on the taste shocks, I exploit the theory of symmetric polynomials to approximate the bias by a function of a small number of elementary symmetric polynomials of the choice probabilities, providing a practical dimensional reduction that makes semiparametric estimation feasible even with moderately many choice categories.

For estimation, I propose two-step sieve plug-in estimators that can be implemented using standard software. The first step estimates the selection equation nonparametrically using sieves; the second step includes the estimated control functions as nonparametric regressors in the partially linear outcome regression, with heteroskedasticity-robust standard errors. I establish \sqrt{n} -consistency and asymptotic normality under an $o_p(n^{-1/4})$ rate condition on the first-stage sieve estimation. Monte Carlo simulations across seven data-generating processes confirm that the sieve estimators achieve near-oracle performance with correct coverage, and the corrections are robust to weak nonlinearity.

I apply the proposed framework to estimate the gender wage gap among college graduates in South Korea using the Graduates Occupational Mobility Survey (GOMS). Three selection architectures are implemented: an ordered model for firm-size sorting, and multinomial models for field (STEM vs. non-STEM) and sector (public vs. private) sorting. The uncorrected female hourly-wage penalty is around 5–6 log points in SMEs, non-STEM jobs, and the private sector, around 4 log points in large firms, and near zero (about 1 log point) in STEM and the public sector. After correction in the ordered model, the large-firm female coefficient turns positive, indicating a small conditional premium rather than a penalty, while the SME penalty is little changed. In the field and sector models the correction is modest. The corrected gap has also narrowed over the sample period (2008–2019). These empirical findings connect to a large literature on the gender wage gap, which has long recognized that selection into employment and across occupations is a first-order con-

cern for measuring the gap.¹ Existing corrections in this literature typically either impose parametric distributional assumptions (Mulligan and Rubinstein, 2008) or settle for partial identification (Blundell et al., 2007, Lee, 2009). The current application adds the intensive-margin sorting layer that these studies abstract from.

This paper also connects to several strands of the econometrics literature. In the binary selection setting, Heckman (1974, 1979) established the foundational control function approach under joint normality, while Chamberlain (1986), Ahn and Powell (1993), Powell, Stock, and Stoker (1989), Newey, Powell, and Walker (1990), Newey (2009), and Das, Newey, and Vella (2003) developed semiparametric/nonparametric alternatives, all requiring exclusion restrictions. Recent work has pursued identification without excluded variables through partial identification (Lee, 2009, Honoré and Hu, 2020), heteroscedasticity (Lewbel, 2007, Klein and Vella, 2010), and functional form variation: Escanciano, Jacho-Chávez, and Lewbel (2016) showed that nonlinearity in the selection mechanism can substitute for an exclusion restriction, and Kim and Lee (2025) applied this to the semiparametric selection model with a linear outcome equation. Pan and Zhang (2024) develop a related strategy using debiased machine learning. My paper generalizes the existing frameworks to multilayered selection, where fundamentally new identification arguments are needed when the bias depends on multiple indices.

For multinomial selection, Lee (1983) coupled a logit selection specification with joint normality, Dubin and McFadden (1984) relaxed the outcome error distribution, and Dahl (2002) introduced a semiparametric polynomial correction. Bourguignon, Fournier, and Gurgand (2007) compared and extended these approaches. All require exclusion restrictions or parametric distributional assumptions. Sheng and Sun (2025) exploit exchangeability and elementary symmetric polynomials for dimensionality reduction in social interaction models; I adapt this device to the multinomial sample selection setting. The Roy model tradition (Roy, 1951, Heckman, 1990, French and Taber, 2011, Heckman and Pinto,

¹See, for example, Neal (2004), Olivetti and Petrongolo (2008), Blau and Kahn (2017). Blau et al. (2024) provide recent evidence that selection into employment substantially affects measured gender wage gaps in the United States, finding that correcting for selection narrows the gap by 15–20% over a four-decade period.

2018) provides the theoretical foundation for sorting across sectors. [D’Haultfœuille and Maurel \(2013\)](#) obtain an identification result for a binary extended Roy model in which selection is driven by both potential earnings and a non-pecuniary cost. Exploiting that additive structure together with continuity of at least one covariate, they point-identify the non-pecuniary component without exclusion restrictions or large-support conditions. Identifying the covariate effects on sector-specific earnings, however, still requires either an exclusion restriction or an identification-at-infinity argument on potential outcomes. Most directly related is [Kroft, Mourifié, and Vayalinkal \(2024\)](#), who extend [Lee \(2009\)](#)’s non-parametric bounds to multilayered settings under conditional selection monotonicity. Their bounds impose minimal structural assumptions but can be wide in practice. My approach trades that nonparametric generality for point identification.

The remainder of the paper is organized as follows. Section 2 introduces the multilayered selection framework and establishes identification. Section 3 presents the estimators and their asymptotic properties. Section 4 evaluates finite-sample performance through simulations. Section 5 applies the method to gender wage gaps among Korean college graduates. Section 6 concludes. Technical proofs and additional simulation and empirical results are provided in the Appendix.

2. THE MULTILAYERED SELECTION MODELS

Consider a population of individuals indexed by i , each characterized by observable covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$, a discrete selection indicator $D_i \in \mathcal{C} := \{0, 1, \dots, K\}$, and potential outcomes $\{Y_{ik}^*\}_{k=1}^K$. D_i encodes two layers of choice: $D_i = 0$ indicates non-participation (e.g., unemployment), while $D_i = k$ for $k \geq 1$ indicates participation in category k . The potential outcomes for each k and the observed outcome Y_i are determined by

$$Y_{ik}^* = \alpha_k + X_i \beta_k + V_{ik}, \quad Y_i = \sum_{k=1}^K \mathbf{1}[D_i = k] \cdot Y_{ik}^*, \quad (2.1)$$

where $\beta_k \in \mathbb{R}^{d_X}$ is a category-specific parameter vector, α_k is a category-specific intercept, V_{ik} is unobserved heterogeneity with $E[V_{ik}|X_i] = 0$, and $\mathbf{1}[\cdot]$ is the indicator function.

Conditional on selection into category k , the expected observed outcome is

$$E[Y_i|X_i = x, D_i = k] = \alpha_k + x\beta_k + \underbrace{E[V_{ik}|X_i = x, D_i = k]}_{\text{selection bias}}, \quad (2.2)$$

where the selection bias captures the systematic difference in unobservable characteristics between individuals who select into category k and the population average. The object of interest is β_k , the effect of covariates on potential outcomes *within* category k .

The key modeling challenge is to specify the selection process generating D_i in a way that (i) allows the selection bias to be characterized by a tractable control function, (ii) permits point identification of β_k without exclusion restrictions, and (iii) accommodates the economic structure of occupational sorting. I consider two main architectures in turn: vertical and horizontal sorting.

2.1. Vertical sorting: ordered selection

2.1.1. Parametric ordered selection

Suppose the categories in \mathcal{C} are vertically differentiated, so that $D_i = k$ if the individual's "quality" index falls in the k -th interval of an ordered partition:

$$D_i = k \quad \text{if} \quad c_k \leq Z_i\gamma + \varepsilon_i < c_{k+1}, \quad (2.3)$$

where Z_i is a row vector of covariates affecting selection, γ is a parameter vector, ε_i is a mean-zero error term, and $-\infty = c_0 < c_1 < \dots < c_K < c_{K+1} = \infty$ are threshold parameters. When $\varepsilon_i \sim N(0, 1)$, this is the standard ordered probit model. The selection bias takes a known parametric form under joint normality of (V_{ik}, ε_i) that are independent of (X_i, Z_i) , with $\text{Corr}(V_{ik}, \varepsilon_i) = \rho_k$ and $\text{Var}(V_{ik})^{1/2} = \sigma_k$ as follows:

$$E[V_{ik}|Z_i = z, D_i = k] = \sigma_k \rho_k \cdot \frac{\phi(c_{k+1} - z\gamma) - \phi(c_k - z\gamma)}{\Phi(c_{k+1} - z\gamma) - \Phi(c_k - z\gamma)} =: \sigma_k \rho_k \cdot \lambda(z\gamma; c_k, c_{k+1}), \quad (2.4)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal p.d.f. and c.d.f. respectively. This generalizes the inverse Mills ratio in Heckman's binary model and provides a category-specific control function $\lambda(z\gamma; c_k, c_{k+1})$ that can be plugged into the outcome regression.

In the Heckman model, it is well known that identification without an exclusion restriction is fragile because the inverse Mills ratio is approximately linear over much of the effective support of $Z\gamma$ (Leung and Yu, 1996). This near-collinearity problem is equally severe, and arguably worse, in the ordered case. The control function $\lambda(z\gamma; c_k, c_{k+1})$ can exhibit a nearly linear relationship with the index $z\gamma$ (see Figure B.1 in Appendix B).

2.1.2. Semiparametric ordered selection

To relax joint normality without an exclusion restriction while retaining a tractable structure, I consider a semiparametric specification in which the distribution of ε_i is parametrically specified but the selection index function is left unrestricted:

$$D_i = k \quad \text{if} \quad c_k \leq h(X_i) + \varepsilon_i < c_{k+1}, \quad (2.5)$$

where $h : \mathcal{X} \rightarrow \mathbb{R}$ is an unknown smooth function. Under this specification, the selection bias conditional on $D_i = k$ becomes a function of the single index $h(x)$:

$$E[V_{ik} | X_i = x, D_i = k] = \frac{\int_{c_k - h(x)}^{c_{k+1} - h(x)} E[V_{ik} | \varepsilon_i = e] f_\varepsilon(e) de}{F_\varepsilon(c_{k+1} - h(x)) - F_\varepsilon(c_k - h(x))} =: \lambda_k(h(x)), \quad (2.6)$$

where f_ε and F_ε denote the p.d.f. and c.d.f. of ε_i . $\lambda_k(\cdot)$ is category-specific because the threshold constants c_k, c_{k+1} are category-specific. Consequently, the conditional mean of the observed outcome takes the partial linear form:

$$E[Y_i | X_i = x, D_i = k] = x\beta_k + \lambda_k(h(x)), \quad (2.7)$$

which is precisely the structure analyzed in Kim and Lee (2025). Under standard regularity conditions therein, β_k and λ_k are identified for each $k = 1, \dots, K$.²

²The regularity conditions require continuous variation in at least one covariate, smoothness of λ_k , no perfect multicollinearity, and a nonlinearity condition on the composite selection probability $p_k(x) = F_\varepsilon(c_{k+1} - h(x)) - F_\varepsilon(c_k - h(x))$. The intercept α_k is not separately identified from λ_k so is normalized to 0.

2.1.3. Nonparametric ordered selection

The semiparametric specification (2.5) restricts all threshold functions to shift in parallel through the common index $h(x)$. This is a substantive restriction as it requires, for example, that a covariate that makes an individual more likely to surpass the threshold into category 2 also makes them more likely to surpass the threshold into category 3, and by the same amount in index units. To relax this restriction, I consider a fully nonparametric ordered selection model following Chesher and Smolinski (2012):

$$D_i = k \quad \text{if} \quad h_k(X_i) \leq U_i < h_{k+1}(X_i), \quad (2.8)$$

where U_i is normalized to $\text{Unif}(0, 1)$, $X_i \perp\!\!\!\perp U_i$, $h_0(x) = 0$, $h_{K+1}(x) = 1$, and $0 < h_1(x) < h_2(x) < \dots < h_K(x) < 1$ for all x in the support. $h_k(\cdot)$ is now free to depend on x in an unrestricted manner. This model nests (2.5) as a special case.

The threshold functions are nonparametrically identified from the choice probabilities. Defining $\pi_j(x) := P[D_i = j | X_i = x]$, we have

$$h_k(x) = \sum_{j=0}^{k-1} \pi_j(x) = P[D_i \leq k-1 | X_i = x], \quad k = 1, \dots, K. \quad (2.9)$$

Under the nonparametric specification, the selection bias conditional on $D_i = k$ becomes:

$$E[V_{ik} | X_i = x, D_i = k] = \frac{\int_{h_k(x)}^{h_{k+1}(x)} E[V_{ik} | U_i = u] du}{h_{k+1}(x) - h_k(x)} =: \lambda_k(h_k(x), h_{k+1}(x)). \quad (2.10)$$

Crucially, the bias function now depends on *two* indices rather than the single index $h(x)$ that arose in the semiparametric case. This doubled index structure fundamentally changes the identification problem. Define the threshold mapping $H_k(x) := (h_k(x), h_{k+1}(x)) \in [0, 1]^2$. The conditional mean of the outcome can then be written as:

$$m_k(x) := E[Y_i | X_i = x, D_i = k] = x\beta_k + \lambda_k(H_k(x)). \quad (2.11)$$

I first establish that identification fails generically when H_k is injective.

PROPOSITION 1—Non-identification under injectivity: *Suppose H_k is injective on the support of X_i conditional on $D_i = k$, i.e., $H_k(x) = H_k(x')$ implies $x = x'$ almost surely. Then β_k is not identified: for every $\beta \in \mathbb{R}^{d_x}$, there exists a measurable function $\tilde{\lambda} : [0, 1]^2 \rightarrow \mathbb{R}$ such that $m_k(x) = x\beta + \tilde{\lambda}(H_k(x))$ almost surely conditional on $D_i = k$.*

Injectivity of H_k allows defining a matching $\tilde{\lambda}$ for any candidate β . Since $H_k : \mathbb{R}^{d_c} \rightarrow \mathbb{R}^2$ is generically injective when $d_c \leq 2$ and generically non-injective when $d_c \geq 3$, identification requires at least three continuous covariates, a sharp contrast to binary selection, where one suffices.³ The following assumption collects the conditions under which identification can be established with three continuous covariates.

ASSUMPTION 1: (i) *At least three variables $X_c := (X_1, X_2, X_3)$ in X are continuously distributed; let x_c denote a realized value. For $k = 1, \dots, K$: (ii) $h_k(x)$ and $h_{k+1}(x)$ are continuous on $\text{supp}(X_i | D_i = k)$ and continuously differentiable with respect to x_c almost everywhere; (iii) $\lambda_k(\cdot, \cdot)$ is continuously differentiable almost everywhere; (iv) the Jacobian matrix $J_k(x) := \partial H_k(x) / \partial x_c \in \mathbb{R}^{3 \times 2}$ has full column rank (rank 2) with probability one; (v) there exist three points $x^{(1)}, x^{(2)}, x^{(3)} \in \text{supp}(X_i | D_i = k)$ such that $\bigcap_{t=1}^3 \text{col}(J_k(x^{(t)})) = \{0\}$. (vi) The variables in X are not perfectly multicollinear. (vii) For each $k = 1, \dots, K$, the image $H_k(\text{supp}(X_i | D_i = k))$ is connected.*

Assumption 1 (ii)–(iii) require smoothness of selection indices and selection bias functions, and (iv) requires that h_k and h_{k+1} respond to the three continuous covariates in linearly independent directions. This rules out the case where h_{k+1} is merely a parallel shift of h_k by a constant (as in the semiparametric model). Assumption 1(v) is the key nonlinearity condition. Each Jacobian $J_k(x)$ has a two-dimensional column space in \mathbb{R}^3 , and hence a one-dimensional left null space. The condition requires that these one-dimensional null spaces, evaluated at three points, collectively span \mathbb{R}^3 . This is generically satisfied when the threshold functions exhibit sufficient nonlinearity. Assumption 1(vii) is a mild topological regularity condition: the selected sample explores the threshold-index space as a single

³A smooth map from \mathbb{R}^{d_c} to \mathbb{R}^2 with full-rank Jacobian is locally injective when $d_c \leq 2$ but has $(d_c - 2)$ -dimensional fibers when $d_c \geq 3$, making injectivity generically impossible.

connected region rather than as several isolated pieces. A simple sufficient condition is that $\text{supp}(X_i | D_i = k)$ itself is connected. Now I establish identification of β_k and λ_k in the following proposition.

PROPOSITION 2—Identification under nonparametric ordered selection: *Let Assumption 1 hold. Then β_k and λ_k are identified for each $k = 1, \dots, K$.*

When both threshold indices take linear form $h_k(x) = f_k(x' \gamma_k)$, the column space of $J_k(x)$ is a fixed plane in \mathbb{R}^3 regardless of x_c , violating Assumption 1(v). Continuous excluded variables can substitute for covariates, but the substitution rate is non-uniform. The first excluded variable replaces one continuous covariate, leaving the requirement $d_c^{\text{covariates}} + d_c^{\text{excluded}} \geq 3$ when $d_c^{\text{excluded}} \in \{0, 1\}$. The second excluded variable, however, replaces *two* continuous covariates: once $d_c^{\text{excluded}} \geq 2$, identification follows from the exclusion-based route in [Das, Newey, and Vella \(2003\)](#), and no continuous covariate is needed. The required number of continuous covariates is therefore three, two, and zero as d_c^{excluded} moves from zero to one to two, reflecting the regime switch from nonlinearity-based identification to exclusion-based identification.

2.2. Horizontal sorting: multinomial selection

When the categories in \mathcal{C} are horizontally differentiated (for example, broadly defined occupations or industries with no natural ordering), the ordered threshold-crossing framework is inappropriate. Instead, I model selection as a utility-maximizing multinomial choice:

$$D_i = k \quad \text{if} \quad u_k(X_i) + \varepsilon_{ik} \geq u_j(X_i) + \varepsilon_{ij} \quad \text{for all } j \neq k, \quad (2.12)$$

where $u_k(X_i)$ is the deterministic utility component for category k , and ε_{ik} captures preference heterogeneity. The individual chooses the category that maximizes utility.⁴ This

⁴The classical Roy model ([Roy, 1951](#)) is a special case of (2.12) in which $u_k(X_i) + \varepsilon_{ik} = Y_{ik}^* = \alpha_k + X_i \beta_k + V_{ik}$, so that selection is driven by potential wages alone. The generalized Roy model ([Heckman, 1990](#)) is the case $u_k(X_i) + \varepsilon_{ik} = Y_{ik}^* - C_{ik}$, allowing a non-pecuniary cost C_{ik} . The present framework accommodates both without requiring utility and potential wages to coincide.

framework is also considered in [Kroft, Mourifié, and Vayalinkal \(2024\)](#) as a parametric special case of their nonparametric multilayered selection model. Given the selection rule (2.12), the selection bias conditional on $D_i = k$ is

$$E[V_{ik}|X_i = x, D_i = k] = E[V_{ik}|u_k(x) - u_j(x) \geq \varepsilon_{ij} - \varepsilon_{ik}, \forall j \neq k] = \lambda_k((\delta_{kj}(x))_{j \neq k}), \quad (2.13)$$

where $\delta_{kj}(x) := u_k(x) - u_j(x)$ for $j \neq k$ and $\lambda_k : \mathbb{R}^K \rightarrow \mathbb{R}$ is an unknown function whose argument is the K -vector of pairwise differences indexed over the non-chosen alternatives. In general, the bias depends on K indices (one for each pairwise comparison), creating a severe curse of dimensionality as the number of categories grows.

An equivalent representation expresses the bias as a function of the choice probabilities rather than the utility differences. This reformulation is not required for identification, but it is practically useful in estimation.⁵ The equivalence, which holds under mild regularity conditions, provides the conceptual foundation for using estimated choice probabilities as control functions in the second stage.

PROPOSITION 3—Selection bias as a function of choice probabilities: *Suppose the preference shocks $(\varepsilon_{i0}, \dots, \varepsilon_{iK})$ have a joint density f_ε that is continuous and strictly positive on \mathbb{R}^{K+1} . Adopting the location normalization $u_0(x) \equiv 0$, the mapping*

$$\Psi : u = (u_1(x), \dots, u_K(x)) \mapsto (p_0(x), \dots, p_K(x)), \quad p_j(x) := P[D_i = j | X_i = x],$$

from the vector of normalized utilities to the choice probability vector is a diffeomorphism from \mathbb{R}^K onto the interior of the K -simplex $\Delta^K = \{p \in \mathbb{R}_+^{K+1} : \sum_{j=0}^K p_j = 1\}$. Consequently, the selection bias (2.13) can be equivalently written as

$$E[V_{ik}|X_i = x, D_i = k] = \tilde{\lambda}_k(p_0(x), \dots, p_K(x)), \quad (2.14)$$

⁵Choice probabilities are compactly supported on $[0, 1]$, whereas utility differences range over \mathbb{R} . Sieve smoothers approximating the selection bias function in the second stage behave substantially better in finite samples with a bounded argument than with an unbounded index.

where $\tilde{\lambda}_k := \lambda_k \circ L_k \circ \Psi^{-1}$ and $L_k : \mathbb{R}^K \rightarrow \mathbb{R}^K$ is the linear bijection $u \mapsto (u_k - u_j)_{j \neq k}$ (with $u_0 \equiv 0$) that converts normalized utilities into the pairwise-difference vector $(\delta_{kj})_{j \neq k}$ on which λ_k was originally defined. The resulting $\tilde{\lambda}_k$ is an unknown function of the choice probability vector.

The idea that choice probabilities can invert latent utility indices is well established in discrete choice and demand analysis.⁶ Proposition 3 establishes a general inversion result for the present static multinomial selection model under arbitrary continuous strictly positive joint densities of the preference shocks. This, in turn, justifies using estimated choice probabilities as control-function arguments in the sample-selection outcome equation and provides a formal foundation for the probability-based correction of Dahl (2002), who derives a single-index reduction under a maintained index-sufficiency assumption.

The conditional mean of the observed outcome now takes the partially linear form with K nonparametric indices:

$$E[Y_i | X_i = x, D_i = k] = x\beta_k + \tilde{\lambda}_k(p_1(x), \dots, p_K(x)), \quad (2.15)$$

where $p_0(x) = 1 - \sum_{j=1}^K p_j(x)$ is determined by the simplex constraint. Identification of β_k in this K -index model follows from the same argument as the nonparametric ordered case (Proposition 2), generalized from two to K indices.

ASSUMPTION 2: (i) At least $K + 1$ components of X_i are continuously distributed; denote the corresponding sub-vector by $x_c := (x_1, \dots, x_{K+1})$. For each k : (ii) $P_K(x) := (p_1(x), \dots, p_K(x))$ is continuous on $\text{supp}(X_i | D_i = k)$ and continuously differentiable in x_c almost everywhere, and $\tilde{\lambda}_k$ is continuously differentiable on $\text{int}(\Delta^K)$; (iii) the Jacobian $J_P(x) := \partial P_K(x) / \partial x_c \in \mathbb{R}^{(K+1) \times K}$ has full column rank K with probability one (so its left null space is one-dimensional); (iv) there exist $K + 1$ points $x^{(1)}, \dots, x^{(K+1)} \in \text{supp}(X_i | D_i = k)$ such that the associated left-null vectors $a(x^{(t)}) \in \mathbb{R}^{K+1}$ span \mathbb{R}^{K+1} ;

⁶For instance, Hotz and Miller (1993) derive CCP inversion in dynamic logit models, Berry (1994) establishes inversion in the multinomial logit demand system, and Berry, Gandhi, and Haile (2013) provide a more general demand-side invertibility result under connected substitutes.

(v) the support of $X_i \mid D_i = k$ is not contained in any affine hyperplane of \mathbb{R}^{d_X} ; (vi) the image $P_K(\text{supp}(X_i \mid D_i = k))$ is connected.

PROPOSITION 4—Identification under general multinomial selection: *Suppose (2.14) holds and Assumption 2 holds. Then β_k and $\tilde{\lambda}_k$ are identified on $P_K(\text{supp}(X_i \mid D_i = k))$ for each k .*

This result establishes a sharp trade-off between structural restrictions and covariate requirements: without additional restrictions, identification requires at least $K + 1$ continuous covariates. When K is moderate, this is feasible; when K is large, the requirement becomes prohibitive. The remainder of this section develops two structural restrictions that reduce the dimensionality of $\tilde{\lambda}_k$ and correspondingly lower the covariate requirement.

2.2.1. Multinomial logit selection

Suppose $u_k(X_i) = X_i \gamma_k$ and $\varepsilon_{i0}, \dots, \varepsilon_{iK}$ are independently and identically distributed as standard Extreme Value Type I (Gumbel). γ_0 is normalized to 0 for identification. I further assume that V_{ik} depends on $(\varepsilon_{i0}, \dots, \varepsilon_{iK})$ only through ε_{ik} :

$$E[V_{ik} \mid \varepsilon_{i0}, \dots, \varepsilon_{iK}, X_i] = E[V_{ik} \mid \varepsilon_{ik}] =: \mu(\varepsilon_{ik}). \quad (2.16)$$

In the occupational sorting context, this assumption means that the worker's unobserved productivity in occupation k only depends on their taste for that occupation. This is natural under an occupation-specific match quality interpretation: a worker with a strong affinity for a particular occupation (ε_{ik} large) tends to also be productive in that occupation (V_{ik} large), because taste and ability for a specific occupation are correlated. Once conditioned on ε_{ik} , the preference shocks for the other occupations ε_{ij} , $j \neq k$, carry no additional information about their productivity in k . While this restriction is substantive, it substantially simplifies the identification analysis.

REMARK 1: The own-shock restriction is natural when unobserved heterogeneity is sector-specific. For example, a worker with a strong taste for STEM occupations likely possesses STEM-relevant latent abilities that make them productive in STEM jobs. Con-

ditional on this STEM-specific taste, the worker's preferences for non-STEM alternatives (e.g., sales, marketing, or management) carry no additional information about their STEM productivity. The restriction is less plausible when unobserved general ability affects both preferences and productivity across all occupations. This restriction will be relaxed later in this section at the cost of additional continuous covariates. In the terminology of [Dubin and McFadden \(1984\)](#), the own-shock restriction corresponds to a block-diagonal covariance structure between the selection and outcome errors, whereas the Dubin-McFadden correction allows unrestricted covariance but requires exclusion restrictions for identification.

Under these assumptions, the conditional distribution of ε_{ik} given $D_i = k$ and $X_i = x$ admits a simple characterization. Let $U_{ij} = x\gamma_j + \varepsilon_{ij}$ denote the utility from category j , and let $M = \max_{0 \leq j \leq K} U_{ij}$. Since $\varepsilon_{ij} \sim \text{EV}(0, 1)$ independently, the cumulative distribution of the maximum is

$$F_M(m|x) = \prod_{j=0}^K \exp\{-e^{-(m-x\gamma_j)}\} = \exp\{-A(x)e^{-m}\},$$

where $A(x) := \sum_{j=0}^K e^{x\gamma_j} = 1 + \sum_{j=1}^K e^{x\gamma_j}$ is the logit denominator. Hence $M|x \sim \text{EV}(\ln A(x), 1)$. Conditional on $D_i = k$ and $X_i = x$, the winner's utility equals the maximum: $U_{ik} = M$. By the well-known property of Gumbel random variables that the conditional distribution of the maximum given which alternative wins depends on x only through $\ln A(x)$, we have $U_{ik}|D_i = k, X_i = x \sim \text{EV}(\ln A(x), 1)$. Since $\varepsilon_{ik} = U_{ik} - x\gamma_k$, a location shift gives

$$\varepsilon_{ik}|D_i = k, X_i = x \sim \text{EV}(\ln A(x) - x\gamma_k, 1). \quad (2.17)$$

The conditional density of ε_{ik} given selection into k thus depends on x only through the scalar index $\nu_k(x) := \ln A(x) - x\gamma_k$. Applying the restriction [\(2.16\)](#):

$$\begin{aligned} \lambda_k(\nu_k(x)) &:= E[V_{ik}|D_i = k, X_i = x] = \int \mu(\varepsilon) f_{\varepsilon_{ik}}(\varepsilon|D_i = k, X_i = x) d\varepsilon \\ &= \int \mu(\varepsilon) \exp\{-(\varepsilon - \nu_k(x))\} \exp\{-e^{-(\varepsilon - \nu_k(x))}\} d\varepsilon \end{aligned}$$

The selection bias reduces to a function of the single index $\nu_k(x) = \ln(1 + \sum_{j=1}^K e^{x\gamma_j}) - x\gamma_k$. The conditional mean of the outcome is therefore

$$E[Y_i|X_i = x, D_i = k] = x\beta_k + \lambda_k(\nu_k(x)), \quad (2.18)$$

which is again a partial linear model with a single-index control function.

REMARK 2—Connection to Dahl’s index sufficiency: Under multinomial logit, $p_k(x) = e^{x\gamma_k}/A(x)$, so $\nu_k(x) = -\ln p_k(x)$, and the inclusive value and the chosen probability are bijectively related. The single-index reduction in (2.18) is therefore equivalent to expressing the selection bias as a function of $p_k(x)$ alone, which is precisely the index-sufficiency assumption maintained by [Dahl \(2002\)](#).

Since (2.18) has the same partially linear structure as the semiparametric ordered model, identification of β_k and λ_k follows from [Kim and Lee \(2025\)](#) under the same regularity conditions. A distinctive feature of the multinomial logit model is that the nonlinearity condition is *automatically* satisfied, even with a linear specification $u_k(x) = x\gamma_k$ for the deterministic utility. The inclusive value $\nu_k(x)$ is inherently nonlinear in x because the log-sum-exp function is convex and not affine whenever at least two γ_j differ (see [Appendix A.1](#)). This contrasts sharply with the ordered case, where identification without exclusion restrictions is fragile. In the multinomial logit model, identification without exclusion restrictions is robust because the nonlinearity is a structural consequence of the multinomial choice mechanism. [Figure B.2](#) in [Appendix B](#) illustrates this inherent nonlinearity for several parameter configurations.

2.2.2. Semiparametric multinomial selection

The preceding analysis relies on two restrictions: the i.i.d. Gumbel assumption on the preference shocks, which implies the independence of irrelevant alternatives (IIA), and the own-shock restriction (2.16). IIA rules out correlation across alternatives, which is implausible when some occupations are closer substitutes than others. The own-shock restriction fails whenever unobserved ability has a general component that affects both preferences

and productivity across occupations. To accommodate richer dependence among the preference shocks, and between the preference shocks and outcome errors, while maintaining a tractable selection bias structure, I introduce an exchangeability condition under the more general nonparametric utility specification (2.12).

ASSUMPTION 3—Exchangeability: *The joint distribution of $(V_{ik}, \varepsilon_{i0}, \dots, \varepsilon_{iK})$ is invariant under permutations of the indices $\{j \neq k\}$. That is, for any permutation π of $\{0, \dots, K\} \setminus \{k\}$,*

$$f(V_{ik}, \varepsilon_{i0}, \dots, \varepsilon_{iK}) = f(V_{ik}, \varepsilon_{i,\pi(0)}, \dots, \varepsilon_{i,\pi(K)}),$$

where the permutation acts only on the indices different from k .

This assumption does not impose IIA, and it replaces the own-shock restriction with a weaker symmetry requirement that allows V_{ik} to depend on all preference shocks $(\varepsilon_{i0}, \dots, \varepsilon_{iK})$, provided this dependence is symmetric in the non-chosen alternatives. Under this assumption, for an individual contemplating category k , the alternative categories are ex ante symmetric in their unobserved preference. The exchangeability condition is more plausible with broadly defined categories (STEM vs. non-STEM) than with finely disaggregated ones (specific occupations within STEM), where nested structures are more natural. Under exchangeability, the selection bias function $\tilde{\lambda}_k(p_0(x), \dots, p_K(x))$ is symmetric in its non-chosen probabilities. The following proposition exploits this symmetry to reduce the dimensionality of the probability-based bias correction.

PROPOSITION 5—Dimensionality reduction under exchangeability: *Suppose (2.14) holds and Assumption 2(ii) and Assumption 3 hold. Then $\tilde{\lambda}_k(p_0, \dots, p_K)$ is a symmetric function of the non-chosen probabilities $(p_j)_{j \neq k}$. For any $\epsilon > 0$ and any compact domain $\mathcal{P} \subset \text{int}(\Delta^K)$, there exists a polynomial Q in the elementary symmetric polynomials of the non-chosen probabilities*

$$e_1 = \sum_{j \neq k} p_j = 1 - p_k, \quad e_2 = \sum_{\substack{i < j \\ i, j \neq k}} p_i p_j, \quad \dots, \quad e_K = \prod_{j \neq k} p_j, \quad (2.19)$$

such that $\sup_{p \in \mathcal{P}} |\tilde{\lambda}_k(p) - Q(e_1, \dots, e_K)| < \epsilon$.

This proposition provides a principled truncation strategy for the selection bias in terms of observable choice probabilities. For a given truncation order $L \leq K$, one can approximate $\tilde{\lambda}_k$ by a function of only (e_1, \dots, e_L) , discarding higher-order elementary symmetric polynomials. When L is small relative to K , this yields a substantial dimensionality reduction. In practice one truncates the polynomial approximation at some order $L \leq K$ in the elementary symmetric polynomials, treating $\tilde{\lambda}_k(p_0, \dots, p_K)$ as if it were exactly a function $\check{\lambda}_k^{(L)}(e_1, \dots, e_L)$ of only L arguments. The conditional mean of the outcome under this L -truncated working model is

$$E[Y_i | X_i = x, D_i = k] \approx x\beta_k + \check{\lambda}_k^{(L)}(e_1(p(x)), \dots, e_L(p(x))).$$

Write $E_L(x) := (e_1(p(x)), \dots, e_L(p(x)))$ for the vector of the first L elementary symmetric polynomials of the non-chosen probabilities. The identification of β_k under this approximation follows from the multi-index structure of the elementary symmetric polynomials. For $\ell = 1$, $e_1(p(x)) = 1 - p_k(x)$ is nonlinear in x whenever $p_k(x)$ is nonlinear, which holds generically; identification then follows from this single-index nonlinearity with one or two continuous covariates. For $\ell \geq 2$, e_ℓ is a polynomial of degree ℓ in the choice probabilities, and under the maintained rank and spanning conditions of the proposition below it supplies the multi-index nonlinearity needed for identification. For $L = 2$ the resulting two-index structure parallels the nonparametric ordered case and requires three continuous covariates. The general pattern is:

PROPOSITION 6—Identification under exchangeability with L -order truncation: *Consider the L -truncated working model.*

- (i) *If $L = 1$ and $e_1(p(x)) = 1 - p_k(x)$ is nonlinear in x , then β_k is identified with one or two continuous covariates, by Propositions 1–3 of [Kim and Lee \(2025\)](#).*
- (ii) *If $L \geq 2$, suppose: (a) there exist $L + 1$ continuous covariates $x_c = (x_1, \dots, x_{L+1})$ such that $E_L(x)$ is continuous on $\text{supp}(X_i | D_i = k)$ and continuously differentiable in x_c almost everywhere, and $\check{\lambda}_k^{(L)}$ is continuously differentiable on an open set con-*

taining $E_L(\text{supp}(X_i | D_i = k))$; (b) the analogues of Assumption 2(iii)–(vi) hold with (P_K, K) replaced by (E_L, L) . Then β_k is identified.

Without structural restrictions, the dimensionality of the bias grows with K , so in practice researchers should use a small number of broadly defined categories or impose the logit or exchangeability restriction.

3. ESTIMATION

This section develops two-step sieve plug-in estimators for each selection architecture and establishes their asymptotic properties. All models in the previous section produce a conditional mean of the form

$$E[Y_i | X_i = x, D_i = k] = x\beta_k + \lambda_k(g_k(x)), \quad (3.1)$$

where $g_k : \mathcal{X} \rightarrow \mathbb{R}^L$ is a vector of L indices and $\lambda_k : \mathbb{R}^L \rightarrow \mathbb{R}$ is an unknown smooth function. The number of indices L and the structure of g_k depend on the selection architecture. The same second stage is used across all architectures and is described first; let $\hat{g}_k(\cdot)$ denote the first-stage estimator obtained from the selection data $\{(D_i, X_i)\}_{i=1}^n$. For each category $k = 1, \dots, K$, use the subsample $\mathcal{I}_k := \{i : D_i = k\}$ with $n_k := |\mathcal{I}_k|$. Approximate $\lambda_k(\cdot)$ by a sieve basis and estimate β_k jointly:

$$Y_i = X_i\beta_k + B_{J_n}^{(L)}(\hat{g}_k(X_i))'\delta_k + \eta_{ik}, \quad i \in \mathcal{I}_k, \quad (3.2)$$

where $B_{J_n}^{(L)}(\cdot)$ is a sieve basis of dimension κ_n for the L -variate function λ_k , and δ_k is the vector of sieve coefficients. The estimator of β_k is the ordinary least squares (OLS) coefficient on X_i from the regression (3.2).

I specify the sieve basis as follows. For $L = 1$ (single-index models), let $B_{J_n}^{(1)}(s) = (B_1(s), \dots, B_{J_n}(s))'$ be a univariate B-spline basis of order r with J_n interior knots placed on the support of $\hat{g}_k(X_i)$. The sieve dimension is $\kappa_n = J_n + r$. For $L = 2$ (nonparametric ordered model), use a tensor product B-spline basis:

$$B_{J_n}^{(2)}(s_1, s_2) = \left(B_{j_1}^{[1]}(s_1) \cdot B_{j_2}^{[2]}(s_2) \right)_{1 \leq j_1 \leq J_n^{[1]} + r, 1 \leq j_2 \leq J_n^{[2]} + r}, \quad (3.3)$$

where $B^{[\ell]}$ denotes a univariate B-spline basis with $J_n^{[\ell]}$ interior knots in the ℓ -th direction. The sieve dimension is $\kappa_n = (J_n^{[1]} + r)(J_n^{[2]} + r)$. For $L \geq 3$, the tensor product construction extends naturally, with dimension $\kappa_n = \prod_{\ell=1}^L (J_n^{[\ell]} + r)$. Define the augmented regressor vector for observation $i \in \mathcal{I}_k$:

$$W_{ik} := \left(X_i, B_{J_n}^{(L)}(\hat{g}_k(X_i))' \right)' \in \mathbb{R}^{d_X + \kappa_n},$$

and write the second-stage regression compactly as $Y_i = W_{ik}'\theta_k + \eta_{ik}$ where $\theta_k = (\beta_k', \delta_k')' \in \mathbb{R}^{d_X + \kappa_n}$. The OLS estimator is

$$\hat{\theta}_k = \left(\sum_{i \in \mathcal{I}_k} W_{ik} W_{ik}' \right)^{-1} \sum_{i \in \mathcal{I}_k} W_{ik} Y_i, \quad (3.4)$$

and $\hat{\beta}_k$ is the subvector of $\hat{\theta}_k$ corresponding to X_i .

In the following subsection, I present details of the first stage estimation procedure for each selection architecture.

3.1. First-stage estimation details

Let $Q_n(x) = (q^{(1)}(x), \dots, q^{(q_n)}(x))'$ be a sieve basis. For the *semiparametric ordered model*, let θ denote the parameter vector $(\alpha, c_1, \dots, c_K)$. The sieve MLE maximizes the ordered choice log-likelihood:

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \sum_{k=0}^K \mathbf{1}[D_i = k] \log [F_{\varepsilon}(c_{k+1} - Q_n(X_i)'\alpha) - F_{\varepsilon}(c_k - Q_n(X_i)'\alpha)], \quad (3.5)$$

yielding $\hat{h}(x) = Q_n(x)'\hat{\alpha}$. With $F_{\varepsilon} = \Phi$, this is a sieve ordered probit. For the *nonparametric ordered model*, define $\tilde{D}_{ik} = \mathbf{1}[D_i \leq k - 1]$ and estimate each threshold function by sieve logistic regression:

$$\hat{\alpha}_k = \arg \max_{\alpha \in \mathbb{R}^{Q_n}} \sum_{i=1}^n \left[\tilde{D}_{ik} \log \Lambda(Q_n(X_i)'\alpha) + (1 - \tilde{D}_{ik}) \log(1 - \Lambda(Q_n(X_i)'\alpha)) \right], \quad (3.6)$$

giving $\hat{h}_k(x) = \Lambda(Q_n(x)' \hat{\alpha}_k)$ and index vector $\hat{g}_k(x) = (\hat{h}_k(x), \hat{h}_{k+1}(x))$.⁷

For the *multinomial logit model*, the workhorse first-stage estimator is a *sieve multinomial logit*: replace the linear index $X_i \gamma_k$ with a flexible sieve approximation $Q_n(X_i)' \alpha_k$ and maximize the MNL log-likelihood

$$(\hat{\alpha}_1, \dots, \hat{\alpha}_K) = \arg \max_{(\alpha_1, \dots, \alpha_K)} \sum_{i=1}^n \sum_{k=0}^K \mathbf{1}[D_i = k] \log \left(e^{Q_n(X_i)' \alpha_k} / \sum_{j=0}^K e^{Q_n(X_i)' \alpha_j} \right), \quad (3.7)$$

yielding $\hat{\nu}_k(x) = \ln \sum_{j=0}^K e^{Q_n(x)' \hat{\alpha}_j} - Q_n(x)' \hat{\alpha}_k$ and $\hat{p}_j(x) = \text{softmax}(Q_n(x)' \hat{\alpha})_j := e^{Q_n(x)' \hat{\alpha}_j} / \sum_{\ell=0}^K e^{Q_n(x)' \hat{\alpha}_\ell}$, with the baseline normalization $\hat{\alpha}_0 = 0$. For the *exchangeability model*, the same sieve MNL is used; the estimated choice probabilities are then used to form the elementary symmetric polynomials of the non-chosen probabilities:

$$\hat{e}_\ell(x) = e_\ell((\hat{p}_j(x))_{j \neq k}), \quad \ell = 1, \dots, L. \quad (3.8)$$

REMARK 3—Sieve MNL first-stage: The softmax link enforces $\hat{p}_k(x) \in (0, 1)$ and $\sum_k \hat{p}_k(x) = 1$, and the log-likelihood (3.7) is globally concave. With a sufficiently rich sieve basis, the sieve MNL is a universal approximator for conditional probability simplices: for any continuous conditional probability vector $p(x)$ on a compact support, there exist sieve coefficients such that $\sup_x |p_k(x) - \text{softmax}(Q_n(x)' \alpha_k)| \rightarrow 0$ as the sieve dimension $q_n \rightarrow \infty$.⁸ Since the sieve MNL approximates the true choice probabilities regardless of the error distribution, the estimator is robust to misspecification of the selection equation.

3.2. Asymptotic results

Here I derive asymptotic properties of the proposed estimators under regularity conditions. I state the assumptions in a unified manner, applying to all selection architectures.

⁷The separate estimation does not automatically enforce $\hat{h}_1(x) < \dots < \hat{h}_K(x)$; violations are rare in practice and can be corrected by rearrangement (Chernozhukov, Fernández-Val, and Galichon, 2010).

⁸Formally, any strictly positive continuous probability vector $p(x)$ can be represented as $p_k(x) = \exp(f_k(x)) / \sum_j \exp(f_j(x))$ for $f_k(x) = \log p_k(x) - \log p_0(x)$ (with $f_0 = 0$). By the Stone-Weierstrass theorem, each f_k can be uniformly approximated by elements of the sieve space, and continuity of the softmax yields uniform approximation of p .

ASSUMPTION 4—Sampling: $\{(Y_i, X_i, D_i)\}_{i=1}^n$ are i.i.d. with $P[D_i = k|X_i] > c_\pi > 0$ a.s. for each $k = 1, \dots, K$ (overlap), and $E[Y_i^4|X_i, D_i = k] < \bar{M} < \infty$ a.s.

Let $\mathcal{H}^m(\mathcal{S})$ denote the Hölder ball of order m on \mathcal{S} : the set of functions whose partial derivatives up to order $\lfloor m \rfloor$ are bounded and whose $\lfloor m \rfloor$ -th partial derivatives are Hölder continuous of order $m - \lfloor m \rfloor$.

ASSUMPTION 5—Bias function smoothness: For each k , $\lambda_{k0} \in \mathcal{H}^{m_\lambda}(\mathcal{G}_k)$ with $m_\lambda \geq 2$ when $L = 1$ and $m_\lambda > L$ when $L \geq 2$, where $\mathcal{G}_k := g_{k0}(\text{supp}(X|D = k)) \subset \mathbb{R}^L$ is compact with nonempty interior.

ASSUMPTION 6—Sieve approximation: The tensor-product B-spline basis of order $r \geq m_\lambda$ has interior knot numbers $J_n^{[\ell]}$ satisfying: (i) $J_n^{[\ell]} \rightarrow \infty$; (ii) $\kappa_n/n_k \rightarrow 0$ where $\kappa_n = \prod_\ell (J_n^{[\ell]} + r)$; (iii) $\kappa_n^2 \log n/n_k \rightarrow 0$; (iv) $n_k (J_n^{[\ell]})^{-2m_\lambda} \rightarrow 0$ (undersmoothing). For $L = 1$, these are jointly satisfied by any $J_n \asymp n^a$ with $a \in (1/(2m_\lambda), 1/2)$, a nonempty range whenever $m_\lambda \geq 2$. For $L = 2$ the joint range is $a \in (1/(2m_\lambda), 1/4)$, which is nonempty if and only if $m_\lambda > 2$; the boundary case $m_\lambda = 2$ admits no valid knot sequence, reflecting the curse of dimensionality in multi-index sieve estimation.

ASSUMPTION 7—First-stage convergence rate: The first-stage estimator satisfies

$$\|\hat{g}_k - g_{k0}\|_\infty := \sup_{x \in \mathcal{X}} \|\hat{g}_k(x) - g_{k0}(x)\| = o_p(n^{-1/4}). \quad (3.9)$$

The following lemma establishes sufficient conditions for the first stage rate condition (Assumption 7) in each selection architecture.

- LEMMA 1: (a) Semiparametric ordered. Under $h \in \mathcal{H}^{m_h}(\mathcal{X})$ with $m_h > d_c/2$ and standard sieve conditions (Chen, 2007), $\|\hat{h} - h\|_\infty = o_p(n^{-1/4})$.
- (b) Nonparametric ordered. If $h_k \in \mathcal{H}^{m_h}(\mathcal{X})$ with $m_h > d_c/2$ and $Q_n \asymp n^{d_c/(2m_h+d_c)}$, then $\max_k \|\hat{h}_k - h_k\|_\infty = o_p(n^{-1/4})$.
- (c) Multinomial logit. \sqrt{n} -consistency of the MLE and Lipschitz continuity of ν_k give $\|\hat{\nu}_k - \nu_{k0}\|_\infty = O_p(n^{-1/2})$.
- (d) Exchangeability. The delta method gives $\|\hat{e}_\ell - e_{\ell 0}\|_\infty = O_p(n^{-1/2})$.

REMARK 4—Sieve multinomial first stage: Parts (c) and (d) assume a parametric MNL with \sqrt{n} -consistent MLE. With a sieve basis, the estimated utility index converges at a non-parametric rate; Assumption 7 remains satisfied when $m_u > d_c/2$ and $Q_n \asymp n^{d_c/(2m_u+d_c)}$, where m_u is the smoothness of the utility index (Chen, 2007).

ASSUMPTION 8—Rank condition: For each k , $\Sigma_k := E[\tilde{X}_{ik}\tilde{X}'_{ik} \mid D_i = k]$ is positive definite, where $\tilde{X}_{ik} := X_i - \Pi_k(g_{k0}(X_i))$ is the residual from projecting X_i onto the closure of $\text{span}\{B_j^{(L)}(g_{k0}(\cdot))\}_{j \geq 1}$ in $L^2(X \mid D = k)$.

This rank condition is guaranteed by the identification conditions of Section 2 and those of Propositions 1–3 in Kim and Lee (2025).

ASSUMPTION 9—Error moments: For each k , let $\varepsilon_{ik} := Y_i - X_i\beta_{k0} - \lambda_{k0}(g_{k0}(X_i))$. Then (i) $E[\varepsilon_{ik} \mid X_i, D_i = k] = 0$ a.s.; (ii) $E[\varepsilon_{ik}^2 \mid X_i, D_i = k] = \sigma_k^2(X_i)$ is bounded and bounded away from zero; (iii) $E[\varepsilon_{ik}^4 \mid X_i, D_i = k] \leq \bar{M}$ a.s.

The following theorem establishes the \sqrt{n} -consistency and asymptotic normality of $\hat{\beta}_k$.

THEOREM 1—Asymptotic normality: Under Assumptions 4–9,

$$\sqrt{n_k} \left(\hat{\beta}_k - \beta_{k0} \right) \xrightarrow{d} N(0, V_k), \quad (3.10)$$

where $V_k := \Sigma_k^{-1} \Omega_k \Sigma_k^{-1}$ and

$$\Sigma_k = E \left[\tilde{X}_{ik} \tilde{X}'_{ik} \mid D_i = k \right], \quad \Omega_k = E \left[\tilde{X}_{ik} \tilde{X}'_{ik} \sigma_k^2(X_i) \mid D_i = k \right],$$

with $\tilde{X}_{ik} = X_i - \Pi_k(g_{k0}(X_i))$ the residual from projecting X_i onto the closure of the L -variate tensor-product sieve space in $L^2(X \mid D = k)$ (with the convention that $L = 1$ gives a univariate sieve).

The first-stage estimation error is asymptotically negligible under (3.9), so the feasible estimator has the same asymptotic distribution as the oracle using the true g_{k0} . Under homoskedasticity, the asymptotic variance simplifies to $V_k = \sigma_k^2 \Sigma_k^{-1}$. The following corollary establishes that the sieve plug-in estimator achieves the semiparametric efficiency bound under homoskedasticity.

COROLLARY 1—Semiparametric efficiency: *Suppose the conditions of Theorem 1 hold and the conditional variance is homoskedastic ($\sigma_k^2(X_i) = \sigma_k^2$ a.s.). Then the oracle estimator $\tilde{\beta}_k$ in (A.9) attains the partially-linear semiparametric efficiency bound $V_k = \sigma_k^2 \Sigma_k^{-1}$ of Chamberlain (1992), Robinson (1988), and the feasible plug-in $\hat{\beta}_k$ attains the same bound because $\sqrt{n_k}(\hat{\beta}_k - \tilde{\beta}_k) = o_p(1)$.*

The FWL residualization $\tilde{X}_{ik} = X_i - \Pi_k(g_{k0}(X_i))$ is the efficient influence function for the partially linear model, and the sieve basis consistently approximates the projection Π_k . Models with fewer indices ($L = 1$) generically yield smaller asymptotic variance than those with more, reflecting the statistical cost of the structural restrictions needed to reduce the dimensionality of the bias function.⁹

From the asymptotic distribution, I derive the consistent variance estimator in the following theorem.

THEOREM 2—Variance: *Under the conditions of Theorem 1, define*

$$\hat{\Sigma}_k := \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \hat{X}_{ik} \hat{X}_{ik}', \quad \hat{\Omega}_k := \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \hat{X}_{ik} \hat{X}_{ik}' \hat{\varepsilon}_{ik}^2, \quad \hat{V}_k := \hat{\Sigma}_k^{-1} \hat{\Omega}_k \hat{\Sigma}_k^{-1}, \quad (3.11)$$

where $\hat{X}_{ik} := X_i - \hat{B}_i'(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}'\mathbf{X}$ and $\hat{\varepsilon}_{ik} := Y_i - X_i\hat{\beta}_k - \hat{B}_i'\hat{\delta}_k$ with $\hat{B}_i := B^{(L)}(\hat{g}_k(X_i))$. Then $\hat{V}_k \xrightarrow{p} V_k$. Under homoskedasticity:

$$\hat{V}_k^{\text{hom}} = \hat{\sigma}_k^2 \hat{\Sigma}_k^{-1}, \quad \hat{\sigma}_k^2 = \frac{1}{n_k - d_X - \kappa_n} \sum_{i \in \mathcal{I}_k} \hat{\varepsilon}_{ik}^2. \quad (3.12)$$

As the estimator is a plain OLS on the augmented design $W_{ik} = (X_i, B^{(L)}(\hat{g}_k(X_i)))'$, the inference object for β_k is the heteroskedasticity-robust sandwich variance \hat{V}_k from (3.11), obtained directly using standard software.

⁹Under heteroskedasticity, $V_k = \Sigma_k^{-1} \Omega_k \Sigma_k^{-1}$ remains valid but $\hat{\beta}_k$ is no longer efficient. Efficient estimation under heteroskedasticity would require weighted least squares with estimated conditional variance $\sigma_k^2(X_i)$.

4. SIMULATIONS

This section provides Monte Carlo evidence on the finite-sample performance of the proposed estimators across selection architectures. The baseline simulations use $n = 5,000$ observations and 500 replications, with cubic splines throughout. For each data-generating process (DGP) and estimator, I report the root mean squared error (RMSE), mean bias, and empirical coverage probability of the 95% confidence interval.

4.1. Ordered selection

Two DGPs test ordered selection models with $K = 2$ occupation categories. Four estimators are compared in the main text: (i) OLS, which ignores selection; (ii) Linear, a parametric selection correction using a linear ordered probit; (iii) Oracle, an infeasible benchmark using the true correction function in the second stage; and (iv) Sieve, the proposed estimator using a nonparametric first stage and a sieve approximation to $\lambda_k(\hat{h}_k(x), \hat{h}_{k+1}(x))$ in the second stage.

DGP1: Two continuous covariates. The selection mechanism follows an ordered threshold-crossing model with a nonlinear index:

$$D_i = k \quad \text{if} \quad c_k \leq 0.5X_i - 0.5X_i^2 + 0.2X_i^3 + 0.5X_iZ_i + Z_i - 0.5Z_i^2 + U_i < c_{k+1},$$

with thresholds $c_1 = -1.5$ and $c_2 = 0.5$. The covariates $(X_i, Z_i) \sim N(\mathbf{0}, I_2)$ are independent, and the errors (U_i, V_{i1}, V_{i2}) are jointly normal with correlation 0.75 between the selection error and each outcome error. The outcome equations are $Y_{i1} = 0.5 + 0.5X_i + 0.25Z_i + V_{i1}$ and $Y_{i2} = 0.6 + 0.7X_i + 0.5Z_i + V_{i2}$.

DGP2: Mixed covariates. This DGP modifies DGP1 by replacing the continuous covariate Z_i with a binary indicator $Z_i = \mathbf{1}[Z'_i > 0]$ where $Z'_i \sim N(0, 1)$, and enriching the selection index with interaction terms $X_i^2Z_i$ and $X_i^3Z_i$:

$$\tilde{h}(X_i, Z_i) = -0.2X_i - 0.5X_i^2 + 0.3X_i^3 + 0.1X_iZ_i + 0.5Z_i - 0.3X_i^2Z_i + 0.2X_i^3Z_i.$$

TABLE I
ORDERED SELECTION: RMSE, BIAS, AND COVERAGE

		DGP1: continuous X, Z				DGP2: continuous X , binary Z			
		OLS	Linear	Oracle	Sieve	OLS	Linear	Oracle	Sieve
Occ. 1	RMSE	0.701	5.581	0.064	0.091	0.177	5.811	0.031	0.060
	Bias	0.700	2.026	0.001	0.005	0.175	5.265	0.002	0.010
	Coverage	0.000	0.901	0.904	0.884	0.000	0.408	0.899	0.922
Occ. 2	RMSE	0.550	0.383	0.136	0.153	0.184	0.827	0.058	0.077
	Bias	0.542	0.110	0.004	0.012	0.177	0.765	0.001	0.004
	Coverage	0.000	0.927	0.926	0.928	0.057	0.308	0.939	0.945

Note: OLS ignores selection. Linear denotes the parametric control-function estimator based on a linear ordered probit first stage. Oracle denotes the infeasible benchmark that uses the true selection correction function in the second stage. Sieve denotes the proposed sieve plug-in estimator, with a nonparametric first stage.

Table I reports the occupation-level RMSE, absolute bias, and coverage across DGPs. In DGP1, the OLS estimator exhibits substantial bias, confirming large selection bias. The Linear estimator fails catastrophically for Occupation 1, reflecting the severe mismatch between the linear index and the true nonlinear index. For Occupation 2, the Linear estimator is less extreme but still substantially biased. The Oracle estimator is nearly unbiased with tight dispersion and the Sieve estimator tracks the oracle very closely, with RMSE only slightly larger than the infeasible benchmark. In DGP2 the binary covariate complicates the within-category support of the first-stage probability vector, but the Sieve estimator remains near-oracle on both occupations. Coverage is at or near the nominal 95% level for Oracle and Sieve in all four cells.

REMARK 5—Effective control-function dimension in ordered selection: The sieve second stage approximates λ_k with a bivariate cubic B-spline tensor basis. DGPs 1–2 carry only two and one continuous covariate respectively, so the identification condition (3 continuous covariates) is not met. Nonetheless, Sieve achieves near-oracle performance in both cases. The mechanism is that $(\hat{h}_k(X), \hat{h}_{k+1}(X))$ is not a genuinely two-dimensional first-stage index in the DGPs: both components are deterministic monotone transformations of

the same scalar index $h(X) + U$, since $h_k(X) = F(c_k - h(X))$. The pair therefore lies on the one-dimensional parametric curve $\{(F(c_k - t), F(c_{k+1} - t)) : t \in \mathbb{R}\} \subset [0, 1]^2$, and any smooth function λ_k evaluated on this curve collapses to a smooth function of $h(X)$ alone. The bivariate basis is functionally equivalent to a univariate sieve in $h(X)$ and hence the effective control-function dimension is $L = 1$.

Additional simulation results with $K = 3$ are reported in Appendix C: Ordered DGP3 ($K = 3$, three continuous covariates) showcases near-oracle Sieve performance at full non-linearity and stress-tests weak nonlinearity by scaling the higher-order terms of the selection index by δ . As the selection index becomes nearly linear, even the Oracle’s variance inflates; the Sieve estimator essentially matches the Oracle for moderate nonlinearity and incurs a finite-sample bias penalty only in the most adversarial near-linear regime, while remaining far better behaved than the parametric Heckman-type estimator throughout.

4.2. Multinomial selection

Four DGPs are considered for multinomial selection: a baseline $K = 2$ design under IIA, two exchangeable designs with $K = 3$, and a non-exchangeable design with $K = 3$. Table II summarizes the key features of the DGPs. Five estimators are compared in the main text: (i) OLS, which ignores selection; (ii) MLogit, using the sieve-estimated inclusive value $\hat{v}_k(x)$ as a single-index control function under the own-shock restriction; (iii) Oracle, an infeasible benchmark using the true inclusive value as the linear correction (DGPs 1–2, where the own-shock restriction holds) or the true choice-probability vector with the Sieve second-stage (DGPs 3–4, where own-shock fails); (iv) Sieve, using the sieve MNL predicted probability vector through cubic B-spline marginals plus pairwise tensor interactions in the second stage; and (v) Exch- $L2$ (only for DGPs 2–4 with $K = 3$), using the first two elementary symmetric polynomials (\hat{e}_1, \hat{e}_2) of the sieve MNL choice probabilities under exchangeability.

TABLE II

SUMMARY OF MULTINOMIAL SELECTION DGPs

DGP	K	Covariates	Preference shocks	IIA	Exch.	Own-shock
1	2	2 continuous	i.i.d. Gumbel	✓	✓	✓
2	3	3 continuous	i.i.d. Gumbel	✓	✓	✓
3	3	3 continuous	equicorrelated normal	—	✓	—
4	3	3 continuous	non-exch. factor model	—	—	—

Notes: ✓ indicates the condition holds; — indicates it is violated. “IIA” refers to the choice-probability ratio property of multinomial logit. “Exch.” is the exchangeability condition (Assumption 3). “Own-shock” is the restriction (2.16) that V_{ik} depends on the preference vector only through ε_{ik} .

DGP1 defines the selection procedure with $K = 2$ following utility maximization, $D_i = \arg \max_{j \in \{0,1,2\}} \{f_j(X_i, Z_i) + \varepsilon_{ij}\}$, where $f_0 \equiv 0$ and $\varepsilon_{ij} \sim \text{Gumbel}(0, 1)$. Both covariates are independently drawn from the standard normal distribution and the utility functions f_j are polynomials of degree 3 in X and degree 2 in Z with interaction terms. The outcome errors satisfy $V_{ik} = \varepsilon_{ik} + \tilde{\varepsilon}_{ik}$ with $\tilde{\varepsilon}_{ik} \sim \text{Gumbel}(0, 1)$ independent, inducing correlation between selection and outcome errors. The outcome parameters are $\beta_1 = (0.5, 0.7)'$ and $\beta_2 = (0.8, 0.5)'$. DGP1 satisfies IIA, exchangeability, and the own-shock restriction.

DGP2 features $K = 3$ and *three* continuous covariates $(X_i, Z_i, W_i) \sim N(\mathbf{0}, I_3)$. The inclusion of three continuous covariates is dictated by the identification theory: by Proposition 6(ii), the $L = 2$ truncation of the elementary symmetric polynomial correction requires $L + 1 = 3$ continuous covariates for identification. The utilities are $U_{ij} = f_j(X_i, Z_i, W_i) + \varepsilon_{ij}$ for $j \in \{0, 1, 2, 3\}$, where the utility functions f_j include quadratic terms and pairwise interactions among the three covariates, and the ε_{ij} are i.i.d. Gumbel. The outcome equation for each k is: $Y_{ik} = \alpha_k + \beta_{k1}X_i + \beta_{k2}Z_i + \beta_{k3}W_i + V_{ik}$, where $(\alpha_1, \beta_{11}, \beta_{12}, \beta_{13}) = (0.4, 0.5, 0.7, 0.3)$, $(\alpha_2, \beta_{21}, \beta_{22}, \beta_{23}) = (0.6, 0.8, 0.5, 0.4)$, and $(\alpha_3, \beta_{31}, \beta_{32}, \beta_{33}) = (0.5, 0.3, 0.9, 0.6)$. The selection bias for each category is approximated using elementary symmetric polynomials of the choice probabilities, truncated at order $L = 2$ as in the L -truncated working model. This DGP tests whether the exchangeability approximation is effective when the number of occupations exceeds two and the

required continuous variation for identification is available. IIA, exchangeability, and the own-shock restriction hold in this DGP.

DGP3 is designed to showcase the practical value of the exchangeability framework when the own-shock restriction and IIA fail but exchangeability holds. The preference shocks are equicorrelated normal, $\varepsilon_{ij} = \sqrt{\rho} c_i + \sqrt{1-\rho} z_{ij}$, $\forall j \in \{0, 1, 2, 3\}$, where $c_i \sim N(0, 1)$ is a common factor, $z_{ij} \sim N(0, 1)$ are independent, and $\rho = 0.5$. This symmetric structure is exchangeable by construction but violates IIA. The covariates, utility functions, and outcome equations are identical to DGP2. The critical departure from DGP2 is in the outcome errors, which introduce self-reinforcing specialization: $V_{ik} = \varepsilon_{ik} + \gamma_{sr} \varepsilon_{ik} (\varepsilon_{ik} - \bar{\varepsilon}_{i,-k}) + \sigma_\eta \eta_{ik}$, where $\bar{\varepsilon}_{i,-k} = K^{-1} \sum_{j \neq k} \varepsilon_{ij}$ is the mean preference shock of competing alternatives, $\gamma_{sr} = 1.0$ controls the specialization intensity, $\sigma_\eta = 0.2$, and $\eta_{ik} \sim N(0, 1)$. Workers whose preference shock ε_{ik} exceeds the competition average receive amplified productivity in occupation k , creating positive selection. This violates the own-shock restriction because V_{ik} depends on all ε_{ij} through $\bar{\varepsilon}_{i,-k}$. However, the dependence on competing shocks is symmetric, preserving exchangeability.

Lastly, DGP4 breaks IIA, exchangeability, and the own-shock restriction all together to assess the boundary of the proposed methods while keeping $K = 3$. The preference shocks follow a single-factor model: $\varepsilon_{ij} = \lambda_j f_i + u_{ij}$, $\forall j \in \{0, 1, 2, 3\}$, where $f_i \sim N(0, 1)$ is a common factor, $u_{ij} \sim N(0, 1)$ are independent. As the loadings $\lambda = (0, 0.3, 0.8, -0.5)$ are *heterogeneous* across alternatives, the factor does not cancel in utility differences: $\text{Cov}(\varepsilon_{ij} - \varepsilon_{ik}, \varepsilon_{il} - \varepsilon_{ik}) = (\lambda_j - \lambda_k)(\lambda_l - \lambda_k)$ depends on the identities of j and l , not just their count, violating exchangeability. Occupation 2 (loading $\lambda_2 = 0.8$) and occupation 3 ($\lambda_3 = -0.5$) have the most dissimilar loadings and thus the weakest substitutability, while occupation 1 ($\lambda_1 = 0.3$) is closer to the outside option ($\lambda_0 = 0$). The covariates, utility specifications, and outcome equations are identical to DGP2–3. The outcome error follows the same self-reinforcing specialization structure as DGP3, with $\gamma_{sr} = 2.0$ and $\sigma_\eta = 0.2$.

Table III reports the RMSE, bias, and coverage averaged across all occupations and coefficients within the four DGPs. DGPs 1–2 show a consistent pattern. The OLS is severely biased, whereas the MLogit estimator (correctly specified) exhibits near-zero bias. The or-

TABLE III

MULTINOMIAL SELECTION: RMSE, BIAS, AND COVERAGE PROBABILITY

	OLS	MLogit	Oracle	Sieve	Exch- $L2$
<i>DGP1: IIA, $K = 2$</i>					
RMSE	0.173	0.062	0.055	0.117	—
Bias	0.159	0.003	0.002	0.010	—
Coverage	0.437	0.930	0.952	0.925	—
<i>DGP2: IIA, $K = 3$, exchangeability</i>					
RMSE	0.207	0.053	0.043	0.093	0.058
Bias	0.201	0.008	0.002	0.016	0.010
Coverage	0.113	0.908	0.950	0.919	0.912
<i>DGP3: Non-IIA (MNP), $K = 3$, exchangeability</i>					
RMSE	0.195	0.074	0.074	0.158	0.081
Bias	0.179	0.010	0.023	0.017	0.009
Coverage	0.281	0.928	0.919	0.930	0.931
<i>DGP4: Non-exchangeable factor model, $K = 3$</i>					
RMSE	0.762	0.221	0.299	0.392	0.233
Bias	0.733	0.048	0.056	0.065	0.039
Coverage	0.132	0.908	0.940	0.920	0.914

Note: OLS ignores selection. MLogit denotes the single-index control-function estimator under the own-shock restriction. Oracle is the infeasible benchmark. Sieve denotes the proposed estimator using the full K -vector control function. Exch- $L2$ denotes the exchangeability-based estimator using the first two elementary symmetric polynomials of choice probabilities.

acle provides a tight benchmark with the smallest RMSE that MLogit nearly matches. The Sieve estimator also removes most of the selection bias but has larger RMSE due to the higher dimensionality of its control function. The semiparametric estimators (Sieve, Exch- $L2$) all achieve coverage rates close to the nominal 95% level. In DGP2 with three covariates, the Exch- $L2$ estimator, which uses a more flexible two-dimensional control function, is almost identical to MLogit. Sieve achieves comparable but slightly inflated RMSE.

DGP3 provides the most informative test of the exchangeability framework. Although MLogit's own-shock restriction is violated in DGP3, both MLogit and the correctly spec-

ified Exch- $L2$ correct most of the bias and outperform Oracle in terms of RMSE. Sieve performs similarly to Oracle with slightly inflated RMSE. With $K = 3$, the second stage for Sieve and Oracle requires $K + 1 = 4$ continuous covariates for identification. Since only three are available, the probability correction remains under-identified for Sieve and Oracle, confirming the value of the structural restrictions in reducing the effective dimensionality of the bias correction. Coverage is comparable across MLogit, Exch- $L2$ and Sieve.

In DGP4 with non-exchangeability, MLogit achieves the lowest RMSE, with Exch- $L2$ a close second. Both substantially outperform Sieve and Oracle, which remain underidentified yet correctly specified. In terms of bias, Exch- $L2$ leads over MLogit, while Sieve and Oracle exhibit larger bias. The performance is heterogeneous across occupations. For occupation 2 ($\lambda_2 = 0.8$, the most extreme loading), both Sieve and Oracle exhibit substantial bias on β_{21} (-0.231 and -0.183 , respectively) with RMSE substantially larger than that of MLogit and Exch- $L2$. The MLogit and Exch- $L2$ approaches deliver stable performance across all occupations because they require fewer identifying covariates (one and three, respectively).

To isolate structural biases from finite-sample variance, supplementary simulations at $n = 100,000$ are reported in Table C.II in Appendix C. With variance largely eliminated, Exch- $L2$ overtakes MLogit (RMSE 0.066 vs. 0.072). Sieve and Oracle remain substantially worse (RMSE 0.115 and 0.108). These results demonstrate the sharp identification requirement on continuous covariates. Appendix C confirms these findings under additional conditions. Sensitivity analysis at $n = 1,000$ and $n = 2,000$ confirms that bias reduction is already substantial at moderate sample sizes. A bootstrap validation exercise confirms that the analytical standard errors are a reasonable approximation to bootstrap standard errors.

The Monte Carlo results demonstrate the near-oracle performance of the proposed sieve estimators whenever the identification conditions are met. When the own-shock restriction is violated but exchangeability holds, the Exch- $L2$ estimator, which is then correctly specified, performs almost identically to the Oracle, and MLogit remains nearly as accurate despite relying on the now-violated own-shock restriction. The structural restrictions are valuable. When IIA, exchangeability, and the own-shock restrictions do not hold, MLogit

and Exch- $L2$ still perform reasonably well. Sieve remains stable in all of these settings, but it pays its own RMSE penalty when the identification conditions are not met.

5. APPLICATION: ENTRY LEVEL GENDER WAGE GAP IN SOUTH KOREA

I apply the proposed framework to estimate the gender wage gap among recent college graduates in South Korea, using the Graduates Occupational Mobility Survey (GOMS). It is well documented that selection into employment (Mulligan and Rubinstein, 2008, Blau and Kahn, 2017) and occupational sorting (Goldin, 2014) are quantitatively important for estimating gender wage gaps. I extend this line of work by explicitly modeling two layers of selection (participation and sorting). South Korea offers a particularly informative setting for multilayered selection. The economy features a pronounced dualism between large conglomerates and small and medium enterprises (SMEs), which motivates the ordered selection framework. The labor market also exhibits horizontal segmentation along occupation field (STEM vs. non-STEM) and sector (public vs. private). Female labor force participation among young graduates remains lower than male participation, and OECD (2024) reports the largest gender wage gap among member countries. These institutional features can generate strong selection on both the extensive and intensive margins.

The GOMS is a nationally representative survey of college graduates. Each cohort is surveyed approximately 18 months after graduation. I pool the 2008–2019 waves and restrict the sample to graduates aged 35 or younger, yielding a full sample of 207,985. The outcome variable is the log hourly wage, constructed as the log of monthly gross earnings divided by total hours (regular plus overtime). The key covariate of interest is a female indicator. The pre-determined controls that enter *both* the selection and the outcome equations are age, college GPA (on a 0–100 scale), parental income (the midpoint of the reported income bracket), a four-year university indicator (versus two-year colleges), major category (seven groups), university founding type (the institution’s ownership category—national, public, private, and so on; six categories), school region (17 administrative units at the city/province level), and survey year fixed effects. The outcome equation additionally controls for job tenure in months and for the realized sorting categories of the other two architectures; these are determined after selection (and are defined only for the employed),

so they do not enter the selection equation. No *pre-determined* covariate is excluded from the outcome equation; identification rests entirely on the nonlinearity of the control function rather than on an exclusion restriction. The three continuous covariates (age, GPA, and parental income) enter the first stage through a flexible sieve specification (cubic penalized regression splines and pairwise tensor products) and enter the outcome equation linearly.

I implement the following architectures for $D_i \in \{0, 1, 2\}$: i) *Ordered (firm size)* where $D_i = 1$ and 2 denote employment at a SME and at a large firm (≥ 300 employees), respectively; ii) *Multinomial (field)* where $D_i = 1$ and 2 denote a non-STEM job and a STEM job, respectively; and iii) *Multinomial (sector)* where $D_i = 1$ and 2 denote public-sector and private-sector employment, respectively. $D_i = 0$ indicates non-participation. The first architecture reflects the hierarchy in the Korean labor market. In the ordered first stage, an ordered probit with a linear index (including cubic terms of continuous covariates and interactions) is estimated for the parametric control function. For nonparametric first stage, I separately estimate $P(D_i \geq 1 | X_i)$ and $P(D_i = 2 | X_i)$ which serve as the control function arguments. Unlike firm size, there is no natural ordering in multinomial architectures. In both multinomial architectures, the first stage estimates a sieve multinomial logit as in (3.7), from which the inclusive values and choice probabilities are constructed.

Table IV reports summary statistics by gender and by the three selection classifications. Several patterns are noteworthy. Women constitute 46.4% of the sample but are underrepresented at large firms and in STEM jobs, and overrepresented in the public sector. Women have modestly higher GPAs but are less likely to hold STEM majors. The raw gender wage gap is substantial: the mean difference in log hourly wages is 12 log points. Across selection categories, large-firm workers earn more than SME workers, STEM workers earn more than non-STEM workers, and private-sector workers earn more than public-sector workers. Non-participants have similar age and parental income to the employed sample, but lower GPAs and fewer four-year university graduates.

For each architecture, I compare several estimators: i) OLS on the selected subsample; ii) a parametric correction: the ordered probit generalized inverse Mills ratio (ordered) or the inclusive value (multinomial) as a single-index control function; and iii) the proposed

TABLE IV
SUMMARY STATISTICS: SAMPLE MEANS

				Ordered sel.		Occ. sel.		Sector sel.		
	All	Male	Female	Non-part.	SME	Large	Non-STEM	STEM	Public	Private
<i>N</i>	207,985	111,428	96,557	76,802	79,405	51,754	94,209	36,950	27,065	102,894
<i>Demographics</i>										
Female (%)	46.4	0.0	100.0	49.1	49.6	37.5	47.6	37.8	57.9	41.4
Age	26.1	27.1	24.9	26.0	26.0	26.6	26.2	26.4	26.0	26.3
<i>Education</i>										
4-yr univ. (%)	74.6	75.5	73.5	75.7	69.8	80.2	74.1	73.3	84.8	70.9
STEM major (%)	48.0	59.6	34.6	44.8	45.5	56.5	33.9	90.5	37.4	53.0
GPA (0–100)	81.6	80.3	83.0	81.3	81.4	82.2	81.8	81.5	83.0	81.4
Semesters	7.3	7.3	7.2	7.2	7.1	7.5	7.2	7.5	7.6	7.2
<i>Family background</i>										
Parents' inc. (10K KRW)	435	428	442	434	420	459	437	431	440	434
<i>Employment</i>										
Employed (%)	69.9	71.5	68.0	18.4	100.0	100.0	100.0	100.0	100.0	100.0
In wage sample (%)	63.1	64.9	60.9	0.0	100.0	100.0	100.0	100.0	100.0	100.0
<i>Job characteristics (wage sample)</i>										
STEM job (%)	28.2	31.8	23.8	—	26.9	30.1	0.0	100.0	21.6	29.7
Large firm (%)	39.5	44.7	33.0	—	0.0	100.0	38.4	42.2	40.5	39.1
Public sector (%)	20.6	15.7	26.7	—	20.3	21.2	22.5	15.8	100.0	0.0
Tenure (mo.)	12.7	13.4	11.7	—	11.8	14.0	12.5	13.2	11.6	13.0
Hours/week	45.8	47.1	44.3	—	45.4	46.5	45.4	46.9	43.3	46.5
<i>Wages (wage sample)</i>										
Log hourly wage	0.03	0.09	−0.04	—	−0.04	0.14	0.01	0.08	0.01	0.04
Log monthly wage	5.30	5.38	5.19	—	5.21	5.43	5.26	5.38	5.21	5.32

Notes: Data from GOMS 2008–2019. Non-participants include unemployed, part-time, self-employed, and those with missing wage data. SME = firms with <300 employees; Large = firms with ≥ 300 employees. STEM job defined by KECO2018 occupation codes 2 (research and engineering) and 4 (health and medical). Public sector = government agencies, government-invested organizations, educational institutions. Private = private companies, foreign companies, corporate bodies. Parents' income is the midpoint of the reported monthly-income bracket (in 10,000 KRW). Job characteristics and wages are conditional on being in the wage sample. Standard deviations omitted for readability.

Sieve estimator, implemented as OLS on a tensor-product cubic B-spline basis in the estimated choice probabilities. In the ordered model, the wage equation for SME workers includes both threshold-probability indices (\hat{p}_1, \hat{p}_2) , while the large-firm equation includes only the upper threshold probability \hat{p}_2 , reflecting the boundary structure of the top category. In the multinomial models, the wage regression uses the inclusive value \hat{v}_k as a single-index control in the MLogit estimator, while the Sieve estimator uses the estimated choice probabilities (\hat{p}_1, \hat{p}_2) as controls. Additionally, an Exch- $L1$ estimator uses the first elementary symmetric polynomial $(1 - \hat{p}_k)$ as a control assuming exchangeability.

5.1. Results

Table V reports the estimated female coefficients across all specifications for both log hourly and log monthly wages. For hourly wage, the results reveal different selection patterns across the three architectures. For firm-size sorting, the OLS gender gap is -0.047 in SMEs and -0.038 in large firms. The Sieve estimator leaves the SME gap little changed (-0.055) but turns the large-firm gap positive ($+0.027$), reversing its sign. The large-firm result is the notable one: a substantial *upward* correction that implies negative selection bias, namely that conditional on working at a large firm, women have lower mean unobserved wage components than men. This pattern can be explained with amenity-based sorting. Large Korean firms offer substantial non-wage benefits (parental leave, regular working hours, and job security) that are particularly valued by women. If women sort into large firms partly for these amenities and are willing to accept employment even with relatively low wage draws, while men in large firms are selected primarily on the wage dimension, then the female workforce at large firms will have systematically lower unobserved productivity than the male workforce. For SMEs the selection correction is small and its sign varies across estimators: the parametric control function attenuates the gap to -0.028 , while the Sieve estimator slightly widens it to -0.055 . This is consistent with a weaker amenity bundle at smaller firms that provides less scope for amenity-wage trade-offs, leaving the SME gap robustly negative at around -0.05 .

In multinomial field sorting, the OLS female penalty in non-STEM occupations (-0.061) is modestly reduced by the MLogit correction (-0.055) but left essentially unchanged by the Sieve estimator (-0.063). In STEM, the OLS gap of -0.014 shifts to -0.021 under Sieve and to -0.026 under MLogit CF. The selection corrections on this margin are small, and the STEM gap in particular remains small in absolute terms across all estimators. Both men and women in STEM have passed through similar meritocratic screening (technical credentials, quantitative aptitude, degree requirements) that operates comparably regardless of gender, leaving less scope for gender-differential compositional effects. The single-index Exch- $L1$ estimator falls close to MLogit CF, plausibly because the single-index correction does not fully account for the selection structure.

The sectoral sorting results reflect the wage structure of the Korean public sector. The public-sector OLS female coefficient is close to 0, consistent with the seniority-based salary schedules that leave little scope for gender-differential pay, and the Sieve estimator barely moves it (to -0.012). Women make up 57.9% of public-sector workers but only 41.4% of private-sector workers, confirming strong gender-differential sorting on this margin even though it translates into little within-sector wage gap. In the private sector, the OLS estimate of -0.058 is essentially unchanged under Sieve (-0.057), whereas the MLogit and Exch- $L1$ single-index corrections move it upward to -0.047 . The remaining gender gap is the largest across all three architectures, consistent with the greater scope for discretionary wage-setting in the private sector. The divergence between the unrestricted Sieve estimate and the single-index MLogit and Exch- $L1$ estimates indicates that the single-index restriction those estimators impose is violated here.¹⁰

For log monthly wage regression, the estimated gender gaps are systematically 3–4 log points larger than the hourly wage gaps. The two are linked by the log identity $\log w^{\text{hourly}} = \log w^{\text{monthly}} - \log h^{\text{total}}$, which implies that the female coefficient in the hourly wage specification differs from that in the monthly wage specification by approximately the gender gap in log total hours. Women in the wage sample work 44.3 hours per week vs.

¹⁰Table G.I in Appendix G reports joint Wald tests on the second-stage control-function sieve basis terms, confirming that the single-index restrictions imposed by MLogit and Exch- $L1$ are empirically binding.

TABLE V

ESTIMATED FEMALE COEFFICIENTS WITH ROBUST STANDARD ERRORS IN PARENTHESES					
Model	Estimator	Log hourly wage		Log monthly wage	
		Cat. 1	Cat. 2	Cat. 1	Cat. 2
Ordered selection (categories: 1. SME / 2. Large)	OLS	-0.047 (0.004)	-0.038 (0.005)	-0.082 (0.003)	-0.074 (0.005)
	Parametric CF	-0.028 (0.006)	-0.008 (0.010)	-0.063 (0.005)	-0.049 (0.009)
	Sieve	-0.055 (0.006)	0.027 (0.007)	-0.083 (0.005)	-0.007 (0.006)
Multinomial selection (categories: 1. non-STEM / 2. STEM)	OLS	-0.061 (0.004)	-0.014 (0.006)	-0.099 (0.003)	-0.037 (0.005)
	MLogit CF	-0.055 (0.004)	-0.026 (0.006)	-0.092 (0.003)	-0.048 (0.005)
	Sieve	-0.063 (0.004)	-0.021 (0.006)	-0.100 (0.004)	-0.045 (0.005)
	Exch-L1	-0.055 (0.004)	-0.027 (0.006)	-0.092 (0.003)	-0.048 (0.005)
Multinomial selection (categories: 1. Public / 2. Private)	OLS	-0.010 (0.007)	-0.058 (0.003)	-0.033 (0.007)	-0.098 (0.003)
	MLogit CF	-0.007 (0.010)	-0.047 (0.003)	-0.027 (0.009)	-0.087 (0.003)
	Sieve	-0.012 (0.010)	-0.057 (0.004)	-0.031 (0.009)	-0.099 (0.004)
	Exch-L1	-0.006 (0.010)	-0.047 (0.003)	-0.026 (0.009)	-0.087 (0.003)

47.1 for men (log-ratio ≈ -0.06), and the bulk of this gap comes from overtime: men report 5.0 overtime hours per week against women's 3.3, while regular hours are similar (42.1 vs. 41.0). Dividing monthly pay by total hours mechanically attributes the hours difference to women's hourly rate and shrinks the gap by roughly the log-hours ratio. Despite the level difference between the two wage measures, the selection-correction patterns are qualitatively the same: the large-firm Sieve correction is strongly upward in both specifications,

turning the hourly gap positive ($-0.038 \rightarrow +0.027$) and nearly eliminating the monthly gap ($-0.074 \rightarrow -0.007$).

The overlap condition (Assumption 4) is verified in Table E.I in Appendix E. Adequate overlap is confirmed and no trimming is required. The first-stage estimates exhibit strong nonlinearity in the continuous covariates. Table F.I in Appendix F reports that the marginal sieve terms for age and GPA are strongly nonlinear and the parental-income term is also nonlinear (effective degrees of freedom above one), with several pairwise tensor interactions entering nonlinearly as well, providing sufficient identifying variation in the absence of exclusion restrictions. Two robustness checks, reported in Appendix D, confirm the main findings. First, augmenting the first-stage with an additional quasi-continuous covariate (semesters completed) yields similar results (Table D.I). Second, estimating all three architectures separately for each year produces estimates that are qualitatively consistent with the pooled results (Table D.II).

5.2. Decomposition analysis

The proposed framework permits a decomposition of the raw gender gap into a structural within-category gap, a within-category covariate-composition term, and a between-category sorting term. For gender g in category k , let \bar{w}_k^g denote the mean log wage, s_k^g the share, and $\hat{\beta}_k$ the estimated female coefficient in the category- k wage regression (so $-\hat{\beta}_k$ is the regression-adjusted within-category gap). The male-female mean log-wage difference can be written as

$$\bar{w}^M - \bar{w}^F = \underbrace{\sum_k s_k^M (-\hat{\beta}_k)}_{\text{structural within}} + \underbrace{\sum_k s_k^M [(\bar{w}_k^M - \bar{w}_k^F) + \hat{\beta}_k]}_{\text{covariate composition}} + \underbrace{\sum_k \bar{w}_k^F (s_k^M - s_k^F)}_{\text{between-category sorting}}.$$

The structural within term is the regression-adjusted gender differential, weighted by male sorting shares; the covariate-composition term collects the part of the raw within-category gap explained by gender differences in covariates (tenure, major, and the like); and the sorting term is the part of the gap generated by women being concentrated in lower-paying categories, valued at female within-category mean wages. Table VI reports the decomposition

TABLE VI
DECOMPOSITION OF THE GENDER WAGE GAP IN LOG HOURLY WAGE

	Ordered (firm size)	Field (STEM)	Sector (public/private)
Raw gap (male – female)	0.121	0.121	0.122
<i>Three-way decomposition of the raw gap</i>			
Structural within-category gap	0.043 (36%)	0.046 (38%)	0.050 (41%)
Covariate composition	0.062 (52%)	0.068 (56%)	0.076 (63%)
Between-category sorting	0.015 (13%)	0.007 (5%)	–0.005 (–4%)
<i>Effect of the selection correction</i>			
Structural within-category gap (Sieve)	0.018	0.049	0.050
Selection component (OLS – Sieve)	0.025	–0.003	0.000

for each architecture. The raw gender gap is approximately 12 log points. The structural within-category gap accounts for 36–41% of it, covariate composition for 52–63%, and *pure* between-category sorting for only 13% (firm size), 5% (occupation field), and –4% (sector). Women do sort disproportionately into lower-paying categories (67% at SMEs versus 55% of men, 76% in non-STEM occupations versus 68%, and 27% in the public sector versus 16%), but because the unconditional wage premia across these categories are modest, the mechanical contribution of that sorting to the aggregate gap is small. Most of the raw gap is a within-category phenomenon.

The bottom panel of Table VI isolates the effect of the selection correction on the structural within-category gap. Replacing the OLS coefficient with the selection-corrected Sieve coefficient changes this component appreciably only in the ordered (firm-size) architecture, where it falls from 0.043 to 0.018 (a selection component of 0.025). This reflects the off-setting category-level corrections documented above: the corrected female coefficient is negative at SMEs (–0.055) but positive at large firms (+0.027), and the two nearly cancel in the male-share-weighted average. In the occupation and sector architectures the selection correction barely moves the structural within-category gap (selection components of –0.003 and 0.000), consistent with the small corrections reported in Table V.

The between-category sorting term has a direct counterfactual reading: it is how much the gap would change if women sorted like men, holding female within-category mean wages fixed. Equalizing firm-size sorting would reduce the gap by 1.5 log points (13% of the raw gap), and equalizing occupation-field sorting by only 0.7 log points (5%). Equalizing sector sorting would slightly *increase* the gap (by 0.5 log points), because the public sector pays less than the private sector on average but exhibits a smaller within-sector gender differential, so reducing women's overrepresentation in public pushes them into a sector with a larger penalty. The modest size of these counterfactuals underscores that cross-category sorting, while real, accounts for a small share of the entry-level gap; the larger pieces are the structural within-category penalty and gender differences in covariates within categories.

TABLE VII
DYNAMICS OF THE FEMALE COEFFICIENT BY SUBPERIOD

Model / Category	Estimator	Log hourly wage			Log monthly wage		
		08–11	12–15	16–19	08–11	12–15	16–19
<i>A. Ordered selection (firm size)</i>							
SME	OLS	-0.062	-0.051	-0.028	-0.111	-0.082	-0.051
	Sieve	-0.062	-0.050	-0.018	-0.097	-0.085	-0.031
Large	OLS	-0.064	-0.012	-0.031	-0.098	-0.052	-0.064
	Sieve	-0.015	0.041	0.061	-0.052	-0.001	0.029
<i>B. Multinomial selection (field)</i>							
Non-STEM	OLS	-0.074	-0.052	-0.052	-0.123	-0.089	-0.077
	Sieve	-0.066	-0.063	-0.053	-0.114	-0.101	-0.076
STEM	OLS	-0.049	-0.015	0.012	-0.066	-0.036	-0.015
	Sieve	-0.045	-0.022	0.011	-0.062	-0.042	-0.015
<i>C. Multinomial selection (sector)</i>							
Public	OLS	-0.020	-0.005	-0.007	-0.057	-0.025	-0.016
	Sieve	-0.009	-0.015	0.000	-0.043	-0.036	-0.005
Private	OLS	-0.078	-0.048	-0.045	-0.124	-0.087	-0.079
	Sieve	-0.051	-0.050	-0.039	-0.101	-0.095	-0.074

5.3. *Dynamics of the gender wage gap*

To examine whether the selection patterns are stable over time, I re-estimate all specifications on three subperiods: 2008–2011, 2012–2015, and 2016–2019. This approach allows both the structural wage parameters and the selection correction to vary freely across periods. The GOMS is a repeated cross-section, hence the ‘dynamics’ documented here reflect cohort-level changes in the entry-level gender wage gap. Changes across subperiods may be driven by compositional shifts in the graduating population, evolving labor market institutions, or macroeconomic conditions affecting cohort-specific labor demand. These sources cannot be disentangled in the present data. Table VII reports the estimated female coefficient from the Sieve estimator alongside the OLS baseline for each subperiod.

Three patterns emerge from the hourly wage results. First, the OLS gap narrows steadily over time in most categories: the SME gap from -0.062 to -0.028 , the large firm gap from -0.064 to -0.031 , the STEM gap from -0.049 to $+0.012$, and the private sector gap from -0.078 to -0.045 . In non-STEM jobs and the public sector, the decline is more modest. Second, the selection-corrected gaps confirm the main findings from Table V. In the ordered model, the large-firm Sieve estimates are far less negative than OLS, turning positive from 2012–2015 onward, while the SME Sieve estimates track OLS closely. In the multinomial models the Sieve corrections are more modest in every period. Third, the corrected gap narrows or is stable over time. In large firms and STEM jobs, the Sieve estimate becomes positive in recent years, while the SME gap narrows toward zero (-0.018 by 2016–2019). The corrected gap is near zero throughout in the public sector, whereas the gap persists in non-STEM jobs and the private sector. The monthly wage results are broadly consistent with hourly wage patterns.

6. CONCLUSION

This paper establishes semiparametric identification of multilayered selection models without exclusion restrictions. The theoretical contribution provides a unified framework for correcting selection bias in settings where individuals first decide whether to participate and then sort into one of several categories, either vertically or horizontally. The key insight is that, when enough continuous covariates are available, nonlinearity in the selec-

tion structure generates sufficient variation to separate the structural outcome parameters from the selection bias. The empirical application shows that selection into firm size in particular materially reshapes the measured gap: the selection correction reverses the sign of the large-firm gender gap, while the corrections on the occupation and sector margins are more modest. A decomposition shows that most of the raw entry-level gap in Korea is a within-category phenomenon, a structural female penalty plus gender differences in covariates, with pure cross-category sorting contributing only modestly. In the most recent cohorts the corrected gap is closed or reversed in large firms and STEM jobs and near zero in the public sector, though it persists in non-STEM occupations and the private sector, suggesting that gender-equity policy should direct attention to the lifecycle dynamics that emerge after market entry.

Several extensions of the framework are worth pursuing. The current analysis treats the selection architecture as given; developing formal specification tests to discriminate between ordered and multinomial selection would strengthen empirical practice. The exchangeability assumption for multinomial selection, while considerably weaker than IIA, may still be restrictive in settings with strongly asymmetric substitution patterns; extending the framework to accommodate richer correlation structures while maintaining tractability is an open challenge. Finally, augmenting the bounds approach for multilayered selection (as in [Kroft, Mourifié, and Vayalinkal \(2024\)](#)) with structural restrictions could provide informative bounds in settings where the point identification conditions are not satisfied.

REFERENCES

- Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58(1-2), 3–29. [4]
- Berman, A. and R. J. Plemmons (1994). *Nonnegative Matrices in the Mathematical Sciences*. SIAM. [3]
- Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* 25(2), 242–262. [12]
- Berry, S. T., A. Gandhi, and P. A. Haile (2013). Connected substitutes and invertibility of demand. *Econometrica* 81(5), 2087–2111. [12]
- Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature* 55(3), 789–865. [4, 31]

- Blau, F. D., L. M. Kahn, N. Boboshko, and M. J. Comey (2024). The impact of selection into the labor force on the gender wage gap. *Journal of Labor Economics* 42(4), 1093–1133. [4]
- Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica* 75(2), 323–363. [4]
- Bourguignon, F., M. Fournier, and M. Gurgand (2007). Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons. *Journal of Economic Surveys* 21(1), 174–205. [4]
- Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics* 32(2), 189–218. [4]
- Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* 60(3), 567–596. [23]
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 5549–5632. Elsevier. [21, 22, 7, 11]
- Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica* 78(3), 1093–1125. [20]
- Chesher, A. and K. Smolinski (2012). IV models of ordered choice. *Journal of Econometrics* 166(1), 33–48. [8]
- Dahl, G. B. (2002). Mobility and the return to education: Testing a Roy model with multiple markets. *Econometrica* 70(6), 2367–2420. [4, 12, 15]
- Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies* 70(1), 33–58. [4, 10]
- D’Haultfoeuille, X. and A. Maurel (2013). Inference on an extended Roy model, with an application to schooling decisions in France. *Journal of Econometrics* 174(2), 95–106. [5]
- de Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). Springer-Verlag. [8]
- Dubin, J. A. and D. L. McFadden (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52(2), 345–362. [4, 14]
- Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2016). Identification and estimation of semiparametric two-step models. *Quantitative Economics* 7(2), 561–589. [4]
- French, E. and C. Taber (2011). Identification of models of the labor market. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4A, pp. 537–617. Elsevier. [4]
- Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review* 104(4), 1091–1119. [31]
- Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica* 42(4), 679–694. [1, 4]
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* 47(1), 153–161. [1, 4]
- Heckman, J. J. (1990). Varieties of selection bias. *The American Economic Review* 80(2), 313–318. [4, 10]
- Heckman, J. J. and R. Pinto (2018). Unordered monotonicity. *Econometrica* 86(1), 1–35. [4]
- Honoré, B. E. and L. Hu (2020). Selection without exclusion. *Econometrica* 88(3), 1007–1029. [4]
- Hotz, V. J. and R. A. Miller (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies* 60(3), 497–529. [12]

- Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics* 31(5), 1600–1635. [10]
- Kim, D. and Y. J. Lee (2025). Point-identifying semiparametric sample selection models with no excluded variable. Working Paper, Simon Fraser University. [4, 7, 15, 17, 22, 5, 6]
- Klein, R. and F. Vella (2010). Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics* 154(2), 154–164. [4]
- Krantz, S. G. and H. R. Parks (2013). *The Implicit Function Theorem: History, Theory, and Applications*. Birkhäuser, Boston, reprint of the 2003 edition. [3]
- Kroft, K., I. Mourifié, and A. Vayalinkal (2024). Lee bounds with multilayered sample selection. NBER Working Paper. [5, 11, 41]
- Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica* 51(2), 507–512. [4]
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies* 76(3), 1071–1102. [4, 5]
- Leung, S.-F. and S. Yu (1996). On the choice between sample selection and two-part models. *Journal of Econometrics* 72(1-2), 197–229. [7]
- Lewbel, A. (2007). Endogenous selection or treatment model estimation. *Journal of Econometrics* 141(2), 777–806. [4]
- McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press. [7]
- Mulligan, C. B. and Y. Rubinstein (2008). Selection, investment, and women’s relative wages over time. *The Quarterly Journal of Economics* 123(3), 1061–1110. [4, 31]
- Neal, D. (2004). The measured black-white wage gap among women. *Journal of Political Economy* 112(S1), S1–S28. [4]
- Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79(1), 147–168. [9, 11]
- Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal* 12(S1), S217–S229. [4]
- Newey, W. K., J. L. Powell, and J. R. Walker (1990). Semiparametric estimation of selection models: Some empirical results. *The American Economic Review* 80(2), 324–328. [4]
- OECD (2024). *OECD Economic Surveys: Korea 2024*. OECD Publishing, Paris. [31]
- Olivetti, C. and B. Petrongolo (2008). Unequal pay or unequal employment? A cross-country analysis of gender gaps. *Journal of Labor Economics* 26(4), 621–654. [4]
- Pan, Z. and Y. Zhang (2024). Locally robust semiparametric estimation of sample selection models without exclusion restrictions. arXiv preprint arXiv:2412.01208. [4]
- Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica* 57(6), 1403–1430. [4]

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* 56(4), 931–954. [23]

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3(2), 135–146. [4, 10]

Schumaker, L. L. (2007). *Spline Functions: Basic Theory* (3rd ed.). Cambridge University Press. [8]

Sheng, S. and X. Sun (2025). Social interactions in endogenous groups. arXiv preprint arXiv:2306.01544. [4]

Online Supplementary Appendix

for “Identification and Estimation of Semiparametric Multilayered Sample Selection Models” by Dongwoo Kim

APPENDIX A: TECHNICAL PROOFS

PROOF OF PROPOSITION 1: Since H_k is injective, it admits a measurable left inverse H_k^{-1} on $H_k(\text{supp}(X_i|D_i = k))$. For any candidate β , define $\tilde{\lambda}(s) := m_k(H_k^{-1}(s)) - H_k^{-1}(s)\beta$. Then for any x in the support, $x\beta + \tilde{\lambda}(H_k(x)) = x\beta + m_k(x) - x\beta = m_k(x)$.

Q.E.D.

PROOF OF PROPOSITION 2: Fix k and suppress the subscript k where no confusion arises. Suppose there exists an observationally equivalent pair (β, λ) such that $m_k(x) = x\beta + \lambda(H_k(x))$ almost surely on $\text{supp}(X|D = k)$. Define $l(x) := x(\beta_k - \beta)$ and $b(s) := \lambda_k(s) - \lambda(s)$ for $s \in [0, 1]^2$. Observational equivalence requires

$$l(x) + b(H_k(x)) = 0 \quad \text{a.s. on } \text{supp}(X|D = k). \quad (\text{A.1})$$

Step 1: Identification of coefficients on (X_1, X_2, X_3) . Let $\Delta_c := (\beta_{k1} - \beta_1, \beta_{k2} - \beta_2, \beta_{k3} - \beta_3)' \in \mathbb{R}^3$. Differentiating (A.1) with respect to $x_c = (x_1, x_2, x_3)$ yields

$$\Delta_c + J_k(x)\nabla b(H_k(x)) = 0, \quad (\text{A.2})$$

where $\nabla b(s) = (\partial b/\partial s_1, \partial b/\partial s_2)' \in \mathbb{R}^2$ and $J_k(x) \in \mathbb{R}^{3 \times 2}$. By Assumption 1(iv), $J_k(x)$ has rank 2, so its left null space is one-dimensional. Define the left-null vector

$$a(x) := \nabla_{x_c} h_k(x) \times \nabla_{x_c} h_{k+1}(x) \in \mathbb{R}^3,$$

where \times denotes the cross product. By construction, $a(x)'J_k(x) = 0$. Pre-multiplying (A.2) by $a(x)'$ yields

$$a(x)'\Delta_c = 0 \quad \text{a.e. on } \text{supp}(X|D = k). \quad (\text{A.3})$$

By Assumption 1(v), there exist $x^{(1)}, x^{(2)}, x^{(3)} \in \text{supp}(X_i | D_i = k)$ such that the vectors $a(x^{(1)}), a(x^{(2)}), a(x^{(3)})$ are linearly independent (since the intersection of the column

spaces is $\{0\}$, the null vectors span \mathbb{R}^3). Evaluating (A.3) at these three points yields three linearly independent equations $a(x^{(t)})' \Delta_c = 0$ for $t = 1, 2, 3$, which implies $\Delta_c = 0$.

Step 2: $b(\cdot)$ is constant on the relevant support. With $\Delta_c = 0$, equation (A.2) becomes $J_k(x) \nabla b(H_k(x)) = 0$. Since $J_k(x)$ has rank 2, this implies $\nabla b(H_k(x)) = 0$ almost everywhere, so b is locally constant on $H_k(\text{supp}(X_i | D_i = k))$. By Assumption 1(vii), this image is connected, so b is globally constant on it; denote this constant by C .

Step 3: Identification of β_k and λ_k . Substituting $\Delta_c = 0$ and $b \equiv C$ back into (A.1) gives $x(\beta_k - \beta) + C = 0$ for all x in the support of $X_i | D_i = k$. If $\beta_k \neq \beta$, then $x'(\beta_k - \beta) = -C$ for all x in the support, which would confine the support of X_i to a hyperplane orthogonal to $\beta_k - \beta$, contradicting Assumption 1(vi). Hence $\beta = \beta_k$ and $C = 0$. Then $\lambda_k(s)$ is identified on $H_k(\text{supp}(X | D = k))$ by $\lambda_k(s) = E[m_k(X) - X\beta_k | H_k(X) = s, D = k]$. *Q.E.D.*

PROOF OF PROPOSITION 3: Under the additive random utility model $U_{ij} = u_j + \varepsilon_{ij}$, normalizing $u_0 = 0$, the choice probabilities $p_j(u) = P[\varepsilon_j + u_j \geq \varepsilon_\ell + u_\ell, \forall \ell \neq j]$ depend on x only through $u = (u_1, \dots, u_K) \in \mathbb{R}^K$. Define $\Psi : \mathbb{R}^K \rightarrow \text{int}(\Delta^K)$ by $\Psi(u) = (p_0(u), \dots, p_K(u))$, and let $\bar{\Psi}(u) = (p_1(u), \dots, p_K(u))$ denote the K -dimensional restriction (since $p_0 = 1 - \sum_{j \geq 1} p_j$). We show $\bar{\Psi}$ is a diffeomorphism from \mathbb{R}^K onto the set $\{p \in \mathbb{R}^K : p_j > 0, \sum_{j=1}^K p_j < 1\}$.

Step 1: Local invertibility via the M-matrix structure. Consider the Jacobian $J := \partial \bar{\Psi} / \partial u \in \mathbb{R}^{K \times K}$, with entries $J_{jl} = \partial p_j / \partial u_l$ for $j, l = 1, \dots, K$. Let f denote the joint density of $(\varepsilon_0, \dots, \varepsilon_K)$ on \mathbb{R}^{K+1} , assumed continuous and strictly positive. Conditioning on ε_j and letting $G_j(\tau_{-j} | \varepsilon_j) := P(\varepsilon_m \leq \tau_m, \forall m \neq j | \varepsilon_j)$ denote the joint conditional CDF of the non- j shocks, we have $p_j(u) = \int f_j(\varepsilon_j) G_j((\varepsilon_j + u_j - u_m)_{m \neq j} | \varepsilon_j) d\varepsilon_j$, where f_j is the marginal density of ε_j . This representation does *not* rely on conditional independence of the non- j shocks: G_j is a genuine joint conditional CDF, not a product of marginals. Because f is strictly positive and continuous on \mathbb{R}^{K+1} , G_j is continuously differentiable in each τ_m with strictly positive partial derivatives on all of \mathbb{R}^K (the conditional marginal densities $\partial_m G_j$ inherit strict positivity from f).

Diagonal entries are strictly positive. Differentiating under the integral sign in u_j , each of the K arguments $\tau_m = \varepsilon_j + u_j - u_m$ increases by one, so

$$\frac{\partial p_j}{\partial u_j} = \int f_j(\varepsilon_j) \sum_{m \neq j} \partial_m G_j((\varepsilon_j + u_j - u_m)_{m \neq j} | \varepsilon_j) d\varepsilon_j > 0,$$

since every summand is non-negative and strictly positive on a set of positive f_j -measure.

Off-diagonal entries are strictly negative. For $l \neq j$, only τ_l depends on u_l , and it decreases by one when u_l increases, so

$$\frac{\partial p_j}{\partial u_l} = - \int f_j(\varepsilon_j) \partial_l G_j((\varepsilon_j + u_j - u_m)_{m \neq j} | \varepsilon_j) d\varepsilon_j < 0.$$

Column sums are strictly positive. From $\sum_{j=0}^K p_j(u) \equiv 1$,

$$\sum_{j=1}^K \frac{\partial p_j}{\partial u_l} = - \frac{\partial p_0}{\partial u_l} > 0 \quad \text{for each } l = 1, \dots, K,$$

where the last inequality follows from the off-diagonal calculation applied with $j = 0$.

The Jacobian J is therefore a Z-matrix (non-positive off-diagonals) with strictly positive column sums. By the characterization of non-singular M-matrices,¹¹ J is a non-singular M-matrix, and in particular is non-singular at every $u \in \mathbb{R}^K$.

Step 2: Properness and global invertibility via Hadamard. Let $u^{(n)} \in \mathbb{R}^K$ be any sequence with $\|u^{(n)}\| \rightarrow \infty$. Write $u^{(n)} = t_n v^{(n)}$ with $t_n = \|u^{(n)}\| \rightarrow \infty$ and $\|v^{(n)}\| = 1$; by compactness, along a subsequence $v^{(n)} \rightarrow v$ with $\|v\| = 1$. Set $v_0 := 0$ and $j^* := \arg \max_{0 \leq j \leq K} v_j$ (ties allowed). For any j with $v_j < \max_{\ell} v_{\ell}$, we have $u_j^{(n)} - u_{j^*}^{(n)} \rightarrow -\infty$ along the subsequence, so $p_j(u^{(n)}) \rightarrow 0$. In particular, at least one component of $\bar{\Psi}(u^{(n)})$ approaches either 0 or the simplex boundary, so $\bar{\Psi}(u^{(n)})$ leaves every compact subset of the open simplex $\{p > 0, \sum p_j < 1\}$. Hence $\bar{\Psi}$ is a proper map. The target set is a nonempty connected, simply connected open subset of \mathbb{R}^K (a convex open simplex interior). By the Hadamard global inverse function theorem, a C^1 proper map between connected smooth manifolds whose target is simply connected and whose Jacobian is everywhere non-singular is a diffeomorphism (e.g., [Krantz and Parks, 2013](#), Ch. 6). This estab-

¹¹A Z-matrix A is a non-singular M-matrix if and only if there exists $v > 0$ with $A'v > 0$. Setting $v = \mathbf{1}$ gives $A'\mathbf{1} = \text{column sums of } A > 0$. See [Berman and Plemmons \(1994\)](#).

lishes the claim for $\bar{\Psi}$, and hence for Ψ via the simplex constraint $p_0 = 1 - \sum_{j \geq 1} p_j$. Composing λ_k with $L_k \circ \Psi^{-1}$, where $L_k : u \mapsto (u_k - u_j)_{j \neq k}$ converts normalized utilities into the pairwise-difference vector on which λ_k was originally defined, yields the representation (2.14) on the image of Ψ ; by properness this image equals the full open simplex. *Q.E.D.*

PROOF OF PROPOSITION 4: Suppose $(\beta_k, \tilde{\lambda}_k)$ and $(\tilde{\beta}, b)$ both rationalize m_k on $\text{supp}(X_i | D_i = k)$. Writing $l(x) := x(\beta_k - \tilde{\beta})$ and $B(s) := \tilde{\lambda}_k(s) - b(s)$ on the image of P_K ,

$$l(x) + B(P_K(x)) = 0 \quad \text{a.s. on } \text{supp}(X_i | D_i = k). \quad (\text{A.4})$$

Step 1: Identification of the coefficients on x_c . Differentiating (A.4) with respect to x_c and writing $\Delta_c := ((\beta_{k,j} - \tilde{\beta}_j))_{j=1}^{K+1} \in \mathbb{R}^{K+1}$,

$$\Delta_c + J_P(x) \nabla B(P_K(x)) = 0. \quad (\text{A.5})$$

By Assumption 2(iii), the left null space of $J_P(x)$ is one-dimensional; denote a left null vector by $a(x) \in \mathbb{R}^{K+1}$. Pre-multiplying (A.5) by $a(x)^\top$ gives $a(x)^\top \Delta_c = 0$ almost everywhere. Evaluated at the $K + 1$ support points of Assumption 2(iv), this yields $K + 1$ independent linear equations and hence $\Delta_c = 0$.

Step 2: B is constant on the image of P_K . With $\Delta_c = 0$, (A.5) reduces to $J_P(x) \nabla B(P_K(x)) = 0$. Since $J_P(x)$ has column rank K , $\nabla B(P_K(x)) = 0$ a.e., so B is locally constant on $P_K(\text{supp}(X_i | D_i = k))$. By Assumption 2(vi), this image is connected, so B is globally constant on it; call this constant C .

Step 3: Identification of the remaining components and $\tilde{\lambda}_k$. Substituting back, (A.4) becomes $x(\beta_k - \tilde{\beta}) + C = 0$ on $\text{supp}(X_i | D_i = k)$. If $\beta_k \neq \tilde{\beta}$, the support would be contained in the affine hyperplane $\{x : x(\beta_k - \tilde{\beta}) = -C\}$, contradicting Assumption 2(v). Hence $\tilde{\beta} = \beta_k$ and $C = 0$, so $B \equiv 0$ on the image of P_K . Identification of $\tilde{\lambda}_k$ on $P_K(\text{supp}(X_i | D_i = k))$ then follows from $\tilde{\lambda}_k(s) = E[m_k(X_i) - X_i \beta_k | P_K(X_i) = s, D_i = k]$. *Q.E.D.*

A.1. Identification under multinomial logit selection

The conditional mean (2.18) has the partial linear structure $E[Y | X = x, D = k] = x\beta_k + \lambda_k(\nu_k(x))$, which is the framework of Kim and Lee (2025). Under standard regular-

ity conditions (continuous variation, smoothness, no multicollinearity), it suffices to verify that the inclusive value $\nu_k(x)$ is nonlinear; identification of β_k and λ_k then follows from Propositions 1–3 of [Kim and Lee \(2025\)](#). The following lemma establishes this nonlinearity, which holds automatically whenever at least two of the utility coefficients γ_j differ.

LEMMA 2: *Suppose $K \geq 2$ and there exist $j_1, j_2 \in \{0, 1, \dots, K\} \setminus \{k\}$ such that $\gamma_{j_1} \neq \gamma_{j_2}$ (where $\gamma_0 = 0$). Then:*

- (a) $\nu_k(x)$ is not an affine function of x .
- (b) For any two continuous covariates X_1, X_2 along which the alternative-specific coefficient differences $\{\gamma_j - \gamma_k\}_j$, projected onto the (x_1, x_2) plane, are not all collinear, the ratio $(\partial\nu_k/\partial x_1)/(\partial\nu_k/\partial x_2)$ is nonconstant in x .

PROOF: (a) The function $g(x) := \ln(\sum_{j=0}^K e^{x\gamma_j})$ is the log-sum-exp of $K + 1$ affine functions. Its Hessian is

$$\nabla^2 g(x) = \sum_{j=0}^K w_j(x) \gamma_j \gamma_j' - \left(\sum_{j=0}^K w_j(x) \gamma_j \right) \left(\sum_{j=0}^K w_j(x) \gamma_j \right)',$$

where $w_j(x) = e^{x\gamma_j} / \sum_{\ell} e^{x\gamma_{\ell}}$ are the choice probabilities. This is the covariance matrix of γ under the probability weights $\{w_j\}$, which is positive semidefinite and equals zero only if all γ_j are identical. Since $\gamma_{j_1} \neq \gamma_{j_2}$, the Hessian is nonzero, so g is convex and not affine. Since $\nu_k(x) = g(x) - x\gamma_k$ and $x\gamma_k$ is affine, ν_k is not affine.

(b) We have $\partial\nu_k/\partial x_{\ell} = \sum_j w_j(x)(\gamma_{j\ell} - \gamma_{k\ell})$ for $\ell = 1, 2$. Note first that this partial is constant in x whenever the ℓ -th coefficient is common across alternatives ($\gamma_{j\ell} = \gamma_{k\ell}$ for all j), since the weights $w_j(x)$ sum to one. Writing $d_j := (\gamma_{j1} - \gamma_{k1}, \gamma_{j2} - \gamma_{k2})$, the gradient $(\partial\nu_k/\partial x_1, \partial\nu_k/\partial x_2) = \sum_j w_j(x) d_j$ is a $w(x)$ -weighted average of the projected coefficient differences. When the $\{d_j\}$ are not all collinear, this weighted average traces a nondegenerate curve as the weights vary with x , so the gradient does not remain on a fixed ray and the ratio of its components is nonconstant. If instead all d_j are collinear—for instance when the alternatives differ only in a third covariate direction, so that $\gamma_{j1} = \gamma_{k1}$ and $\gamma_{j2} = \gamma_{k2}$ for all j —both partials are constant and the ratio is constant or undefined; this is the case excluded by the hypothesis. *Q.E.D.*

PROOF OF PROPOSITION 5: By Assumption 3, $\lambda_k((\delta_{kj})_{j \neq k})$ is symmetric in its non-chosen arguments. By Proposition 3, $\tilde{\lambda}_k := \lambda_k \circ L_k \circ \Psi^{-1}$ is a well-defined function on $\text{int}(\Delta^K)$, where $L_k : u \mapsto (u_k - u_j)_{j \neq k}$ converts normalized utilities into pairwise differences. The map $L_k \circ \Psi^{-1}$ is equivariant under permutations of non-chosen indices: a joint ε -density that is exchangeable in non-chosen indices yields choice probabilities satisfying $p_{\pi(j)}(\pi u) = p_j(u)$ for any permutation π of non-chosen indices, so permuting the non-chosen probabilities corresponds to permuting the non-chosen utility differences. Composing with the symmetry of λ_k , $\tilde{\lambda}_k(p_0, \dots, p_K)$ is symmetric in $(p_j)_{j \neq k}$.

Fix the compact $\mathcal{P} \subset \text{int}(\Delta^K)$ from the proposition's hypothesis, and let $\mathcal{P}^* := \bigcup_{\pi} \pi(\mathcal{P})$ denote its orbit under permutations of the non-chosen indices. As a finite union of compact sets, \mathcal{P}^* is compact, and by construction it is invariant under such permutations. By Assumption 2(ii), $\tilde{\lambda}_k$ is continuously differentiable on $\text{int}(\Delta^K)$, and so its restriction to \mathcal{P}^* is continuous. The Weierstrass approximation theorem then yields a polynomial P_n in the K non-chosen probability coordinates such that $\sup_{p \in \mathcal{P}^*} |\tilde{\lambda}_k(p) - P_n(p)| < \varepsilon$. Symmetrize by averaging: define $Q(p) := (K!)^{-1} \sum_{\pi} P_n(\pi p)$, where π runs over permutations of non-chosen indices. By the symmetry of $\tilde{\lambda}_k$ in non-chosen coordinates and the permutation invariance of \mathcal{P}^* ,

$$\sup_{p \in \mathcal{P}} |\tilde{\lambda}_k(p) - Q(p)| \leq (K!)^{-1} \sum_{\pi} \sup_{p \in \mathcal{P}^*} |\tilde{\lambda}_k(\pi p) - P_n(\pi p)| < \varepsilon.$$

By Newton's fundamental theorem of symmetric polynomials, Q can be expressed as a polynomial in the elementary symmetric polynomials (e_1, \dots, e_K) of the non-chosen probabilities, completing the result. *Q.E.D.*

PROOF OF PROPOSITION 6: The proof of (i) follows directly from Propositions 1–3 of [Kim and Lee \(2025\)](#) applied to the single-index partial linear model with index $e_1(p(x)) = 1 - p_k(x)$. The proof of (ii) follows the same structure as Proposition 2, generalized from two indices to L indices. For (ii), the argument is identical to that of Proposition 4 under the substitution $(P_K, K) \mapsto (E_L, L)$ posited in (b), with the $L + 1$ continuous covariates $x_c = (x_1, \dots, x_{L+1})$ of condition (a) as the differentiating coordinates. If $(\beta, \check{\lambda}^{(L)})$ is observationally equivalent, set $\Delta_c := (\beta_{k,j} - \beta_j)_{j=1}^{L+1}$ and $b := \check{\lambda}_k^{(L)} - \check{\lambda}^{(L)}$, so

that $x'_c \Delta_c + b(E_L(x)) = 0$ a.s. Differentiating in x_c and pre-multiplying by a left-null vector of the rank- L Jacobian $J_E(x)$ gives $\Delta_c = 0$ at the $L + 1$ spanning points of the analogue of Assumption 2(iv); then $\nabla b \equiv 0$ on the connected image $E_L(\text{supp}(X_i | D_i = k))$ forces $b \equiv C$; and the no-hyperplane analogue of (v) yields $\beta = \beta_k$, $C = 0$. Identification of $\check{\lambda}_k^{(L)}$ follows from $\check{\lambda}_k^{(L)}(s) = E[m_k(X) - X\beta_k | E_L(X) = s, D = k]$. *Q.E.D.*

PROOF OF LEMMA 1: (a) follows from Theorem 3.1 of [Chen \(2007\)](#) applied to the sieve MLE of the ordered choice model. The sup-norm convergence rate for sieve estimators with d_c continuous regressors and smoothness m_h is $O_p(Q_n^{-m_h/d_c} + \sqrt{Q_n \log n/n})$, the sum of a sieve-approximation bias and an estimation-error term. Setting $Q_n \asymp n^{d_c/(2m_h+d_c)}$ balances the two terms and yields the rate $O_p(n^{-m_h/(2m_h+d_c)}(\log n)^c)$. This is $o(n^{-1/4})$ when $m_h/(2m_h + d_c) > 1/4$, i.e., $m_h > d_c/2$. In the designs considered, where $d_c \leq 3$, $m_h \geq 2$ suffices.

(b) The argument parallels (a), now applied to the sieve logistic regression estimator (3.6) for each k . Since the threshold function $h_k(x) = \Lambda(h_k^*(x))$ where $h_k^*(x) = \log(h_k(x)/(1 - h_k(x)))$ is the log-odds, the sieve logistic estimator targets h_k^* and converts to h_k via Λ . The smoothness of Λ preserves the sup-norm rate. With the $d_c = 3$ continuous covariates required to identify the nonparametric ordered model, $m_h \geq 2$ satisfies the general condition $m_h > d_c/2 = 3/2$.

(c) is standard; see [McFadden \(1972\)](#). The uniform convergence over x follows from $\nu_k(x)$ being a smooth function of $(\gamma_1, \dots, \gamma_K)$ and x , combined with the compactness of \mathcal{X} .

(d) follows from (c) by the continuous mapping theorem, since the elementary symmetric polynomials of the choice probabilities are smooth functions of the probability vector, which is itself a smooth function of the sieve parameters. *Q.E.D.*

LEMMA 3—Tensor product B-spline perturbation: *For tensor product B-splines of order r with $J_n^{[\ell]}$ interior knots in dimension ℓ ,*

$$\|B^{(L)}(s) - B^{(L)}(s')\| \leq C_r \left(\max_{\ell} J_n^{[\ell]} \right) \|s - s'\|$$

for a constant C_r depending only on r .

PROOF: The tensor product basis is $B_{\mathbf{j}}^{(L)}(s) = \prod_{\ell=1}^L B_{j_\ell}^{[\ell]}(s_\ell)$. For fixed ℓ , $B_{j_\ell}^{[\ell]}(\cdot)$ is Lipschitz with constant $C_r J_n^{[\ell]}$ (the derivative of a B-spline of order r is bounded by C_r times the inverse knot spacing). Each factor $B_{j_\ell}^{[\ell]}$ is bounded by 1 so the product rule gives

$$\begin{aligned} |B_{\mathbf{j}}^{(L)}(s) - B_{\mathbf{j}}^{(L)}(s')| &\leq \sum_{\ell=1}^L |B_{j_\ell}^{[\ell]}(s_\ell) - B_{j_\ell}^{[\ell]}(s'_\ell)| \prod_{\ell' \neq \ell} \max(|B_{j_{\ell'}}^{[\ell']}(s_{\ell'})|, |B_{j_{\ell'}}^{[\ell']}(s'_{\ell'})|) \\ &\leq \sum_{\ell=1}^L C_r J_n^{[\ell]} |s_\ell - s'_\ell| \leq C_r L \left(\max_{\ell} J_n^{[\ell]} \right) \|s - s'\|. \end{aligned}$$

Squaring and summing over the multi-index \mathbf{j} (noting that at most $(r+1)^L$ tensor product basis functions are nonzero at any point), the result follows. For $L=1$ this reduces to $\|B^{(1)}(s) - B^{(1)}(s')\| \leq C_r J_n |s - s'|$. *Q.E.D.*

PROOF OF THEOREM 1: Throughout, fix category k and write n for n_k , g_0 for g_{k0} , β_0 for β_{k0} , and so on. The sum \sum_i is taken over $i \in \mathcal{I}_k$.

Step 1: Setup and decomposition. Write the model as $Y_i = X_i \beta_0 + \lambda_0(g_0(X_i)) + \varepsilon_i$ where $E[\varepsilon_i | X_i, D_i = k] = 0$. By Assumption 5, there exists a sieve coefficient vector $\delta_0 \in \mathbb{R}^{\kappa_n}$ such that

$$\sup_{s \in \mathcal{G}_k} \left| \lambda_0(s) - B^{(L)}(s)' \delta_0 \right| \leq C \cdot \left(\max_{\ell} J_n^{[\ell]} \right)^{-m_\lambda}, \quad (\text{A.6})$$

by the standard approximation theory for tensor product B-splines (Schumaker, 2007, de Boor, 2001). Define the approximation residual $r_i := \lambda_0(g_0(X_i)) - B^{(L)}(g_0(X_i))' \delta_0$, so that $Y_i = X_i \beta_0 + B^{(L)}(g_0(X_i))' \delta_0 + r_i + \varepsilon_i$. Let $W_i^0 = (X_i, B^{(L)}(g_0(X_i))')' \in \mathbb{R}^{d_X + \kappa_n}$ denote the infeasible augmented regressor, and $\hat{W}_i = (X_i, B^{(L)}(\hat{g}(X_i))')'$ the feasible counterpart. Define $\theta_0 = (\beta_0', \delta_0')'$, $\tilde{\theta} = (\mathbf{W}_0' \mathbf{W}_0)^{-1} \mathbf{W}_0' \mathbf{Y}$ (oracle), and $\hat{\theta} = (\hat{\mathbf{W}}' \hat{\mathbf{W}})^{-1} \hat{\mathbf{W}}' \mathbf{Y}$ (feasible), where \mathbf{W}_0 , $\hat{\mathbf{W}}$, and \mathbf{Y} are the stacked matrices.

Step 2: Oracle estimator asymptotics. The oracle estimator satisfies

$$\tilde{\theta} - \theta_0 = \left(\frac{1}{n} \sum_i W_i^0 (W_i^0)' \right)^{-1} \frac{1}{n} \sum_i W_i^0 (\varepsilon_i + r_i). \quad (\text{A.7})$$

Let $P_n^0 := \mathbf{B}_0(\mathbf{B}'_0\mathbf{B}_0)^{-1}\mathbf{B}'_0$ denote the projection onto the column space of the sieve basis $\mathbf{B}_0 = (B^{(L)}(g_0(X_i)))'_{i \in \mathcal{I}_k}$, and let $M_n^0 = I_n - P_n^0$. By Frisch–Waugh–Lovell theorem,

$$\tilde{\beta} - \beta_0 = \left(\frac{\mathbf{X}'M_n^0\mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}'M_n^0(\boldsymbol{\varepsilon} + \mathbf{r})}{n}, \quad (\text{A.8})$$

where \mathbf{X} is the $n \times d_X$ covariate matrix for the selected subsample. I analyze the two terms separately. The approximation bias is $o(n^{-1/2})$. By (A.6), $\|r_i\| \leq C(\max_\ell J_n^{[\ell]})^{-m_\lambda}$ for all i , so $\left\| \frac{\mathbf{X}'M_n^0\mathbf{r}}{n} \right\| \leq \frac{\|\mathbf{X}\| \|\mathbf{r}\|}{n} \leq C \left(\max_\ell J_n^{[\ell]} \right)^{-m_\lambda}$. By Assumption 6(iv), $\sqrt{n} \cdot (\max_\ell J_n^{[\ell]})^{-m_\lambda} \rightarrow 0$. For the stochastic term, standard arguments (Newey, 1997) yield $\frac{\mathbf{X}'M_n^0\boldsymbol{\varepsilon}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_i \tilde{X}_i \varepsilon_i + o_p(1)$, where $\tilde{X}_i := X_i - \Pi(g_0(X_i))$ is the population projection residual. By the Lindeberg–Feller CLT, $\frac{1}{\sqrt{n}} \sum_i \tilde{X}_i \varepsilon_i \xrightarrow{d} N(0, \Omega_k)$. Combining with $n^{-1}\mathbf{X}'M_n^0\mathbf{X} \xrightarrow{p} \Sigma_k$ (by the law of large numbers and the consistency of the sieve projection), we obtain

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma_k^{-1}\Omega_k\Sigma_k^{-1}). \quad (\text{A.9})$$

Step 3: Generated regressor negligibility. I show that $\sqrt{n}(\hat{\beta} - \tilde{\beta}) = o_p(1)$. By the FWL theorem applied to the feasible estimator, with $\hat{P}_n = \hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}'$ and $\hat{M}_n = I - \hat{P}_n$,

$$\hat{\beta} - \beta_0 = \left(\frac{\mathbf{X}'\hat{M}_n\mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}'\hat{M}_n(\mathbf{Y} - \mathbf{X}\beta_0)}{n}, \quad (\text{A.10})$$

where $\mathbf{Y} = \mathbf{X}\beta_0 + \mathbf{B}_0\delta_0 + \mathbf{r} + \boldsymbol{\varepsilon}$. Therefore,

$$\hat{\beta} - \tilde{\beta} = \left(\frac{\mathbf{X}'\hat{M}_n\mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}'(\hat{M}_n - M_n^0)(\mathbf{B}_0\delta_0 + \mathbf{r} + \boldsymbol{\varepsilon})}{n} + \text{second-order terms.}$$

The dominant term is

$$R_n := \frac{\mathbf{X}'(\hat{M}_n - M_n^0)\boldsymbol{\varepsilon}}{n}. \quad (\text{A.11})$$

To bound R_n , note that $\hat{M}_n - M_n^0 = P_n^0 - \hat{P}_n$, and the operator norm of the projection difference is bounded by

$$\|P_n^0 - \hat{P}_n\|_{\text{op}} \leq C \cdot \frac{\|\mathbf{B}_0 - \hat{\mathbf{B}}\|_F}{\sigma_{\min}(\mathbf{B}_0)},$$

where σ_{\min} denotes the minimum singular value. By Lemma 3,

$$\|\mathbf{B}_0 - \hat{\mathbf{B}}\|_F \leq \sqrt{n} \cdot \max_i \|B^{(L)}(g_0(X_i)) - B^{(L)}(\hat{g}(X_i))\| \leq C_r \sqrt{n} \left(\max_{\ell} J_n^{[\ell]} \right) \|\hat{g} - g_0\|_{\infty}. \quad (\text{A.12})$$

By standard results on the minimum singular value of sieve design matrices (Huang, 2003), the restricted eigenvalue condition $\sigma_{\min}(\mathbf{B}_0/\sqrt{n}) \geq c > 0$ holds with probability approaching one under Assumption 6(iii), which requires that the sieve basis is not nearly collinear on the support of the data. For tensor product B-splines on a compact domain with bounded density, this follows from the local support property of B-splines, which ensures that the Gram matrix $\mathbf{B}'_0 \mathbf{B}_0/n$ converges to a banded positive definite matrix. Hence $\sigma_{\min}(\mathbf{B}_0) = \sqrt{n} \sigma_{\min}(\mathbf{B}_0/\sqrt{n}) \geq c\sqrt{n}$ with probability approaching one, so the \sqrt{n} factor in (A.12) cancels against the \sqrt{n} in the denominator: $\|P_n^0 - \hat{P}_n\|_{\text{op}} = O_p \left(\left(\max_{\ell} J_n^{[\ell]} \right) \|\hat{g} - g_0\|_{\infty} \right)$.

For the term $R_n = \mathbf{X}'(P_n^0 - \hat{P}_n)\boldsymbol{\varepsilon}/n$, a worst-case operator-norm bound is not sharp enough; the required extra $n^{-1/2}$ comes from $\boldsymbol{\varepsilon}$ being conditionally mean zero. Conditioning on (\mathbf{X}, \hat{g}) , $E[R_n | \mathbf{X}, \hat{g}] = 0$ because $E[\varepsilon_i | X_i, D_i = k] = 0$, and

$$\text{Var}(R_n | \mathbf{X}, \hat{g}) = \frac{1}{n^2} \mathbf{X}'(P_n^0 - \hat{P}_n) \Omega_{\varepsilon} (P_n^0 - \hat{P}_n)' \mathbf{X} \preceq \frac{\bar{\sigma}^2}{n^2} \|\mathbf{X}\|_{\text{op}}^2 \|P_n^0 - \hat{P}_n\|_{\text{op}}^2,$$

where $\Omega_{\varepsilon} = \text{diag}(\sigma_k^2(X_i)) \preceq \bar{\sigma}^2 I$ by the bounded conditional variance in Assumption 9. Hence, using $\|\mathbf{X}\|_{\text{op}}/\sqrt{n} = O_p(1)$ and the projection bound above,

$$\|R_n\| = O_p \left(\frac{\|\mathbf{X}\|_{\text{op}}}{n} \|P_n^0 - \hat{P}_n\|_{\text{op}} \right) = O_p \left(\frac{1}{\sqrt{n}} \left(\max_{\ell} J_n^{[\ell]} \right) \|\hat{g} - g_0\|_{\infty} \right).$$

Therefore,

$$\sqrt{n} \|R_n\| = O_p \left(\left(\max_{\ell} J_n^{[\ell]} \right) \|\hat{g} - g_0\|_{\infty} \right) = O_p \left(\left(\max_{\ell} J_n^{[\ell]} \right) \cdot o_p(n^{-1/4}) \right).$$

Under Assumption 6(i)–(iii), $\max_{\ell} J_n^{[\ell]} = O(n^{1/(2m_{\lambda}+L)+\epsilon})$ for the optimal rate, and the product $J_n^{[\ell]} \cdot n^{-1/4} \rightarrow 0$ when $1/(2m_{\lambda} + L) < 1/4$, i.e., $m_{\lambda} > (4 - L)/2$. For $L = 2$ and $m_{\lambda} \geq 2$, this is satisfied. Therefore $\sqrt{n} R_n = o_p(1)$.

Two terms remain. For the approximation residual, $\|\mathbf{r}\|_{\infty} = O((\max_{\ell} J_n^{[\ell]})^{-m_{\lambda}})$, so the same projection bound gives $\sqrt{n} \|\mathbf{X}'(\hat{M}_n - M_n^0)\mathbf{r}/n\| = O_p(\sqrt{n} (\max_{\ell} J_n^{[\ell]})^{1-m_{\lambda}} \|\hat{g} -$

$g_0\|_\infty) = o_p(1)$ under Assumption 6(iv). The term in $\mathbf{B}_0\delta_0$ is the genuine generated-regressor bias and requires more than the projection bound. Because $\mathbf{B}_0\delta_0$ lies in the column space of \mathbf{B}_0 , the oracle projection annihilates it exactly ($M_n^0\mathbf{B}_0\delta_0 = 0$), so the term reduces to $\mathbf{X}'\hat{M}_n\mathbf{B}_0\delta_0/n = \mathbf{X}'(I - \hat{P}_n)(\mathbf{B}_0 - \hat{\mathbf{B}})\delta_0/n$, using $\hat{M}_n\hat{\mathbf{B}} = 0$. Writing $\mathbf{X}'(I - \hat{P}_n) = (\hat{M}_n\mathbf{X})' = \tilde{\mathbf{X}}'$, this is an inner product of the estimated projection residual \tilde{X}_i with $(B^{(L)}(g_0(X_i)) - B^{(L)}(\hat{g}(X_i)))'\delta_0 \approx \lambda'_0(g_0(X_i))(g_0(X_i) - \hat{g}(X_i))$, a near-function of the indices. Its \sqrt{n} -negligibility under the sup-norm rate $\|\hat{g} - g_0\|_\infty = o_p(n^{-1/4})$ together with the orthogonality $E[\tilde{X}_i | g_0(X_i), D_i = k] = 0$ (which makes the leading bias second order) is the standard two-step sieve generated-regressor result (Newey, 1997, Chen, 2007); it is this rate-and-orthogonality mechanism, rather than a moment-level Neyman orthogonality, that the Remark below refers to. This establishes $\sqrt{n}(\hat{\beta} - \tilde{\beta}) = o_p(1)$.

Step 4: Conclusion. Combining Steps 2 and 3:

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sqrt{n}(\tilde{\beta} - \beta_0) + o_p(1) \xrightarrow{d} N(0, \Sigma_k^{-1}\Omega_k\Sigma_k^{-1}).$$

This completes the proof. *Q.E.D.*

REMARK 6: Negligibility of the generated regressor is a sieve-rate phenomenon, not moment-level Neyman orthogonality. The naive Robinson moment $m(\beta, g) = E[\tilde{X}_i(Y_i - X_i\beta - \lambda(g(X_i)))]$ has Gateaux derivative $\partial_g m \cdot \Delta g = -E[\tilde{X}_i\lambda'(g_0(X_i))\Delta g(X_i)]$, which vanishes for $\sigma(g_0)$ -measurable directions Δg (by the projection property $E[\tilde{X}_i | g_0(X_i)] = 0$) but is non-zero for general X -measurable directions. For perturbations of the form $\hat{g} - g_0$ that lie in $\sigma(X)$ rather than $\sigma(g_0)$, the leading first-order term is bounded by $C\|\hat{g} - g_0\|_{L^2}$, and Step 3 absorbs it into the $o_p(1)$ remainder using the rate condition $\|\hat{g} - g_0\|_{\text{sup}} = o_p(n^{-1/4})$ together with sieve growth $J_n = o(n^{1/4})$ and Hölder smoothness of λ_k .

PROOF OF THEOREM 2: The consistency of $\hat{\Sigma}_k$ follows from the uniform law of large numbers applied to $\hat{X}_{ik}\hat{X}'_{ik}$, combined with $\hat{X}_{ik} - \tilde{X}_{ik} = o_p(1)$ uniformly, which follows from the consistency of the sieve projection. For $\hat{\Omega}_k$, we additionally need $\hat{\varepsilon}_{ik}^2 \xrightarrow{p} \sigma_k^2(X_i)$ in an appropriate sense. By the consistency of $\hat{\beta}_k$ and $\hat{\delta}_k$, and the uniform convergence of the sieve approximation, $\hat{\varepsilon}_{ik} = \varepsilon_{ik} + o_p(1)$. The bounded fourth moment (Assumption 9(iii)) ensures uniform integrability, completing the argument. *Q.E.D.*

APPENDIX B: CONTROL FUNCTION LINEARITY AND NONLINEARITY

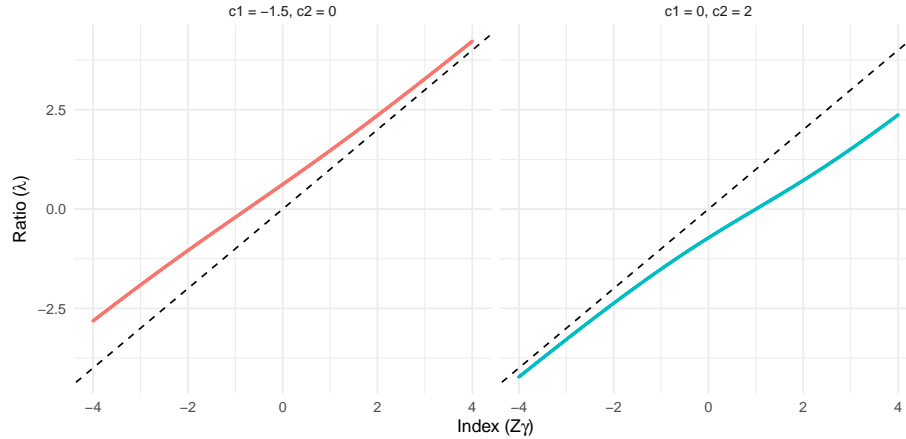


FIGURE B.1.—Near-linearity of the ordered probit control function. Each panel plots the generalized inverse Mills ratio $\lambda(z\gamma; c_1, c_2) = [\phi(c_2 - z\gamma) - \phi(c_1 - z\gamma)] / [\Phi(c_2 - z\gamma) - \Phi(c_1 - z\gamma)]$ against the linear index $z\gamma$ (solid) alongside the 45-degree line (dashed) for two threshold configurations. The ratio is approximately linear over much of the support, illustrating why identification without an exclusion restriction is fragile under the Gaussian ordered selection specification with a linear selection index.

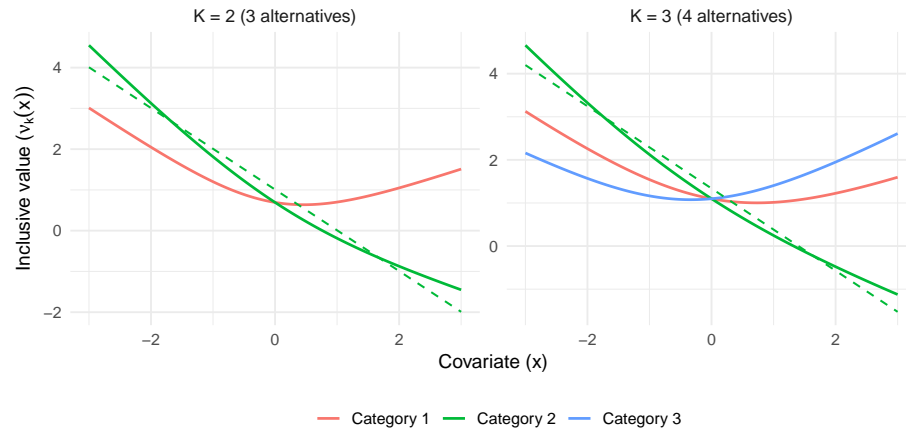


FIGURE B.2.—Inherent nonlinearity of the multinomial logit inclusive value. Each panel plots the inclusive value $\nu_k(x) = \log(\sum_{j \neq k} \exp(-x \cdot \delta_{kj}))$ against the covariate x for different numbers of alternatives. The log-sum-exp structure produces a convex function that is inherently nonlinear in x , providing the identifying variation needed for selection correction without exclusion restrictions. The dashed line shows the best linear approximation for Category 2, whose curvature is less pronounced than the other categories; the visible departure confirms that nonlinearity is present even in this case.

APPENDIX C: ADDITIONAL SIMULATION RESULTS

Ordered DGP3: $K = 3$ with three continuous covariates. To move beyond the $K = 2$ ordered designs, this DGP uses $K = 3$ (non-participation and 3 occupations) with three continuous covariates (X, Z, W) and a selection index whose higher-order terms are scaled by a nonlinearity parameter δ :

$$h(X, Z, W) = 0.5X + 0.3Z + 0.8W + \delta \cdot (-0.5X^2 + 0.2Z^2 - 0.4XZ) + \varepsilon,$$

with thresholds $c_1 = -1.0$, $c_2 = 0.5$, $c_3 = 1.5$ and outcome-error correlations $\rho_k \in \{0.3, 0.5, 0.7\}$ with the selection error, so the selection bias is progressively stronger for higher categories. The full-nonlinearity baseline is $\delta = 1$; smaller values scale the higher-order terms toward zero and probe progressively weaker nonlinearity, the regime in which identification without exclusion restrictions is most fragile. At $\delta = 0.1$ the latent index is nearly linear in the covariates, so any smooth function of the choice probabilities, including everything in the cubic B-spline tensor basis, approximates a linear function of (X, Z, W) ; the CF basis and the linear regressors then occupy nearly overlapping column space in the second-stage OLS, inducing severe multicollinearity that the bivariate basis cannot escape. Table C.I reports results across $\delta \in \{1, 0.7, 0.3, 0.1\}$ for all three occupations.

At full nonlinearity ($\delta = 1$) the Sieve estimator achieves near-oracle performance across all three occupations and nine coefficients, with coverage between 90% and 94%. More generally, for moderate-to-strong nonlinearity ($\delta \in \{0.3, 0.7, 1\}$) Sieve essentially matches the Oracle in every occupation: bias under 0.04 against the Oracle's 0.001–0.004, with RMSE only modestly above. At $\delta = 0.1$, however, Sieve bias inflates by an order of magnitude in every cell (to 0.101 in Occupation 1, 0.122 in Occupation 2, and 0.170 in Occupation 3) while the Oracle's bias remains negligible (0.006–0.026). With $h(X, Z, W)$ nearly linear, the control function basis spans a function class approximately collinear with the linear regressors, so the second-stage design matrix is near-rank-deficient: the partialled-out variation in (X, Z, W) shrinks toward zero and any noise in the first-stage probability estimates is amplified into finite-sample bias on $\hat{\beta}_k$. The pathology worsens with ρ because the true correction has greater curvature for the higher-bias occupations and an ill-conditioned

TABLE C.I

ORDERED SELECTION WITH $K = 3$: WEAK-TO-STRONG NONLINEARITY

	Average RMSE				Average Bias				Average Coverage			
	$\delta=1$	$\delta=0.7$	$\delta=0.3$	$\delta=0.1$	$\delta=1$	$\delta=0.7$	$\delta=0.3$	$\delta=0.1$	$\delta=1$	$\delta=0.7$	$\delta=0.3$	$\delta=0.1$
<i>Occupation 1 ($\rho = 0.3$, bivariate CF)</i>												
OLS	0.112	0.127	0.157	0.165	0.105	0.122	0.153	0.162	0.337	0.188	0.058	0.027
Linear	9.161	5.769	3.124	2.562	6.094	3.726	1.417	0.424	0.790	0.830	0.910	0.962
Oracle	0.039	0.040	0.072	0.197	0.003	0.002	0.002	0.007	0.947	0.945	0.935	0.933
Sieve	0.045	0.049	0.080	0.165	0.007	0.005	0.020	0.101	0.905	0.913	0.933	0.875
<i>Occupation 2 ($\rho = 0.5$, bivariate CF)</i>												
OLS	0.188	0.217	0.269	0.296	0.181	0.212	0.266	0.294	0.233	0.095	0.002	0.000
Linear	21.705	18.997	11.545	10.043	2.754	5.577	3.820	0.073	0.905	0.910	0.910	0.953
Oracle	0.051	0.056	0.087	0.252	0.001	0.002	0.001	0.026	0.945	0.942	0.952	0.938
Sieve	0.055	0.064	0.107	0.296	0.001	0.004	0.020	0.122	0.935	0.937	0.932	0.923
<i>Occupation 3 ($\rho = 0.7$, univariate CF)</i>												
OLS	0.198	0.233	0.276	0.304	0.188	0.225	0.272	0.300	0.303	0.178	0.017	0.003
Linear	0.347	0.296	0.260	0.204	0.058	0.085	0.110	0.028	0.932	0.928	0.887	0.930
Oracle	0.055	0.064	0.100	0.158	0.003	0.003	0.004	0.006	0.955	0.935	0.935	0.940
Sieve	0.059	0.068	0.124	0.288	0.004	0.005	0.037	0.170	0.937	0.920	0.902	0.845

Note: $n = 5,000$, $R = 200$ replications. Each cell averages the metric across the three slope coefficients within the indicated occupation. The selection index scales its higher-order terms by δ ; $\delta = 1$ is the full-nonlinearity baseline and smaller δ approaches linearity. Sieve uses a bivariate control function for the two interior occupations and a univariate control function for the top occupation. “Linear” uses a misspecified linear ordered probit; “Oracle” uses the correctly specified nonlinear-index ordered probit.

design captures that curvature poorly. Even the Oracle pays a variance penalty at low δ : its bias is stable across the grid but its RMSE inflates from 0.040–0.064 at $\delta = 0.7$ to 0.158–0.252 at $\delta = 0.1$, because the first-stage coefficients on X^2 , Z^2 , and XZ are themselves estimated with noise when those terms are close to zero. The penalty is intrinsic to nonlinearity-based identification.

The parametric Linear correction inherits the shape distinction between interior and top categories. For the top category, the inverse Mills ratio’s slope-varying curvature still provides identification away from the linear regressors even when its underlying linear or-

dered probit is misspecified, and Linear’s RMSE in Occupation 3 stays bounded (0.20–0.35 across the δ grid). For the interior categories, the IMR has nearly constant slope, so its column is near-collinear with the linear regressors at every δ ; combined with the misspecification, this produces catastrophic RMSE for Linear (2.6 at $\delta = 0.1$ in Occupation 1, up to 21.7 at $\delta = 1$ in Occupation 2). Sieve avoids the interior pathology: its RMSE is one to two orders of magnitude smaller than Linear’s in Occupations 1 and 2 across the entire δ grid.

Multinomial DGP4 at $n = 100,000$. Table C.II reports the supplementary large-sample simulation summarized in Section 4. With the variance largely eliminated, the under-identification of the multivariate sieve correction at $K = 3$ with only three continuous covariates becomes clearly visible: Sieve and Oracle RMSE remain roughly twice that of MLogit and Exch- $L2$, confirming that this gap is structural rather than a small-sample phenomenon.

TABLE C.II

MULTINOMIAL DGP4 (NON-EXCHANGEABLE FACTOR MODEL, $K = 3$): LARGE-SAMPLE SUMMARY

	STATISTICS				
	OLS	MLogit	Oracle	Sieve	Exch- $L2$
Average RMSE	0.737	0.072	0.108	0.115	0.066
Average Bias	0.736	0.047	0.084	0.043	0.036
Average Coverage	0.000	0.744	0.644	0.896	0.817

Note: RMSE, absolute bias, and 95% CI coverage averaged across 3 occupations \times 3 coefficients. $n = 100,000$, $R = 100$ replications.

Sensitivity to sample size. To assess small-sample behavior across estimators, Table C.III reports results for multinomial DGP1 at $n \in \{1,000, 2,000, 5,000\}$. At $n = 1,000$, the MLogit RMSE (0.120) is comparable to OLS (0.114), reflecting the variance cost of estimating the sieve first stage with limited data; the MLogit bias (0.009), however, is already an order of magnitude smaller than OLS (0.054), and coverage is at the nominal level (0.93 vs. 0.93 for OLS). By $n = 2,000$ the MLogit RMSE pulls ahead of OLS, and at $n = 5,000$ the advantage is clear. MLogit coverage stays at 92–94% across all sample sizes, while OLS

coverage deteriorates from 0.93 at $n = 1,000$ to 0.75 at $n = 5,000$ as bias becomes dominant relative to the standard error. The Oracle tracks MLogit closely (RMSE 0.099 falling to 0.047 with bias under 0.01 throughout), confirming that the feasible MLogit gives up little efficiency to the infeasible benchmark even at small n . Sieve carries a larger variance penalty at small n (RMSE 0.260 at $n = 1,000$, falling to 0.109 at $n = 5,000$) because the bivariate control function needs more data to be well-conditioned, but its bias remains small at every n and coverage stays near nominal (0.90–0.93).

TABLE C.III

DGP1: SENSITIVITY TO SAMPLE SIZE

	$n = 1,000$			$n = 2,000$			$n = 5,000$		
	RMSE	Bias	Cov.	RMSE	Bias	Cov.	RMSE	Bias	Cov.
OLS	0.114	0.054	0.93	0.093	0.048	0.85	0.074	0.054	0.75
MLogit	0.120	0.009	0.93	0.086	0.009	0.94	0.055	0.001	0.92
Oracle	0.099	0.003	0.96	0.075	0.007	0.94	0.047	0.001	0.94
Sieve	0.260	0.033	0.90	0.167	0.013	0.93	0.109	0.009	0.92

Note: RMSE, absolute bias, and 95% CI coverage averaged across coefficients within Occupation 1. $R = 200$ replications.

Bootstrap validation. A bootstrap validation exercise ($R = 200$ Monte Carlo draws, $B = 100$ bootstrap replications within each draw) confirms that the heteroskedasticity robust standard errors used throughout are a reasonable approximation to the bootstrap standard errors. Table C.IV compares the analytical and bootstrap standard errors for the four MLogit slope coefficients in multinomial DGP1. The average robust SE is uniformly about 91% of the average bootstrap SE (ratios 0.90–0.91 across coefficients), indicating that the analytical SEs are slightly smaller but of the same order. The bootstrap SE in turn essentially matches the Monte Carlo standard deviation of the point estimates, validating both as faithful measures of sampling dispersion. The 95% confidence-interval coverage is similar under the two SE estimates (91–96% under robust SE vs. 93–96% under bootstrap SE).

TABLE C.IV

BOOTSTRAP VALIDATION OF ROBUST STANDARD ERRORS (MULTINOMIAL DGP1, MLOGIT ESTIMATOR)

Coefficient	True	MC SD	HC SE	Boot SE	HC/Boot	Cov HC	Cov Boot
β_{11} (X , Occ. 1)	0.50	0.056	0.051	0.056	0.91	0.910	0.940
β_{12} (Z , Occ. 1)	0.70	0.054	0.051	0.056	0.91	0.930	0.955
β_{21} (X , Occ. 2)	0.80	0.079	0.070	0.077	0.90	0.910	0.930
β_{22} (Z , Occ. 2)	0.50	0.064	0.059	0.064	0.91	0.955	0.955

Note: Multinomial DGP1, MLogit estimator, $n = 5,000$, $R = 200$ Monte Carlo draws, $B = 100$ bootstrap replications per draw. “MC SD” is the standard deviation of the point estimate across MC draws. “HC SE” and “Boot SE” are the average analytical and bootstrap standard errors. “HC/Boot” is the ratio of average HC to average bootstrap SE. Coverage columns report the empirical coverage of nominal 95% confidence intervals constructed from each SE estimate.

APPENDIX D: ROBUSTNESS TABLES

D.1. *Enriched selection equation*

As a further robustness check, I augment the first-stage selection equation with semesters completed as an additional quasi-continuous covariate, exploiting the variation in time to degree across students. This variable is available from the 2010 wave onward, restricting the sample to 2010–2019 cohorts. The additional covariate provides richer nonparametric variation in the selection equation, yielding a more flexible sieve specification with four quasi-continuous covariates (age, GPA, parental income, and semesters completed) entering with quadratic terms and all pairwise interactions. Table D.I reports the estimated female coefficient for all three architectures. The results on the 2010–2019 subsample with the enriched selection equation are broadly consistent with the full-sample estimates. In the ordered model, the Sieve estimator yields -0.040 for SMEs and $+0.018$ for large firms, reproducing the sign reversal documented in Table V. In the occupation-type model, the Sieve estimate for non-STEM (-0.060) is close to the full-sample estimate (-0.063), and the STEM estimate is near zero (-0.007) as in the full sample. In the sector model, the public-sector gap is again essentially zero under Sieve ($+0.002$) and the private-sector gap remains negative (-0.045), preserving the qualitative pattern of the full-sample sector re-

sults. The consistency across estimators and sample restrictions confirms the robustness of the selection patterns documented in Table V.

TABLE D.I
ROBUSTNESS: ENRICHED SELECTION EQUATION (2010–2019 SUBSAMPLE)

Model	Estimator	Category	
		SME / non-STEM / Public	Large firm / STEM / Private
A. Ordered selection (firm size)	OLS	−0.046	−0.045
	Parametric CF	−0.013	−0.023
	Semiparametric (Sieve)	−0.040	0.018
B. Multinomial selection (field)	OLS	−0.060	−0.013
	MLogit CF	−0.056	−0.014
	Semiparametric (Sieve)	−0.060	−0.007
	Exch- <i>L1</i>	−0.061	−0.014
C. Multinomial selection (sector)	OLS	−0.013	−0.053
	MLogit CF	−0.016	−0.043
	Semiparametric (Sieve)	0.002	−0.045
	Exch- <i>L1</i>	−0.019	−0.034

Note: Second-stage wage equations as in Table V, except that the firm-size and occupation equations omit the public-sector indicator; the first-stage selection equation adds semesters completed. Sample restricted to 2010–2019 GOMS cohorts.

D.2. Year-by-year estimation

The subperiod analysis in Table VII imposes a common selection mechanism across multiple survey years. To verify that this restriction does not distort the year-specific estimates, I re-estimate all three architectures separately for each year 2008–2019. This approach allows both the selection equation and the outcome equation to vary freely across years, at the cost of smaller within-year samples (approximately 2,500–6,300 wage workers per category per year). Table D.II reports the OLS and Sieve female coefficients from the year-by-year estimation alongside the pooled estimates.

TABLE D.II

YEAR-BY-YEAR ESTIMATION: FEMALE COEFFICIENT (LOG HOURLY WAGE)

Year	A. Ordered (firm size)				B. Occupation (STEM)				C. Sector			
	SME		Large		non-STEM		STEM		Public		Private	
	OLS	Sieve	OLS	Sieve	OLS	Sieve	OLS	Sieve	OLS	Sieve	OLS	Sieve
2008	-0.035	-0.039	-0.042	-0.035	-0.051	-0.039	-0.036	-0.029	0.035	0.031	-0.070	-0.060
2009	-0.062	-0.072	-0.078	-0.024	-0.083	-0.078	-0.046	-0.046	-0.037	-0.030	-0.080	-0.075
2010	-0.083	-0.120	-0.092	-0.001	-0.099	-0.093	-0.067	-0.085	-0.028	-0.017	-0.104	-0.073
2011	-0.047	-0.038	-0.037	0.019	-0.047	-0.034	-0.037	-0.015	-0.029	-0.001	-0.044	-0.028
2012	-0.030	-0.017	-0.077	-0.054	-0.052	-0.055	-0.073	-0.075	0.000	0.003	-0.065	-0.065
2013	-0.062	-0.082	-0.002	0.001	-0.056	-0.057	-0.021	-0.029	-0.036	-0.041	-0.045	-0.044
2014	-0.041	0.016	0.015	0.023	-0.046	-0.054	0.022	0.024	0.041	0.026	-0.043	-0.050
2015	-0.070	-0.048	0.000	0.100	-0.060	-0.060	-0.013	-0.029	-0.025	-0.034	-0.046	-0.056
2016	-0.018	-0.004	-0.022	0.054	-0.050	-0.051	0.028	0.023	0.028	0.041	-0.049	-0.050
2017	-0.044	-0.026	-0.022	0.053	-0.059	-0.057	0.013	0.015	-0.021	-0.012	-0.051	-0.050
2018	-0.025	-0.028	-0.035	0.065	-0.047	-0.051	0.007	0.007	-0.015	-0.024	-0.037	-0.031
2019	-0.022	0.023	-0.050	0.048	-0.050	-0.040	-0.016	-0.007	-0.029	-0.024	-0.038	-0.036

Notes: All specifications are identical to Table V except that year fixed effects are omitted. Sample sizes per year range from 14,000–17,000 (full) and 2,500–6,300 (per wage category).

The ordered model results are qualitatively consistent with the subperiod estimates: the Sieve female coefficient for large firms is positive in every year from 2013 onward (and in 2011), with 2010 and 2012 the only exceptions (a near-zero -0.001 and a -0.054), confirming that the sign reversal documented in Table V is not driven by a single cohort. The year-by-year estimates are more volatile (ranging from -0.054 to $+0.100$) than the subperiod estimates in Table VII (-0.015 to $+0.061$), reflecting the smaller within-year samples. For occupation-field sorting, the gender gap estimates tend to shrink toward zero: the non-STEM Sieve estimate moves from -0.078 in 2009 to -0.040 in 2019, while the STEM coefficient is moderately negative in 2008–2013 (between -0.015 and -0.085) and oscillates around zero from 2014 onward. The public-sector gap oscillates around zero in every year, confirming that the regulated salary structure eliminates the gender wage gap within the public sector regardless of which year's selection mechanism is used. The

private-sector gap narrows from -0.075 in 2009 to -0.036 in 2019, indicating that the convergence documented in the subperiod analysis is robust to year-by-year estimation.

APPENDIX E: OVERLAP DIAGNOSTICS

Table E.I reports quantiles of the estimated selection probabilities $\hat{P}(D_i = k|X_i)$ from the ordered probit and sieve MNL first stages, separately by gender. For the ordered probit (Panel A), the minimum estimated probability is 0.038 for large firms, and the 5th percentile exceeds 0.12 in all categories, indicating strong overlap. Women have higher non-participation probabilities (median 0.40 vs. 0.32 for men) and lower large-firm probabilities (median 0.21 vs. 0.28). For the sieve MNL (Panels B and C), the flexible B-spline specification produces some near-zero probabilities at the extremes, but the 5th percentiles are well above zero for the main categories. In the occupation-type model, women have substantially lower STEM probabilities (median 0.04 vs. 0.21 for men), reflecting differential major and occupation choices. In the sector model, women have higher public-sector probabilities (median 0.13 vs. 0.09), reflecting the attraction of regulated public-sector employment.

TABLE E.I

DISTRIBUTION OF ESTIMATED SELECTION PROBABILITIES

		A. Ordered probit					B. MNL (field)					C. MNL (sector)				
		Min	p_5	Med	p_{95}	Max	Min	p_5	Med	p_{95}	Max	Min	p_5	Med	p_{95}	Max
$P(D=0 X)$	Male	0.082	0.212	0.321	0.469	0.765	0.000	0.227	0.336	0.479	0.997	0.000	0.224	0.338	0.484	1.000
	Female	0.060	0.253	0.398	0.537	0.759	0.000	0.218	0.385	0.516	0.856	0.000	0.221	0.387	0.518	0.824
$P(D=1 X)$	Male	0.197	0.363	0.395	0.399	0.400	0.000	0.247	0.413	0.689	0.978	0.000	0.038	0.085	0.268	0.746
	Female	0.201	0.336	0.386	0.399	0.400	0.062	0.128	0.511	0.672	0.982	0.000	0.054	0.131	0.490	0.779
$P(D=2 X)$	Male	0.038	0.166	0.281	0.403	0.634	0.000	0.023	0.207	0.441	0.960	0.000	0.326	0.566	0.698	1.000
	Female	0.040	0.127	0.215	0.351	0.694	0.000	0.020	0.039	0.647	0.824	0.089	0.180	0.463	0.617	1.000

Notes: Estimated selection probabilities from the ordered probit (Panel A), sieve multinomial logit for field (Panel B), and sieve multinomial logit for sector (Panel C). $D = 0$: non-participation; $D = 1$: SME / non-STEM / public; $D = 2$: large firm / STEM / private. $N = 189,589$ (ordered and occupation-type) and 188,459 (sector) after dropping observations with missing covariates.

APPENDIX F: FIRST-STAGE SELECTION ESTIMATION

This appendix details the first-stage estimation. In the ordered selection model, I use a *generalized additive model* (GAM), which extends a generalized linear model by allowing the linear index to depend on continuous covariates through unknown smooth functions estimated nonparametrically. The probit GAM for $P(D_i \geq k | X_i)$ takes the form

$$\Phi^{-1}(P(D_i \geq k | X_i)) = \alpha_k + Z_i' \gamma_k + \sum_{\ell} f_{k\ell}(X_{i\ell}) + \sum_{\ell < m} g_{k\ell m}(X_{i\ell}, X_{im}),$$

where Z_i collects the categorical and binary covariates entering the index linearly (female, four-year university, major category, university founding type, 17 province/city-level school-region indicators, and year fixed effects), all of which are determined at or before graduation and are therefore available for participants and non-participants alike, and each marginal smooth term $f_{k\ell}$ and each tensor smooth term $g_{k\ell m}$ is an unknown nonparametric function approximated by penalized cubic regression splines. A marginal smooth $f(x) = \sum_{b=1}^{k_b} \theta_b B_b(x)$ is a linear combination of k_b cubic B-spline basis functions B_b on the support of x . Its spline coefficients θ are shrunk toward those of a straight line by a roughness penalty $\lambda \int [f''(x)]^2 dx$ on the integrated squared second derivative. The smoothing parameter λ controls flexibility: $\lambda \rightarrow \infty$ collapses the smooth to a straight line, $\lambda \rightarrow 0$ permits an essentially interpolating curve. A tensor smooth $g(x_1, x_2)$ is the bivariate analogue on a tensor-product B-spline basis with margin-specific penalties.

I implement this in R using the `mgcv` package. The probit GAM is fit by `mgcv::gam` with marginal smooths `s(age, bs="cr", k=10)`, `s(gpa, bs="cr", k=10)`, and `s(income, bs="cr", k=5)`, plus pairwise tensor interactions `ti(age, gpa, bs="cr", k=c(5, 5))` and analogues for the other two pairs of continuous covariates; `bs="cr"` specifies a cubic regression spline basis and `k` sets the basis dimension. The argument `family = binomial(link = "probit")` specifies the binary probit GAM, which `mgcv` fits by penalized iteratively reweighted least squares. The smoothing parameters $\{\lambda_{k\ell}\}$ are selected by restricted maximum likelihood (REML), a marginal-likelihood criterion that profiles the fixed-effect parameters out of the likelihood and

chooses each smooth’s flexibility to maximize the resulting profile criterion. This gives a fully data-driven balance between bias and variance and avoids ad-hoc tuning.

Table [F.I](#) reports the effective degrees of freedom (edf) selected for each smooth term in the ordered selection model. The edf of a fitted smooth equals the trace of the smoother matrix (the linear operator that maps the response to its fitted values) and lies between 0 (smooth penalized to zero) and the maximum basis dimension. An edf of 1 corresponds to a straight line, and values close to the maximum indicate the data demand essentially the full flexibility of the basis. For the ordered architecture, separate probit GAMs estimate $P(D_i \geq 1 | X_i)$ and $P(D_i = 2 | X_i)$. Both equations exhibit substantial first-stage nonlinearity. The marginal smooths in age and GPA reach essentially the full basis flexibility (edf 6.7–8.7 of 9), and the parental-income marginal saturates its smaller basis (edf 3.0–3.6 of 4). The tensor interactions are estimated at intermediate flexibility, with $\text{ti}(\text{age}, \text{GPA})$ and $\text{ti}(\text{GPA}, \text{income})$ taking moderate edf values; only $\text{ti}(\text{age}, \text{income})$ in $P(D_i \geq 1 | X_i)$ is shrunk close to a near-linear surface (edf 1.8). The first stage therefore demands genuine nonparametric flexibility, providing identifying variation in the absence of an exclusion restriction in the ordered selection model.

For multinomial selection models, the first stage uses a sieve MNL, not a penalized GAM, so a comparable edf summary does not apply. The sieve MNL first stage is estimated by `nnet::multinom` in R with cubic B-spline bases for age, GPA, and parental income, and pairwise tensor product interactions. Because there is no penalization, the natural nonlinearity diagnostic is the joint significance of each smooth-term group rather than an edf. Table [F.II](#) reports joint Wald tests on each marginal basis and each tensor interaction. Each test asks whether the indicated subset of multinomial logit coefficients is jointly zero across both non-base equations; under the null, the statistic is asymptotically χ_q^2 with q equal to the number of restrictions.

TABLE F.I

FIRST-STAGE GAM SMOOTH TERM DIAGNOSTICS (ORDERED SELECTION)

Smooth term	$P(\text{employed})$		$P(\text{large firm})$	
	edf	max	edf	max
$s(\text{age})$	8.0	9	8.7	9
$s(\text{GPA})$	6.7	9	8.6	9
$s(\text{income})$	3.6	4	3.0	4
$\text{ti}(\text{age}, \text{GPA})$	10.1	16	8.9	16
$\text{ti}(\text{age}, \text{income})$	1.8	16	8.6	16
$\text{ti}(\text{GPA}, \text{income})$	8.5	16	6.4	16

Note: “edf” is the effective degrees of freedom selected by REML; “max” is the maximum edf allowed by the basis dimension ($k - 1$ for marginal terms, $(k_1 - 1)(k_2 - 1)$ for tensor interactions). First-stage GAMs use $k = 10$ for age and GPA, $k = 5$ for parental income (which has limited variation from ordinal brackets), and $k = 5$ for tensor interactions. An edf close to the maximum indicates the smooth is constrained by the basis dimension. $N = 189,589$.

TABLE F.II

FIRST-STAGE JOINT WALD TESTS ON THE MULTINOMIAL SIEVE BASIS (MODELS B AND C)

Smooth term	q	STEM/non-STEM		public/private	
		Wald	p	Wald	p
age (marginal)	10	4.8	0.90	4.2	0.94
GPA (marginal)	10	3.6	0.97	4.8	0.91
income (marginal)	8	3.7	0.88	3.3	0.91
age \times GPA	50	79.5	0.005	107.3	4.7×10^{-6}
age \times income	40	76.0	5.2×10^{-4}	72.6	0.001
GPA \times income	40	39.3	0.50	34.1	0.73
All marginals	28	13.1	0.99	13.6	0.99
All tensors	130	200.7	6.9×10^{-5}	221.3	1.1×10^{-6}
All spline + tensor	158	1360.0	$< 10^{-16}$	1604.8	$< 10^{-16}$

Note: Joint Wald statistic on the indicated subset of MNL coefficients across both non-base equations, with q degrees of freedom from the asymptotic χ_q^2 null. “Marginal” refers to the $\text{bs}(x, df)$ basis terms ($df = 5$ for age and GPA, $df = 4$ for parental income); “ $x \times y$ ” refers to the corresponding pairwise tensor interaction basis. The bottom three rows aggregate the per-term tests. The aggregate “All spline + tensor” Wald rejects the linear-categorical-only baseline at the χ_{158}^2 floor of machine precision in both architectures.

The joint test on *all* 158 spline and tensor basis coefficients rejects the linear-categorical-only baseline at the floor of machine precision in both architectures: the multinomial first stage is unambiguously and highly nonlinear. The variation lives in the tensor interactions rather than in the marginal smooths. The marginal $\text{bs}(\cdot)$ basis terms alone are not jointly significant after controlling for the pairwise tensors, while the tensor block is significant at $p < 10^{-4}$. This is expected: tensor-product splines span functions that include the main effects, so the marginal-only Wald in the presence of the tensors loses power. The aggregate test is the appropriate diagnostic for first-stage flexibility.

APPENDIX G: SECOND-STAGE CONTROL FUNCTION DIAGNOSTICS

The second stage uses fixed-basis OLS so I report joint heteroskedasticity-robust Wald tests on the second-stage control-function basis. Table G.I reports three nested Wald tests for each architecture-cell pair. Five of the six cells use a bivariate ($L = 2$) CF: the SME (interior) cell of the ordered architecture and all four cells of the two multinomial architectures. The remaining cell, Large (top) of the ordered architecture, uses a univariate ($L = 1$) CF by construction.

(i) *Full CF basis*. For the bivariate-CF cells the basis is $\text{bs}(\hat{p}_1, df = 6) + \text{bs}(\hat{p}_2, df = 6) + \text{bs}(\hat{p}_1, df = 4) : \text{bs}(\hat{p}_2, df = 4)$ (a 28-dimensional basis). For the Large cell it is just $\text{bs}(\hat{p}_2, df = 6)$ (a 6-dimensional basis). The null is that every basis coefficient is zero; rejection says selection bias is empirically present in the cell.

(ii) *Orthogonal subspace* (bivariate-CF cells only). The null is that the coefficients on all bivariate-basis terms other than those of $\text{bs}(\hat{p}_{\text{own}}, df = 6)$ are jointly zero ($q = 22$), i.e., that the bias depends only on the own-category probability through a single index. In the multinomial architectures rejection refutes the single-index restrictions used by MLogit and Exch-L1. In the ordered SME cell, rejection refutes the analogous single-index restriction underlying the parametric ordered-probit correction. The Large cell is omitted from this test because its CF is already a single index in \hat{p}_{own} by construction.

(iii) *Tensor interaction* (bivariate-CF cells only). The null is that the bivariate tensor $\text{bs}(\hat{p}_1, df = 4) : \text{bs}(\hat{p}_2, df = 4)$ is jointly zero ($q = 16$), i.e., that the bias correction is ad-

ditively separable in the two probabilities as $f_1(\hat{p}_1) + f_2(\hat{p}_2)$. Rejection says the bias is genuinely two-dimensional and that an additive correction would not suffice. Again, the test does not apply to the Large cell since its CF has no tensor component.

The tests are nested: rejection of (iii) implies rejection of (ii), which in turn implies rejection of (i). The full CF basis is jointly significant at the floor of machine precision in every cell of every architecture, confirming that selection bias is empirically present everywhere. The orthogonal subspace is also jointly significant at $p < 10^{-16}$ in every cell, indicating that the single-index restrictions are empirically binding in all three architectures. The tensor interactions are also strongly significant across all cells, indicating the selection bias is not additively separable in the control functions.

TABLE G.I

JOINT HETEROSKEDASTICITY-ROBUST WALD TESTS ON THE SECOND-STAGE SIEVE BASIS

	Log hourly wage			Log monthly wage		
	Full CF ($q=28$)	Orth. ($q=22$)	Tensor ($q=16$)	Full CF ($q=28$)	Orth. ($q=22$)	Tensor ($q=16$)
<i>A. Ordered selection (firm size)</i>						
SME (interior)	522.8	249.3	78.0	824.3	357.8	99.1
Large (top) ^a	527.4	—	—	749.8	—	—
<i>B. Multinomial selection (occupation)</i>						
non-STEM	664.2	443.2	249.0	900.6	564.5	268.2
STEM	227.4	186.2	92.5	353.8	299.6	100.1
<i>C. Multinomial selection (sector)</i>						
Public	228.3	163.4	48.0	358.1	221.0	50.8
Private	536.8	293.3	158.2	823.9	497.5	286.9

Note: Each cell reports the joint Wald statistic for the indicated subset of the second-stage Sieve basis; q in the column header is the number of restrictions. “Full CF” tests that all second-stage CF basis terms are jointly zero. “Orth.” tests the subspace of the bivariate basis orthogonal to a single-index function of the own-category probability. “Tensor” tests the bivariate tensor interaction in isolation. ^aThe Large (top) category uses a univariate ($L = 1$) control function, so the orthogonal-subspace and tensor tests do not apply.