Distinguishing Biases from Personal Preferences: An Honest Machine Learning Approach

Mahyar Habibi* Zahra Khanalizadeh† Negar Ziaeian‡

November 6, 2026

Abstract

Estimating individual-level discrimination from observational data in bilateral evaluation settings, such as matching platforms, is challenging when unobserved evaluator preferences and item characteristics confound the relationship between group membership and outcomes. This problem is especially difficult when rich item-level covariates are unavailable and researchers observe only the interaction outcome matrix. We develop a methodology that combines Collaborative Filtering with Causal Machine Learning to estimate evaluator-specific bias while accounting for latent preference and match-quality differences. The first stage introduces Honest Collaborative Filtering (HCF), which learns evaluator preference and item characteristic embeddings from the observed outcome matrix using an honesty restriction designed to reduce contamination from potentially biased evaluations. The second stage applies Double Machine Learning (DML) to estimate evaluator-specific bias parameters while flexibly controlling for these learned embeddings. Monte Carlo simulations show that the proposed HCF+DML estimator reduces bias and RMSE relative to naive OLS under latent confounding and realistic first-stage misspecification, including misspecification of the embedding dimension. We apply the method to nearly 137,000 film reviews to estimate critic-level gender bias. Naive OLS benchmarks suggest that 26% of critics exhibit statistically significant favoritism toward female-directed films. Our proposed HCF+DML estimator reduces this figure to 2% after accounting for latent critic preferences and film characteristics. For comparison, a leave-one-out crowd benchmark,

*Lyft; mhyrhabibi@gmail.com.

†University of Washington; zkhnl@uw.edu.

‡University of Warwick; negar.ziaeian-ghasemzadeh@warwick.ac.uk.

which compares each critic's ratings to the average rating from other critics for the same film, finds no statistically significant critic-level deviations. These comparisons show that estimates of individual-level bias are highly sensitive to how evaluator preferences and item quality are controlled for. The application illustrates how preference sorting can be mistaken for discrimination in observational interaction data, with implications for other bilateral evaluation settings such as hiring, lending, and peer review.

1 Introduction

Discrimination in employment, credit, housing, and other markets remains a central economic and policy concern. In many settings, however, regulators and researchers need more than an estimate of average discrimination: they need to quantify bias at the level of individual agents—such as whether particular firms, hiring managers, or loan officers systematically treat otherwise comparable people differently based on protected attributes like gender, race, or age. Identifying such bias is challenging in observational data because protected attributes may be correlated with unobserved determinants of decisions, such as productivity, risk, or match quality, making it difficult to separate discrimination from omitted characteristics. As a result, much of the empirical literature has turned to correspondence studies (Bertrand and Mullainathan, 2004; Bertrand and Duflo, 2017) and blind versus non-blind evaluations Blank (1991), where group membership can be randomized or concealed, and discrimination can be identified cleanly. Most studies using these designs consequently emphasize average treatment effects, rather than isolating which specific decision-makers consistently exhibit bias. For audit-style applications of fairness in matching markets—hiring committees, lending officers, peer reviewers, content-moderation panels—the practitioner needs more than a population-level treatment effect: they need a defensible per-decision-maker statistic that is robust to taste-driven sorting, which is precisely what we develop below.

This paper develops a method to estimate *individual-level discrimination* from observational data in *bilateral interaction* settings. By bilateral, we mean a setting with many evaluators and many targets, where evaluators assess several targets, and each target is assessed by several evaluators.We focus on a setting like film criticism, where each critic reviews many films and each film receives reviews from many critics. In this environment, an observed review score combines at least three forces. First, critics differ in (often unobserved) tastes-for example, some critics consistently prefer slower, arthouse films while others favor mainstream genres. Second, films differ in

latent quality and other attributes that directly influence reviews, even if they are imperfectly measured in the data. Third, critics may apply different standards to films associated with a protected attribute. For instance, they may systematically rate films directed by women differently than otherwise similar films directed by men. Separating the individual-level bias from tastes and attributes is the paper's core methodological challenge.

Our central contribution is a new estimator that combines collaborative filtering (CF) with double/debiased machine learning (DML) to estimate individual-level discrimination from observational data, especially in bilateral evaluation settings where key confounders are latent and rich covariates are missing or unavailable due to privacy constraints. CF is widely used in recommender systems: it infers a small set of underlying "taste" and "feature" dimensions for users and items from incomplete rating data, where most user–item pairs are never observed (Koren et al., 2009). We repurpose this machinery for causal inference in bilateral evaluation settings. The CF produces embeddings that summarize latent evaluator preferences and latent item characteristics-key confounders that would otherwise bias estimates of discrimination in observational data. After extracting these preference and characteristic components, we use DML (Chernozhukov et al., 2018) to estimate evaluator-specific treatment effects (i.e., unit-level discrimination) while flexibly controlling for the high-dimensional embeddings and other observables. We use DML to obtain valid inference after orthogonalizing with respect to learned nuisance components, and we combine it with cross-fitting / sample splitting to account for the fact that those nuisances (including CF embeddings) are estimated.

A key challenge is that standard collaborative filtering can unintentionally absorb the protected attribute, and the very discrimination signal we want to estimate, into the embeddings themselves. If we fit CF on the full set of observed outcomes, the resulting "preference" factors may partly reflect whether an observation is treated (e.g., associated with a female versus male director), so the representation is no longer a clean summary of tastes and characteristics; it is contaminated by treatment. To avoid this problem, we introduce *Honest Collaborative Filtering* (HCF), inspired by the honest estimation principle in Athey and Imbens (2016). Under HCF, the embeddings are learned from a restricted sample that excludes treated items. Concretely, we train CF only on untreated items (for example, ratings of male-directed films) to learn embeddings that capture critics' tastes without directly incorporating the treatment signal, and then we hold these embeddings fixed when estimating unit-level discrimination with DML.

Existing work at the intersection of causal inference and recommender systems often

asks how causal tools can improve recommendations, by addressing selection bias, estimating the effects of exposure, or evaluating counterfactual rankings. In this paper, we take the reverse approach: we use recommendation methods for causal inference. Collaborative filtering is not the object we aim to optimize; instead, it is a measurement step that extracts latent controls from sparse bilateral interaction data, which we then feed into a separate causal estimator targeting a unit-level treatment effect—*discrimination*. This paper is also distinct from the algorithmic fairness literature, which studies whether machine learning systems generate disparate outcomes and derives constraints or impossibility results for fair prediction (Kleinberg et al., 2017; Chouldechova, 2017). Here, machine learning is used instead as a tool to detect discriminatory *human* decision-makers.

Our key contribution is an *Honest Collaborative Filtering + DML* estimator that identifies unit-level discrimination from observational bilateral interaction data. The approach introduces an honest representation-learning design for collaborative filtering that prevents the protected attribute from being absorbed into the learned preference embeddings. We then integrate these embeddings into a DML method to obtain valid unit-level causal estimates and inference in high-dimensional settings (Chernozhukov et al., 2018). Taken together, the method enables discrimination measurement with minimal inputs: who interacted with whom, and the resulting outcome.

We evaluate the method in both controlled simulations and an empirical application. In Monte Carlo experiments that simulate a gender-differentiated evaluation process with sparse matching, latent preferences, and latent item characteristics, we compare our estimator to a benchmark OLS approach that does not adequately control for latent confounding. We study performance across confounding regimes and show how honesty in the embedding stage affects bias and variance of unit-level estimates. We then apply the method to film ratings data from Metacritic, using 136,847 critic-film ratings across 205 critics and 7,384 films to estimate critic-level bias toward female-directed films. The empirical results illustrate why latent preference control matters. Naive approaches can attribute systematic taste differences and selection patterns to discrimination, producing inflated counts of "significant" decision-makers. Our method sharply reduces these false positives and isolates the subset of evaluators whose rating behavior is consistent with a causal effect of director gender. For audit and accountability applications this matters concretely: the gap between naive flagging (roughly one in four critics) and our preferred per-critic estimate (roughly one in fifty) is precisely the false-positive load that would land on the wrong decision-makers if a regulator or platform used a raw comparison as a discrimination filter.

The remainder of the paper is organized as follows. Section 2 situates our contribution in the literatures on discrimination, causal machine learning, and collaborative filtering. Section 3 presents the model, the HCF construction, and the DML estimation. Section 4 presents the application to film critics. Section 5 reports simulation results. Section 6 concludes.

2 Literature Review

This paper sits at the intersection of three main literatures. First, it contributes to the economics of discrimination by developing tools to measure discrimination at the decision-maker level rather than only on average. Second, it builds on causal machine learning, especially double/debiased machine learning and honest estimation ideas, to recover unit-level causal effects with valid inference in high-dimensional settings. Third, it draws on collaborative filtering and matrix factorization from recommender systems, repurposing these methods as a way to learn latent preferences and latent item attributes from sparse interaction data that can then be used as controls for causal identification.

The economic literature finds that discrimination is rooted in two foundational frameworks: taste-based discrimination, which models prejudice as a preference that decision-makers are willing to pay to indulge (Becker, 1957), and statistical discrimination, where differential treatment arises from inference under incomplete information (Arrow, 1973; Phelps, 1972). Empirically, the most evidence has come from field experiments, especially correspondence studies and related audit designs, which randomize or conceal group membership to cleanly identify differential treatment (Bertrand and Mullainathan, 2004; Bertrand and Duflo, 2017). While these designs provide credible estimates of average discrimination, they are costly, constrained in many settings, and typically not built to identify which specific decision-makers discriminate. Recent work has moved toward decision-maker-level discrimination using large-scale audits and empirical Bayes approach (Kline and Walters, 2021; Kline et al., 2022, 2023), as well as quasi-experimental designs that leverage (quasi-)random assignment of cases to decision-makers (Arnold et al., 2018). Our contribution complements these advances by providing a method to estimate unit-level discrimination using observational bilateral interaction data by exploiting the interaction structure to control for latent preferences and latent item attributes.

In causal machine learning, a central goal is to use flexible prediction tools while still obtaining valid estimates and inference for causal parameters. Double/Debiased Machine Learning (DML) provides a leading framework for doing so by combining

Neyman-orthogonal moment conditions with cross-fitting, allowing high-dimensional nuisance components to be learned with modern machine learning methods without invalidating inference on the causal parameter of interest (Chernozhukov et al., 2018). A related literature studies heterogeneous treatment effects and emphasizes "honest" estimation—separating the data used to choose a model or partition from the data used to estimate effects—starting with causal trees (Athey and Imbens, 2016) and extending to causal forests (Wager and Athey, 2018). This paper contributes to this literature by adapting these ideas to sparse bilateral interaction data: we treat the collaborative-filtering embeddings as high-dimensional nuisance controls inside a DML method, and we introduce an "honest" representation-learning design to prevent the protected attribute from being absorbed into the representation used for causal estimation (Athey and Imbens, 2016).

Collaborative filtering and matrix factorization are central tools in recommender systems, especially after the Netflix Prize highlighted their practical success. In this literature, matrix factorization decomposes a sparse user–item outcome matrix into low-dimensional latent factors that summarize users' preferences and items' characteristics, with regularization used to improve prediction and prevent overfitting (Koren et al., 2009). Our paper repurposes this machinery for causal inference rather than prediction: in bilateral evaluation data, the latent factors learned by collaborative filtering provide a compact way to capture unobserved evaluator tastes and item attributes that would otherwise confound observational estimates of discrimination. We use collaborative filtering embeddings as controls inside a causal framework to recover unit-level treatment effects.

The matching-evaluation setting we study shares its primitive, a partially observed $N \times J$ interaction matrix, with the interactive fixed effects (IFE) literature in econometrics. Bai (2009) shows how to consistently estimate panel slope coefficients in the presence of a bilinear latent factor term $\lambda_i' f_j$, and Moon and Weidner (2015) extend the theory to weakly stationary regressors; Pesaran (2006) develops a complementary common-correlated-effects estimator. Closer to our setting, Athey et al. (2021) develop a matrix-completion estimator for panel causal inference, and Griffith and Peng (2023) identify latent factor structure in sparse networked data. Our HCF+DML approach differs in three ways. (i) We do not assume the latent preference term is bilinear in observed dimensions; the HCF first stage learns a flexible item-side embedding $\hat{C}_i$ and a per-evaluator nonlinear map $\hat{g}_j(\hat{C}_i, \hat{P}_j)$ rather than imposing $\lambda_i' f_j$. (ii) Unlike IFE we target per-unit treatment heterogeneity θ_j rather than a common slope, which is the relevant quantity when individual-level discrimination is the object of interest. (iii) We obtain Neyman-orthogonal inference for θ_j via DML

cross-fitting, so first-stage estimation error in the embeddings enters at second order. The closest method conceptually is Athey et al. (2021)'s matrix-completion estimator combined with sample splitting; we add an *honest* restriction (training embeddings only on untreated cells) to prevent the treatment signal from being absorbed into the latent representation.

3 Conceptual Framework and Methodology

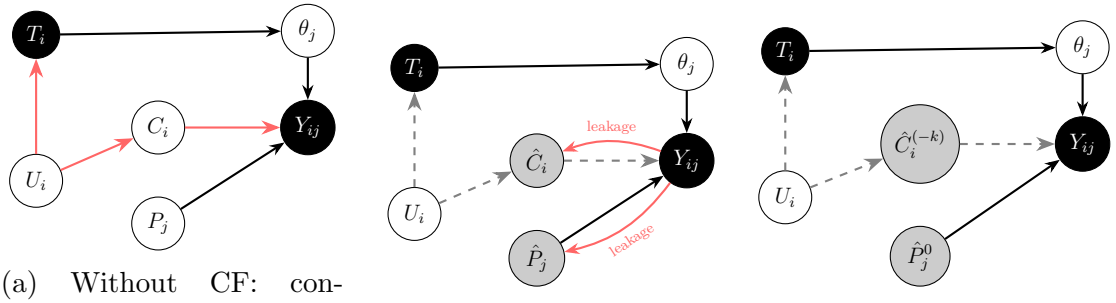
The conceptual framework of this study is based on interactions between two distinct sets of entities: individuals $i \in I$ who are evaluated, and evaluators $j \in J$ who assess them. There is a many-to-many relationship between the two sets: each individual i may be evaluated by multiple evaluators, and each evaluator j may assess multiple individuals. This setting is analogous to the job market, where applicants apply to multiple employers and employers receive applications from multiple candidates. Following this analogy, we refer to individuals $i \in I$ as applicants and entities $j \in J$ as employers throughout this section, though the framework applies equally to other bilateral evaluation settings such as film critics reviewing movies or teachers grading students.

Each applicant is associated with a trait $T_i \in \{0, 1\}$ (e.g., gender or race) that may be subject to bias from employers. Beyond potential biases, employers have preferences P_j over applicant characteristics C_i that may be correlated with the trait. The researcher observes only the trait T_i and outcome Y_{ij} (e.g., hiring decision or rating), but not the latent preferences or characteristics. The goal is to determine whether and which employers exhibit bias in their evaluations, beyond what is attributable to legitimate preferences over applicant characteristics.

3.1 The Core Empirical Challenge

Figure 1 illustrates the empirical challenge using causal diagrams (Hernán and Robins, 2020). The treatment T_i (e.g., applicant gender) affects the outcome Y_{ij} through two channels. The first is the *direct* channel through evaluator bias θ_j: evaluators may systematically favor or disfavor individuals based on group membership, independent of qualifications. The second is the *indirect* channel through characteristics C_i: if treated individuals differ systematically in their latent characteristics—for instance, if female applicants cluster in certain fields or female directors concentrate in certain genres—then evaluators with preferences P_j over these characteristics will rate treated individuals differently even absent any bias.

In figure 1 , black nodes (T, Y) denote variables observed by the researcher, while white nodes (U, C, P, θ) represent latent or unobserved quantities. Gray nodes indicate variables that are controlled for in estimation. The unobserved factor U represents systematic differences in how treated and untreated individuals sort into different types of work (genres, fields, industries), creating a backdoor path $T \leftarrow U \rightarrow C \rightarrow Y$ that confounds naive estimates of bias. Our goal is to estimate θ_j for individual evaluators $j \in J$ by blocking this confounding path while avoiding post-treatment contamination.



(a) Without CF: confounded

(b) Naive CF: leakage risk

(c) Honest CF: identified

Figure 1: **Causal structure and identification.** Observed variables are shown in black, unobserved in white, and controlled variables in gray. T_i denotes the item-level treatment (e.g., female applicant), Y_{ij} the outcome (e.g., rating or callback by evaluator j for item i), θ_j the evaluator-specific bias, C_i the latent characteristics of item i, P_j the evaluator's preferences, and U_i unobserved item-level factors (e.g., field or genre) that influence both T_i and C_i. **(a)** Without controlling for latent factors, the backdoor path $T_i \leftarrow U_i \rightarrow C_i \rightarrow Y_{ij}$ confounds the effect of T_i on Y_{ij}, biasing naive estimates of θ_j. **(b)** Naive collaborative filtering estimates $(\hat{C}_i, \hat{P}_j)$ from outcomes including Y_{ij}. Because Y_{ij} is itself affected by θ_j, the learned embeddings can absorb evaluator j's discriminatory bias: leakage into $\hat{P}_j$ risks learning discrimination as "preferences," and leakage into $\hat{C}_i$ contaminates each evaluator's control variables used to estimate θ_j. **(c)** Honest CF estimates $\hat{P}_j^0$ using untreated observations only and cross-fits item embeddings $\hat{C}_i^{(-k)}$ by excluding all evaluators in fold k when constructing $\hat{C}_i$, eliminating treatment contamination from any evaluator $j \in J_k$.

Formally, the researcher's objective is to estimate θ_j for each evaluator $j \in J$ from the following model:

$$Y_{ij} = \theta_j T_i + g(C_i, P_j) + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the outcome, T_i is the treatment indicator, $g(\cdot)$ captures how the match between applicant characteristics C_i and employer preferences P_j affects outcomes,

and ε_{ij} is an idiosyncratic error. The challenge is that neither C_i, P_j, nor $g(\cdot)$ is observed. How can one account for unobserved characteristics and preferences when no direct measurements are available?

3.2 Collaborative Filtering

The answer lies in collaborative filtering (CF), a technique widely used in machine learning for recommendation systems. CF exploits the insight that if person A agrees with person B on some items, A is likely to share B's opinion on other items as well. In essence, CF constructs latent representations of users' preferences and items' characteristics from observed interaction patterns, then uses these representations to predict unobserved interactions.

We employ Regularized Matrix Factorization (RMF), a fundamental CF method requiring minimal data (Koren et al., 2009). RMF begins with the outcome matrix R of dimension $|I| \times |J|$, where entry Y_{ij} records evaluator j's assessment of individual i, with missing entries where no evaluation occurred. This matrix is typically sparse, as evaluators assess only a fraction of all individuals. RMF decomposes the high-dimensional sparse matrix R into two lower-dimensional matrices: C of dimension $|I| \times d$ (individual characteristics) and P of dimension $|J| \times d$ (evaluator preferences), where the embedding dimension $d \ll |I|, |J|$ is a hyperparameter. The objective is for the product CP' to approximate the observed entries in R. Using mean squared error loss with regularization on the parameter magnitudes, the loss function is:

$$L = \frac{1}{|M|} \sum_{(i,j) \in M} \left(Y_{ij} - C_i P_j'\right)^2 + \lambda \left(\frac{1}{|I|} \sum_i \|C_i\|_2^2 + \frac{1}{|J|} \sum_j \|P_j\|_2^2 \right). \tag{2}$$

where M denotes the set of observed entries and λ is a regularization parameter. Regularized matrix factorization operates on the *observed* entries of R only; missing entries are simply omitted from the loss in Equation (2), a standard property of low-rank matrix completion methods (Mazumder et al., 2010; Koren et al., 2009). We optimize the Mean Squared Error (MSE) with mean L_2 regularization. Normalizing by the number of observations $|M|$, items $|I|$, and evaluators $|J|$ ensures the regularization coefficient λ is invariant to dataset size.

The key property of RMF is that it distills information from the outcome matrix into compact latent representations, where P captures evaluators' latent preferences and C captures individuals' latent characteristics. Unlike standard recommender system

implementations which often separate scalar bias terms for evaluators and items (Koren et al., 2009), we deliberately omit explicit intercepts in Equation 2. This design choice forces the latent embeddings P and C to capture the full variation in ratings, including baseline evaluator leniency and item quality, ensuring they act as comprehensive controls in the subsequent Double Machine Learning stage. In constructing these latent spaces, the model positions similar evaluators, and similar individuals, close to one another, allowing RMF to infer rich representations of both parties based solely on observed evaluation outcomes.

3.3 Honest Collaborative Filtering

Applying standard CF to study discrimination is problematic because it may incorporate evaluator biases into the learned embeddings. If a biased evaluator systematically rates treated individuals lower, this pattern will be reflected in both the evaluator's preference vector and the individuals' characteristic vectors, contaminating the controls we seek to construct.

To address this, we propose *honest collaborative filtering*, an approach inspired by the "honest trees" of Athey and Imbens (2016). The method has two components:

First, to ensure that trait-based biases do not contaminate evaluator preferences, we estimate preferences using only outcomes from the untreated group. Specifically, we factorize the restricted matrix $R_0 = \{Y_{ij} : T_i = 0\}$, which includes only evaluations of individuals with $T = 0$. This yields preference estimates $\hat{P}^0$ that reflect evaluators' tastes over characteristics, uncontaminated by any bias toward the treated group.

Second, to ensure that an evaluator's own bias does not contaminate the characteristic estimates used as controls for that evaluator, we employ cross-fitting. We partition the set of evaluators J into K disjoint subsets $J_1, \ldots, J_K$. For each subset J_k, we construct the outcome matrix R_{-k} using only evaluations from evaluators *not* in J_k, then factorize R_{-k} to obtain characteristic estimates $\hat{C}_i^{(-k)}$. When estimating bias for evaluators in fold J_k, we use $\hat{C}^{(-k)}$ as controls. This ensures that no evaluator $j \in J_k$ has her own ratings influencing the characteristic estimates used in her regression.

This design relies on sufficient common support between treated and untreated items in the latent characteristic space. Assumption 4 ensures that, for any configuration of latent characteristics and preferences, both treated and untreated items are observed with positive probability. For the honest embedding construction, we require a stronger implication: the *marginal* distribution of characteristics among untreated

items must cover the regions of the characteristic space occupied by treated items. This condition ensures that preferences learned from R_0 remain informative when applied to treated observations.

More precisely, if evaluator j's own ratings were used to estimate $\hat{C}_i$, then the residual ε_{ij} would be mechanically correlated with $\hat{C}_i$ through the matrix factorization objective, violating the independence condition required for DML consistency. Preferences do not require analogous cross-fitting because $\hat{P}_j^0$ is estimated from the untreated matrix R_0, and the DML regression for evaluator j uses both treated and untreated observations: the treated observations in j's regression are never used in estimating $\hat{P}_j^0$, so the relevant independence is preserved for those observations.

With these honest embeddings, Equation (1) becomes:

$$Y_{ij} = \theta_j T_i + g(\hat{C}_i^{(-k)}, \hat{P}_j^0) + \varepsilon_{ij} \tag{3}$$

where the embeddings $\hat{P}^0$ and $\hat{C}^{(-k)}$ are constructed to avoid the contamination illustrated in Figure 1b.

3.4 Double Machine Learning Estimation

Two challenges remain in estimating Equation (3). First, the function $g(\cdot)$ relating embeddings to outcomes is unknown and must be estimated from data. Second, the treatment indicator T_i may itself be correlated with the learned embeddings $(\hat{C}_i^{(-k)}, \hat{P}_j^0)$. This makes it important to account for treatment-selection patterns in addition to flexibly modeling outcomes.

To address these challenges, we adopt the Double/Debiased Machine Learning (DML) framework of Chernozhukov et al. (2018). Consider the partially linear model:

$$\begin{aligned} Y &= \theta_0 D + g_0(X) + U, \quad \mathbb{E}[U \mid X, D] = 0, \\ D &= m_0(X) + V, \quad \mathbb{E}[V \mid X] = 0. \end{aligned}$$

where D is the treatment, Y is the outcome, and X is a vector of controls. The first equation models the outcome as a function of treatment and controls; the second models treatment assignment as a function of controls. Under conditional exogeneity, treatment is as good as random conditional on X, so θ_0 has a causal interpretation.

When X is high-dimensional, flexible estimation of the nuisance functions is needed, and standard low-dimensional parametric approaches may be restrictive. A naive approach using regularized machine learning directly in the outcome equation can

yield biased estimates of θ_0, because regularization error in the nuisance component can contaminate the treatment coefficient. Chernozhukov et al. (2018) address this problem through orthogonalization and cross-fitting. In the residual-on-residual implementation, we estimate the nuisance functions $\ell_0(X) = E[Y \mid X]$ and $m_0(X) = E[D \mid X]$ on auxiliary folds and evaluate them on held-out folds. We then compute the residualized treatment and outcome:

$$\hat{V}_i = D_i - \hat{m}_0(X_i), \qquad \tilde{Y}_i = Y_i - \hat{\ell}_0(X_i).$$

The debiased estimator is:

$$\hat{\theta}_0 = \frac{\sum_i \hat{V}_i \tilde{Y}_i}{\sum_i \hat{V}_i^2}.$$

This is simply OLS of the residualized outcome on the residualized treatment. Orthogonalization makes the estimate less sensitive to regularization error in the nuisance functions, while cross-fitting reduces overfitting bias from using flexible machine learning methods.

We adapt this framework to estimate evaluator-level bias parameters. The model becomes:

$$\begin{aligned} Y_{ij} &= \theta_j \, T_i + g(\hat{C}_i^{(-k)}, \hat{P}_j^0) + \varepsilon_{ij} \\ T_i &= m(\hat{C}_i^{(-k)}, \hat{P}_j^0) + \epsilon_{ij} \end{aligned} \tag{4}$$

The first equation models outcomes among observed evaluator-individual pairs. The second equation models treatment status conditional on the learned embeddings among those observed pairs. For each evaluator j, we estimate θ_j using the DML procedure applied to j's subset of observations. Because estimation is performed separately for each evaluator, $\hat{P}_j^0$ is fixed within evaluator j's regression and indexes the evaluator-specific nuisance functions; within-evaluator identification of θ_j comes from variation across items in $\hat{C}_i^{(-k)}$ and T_i. The honest restriction on $\hat{P}^0$ estimation (factorizing R_0 only) thus protects the conditioning set in the population model rather than supplying within-regression variation; identification is driven by the cross-fit $\hat{C}_i^{(-k)}$ and per-evaluator stratification.

Algorithm 1 summarizes the complete methodology.

Algorithm 1 Honest Collaborative Filtering with Double Machine Learning

1. **Estimate honest preferences:** Construct $R_0 = \{Y_{ij} : T_i = 0\}$ and factorize to obtain $\hat{P}^0$.
2. **Partition evaluators:** Split J into K disjoint folds $J_1, \ldots, J_K$.
3. **Cross-fit characteristics:** For each fold $k \in \{1, \ldots, K\}$:
   (a) Construct R_{-k} using ratings from evaluators not in J_k;
   (b) Factorize R_{-k} to obtain $\hat{C}^{(-k)}$.
4. **Estimate bias via DML:** For each evaluator $j \in J_k$, estimate θ_j from Equation (4) using embeddings $(\hat{C}^{(-k)}, \hat{P}^0_j)$.

3.5 Assumptions

We now state the assumptions under which our estimator of evaluator-level bias parameters θ_j admits a causal interpretation. Let Y_{ij} denote the outcome when evaluator j assesses individual i, let $T_i \in 0, 1$ indicate membership in the treated group, let $C_i \in \mathbb{R}^d$ denote latent characteristics of individual i, and let $P_j \in \mathbb{R}^d$ denote the latent preferences of evaluator j.

Assumption 1 (Preference Stability)**.** *The evaluator preference vector P_j that enters the match quality function $g(C_i, P_j)$ is invariant to treatment status. That is, the same P_j governs evaluator j's valuation of characteristics C_i for both treated ($T_i = 1$) and untreated ($T_i = 0$) items. This ensures that preference embeddings estimated from untreated observations alone remain valid controls when applied to the full sample.*

This assumption requires that critics' preferences operate identically when evaluating male- and female-directed films. It would be violated if, for example, critics apply different evaluative criteria by director gender, so that the same latent preference vector P_j maps into different dimensions of perceived quality when $T_i = 1$ versus $T_i = 0$. Under Assumption 1, preference embeddings estimated from male-directed films alone provide valid controls for the full sample.

Assumption 2 (Additive Discrimination)**.** *The conditional expectation of outcomes follows an additive structure:*

$$E[Y_{ij} \mid T_i, C_i, P_j] = \theta_j \cdot T_i + g(C_i, P_j) \tag{5}$$

where θ_j is evaluator j's bias parameter and $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ captures the match quality between individual characteristics and evaluator preferences.

This assumption embeds our causal parameter within a collaborative filtering framework. The function $g(C_i, P_j)$ represents how well individual i's characteristics align

with evaluator j's preferences. The additive separability implies that bias operates as a constant shift in outcomes for treated individuals, independent of their characteristics.

Assumption 3 (Conditional Independence)**.** *For each evaluator $j \in J$, conditional on latent characteristics C_i and evaluator preferences P_j, treatment assignment is independent of the idiosyncratic error:*

$$T_i \perp \varepsilon_{ij} \mid C_i, P_j. \tag{6}$$

Here ε_{ij} is defined by the structural equation

$$Y_{ij} = \theta_j T_i + g(C_i, P_j) + \varepsilon_{ij}.$$

This assumption is the central condition under which we interpret θ_j as reflecting evaluator-level bias rather than residual differences in match quality. It requires that, after conditioning on the latent determinants of match quality captured by C_i and P_j, treatment status contains no additional information about the unobserved, idiosyncratic component of the evaluation. The restriction imposes no structure on heterogeneity in θ_j, so evaluator-specific bias may be arbitrary, provided that confounding which links treatment to outcomes through unobserved determinants of C_i is absorbed by the latent factors (C_i, P_j). In the movie-critic setting, conditional on (C_i, P_j), whether a film is female-directed does not predict ε_{ij}.

Assumption 4 (Overlap)**.** *For all (C_i, P_j) in the support of (C_i, P_j),*

$$0 < \Pr(T_i = 1 \mid C_i, P_j) < 1. \tag{7}$$

Overlap ensures that for any joint configuration of latent item characteristics and evaluator preferences, we observe both treated and untreated items among comparable evaluator–item pairs, enabling comparisons that hold fixed both C_i and P_j while varying only treatment status T_i.

Assumption 5 (Ignorable Selection)**.** *Conditional on treatment and latent factors, selection into observation is independent of potential outcomes:*

$$Y_{ij}(t) \perp\!\!\!\perp \mathbf{1}[(i,j) \in M] \mid T_i, C_i, P_j \quad \text{for } t \in \{0,1\} \tag{8}$$

where M denotes the set of observed evaluator-individual pairs.

This assumption requires that which pairs we observe depends only on preferences and characteristics, not on the potential outcome itself. In practice, selection may partly depend on expected enjoyment—critics may avoid films they expect to dislike. If this selection operates through the same latent factors (C_i, P_j) that drive ratings, Assumption 5 is satisfied. The assumption is violated if critics select based on factors that predict ratings beyond what (C_i, P_j) capture, such as private information about a film's quality from advance screenings. We assess robustness to MNAR selection in Appendix A.6, where the method maintains its advantage even under selection driven by outcome-relevant unobservables.

The DML framework additionally requires that the nuisance functions g and m converge sufficiently fast; specifically, the product of their estimation errors must be $o_p(n^{-1/2})$, the standard rate condition of Chernozhukov et al. (2018). When both nuisance estimators converge at the same rate, this reduces to each achieving $o_p(n^{-1/4})$. For the collaborative filtering stage, Koltchinskii et al. (2011) establish that low-rank matrix completion achieves consistent recovery of latent factors under standard incoherence and sampling conditions, providing the theoretical basis for embedding convergence.

We acknowledge several caveats. First, Assumption 3 can fail if relevant confounders are largely orthogonal to the low-rank structure that collaborative filtering recovers. Second, results depend on embedding quality: if the true latent structure is poorly approximated at the chosen embedding dimension, or if the honest design leaves too little data for precise embedding estimation, estimates can become biased or noisy. Third, our framework targets additive evaluator-specific parameters θ_j and does not capture discrimination that varies with item characteristics. Extending the model to allow item-dependent bias, for example $\theta_j(C_i)$, is a natural direction for future work.

4 Movie Critics' Bias Toward Female-Directed Movies

The ensuing section presents an empirical application of the established methodology using a real-world dataset. This approach is instrumental in demonstrating the framework's efficacy to discern micro-level discrimination in practical settings.

4.1 Data Description

A dataset comprising film reviews from professional critics was constructed using data from Metacritic.com, a review aggregator website. Metacritic collects reviews from approximately 100 sources, assigning ratings on a uniform scale of 0-100. These

ratings were transformed to a 0-1 scale for this analysis[1]. The objective is to apply the methodology described in Section 3 to explore potential discriminatory patterns in critics' reviews of movies directed by women.

Metacritic provides detailed information, such as the names of film directors. The gender of directors was deduced using their first names and the *gender-guesser* Python library[2], a prevalent tool for name-based gender inference. Entries were removed if *gender-guesser* was unable to make a prediction (name not found in its database) or if the name was non-specific to a particular gender. This gender identification procedure was verified for accuracy against a Wikipedia directory of female directors[3], with a misclassification rate under 5%. To maintain simplicity in the analysis, films with more than one director were excluded, as over 95% of movies in the dataset had a single director.

Table 1 offers a summary of statistics for selected variables in the dataset. Data was collected for films released from 1990 to 2021 and having at least ten critic reviews on Metacritic. We further restrict the sample to critics who had reviewed a minimum of 30 films directed by female directors, and to single-director films matched to movie-level covariates (year, runtime, six major genre dummies, and budget). The final analysis sample contains 136,847 reviews from 205 critics, spanning 7,384 films. Films directed by women constitute 12.4% of the analysis dataset. Each film is reviewed, on average, by 18.5 analysis-sample critics, with an average rating of 0.62 on a 0–1 scale.

Table 1: Summary Statistics of Selected Variables

	Count	Mean	Std.	Min	Med	Max
Year	7,384	2009.3	7.38	1990	2010	2021
Critic Rating	136,847	0.623	0.211	0.00	0.63	1.00
Films' # of Reviews	7,384	18.5	9.08	1	17	47

Notes: The table shows summary statistics for the selected variables in the data. The data is limited to reviews from critics who have evaluated at least 30 movies directed by women.

[1] In cases where an explicit rating is absent, Metacritic's evaluators assign a score reflecting their assessment of the article.

[2] https://pypi.org/project/gender-guesser/

[3] https://en.wikipedia.org/wiki/List_of_female_film_and_television_directors

4.2 Estimation

As underscored earlier in this study, the estimation of discrimination or favoritism at the individual level is of considerable significance for several reasons. Chief among them is the potential for aggregate-level bias estimates to be misleading. To illustrate, consider a hypothetical scenario in our context: a seemingly minor bias against movies directed by women could arise either from a general absence of discrimination among critics or from the presence of two distinct groups of reviewers, one disproportionately critical and the other overly favorable towards female-directed films. While both scenarios lead to similar estimates of aggregate-level bias, they depict starkly different realities of micro-level discrimination.

In the assessment of individual-level biases or favoritism, the role of personal preferences among decision-makers is pivotal. For instance, in this analysis, a critic's preference for particular genres or themes, if they are more frequently found in films directed by either gender, might inadvertently color their reviews. This genre or theme preference could manifest as apparent gender bias in reviews, while it truly stems from the critic's own cinematic tastes. Overlooking these personal preferences risks incorrectly categorizing critics as biased.

Therefore, we implement the method outlined in Section 3 to estimate individual-level bias/favoritism regarding critics' evaluation of female-directed films. The approach involves estimating the following Double Machine Learning (DML) model

$$\begin{aligned} r_{i,j} &= \theta_j \, FD_i + g(\hat{C}_i^{(-k)}, \hat{P}_j^0) + \varepsilon_{ij} \\ FD_i &= m(\hat{C}_i^{(-k)}, \hat{P}_j^0) + \epsilon_{ij} \end{aligned} \tag{9}$$

Here, $r_{i,j}$ represents the rating given by critic j to film i, while FD_i is a binary indicator denoting whether film i was directed by a woman. The parameter θ_j is critic j's bias/favoritism toward films directed by women.

In this dataset, typically, critics review only a limited selection of films, and correspondingly, each film is assessed by a small group of critics. With a total of 7,384 films and 205 critics, the analysis sample comprises 136,847 observed ratings, indicating that under 10% of all possible ratings are recorded. We denote by $r_{i,j}$ the rating critic j assigns to film i; when $r_{i,j}$ is unobserved, critic j simply did not review film i. This *sparsity* is a common feature in similar contexts where each item or application is evaluated by only a fraction of potential reviewers. Notably, the use of Collaborative Filtering in industrial settings is intended to predict ratings that a user might assign to items they have not yet reviewed (such as books, music, or movies) and to recommend items likely to be highly rated by the user.

As outlined in Algorithm 1, the procedure begins by applying regularized matrix factorization (RMF) to the rating matrix $R_{I\times J}$, restricted to films directed by men, to obtain evaluator embeddings $\hat{P}^0_{J\times d}$. RMF initializes the latent matrices randomly and minimizes the regularized objective in Equation (2) using gradient-based optimization. We fix the regularization parameter at $\lambda = 0.01$. We select the embedding dimension d by cross-validation, evaluating $d \in \{10, 25, 50, 100, 200\}$ and choosing the value that minimizes the validation mean squared error. This application aims to recover $\hat{P}^0$, the matrix of critics' latent preferences, deliberately isolated from their evaluations of films directed by women.

Upon deriving $\hat{P}^0$, the second step of Algorithm 1 involves randomly dividing critics into $K = 10$ folds. For each fold k, RMF is then applied to the ratings matrix R_{-k}, comprising ratings data from critics not in fold k, to generate $\hat{C}^{(-k)}$. Here, $\hat{C}^{(-k)}$ are the film-characteristic embeddings, isolated from the ratings by critics in fold k.

In the final step, $\hat{P}^0$ and the $\{\hat{C}^{(-k)}\}_{k=1}^K$ matrices are used to obtain DML estimates for the model in Equation 9. For the nuisance function $m(\cdot)$, we use a binary logistic classifier with ℓ_2 regularization, with the regularization strength set to the default value in *scikit-learn*. For the outcome model $g(\cdot)$, we use a Random Forest regressor and follow the default hyperparameter choices recommended for partially linear regression in the `DoubleML` package (Bach et al., 2024, 2022). We employ 5-fold cross-fitting for valid inference. Standard errors are computed using the influence function approach standard in DML, which treats embeddings as fixed. In principle, a bootstrap procedure re-estimating embeddings in each replication would account for first-stage uncertainty. Since our primary goal is to clarify the proposed methodology, we do not perform hyperparameter tuning or a systematic comparison across learners; in practice, both tuning and alternative learners provide useful robustness checks.

To compare the proposed method with conventional approaches, we also estimate three nested ordinary-least-squares (OLS) specifications. Each specification is run separately by critic; critic fixed effects are therefore absorbed mechanically by the per-critic stratification, and all specifications use HC3 robust standard errors.

$$\begin{aligned}
\text{(i) Naive OLS:} \quad & r_{i,j} = \alpha_j + \beta_j^{(1)}\, FD_i + e_{i,j} \\
\text{(ii) OLS with controls:} \quad & r_{i,j} = \alpha_j + \beta_j^{(2)}\, FD_i + X_i^\top \gamma_j + e_{i,j} \\
\text{(iii) OLS with controls + LOO movie quality:} \quad & r_{i,j} = \alpha_j + \beta_j^{(3)}\, FD_i + X_i^\top \gamma_j + \eta_j\, \overline{r}_{i,-j} + e_{i,j}.
\end{aligned} \tag{10}$$

Here FD_i indicates whether film i was directed by a woman, and X_i collects all available pre-treatment film covariates: release year, log(runtime), dummies for the six major genres (Drama, Comedy, Thriller, Romance, Action, Horror), $\log(1 + \text{budget}_i)$, and an indicator for budget-missing.

The regressor $\overline{r}_{i,-j}$ in specification (iii) is the *leave-one-out* (LOO) mean Metacritic score for film i, computed from every other critic who rated it:

$$\overline{r}_{i,-j} \;=\; \frac{1}{|\mathcal{C}_i \setminus \{j\}|} \sum_{j' \in \mathcal{C}_i \setminus \{j\}} r_{i,j'}, \tag{11}$$

where $\mathcal{C}_i$ is the set of all single-director Metacritic critics who reviewed film i in the wider Metacritic catalog, not only the 205 critics in our analysis sample. The construction excludes critic j's own score, so $\overline{r}_{i,-j}$ does not mechanically include the dependent variable $r_{i,j}$; the denominator is at least 9 per cell because films enter the analysis only when at least ten critics reviewed them.

Including $\overline{r}_{i,-j}$ as a regressor controls for film quality using an average review or public score. We therefore treat specification (iii) as the strongest crowd-based OLS comparator we report. A literal two-way fixed-effects specification with both critic and movie fixed effects is not identified for θ_j, because FD_i is constant within film and therefore collinear with movie fixed effects. The LOO crowd mean is a feasible proxy for film-level consensus quality that preserves identification of the treatment coefficient.

This benchmark should be interpreted as a crowd-deviation specification rather than a gold-standard adjustment for true film quality. The leave-one-out mean captures how other critics rated the same film, which may itself reflect systematic gender-related differences in evaluation. If the broader reviewer pool collectively over- or under-rates female-directed films, that component is absorbed into $\overline{r}_{i,-j}$. Specification (iii) therefore asks whether critic j deviates from the reviewing crowd, not whether critic j is unbiased relative to true film quality. The parameter $\beta_j^{(s)}$ in specification s is critic j's estimated bias under that specification.

4.3 Results

We estimate critic-level bias parameters using both naive OLS and HCF+DML, then compare the distributions, magnitudes, and statistical significance of the resulting estimates.

4.3.1 Distribution of Bias Estimates

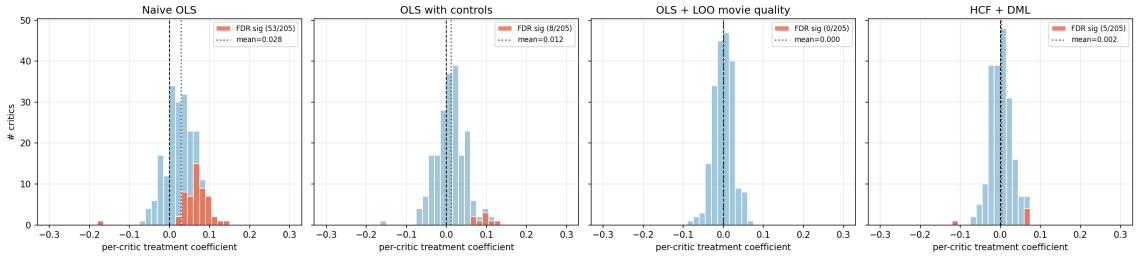
Figure 2a displays the distribution of estimated bias coefficients across the three OLS specifications in Equation (10) and the HCF+DML model. Naive OLS produces a right-skewed distribution centered at 0.028, with 53 critics (26%) FDR-significant under Benjamini–Hochberg at $\alpha = 0.10$. Adding the full set of pre-treatment film controls (year, log runtime, six genre dummies, log(1 + budget), and a budget-missing indicator) attenuates the mean to 0.012 and cuts significant critics to 8. Our strongest OLS comparator further adds the leave-one-out mean Metacritic score as a film-quality proxy; this produces a mean indistinguishable from zero (0.0003) and *no* FDR-significant critics. HCF+DML lies between the controls and LOO columns on mean magnitude (0.002) but identifies 5 FDR-significant critics (4 positive, 1 negative), more than the LOO comparator. The reason is conceptual rather than technical: LOO measures *within-crowd* deviation, while HCF+DML measures *within-taste* deviation. We return to this distinction below.

Table 2 summarizes the estimation results across the four specifications. The mean naive OLS estimate is 0.028; this falls to 0.012 once the full set of pre-treatment film controls is added, to 0.0003 under the LOO movie-quality control, and to 0.002 under HCF+DML—a 92% attenuation relative to naive OLS. The standard deviation of estimates also compresses, from 0.040 (naive) to 0.025 (LOO) and 0.028 (HCF+DML), indicating that controlling for confounding tightens the distribution of estimated biases.

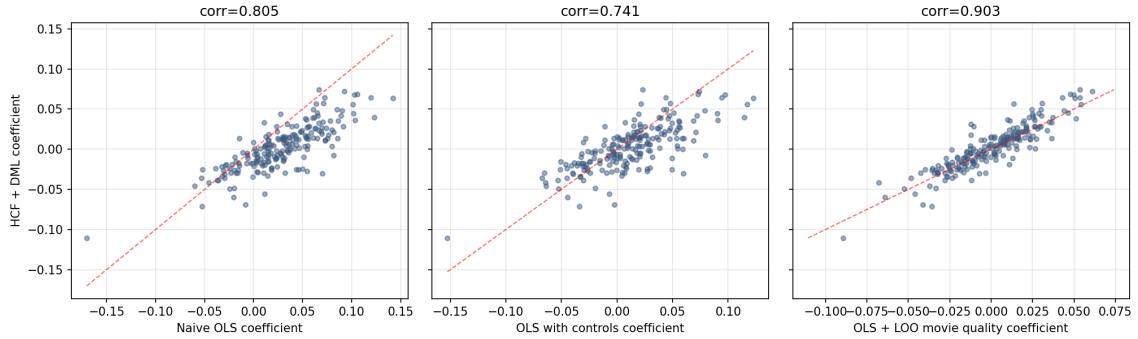
4.3.2 Statistical Significance and Multiple Testing

Given that we test 205 critics simultaneously, standard hypothesis testing at conventional significance levels would produce substantial false positives. We therefore apply the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR) at 10%.

Table 2 summarizes significance by specification. Naive OLS flags 53 critics (26%) as biased, 52 of them positive; this is the figure that without controls would describe "critic gender bias" in this dataset. Adding the full set of pre-treatment film controls shrinks this to 8 critics (4%). The LOO movie-quality OLS—the strongest crowd-based comparator we report—flags zero critics: once each film's mean rating from *other* critics is netted out, no individual critic in our sample stands out as systematically more or less generous than peers. HCF+DML, which conditions on each critic's own taste rather than on the crowd, flags 5 critics (4 positive, 1 negative). Figure 2b plots each OLS column against the HCF+DML estimates; per-critic



(a) Distribution of critic-level bias estimates under four specifications. Per-method mean is the dotted blue line; orange bars are critics significant after Benjamini–Hochberg correction at $\alpha = 0.10$. Significant-critic counts decline monotonically from 53 (naive OLS) to 8 (OLS with controls) to 0 (OLS + LOO movie quality), with HCF+DML at 5.



(b) Per-critic estimates: each OLS specification (horizontal) vs HCF+DML (vertical). Dashed red line is 45°. Correlations vs HCF+DML are 0.81 (naive), 0.74 (OLS with controls), and 0.90 (OLS + LOO movie quality). The LOO panel is the tightest cloud, indicating that LOO and HCF+DML rank critics nearly identically.

Figure 2: Distribution and ordering of critic-level bias estimates across four specifications.

Table 2: Summary of Critic-Level Bias Estimates Across Specifications

Statistic	Naive OLS	OLS with controls	OLS + LOO movie quality	HCF+DML
Number of critics	205	205	205	205
Mean estimate	0.028	0.012	0.000	0.002
Median estimate	0.029	0.013	0.002	0.002
Std. deviation	0.040	0.038	0.025	0.028
Significant at FDR = 0.10	53 (26%)	8 (4%)	0 (0%)	5 (2%)
Positive	52	8	0	4
Negative	1	0	0	1

Notes. Each column reports a per-critic regression of the 0–1 rating on the female-direction indicator, with the listed regressors stacked on top of critic fixed effects (absorbed mechanically). "OLS with controls" includes year, log(runtime), six major-genre dummies, $\log(1 + \text{budget})$, and a budget-missing indicator. "OLS + LOO movie quality" adds the leave-one-out mean Metacritic score for the film (computed over the wider single-director catalog, denominator ≥ 9). Statistical significance is Benjamini–Hochberg at FDR = 0.10. Positive estimates indicate favoritism toward female-directed films.

ordering correlations are 0.81 (naive OLS), 0.74 (OLS with controls), and 0.90 (OLS + LOO movie quality). The LOO panel is the tightest cloud: LOO and HCF+DML produce nearly the same ranking of critics, but their location parameters differ for a structural reason explained next.

Two distinct estimands sit behind these specifications. The LOO column targets *within-crowd deviation*: "given how the rest of the critic community rated this film, is critic j systematically more generous to female-directed films?" Because the LOO benchmark compares each critic to a reference point constructed from the same reviewer population, it is naturally centered on deviations from the critic community rather than on deviations from true film quality. As a result, shared population-level rating differences can be absorbed into the LOO quality proxy. HCF+DML targets *within-taste deviation*: "given critic j's own latent preference profile, learned from her male-directed-film reviews alone and the learned latent characteristics of films, is she more generous to female-directed films than that profile predicts?" Its reference point is the critic's own predicted taste, not the crowd, so it is not mechanically centered on the reviewing crowd and can, in principle, reveal both per-critic outliers and average shifts in evaluator-specific bias. For audit-style applications of fairness—"does this decision-maker treat protected and unprotected applicants differently than her own preference profile predicts?"—the within-taste estimand is the relevant one;

for peer-deviation applications, the within-crowd estimand is.

Naive comparisons suggest modest favoritism toward female-directed films, but this pattern collapses once preference- and quality-driven sorting are addressed. The mean bias estimate falls from 0.028 under naive OLS to 0.002 under HCF+DML, a 92% reduction, and the count of FDR-significant critics drops from 53 (26%) to 5 (2%), a 91% reduction in flagged decision-makers. The strongest crowd-based OLS comparator, which controls for LOO movie quality, drives both quantities to zero. This result should be interpreted as evidence that no critic systematically deviates from the reviewing crowd after conditioning on the crowd's own rating of each film, not as evidence that the crowd rating is an unbiased measure of film quality.

These results indicate that raw rating gaps and covariate-adjusted OLS substantially overstate gender-based favoritism among critics, in part because director gender is correlated with film types and with critic–film matching. The sample restriction to critics with at least 30 reviews of female-directed films may also select critics with favorable dispositions, so the near-zero HCF+DML estimates should be interpreted as, if anything, a lower bound on population-level bias.

A natural interpretation of the attenuation is industry structure and sorting. Female-directed films are not randomly allocated across projects and may be disproportionately concentrated in genres and formats that some critics tend to rate highly. Genres such as action and horror receive systematically lower scores in our data, so critics who prefer character-driven independent dramas may rate female-directed films more highly, even absent gender-based favoritism (see A.2). Our methodology addresses this problem by learning latent representations of critic preferences and film characteristics from the rating matrix. The honest design estimates preference embeddings using only male-directed films, limiting the most direct channel through which treatment information enters the preference profile, and DML then controls for the match between critic preferences and film characteristics when estimating θ_j.

A remaining concern is that embeddings could absorb part of the treatment effect if some latent characteristics are caused by, or constitutive of, director gender. However, the honest design blocks the most direct route for treatment information to enter preference embeddings, and simulations with known nonzero θ_j show that HCF+DML recovers true bias rather than mechanically shrinking estimates toward zero. More broadly, aggregate demographic gaps mix evaluator bias, differences in item characteristics, and preference-based sorting. Targeting decision-makers based on raw gaps can therefore be misleading; the same logic applies to other bilateral evaluation settings such as hiring, peer review, and performance evaluation.

5 Simulation Study

5.1 Overview

This section presents a Monte Carlo simulation study to validate the HCF+DML methodology under controlled conditions where ground truth is known. We generate synthetic data that mirrors the key features of our empirical application: bilateral interactions between units and items, confounding between treatment and latent characteristics, and a sparse observation structure. The simulation demonstrates that HCF+DML recovers individual-level bias parameters with substantially lower bias than naive OLS estimation, even under deliberate misspecification of the embedding dimension.

5.2 Data Generating Process

We simulate an employer-applicant labor market setting that parallels the movie critic application. The outcome model follows:

$$Y_{ij} = \beta \cdot \cos(C_i, P_j) + T_i \cdot \theta_j + \varepsilon_{ij} \tag{12}$$

where $Y_{ij} \in [0, 1]$ is the (normalized continuous) evaluation of applicant i by employer j, $C_i \in \mathbb{R}^d$ represents applicant characteristics, $P_j \in \mathbb{R}^d$ captures employer preferences, $T_i \in \{0, 1\}$ indicates treatment status (gender), $\theta_j \sim N(0, 1)$ is the employer-specific bias parameter of interest, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is idiosyncratic noise.

The match quality function $g(C_i, P_j) = \beta \cdot \cos(C_i, P_j)$ captures how well an applicant's characteristics align with an employer's preferences, analogous to how a film's characteristics match a critic's tastes.

Data Generating Process Summary. We generate synthetic bilateral interaction data as follows:

1. **Latent spaces:** Employer preferences $P_j \in \mathbb{R}^{10}$ are drawn uniformly from $[-1, 1]^{10}$. Applicant characteristics $C_i \in \mathbb{R}^{10}$ are drawn from a Gaussian mixture model with 100 cluster centers, each with standard deviation 0.1.

2. **Treatment and confounding:** Treatment $T_i \in \{0, 1\}$ indicates gender, with equal probability. Confounding is introduced by shifting treated (female) characteristics: $C_i|T_i = 1 \sim C_i|T_i = 0 + \delta \cdot \mathbf{e}$, where $\mathbf{e} = (1, 1, 1, 0, \ldots, 0)^\top$ shifts the first three dimensions.

3. **Outcomes:** Ratings follow $Y_{ij} = \beta \cos(C_i, P_j) + T_i \cdot \theta_j + \varepsilon_{ij}$, where $\cos(\cdot, \cdot)$ denotes cosine similarity, $\theta_j \sim N(0, 1)$ are employer-specific bias parameters, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is idiosyncratic noise.
4. **Observation pattern:** Each dyad (i, j) is observed independently with probability $\lambda = 0.2$, yielding approximately 20% matrix density.
5. **Estimation:** We deliberately misspecify the embedding dimension ($d_{\text{model}} = 8$ vs. true $d = 10$) to test robustness.

5.2.1 Confounding Structure

The identification challenge arises because treatment correlates with characteristics. We induce confounding by shifting treated applicants' characteristics:

$$C_i \mid T_i = 1 \sim C_i \mid T_i = 0 + \delta \cdot \mathbf{e} \tag{13}$$

where $\delta = 0.5$ controls confounding strength and $\mathbf{e}$ shifts the first three dimensions. This models scenarios where female-directed films systematically differ in style, genre, or production characteristics from male-directed films. Under this confounding, naive OLS conflates true bias θ_j with preference-based differences, while HCF+DML aims to separate these effects.

5.2.2 Simulation Parameters

Table 5 in Appendix A.3 provides a comprehensive list of the parameter values, including sample sizes and noise levels. Notably, we set the embedding dimension to $d_{\text{model}} = 8$ while the true latent dimension is $d = 10$; this allows us to evaluate the robustness of the HCF+DML estimator to the type of model misspecification likely encountered in real-world empirical applications.

Figure 3 visualizes the latent space structure underlying our simulation. Panel (A) plots the first two dimensions of applicant characteristics. Male applicants (blue) and female applicants (orange) are drawn from the same Gaussian mixture model, but we introduce confounding by shifting female characteristics by $\delta = 0.5$ in the first three dimensions. This shift, indicated by the arrow connecting group means, creates systematic differences between treatment groups.

Panel (B) displays the 100 simulated employers in the preference space, with color indicating their true bias parameter θ_j. By construction, bias is independent of preferences: employers with similar preferences may have very different biases, and

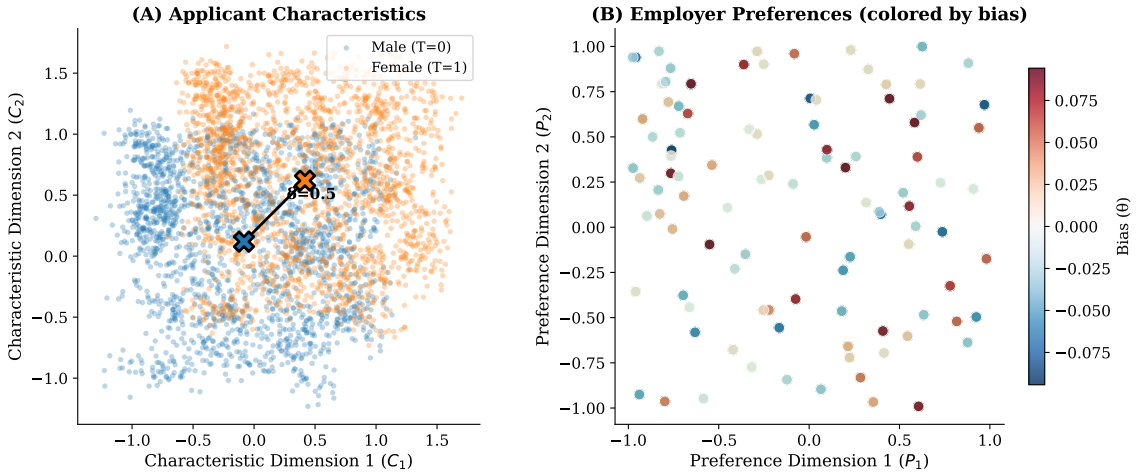


Figure 3: Simulated latent spaces illustrating the confounding structure. Panel (A) shows the first two dimensions of applicant characteristics. The arrow indicates the confounding shift $\delta = 0.5$ applied to female characteristics, creating systematic differences between treatment groups. Panel (B) displays employer preferences in the same latent space, with each point representing one employer colored by their true bias parameter θ_j. Bias parameters are drawn independently from preferences.

vice versa. However, because applicant characteristics differ systematically by gender (Panel A), employers whose preferences align with male-typical characteristics will rate male applicants higher on average, even absent any bias. Naive OLS estimation cannot distinguish this preference-based channel from true discriminatory bias, whereas HCF+DML controls for the confounding path by including learned embeddings of both preferences and characteristics.

5.3 Estimation

For each employer j, we estimate the bias parameter θ_j using two approaches.

OLS (Naive Baseline). For each employer, we regress outcomes on treatment status without controlling for preferences or characteristics:

$$Y_{ij} = \alpha_j + \beta_j T_i + e_{ij} \tag{14}$$

The OLS estimate $\hat{\beta}_j$ is biased under confounding because it captures both true bias and preference-based differences.

HCF+DML We estimate employer bias following Algorithm 1: We first train the Honest Collaborative Filtering model to extract embeddings $\hat{P}_0$ (employer preferences trained on control items only) and $\hat{C}^{(k)}$ (cross-fitted item characteristics). We then estimate employer-specific effects using the partial linear regression model with 5-fold DML cross-fitting.

5.4 Results

Figure 4 displays estimated versus true bias parameters for both methods. Under confounding, OLS estimates exhibit systematic bias: employers with certain preference profiles (those who prefer characteristics more common in the control group) appear more biased than they truly are. HCF+DML substantially reduces this bias by controlling for the confounding path through the learned embeddings.

Table 3 shows that HCF+DML more accurately recovers evaluator-level heterogeneity in discrimination than OLS. The correlation between estimated and true θ is 0.95 under HCF+DML versus 0.81 under OLS, indicating tighter alignment with the true cross-sectional pattern of bias. HCF+DML also reduces average bias by about 67%. We also report *relative bias*, defined as the mean estimation bias divided by the standard deviation of the true bias parameters, SD(θ) (in our simulations,

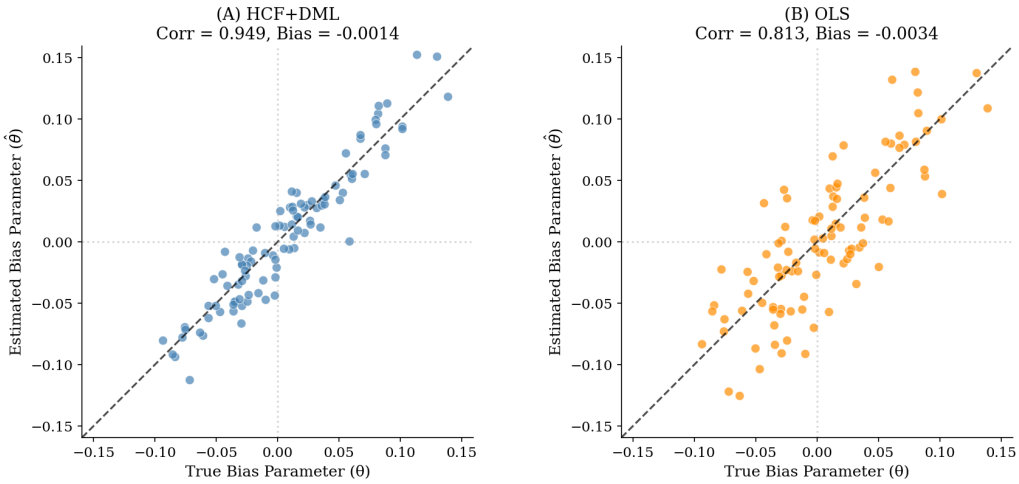


Figure 4: Scatter plots of estimated vs true theta, DML and OLS

SD(θ) $\approx$ 0.05). Under this scaling, HCF+DML exhibits a relative bias of 3% of a standard deviation, whereas naive OLS exhibits a relative bias of 7% of a standard deviation. Consistent with these gains, HCF+DML substantially lowers overall estimation error, with RMSE 0.018 compared to 0.036 for OLS, implying roughly a 50% reduction in RMSE. Together, the higher correlation, lower absolute bias and relative, and lower RMSE suggest that HCF+DML provides a more reliable measure of individual-level discrimination in this simulation setting.

Table 3: Estimation Performance

Metric	HCF+DML	OLS
Correlation with true θ	0.95	0.81
Mean Bias	-0.001	-0.003
Relative Bias	-0.03	-0.07
RMSE	0.018	0.036

Notes: Relative bias is computed as Mean bias/SD(θ_{true}). In this simulation, SD(θ_{true}) = 0.051.

The simulation deliberately misspecifies the embedding dimension (d=8 vs true d=10). Despite this misspecification, HCF+DML maintains its advantage over OLS in terms of bias, demonstrating practical robustness. This is important for applications where the true latent dimensionality is unknown.

Importantly, the simulation provides direct evidence against the concern that HCF+DML mechanically attenuates true effects. The data generating process includes nonzero

bias parameters by construction ($\theta_j \sim \mathcal{N}(0,1)$), and the method recovers these with high fidelity: the 0.95 correlation and near-zero mean bias confirm that the honest embedding design removes confounding without absorbing the causal signal. Under the no-confounding scenario (Appendix A.4), HCF+DML achieves 0.96 correlation with true parameters, further confirming that the method preserves genuine treatment effects when they exist.

Further robustness exercises are reported in the appendix. Appendix A.4 varies confounding strength from $\delta = 0$ (no confounding) to $\delta = 0.6$ (strong confounding), showing that HCF+DML maintains correlation above 0.93 and stable RMSE across all regimes while OLS performance degrades sharply. Appendix A.5 replaces the uniform shift design with a more challenging separate-clusters structure, where male and female characteristics originate from distinct Gaussian mixture components; HCF+DML continues to outperform OLS, confirming robustness to the functional form of confounding. Appendix A.6 introduces MNAR selection mechanisms of increasing severity—selection on match quality, treatment status, and outcome-relevant unobservables—and finds that HCF+DML retains its advantage in all cases, with the largest gains under the most challenging selection environment.

5.5 Discussion

The simulation study validates that HCF+DML successfully recovers individual-level bias parameters under confounding between treatment and latent characteristics. Several findings emerge.

First, HCF+DML substantially reduces estimation error compared to naive OLS when confounding is present. Under moderate confounding, HCF+DML achieves 17% lower RMSE; under strong confounding, this advantage grows to 52%. The method maintains stable performance across confounding levels while OLS deteriorates markedly, confirming that the embedding-based controls effectively address omitted variable bias.

Second, HCF+DML preserves ranking accuracy. Correlation between estimated and true bias parameters remains consistently high across all scenarios, indicating that the method reliably identifies which evaluators exhibit the strongest positive or negative bias. This is particularly relevant for applications where the goal is to flag outliers or compare relative bias across units rather than estimate exact magnitudes.

Third, the method is robust to realistic complications. Performance remains strong under deliberate misspecification of the embedding dimension, alternative confound-

ing structures (Appendix A.5), and MNAR selection mechanisms that depend on match quality, treatment status, or unobservables. This robustness is important for empirical applications where the true data-generating process is unknown.

Fourth, when confounding is absent, OLS outperforms HCF+DML in terms of efficiency. This reflects the cost of the honest estimation design, which restricts preference embeddings to control observations only. Since the presence of confounding is typically unknown in practice, HCF+DML provides insurance against potentially severe bias at the cost of modest efficiency loss when confounding happens to be negligible.

Two methodological considerations merit attention. For inference, we apply the Benjamini-Hochberg procedure to control the false discovery rate when testing multiple evaluators simultaneously. This is essential given the large number of hypothesis tests (one per evaluator) and the exploratory nature of identifying biased individuals. Additionally, standard errors from the DML procedure treat the learned embeddings as fixed, ignoring estimation uncertainty in the first stage. This "generated regressor" problem can lead to understated standard errors and inflated rejection rates. In our empirical application, we address this through bootstrap procedures that re-estimate embeddings in each replication, providing valid inference that accounts for the full estimation pipeline.

Our model estimates an evaluator-specific shift θ_j that is constant across items. If discriminatory effects vary with film characteristics, for example if a critic penalizes female-directed action films but not female-directed dramas, then the estimated θ_j should be interpreted as a weighted average of heterogeneous effects over the distribution of evaluated items. Extending the framework to allow item-dependent bias, e.g. $\theta_j(C_i)$, is a natural direction for future work.

6 Conclusion

This paper develops a methodology for estimating individual-level bias in bilateral evaluation settings where treatment status correlates with unobserved characteristics. We combine Honest Collaborative Filtering, which extracts latent representations of evaluator preferences and item characteristics from observed ratings, with Double Machine Learning to estimate evaluator-specific bias parameters while controlling for these learned embeddings. The "honest" design ensures that preference embeddings are estimated using only control-group items, preventing treatment effects from contaminating the estimates.

Simulations demonstrate that HCF+DML substantially outperforms naive OLS under confounding, reducing RMSE by up to 50% while maintaining high correlation with true parameters. The method proves robust to embedding misspecification, alternative confounding structures, and non-random selection.

Applied to 136,847 critic-film ratings across 205 critics, the methodology overturns naive conclusions about critic bias. Raw comparisons suggest 26% of critics exhibit significant favoritism toward female-directed films. Adding the full set of pre-treatment film controls reduces this to 4%; an even more demanding crowd-deviation benchmark that controls for the film's average rating from other critics flags zero individual critics. Under HCF+DML, which conditions on each critic's own latent taste rather than on peer consensus, 2% of critics remain individually flagged. The apparent pro-female pattern in raw ratings largely reflects critics' preferences for film characteristics, including genre, rather than director-gender favoritism.

The framework applies broadly to hiring, peer review, and other evaluation settings where treatment groups differ in characteristics that evaluators legitimately value. By extracting information from bilateral rating patterns, the approach separates genuine bias from preference-based differences, a distinction that naive methods cannot make.

A Appendix

A.1 Systematic Differences Across Critics

Figure 5 displays critic-level estimates sorted by the HCF+DML point estimate, with 95% confidence intervals shown for the HCF+DML estimates. OLS estimates (red crosses) are systematically higher than HCF+DML estimates (blue points) across nearly the entire distribution. This pattern indicates that the attenuation from OLS to HCF+DML is not driven by a few outliers but reflects a pervasive shift: controlling for film characteristics reduces estimated favoritism for the vast majority of critics.

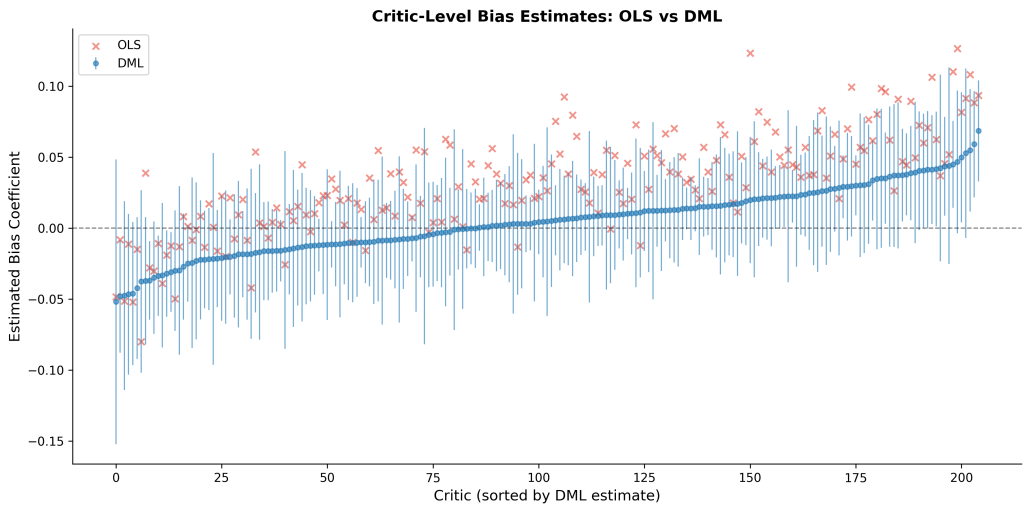


Figure 5: Critic-Level Bias Estimates: OLS vs HCF+DML

Critics are sorted by HCF+DML estimate (blue points) from most negative to most positive. Vertical bars show 95% confidence intervals for HCF+DML estimates. Red crosses show corresponding OLS estimates for the same critics. OLS estimates are consistently higher than HCF+DML estimates, indicating systematic overestimation of favoritism when film characteristics are not controlled.

A.2 Genre Concentration and Rating Differences by Director Gender

Our discussion in Section 4.3 interprets the attenuation of estimated bias under HCF+DML by appealing to two stylized facts: (i) female directors are concentrated in genres that critics tend to rate highly, and (ii) genres such as action and horror receive systematically lower scores than other films. Both claims are observable in our analysis sample. Table 4 reports the relevant statistics.

Table 4: Genre Concentration and Mean Scores by Director Gender

	Share of films in genre			Mean rating (0–1 scale)		
Genre	Male-dir.	Female-dir.	Diff (pp)	In genre	Other films	Gap
Drama	65.8%	78.3%	+12.4	0.655	0.549	+0.106
Comedy	35.2%	34.1%	−1.1	0.593	0.636	−0.043
Thriller	30.7%	16.4%	−14.2	0.596	0.632	−0.036
Romance	20.1%	26.0%	+5.9	0.611	0.624	−0.013
Action	17.9%	6.8%	−11.1	0.570	0.633	−0.064
Horror	10.2%	7.0%	−3.1	0.559	0.627	−0.068

Notes. Sample: 6,788 single-director films (5,876 male-directed, 912 female-directed) reviewed by the 205 analysis-sample critics. Genre dummies are not mutually exclusive: the average film carries 1.78 genre tags. "Share" columns are the fraction of films of each gender that carry the genre tag (e.g., 78.3% of female-directed films are tagged Drama). "Diff" is female share minus male share, in percentage points. "Mean rating" columns are the per-review average score on a 0–1 scale, computed at the review level over films in (or not in) the genre, pooling both genders. "Gap" is the difference. Aggregated, films tagged Action or Horror have a mean rating of 0.569 across 32,045 reviews, versus 0.639 across 94,743 reviews for films in neither category—a gap of −0.070, or 7 percentage points.

Two patterns stand out. First, female directors in our sample are disproportionately concentrated in Drama (+12.4 percentage points), the highest-rated genre on average (+10.6 points relative to non-Drama), and in Romance (+5.9 pp). They are markedly under-represented in Action (−11.1 pp) and Thriller (−14.2 pp), and modestly under-represented in Horror (−3.1 pp). Second, the genres in which female directors are over-represented score higher and those in which they are under-represented score lower, with Action and Horror jointly receiving ratings about 7 percentage points below other films.

These two patterns together rationalize why naive critic-level OLS estimates favor

female-directed films even in the absence of evaluator bias: critics' generic preference for Drama mechanically pushes up the mean rating of female-directed films because more of them happen to be Dramas. The HCF+DML estimator addresses this by conditioning on latent characteristics that capture genre composition and other film-level taste-relevant attributes, so that $\hat{\theta}_j$ reflects only the residual gender-direction signal after this composition is partialled out.

A.3 Simulation Parameters

Table 5 details the parameterization of the Monte Carlo simulation used to validate the HCF+DML framework.

Table 5: Simulation Parameters

Parameter	Value	Description
J	100	Number of units (employers) to estimate
N_{male}	5,000	Control group size
N_{female}	5,000	Treatment group size
d	10	True latent space dimensionality
d_{model}	8	Embedding dimension (deliberately misspecified)
n_{clusters}	100	Number of characteristic cluster centers
λ	0.2	Application probability (sparsity)
β	5.0	Weight of match quality in outcome
σ	1.0	Standard deviation of outcome noise
δ	0.5	Magnitude of treatment-characteristic shift

We simulate a labor market environment with $J = 100$ employers and 10,000 total applications, split evenly between groups. To test the robustness of the estimator against imperfect representation learning, we deliberately set the model's embedding dimension ($d_{\text{model}} = 8$) lower than the true latent dimensionality ($d = 10$). The parameter $\delta = 0.5$ introduces a treatment-characteristic shift, creating the confounding bias that the DML procedure is designed to purge.

A.4 Sensitivity to Confounding Strength

We examine how estimation performance varies with confounding strength by simulating three scenarios with different degrees of correlation between treatment and item characteristics. Recall that confounding arises when the treatment variable (e.g., director gender) is correlated with characteristics that also affect outcomes through evaluator preferences. This correlation generates omitted variable bias in naive OLS estimation, which does not control for characteristics.

We implement confounding by shifting female item characteristics relative to male characteristics by δ along the first three latent dimensions:

1. **No Confounding** ($\delta = 0$): Item characteristics are independent of treatment. Male and female characteristic distributions fully overlap, and OLS faces no omitted variable bias.
2. **Moderate Confounding** ($\delta = 0.3$): Female characteristics are shifted moderately, creating partial separation between treatment groups in the latent space.
3. **Strong Confounding** ($\delta = 0.6$): Female characteristics are shifted substantially, generating pronounced differences between treatment groups.

For the no-confounding baseline, we use the correct embedding dimension ($d = 10$) to establish best-case performance. For the confounded scenarios, we deliberately misspecify the embedding dimension ($d = 8$) to test robustness under realistic conditions where the true latent dimensionality is unknown.

Table 6: Estimation Performance Across Confounding Levels

Scenario	Method	Correlation	Bias	Rel. Bias	RMSE
No Confounding ($\delta = 0$)	HCF+DML	0.963	−0.0025	-0.05	0.017
	OLS	0.997	−0.0006	-0.012	0.004
Moderate ($\delta = 0.3$)	HCF+DML	0.950	−0.0016	-0.032	0.017
	OLS	0.916	−0.0022	-0.044	0.022
Strong ($\delta = 0.6$)	HCF+DML	0.939	−0.0013	-0.026	0.020
	OLS	0.764	−0.0039	-0.078	0.042

Notes: Confounding is introduced by shifting female item characteristics by δ along three latent dimensions. Embedding dimension is $d = 10$ for no confounding and $d = 8$ (misspecified) for confounded scenarios.

Table 6 reports estimation performance across scenarios. Several patterns emerge.

First, when confounding is absent, OLS performs exceptionally well, achieving near-perfect correlation (0.997) with minimal RMSE. HCF+DML performs slightly worse in this setting due to the additional variance introduced by the honest estimation procedure, which splits the sample for embedding estimation. This confirms that HCF+DML incurs a modest efficiency cost when its key identifying assumption of treatment-characteristic correlation is not satisfied.

Second, as confounding increases, OLS performance deteriorates substantially while HCF + DML remains stable. Under strong confounding, OLS correlation with true bias parameters drops to 0.764 and RMSE more than doubles relative to the no-confounding baseline. In contrast, HCF+DML maintains correlation above 0.93 and stable RMSE across all scenarios. The RMSE ratio between methods grows from 0.24 (no confounding) to 0.48 (strong confounding), indicating that the relative advantage of HCF+DML increases with confounding severity.

Figure 6 visualizes these patterns. Under no confounding, both methods produce estimates tightly clustered around the 45-degree line. As confounding increases, OLS estimates become increasingly dispersed, particularly in the strong confounding case where substantial deviations from true values are evident. HCF+DML estimates remain well-calibrated across all scenarios.

Figure 7 summarizes these trends. The correlation between estimated and true bias parameters remains above 0.93 for HCF+DML regardless of confounding strength, while OLS correlation declines from 0.997 to 0.764. Similarly, HCF+DML RMSE remains stable around 0.02, while OLS RMSE increases from 0.004 to 0.042.

These results demonstrate that HCF+DML is robust to confounding and maintains reliable performance even under embedding dimension misspecification. When confounding is absent, practitioners may prefer OLS for its simplicity and efficiency. However, since the presence and magnitude of confounding is typically unknown in empirical applications, HCF+DML provides a robust alternative that guards against potentially severe bias.

Appendix A.5 extends this analysis to alternative confounding structures, confirming that these findings generalize beyond the shift design studied here.

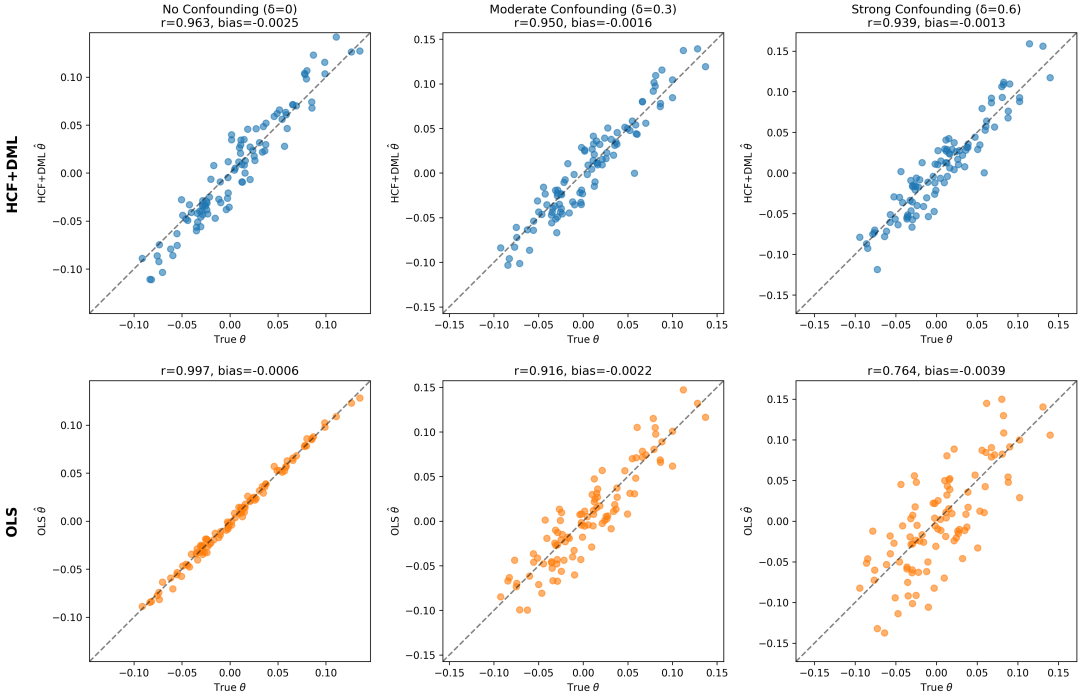


Figure 6: Estimated versus true bias parameters across confounding levels. Each column represents a confounding scenario (no, moderate, strong). Top row: HCF+DML estimates. Bottom row: OLS estimates. Under strong confounding, OLS estimates exhibit substantial dispersion around the 45-degree line, while HCF+DML estimates remain tightly clustered.

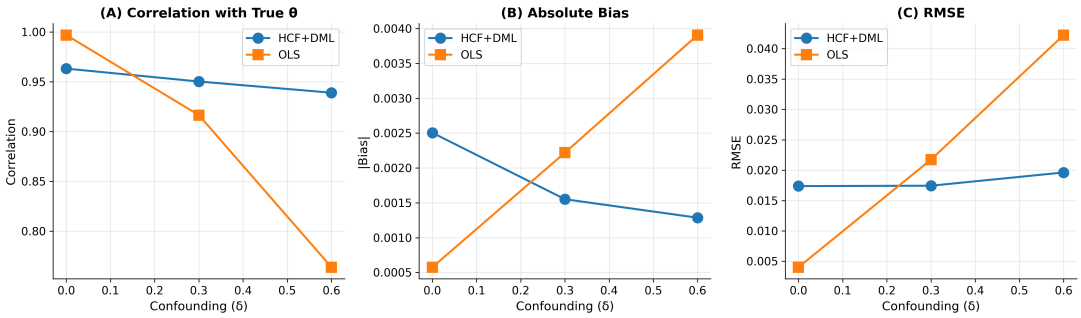


Figure 7: Performance metrics as a function of confounding strength. Left panel: correlation with true parameters. Right panel: RMSE. HCF+DML (blue) maintains stable performance across confounding levels, while OLS (orange) degrades substantially as confounding increases.

A.5 Sensitivity to Confounding Structure

In Section 5, we study a stylized shift design where female characteristics are systematically shifted in a subset of dimensions. This design captures realistic scenarios where female-directed films may systematically differ from male-directed films along observable dimensions (e.g., lower budgets, different genre distributions, varying production scales). However, confounding in practice may arise through more complex mechanisms.

In this appendix, we examine HCF+DML performance under an alternative confounding structure where male and female characteristics originate from *different cluster centers* in the latent space. This design reflects the possibility that male and female directors work in distinct "niches" or stylistic traditions, creating heterogeneous, non-directional confounding patterns rather than a uniform shift.

Data Generating Process

We modify the main simulation DGP as follows. Rather than drawing male and female characteristics from the same Gaussian mixture with a location shift, we generate separate cluster centers for each group:

$$\mu_k^{(M)} \sim \text{Uniform}(-1,1)^d, \quad k = 1, \ldots, N_C^{(M)} \tag{15}$$

$$\mu_k^{(F)} \sim \text{Uniform}(-1,1)^d, \quad k = 1, \ldots, N_C^{(F)} \tag{16}$$

$$C_i | T_i = 0 \sim \sum_{k=1}^{N_C^{(M)}} \frac{1}{N_C^{(M)}} \mathcal{N}(\mu_k^{(M)}, \sigma^2 I_d) \tag{17}$$

$$C_i | T_i = 1 \sim \sum_{k=1}^{N_C^{(F)}} \frac{1}{N_C^{(F)}} \mathcal{N}(\mu_k^{(F)}, \sigma^2 I_d) \tag{18}$$

where $N_C^{(M)} = N_C^{(F)} = 50$ cluster centers are drawn independently for male and female characteristics. The random positioning of cluster centers creates regions where male and female characteristics overlap substantially, alongside regions dominated by one group. This heterogeneity makes identification more challenging than the uniform shift design.

Table 7 summarizes the simulation parameters.

Table 7: Simulation Parameters: Complex Confounding Structure

Parameter	Value
Employers (J)	100
Male items	5,000
Female items	5,000
Latent dimension (d)	10
Male characteristic clusters	50
Female characteristic clusters	50
Cluster std (σ)	0.15
Observation rate (λ)	0.2
Match coefficient (β)	5.0
Bias std (σ_θ)	1.0
Model embedding dimension	8

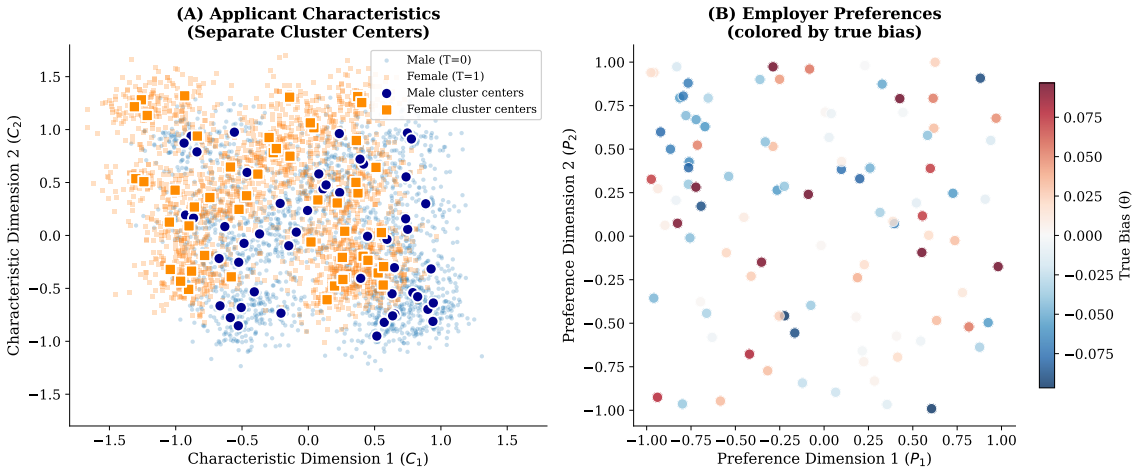


Figure 8: Latent spaces under complex confounding (separate cluster centers). Panel (A) shows applicant characteristics with male (blue) and female (orange) drawn from different Gaussian mixture components. Large markers indicate cluster centers. Panel (B) displays employer preferences colored by true bias θ_j.

Confounding Structure Visualization

Figure 8 visualizes the latent space structure under this alternative DGP. Panel (A) displays the first two dimensions of applicant characteristics, with male and female observations shown alongside their respective cluster centers. Unlike the shift design, confounding here arises from the spatial separation of cluster centers rather than a systematic directional shift. Some regions contain primarily male characteristics, others primarily female, with substantial overlap where cluster centers happen to be proximate. Panel (B) shows employer preferences colored by their true bias parameter θ_j. As in the main simulation, bias is independent of preferences by construction.

Table 8 reports estimation performance under the complex confounding structure. HCF+DML maintains its advantage over naive OLS estimation, achieving higher correlation with true bias parameters and substantially lower systematic bias. The method successfully recovers individual-level bias even when confounding patterns are heterogeneous across the characteristic space.

Table 8: Estimation Performance: Complex Confounding Structure

Method	Correlation	Mean Bias	Rel. Bias	RMSE	MAE
HCF+DML	0.9272	0.0004	0.007	0.0217	0.0177
OLS	0.8310	0.0055	0.102	0.0330	0.0264

Notes: Relative bias is computed as Mean bias/SD(θ_{true}). In this simulation, SD(θ_{true}) = 0.054.

Figure 9 compares estimated versus true bias parameters for both methods. The DML estimates cluster more tightly around the 45-degree line, indicating better recovery of individual-level heterogeneity. OLS estimates exhibit greater dispersion and systematic deviation from ground truth due to uncontrolled confounding.

These results demonstrate that HCF+DML is robust to the functional form of confounding. Under the more challenging separate-clusters design, where confounding is heterogeneous and non-directional, the method continues to outperform naive OLS estimation in recovering true individual-level bias parameters. This robustness is important for empirical applications where the precise nature of confounding between treatment and latent characteristics is unknown.

The key insight is that collaborative filtering embeddings capture the relevant structure of the characteristic space regardless of whether confounding manifests as a uniform shift or through more complex distributional differences. By learning embed-

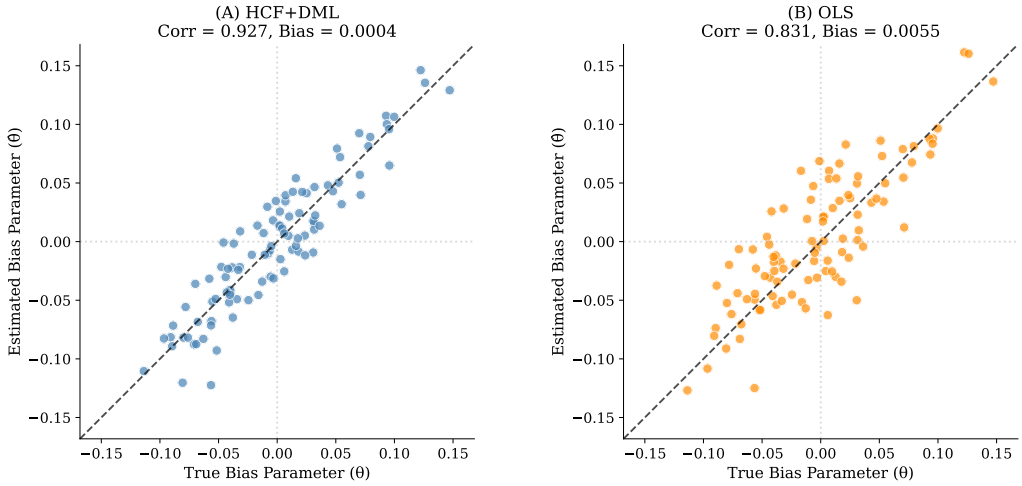


Figure 9: Estimated versus true bias parameters under complex confounding. Panel (A): HCF+DML estimates. Panel (B): OLS estimates. Dashed line indicates perfect recovery.

dings that predict outcomes, the method implicitly controls for whatever characteristic patterns drive rating variation, enabling valid causal inference about individual-level bias.

A.6 Sensitivity to Selection Mechanism

Our baseline simulation assumes random observation patterns. In practice, which evaluator-item pairs we observe may depend on characteristics that also affect outcomes, creating missing-not-at-random (MNAR) selection. We examine three increasingly complex selection mechanisms to assess robustness.

This appendix is designed to isolate the role of selection. We therefore hold the underlying data-generating process fixed across cases, including the latent preference and characteristic embeddings (P, C), the bias parameters θ, and the treatment assignment rule. In particular, the level of confounding induced by the distribution of C across treatment groups is identical across the three selection cases and we only vary observation mechanism.

Let $S_{ij} \in \{0, 1\}$ indicate whether evaluator j's rating of item i is observed. We model selection as:

$$\Pr(S_{ij} = 1) = \sigma(\alpha_0 + \alpha_1 \cdot \text{match}_{ij} + \alpha_2 \cdot T_i + \alpha_3 \cdot U_{ij}) \tag{19}$$

where $\sigma(\cdot)$ is the logistic function, $\text{match}_{ij} = \beta \cos(C_i, P_j)$ is the match quality, T_i is the treatment indicator, and U_{ij} is an unobservable that may also affect outcomes. We calibrate α_0 to achieve approximately 10% observed entries. Table 9 describes the three cases.

Table 9: MNAR Selection Cases

Case	Description	α_1	α_2	α_3
1	Match-only	0.5	0	0
2	Match + Treatment	0.5	0.3	0
3	Match + Treatment + Unobservables	0.5	0.3	0.4

Case 1 represents selection based purely on match quality: evaluator-item pairs with higher latent compatibility are more likely to be observed. This captures settings where interaction is driven by fit or interest, such as firms interviewing applicants who appear to match a job's requirements, or critics reviewing films that align with their established tastes (e.g., genre or style).

Case 2 adds differential selection by treatment status, allowing treated items to be observed at different rates even after conditioning on match quality. This corresponds to environments where protected attributes influence exposure or assignment, such

Table 10: Estimation Performance Under MNAR Selection

Case	Method	Correlation	Bias	Rel. Bias	RMSE
1	HCF+DML	0.817	−0.0047	-0.09	0.0236
	OLS	0.773	−0.0057	-0.11	0.0262
2	HCF+DML	0.801	−0.0064	-0.12	0.0247
	OLS	0.764	−0.0087	-0.17	0.0270
3	HCF+DML	0.791	−0.0039	-0.08	0.0250
	OLS	0.703	−0.0076	-0.15	0.0296

as gendered sorting into job ladders, editorial decisions that differentially allocate reviews, or platforms that route certain items to evaluators at systematically different rates.

Case 3 introduces an unobservable U_{ij} that affects both selection and outcomes (Y_{ij} includes γU_{ij} with $\gamma = 0.5$), creating the most challenging identification environment. This reflects cases where observation depends on factors only partially captured by match quality or observed characteristics, such as time-varying shocks to demand, publicity, or access that simultaneously increase the likelihood of being reviewed and shift ratings.

Table 10 reports estimation performance across selection cases. HCF+DML consistently outperforms OLS, with the advantage most pronounced under Case 3 where selection depends on outcome-relevant unobservables. Because the underlying (P, C) structure is held fixed, these differences are attributable to the selection mechanism rather than changes in confounding strength. The bias reduction ranges from 18% (Case 1) to 49% (Case 3), and HCF+DML achieves higher correlation with true bias parameters in all cases.

These results suggest that HCF+DML is robust to realistic forms of MNAR selection. By controlling for match quality through learned embeddings, the method partially addresses selection that operates through the same channel. The improvement under Case 3 indicates that even when unobservables drive both selection and outcomes, the embedding-based controls provide meaningful bias reduction relative to naive estimation.

References

Arnold, D., Dobbie, W. and Yang, C. S. (2018), 'Racial bias in bail decisions', *Quarterly Journal of Economics* **133**(4), 1885–1932.

Arrow, K. J. (1973), The theory of discrimination, *in* O. Ashenfelter and A. Rees, eds, 'Discrimination in Labor Markets', Princeton University Press, Princeton, NJ, pp. 3–33.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. (2021), 'Matrix completion methods for causal panel data models', *Journal of the American Statistical Association* **116**(536), 1716–1730.

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

Bach, P., Chernozhukov, V., Klaassen, S., Kurz, M. S. and Spindler, M. (2024), 'DoubleML – double machine learning in python'. BSD-3-Clause License. Documentation: https://docs.doubleml.org/stable/index.html.
URL: *https://github.com/DoubleML/doubleml-for-py*

Bach, P., Chernozhukov, V., Kurz, M. S. and Spindler, M. (2022), 'DoubleML – An object-oriented implementation of double machine learning in Python', *Journal of Machine Learning Research* **23**(53), 1–6.
URL: *http://jmlr.org/papers/v23/21-0862.html*

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279.

Becker, G. S. (1957), *The Economics of Discrimination*, University of Chicago Press, Chicago.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)* **57**(1), 289–300.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1995.tb02031.x*

Bertrand, M. and Duflo, E. (2017), 'Field experiments on discrimination', *Handbook of economic field experiments* **1**, 309–393.

Bertrand, M. and Mullainathan, S. (2004), 'Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination', *American economic review* **94**(4), 991–1013.

Blank, R. M. (1991), 'The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review', *The American Economic Review* pp. 1041–1067.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**(1), C1–C68.
URL: *https://academic.oup.com/ectj/article/21/1/C1/5056401*

Chouldechova, A. (2017), 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big Data* **5**(2), 153–163.

Griffith, A. and Peng, S. (2023), Identification of network structure in the presence of latent, unobserved factors: A new result using Turán's theorem. Revise and Resubmit, Quantitative Economics.

Hernán, M. A. and Robins, J. M. (2020), *Causal Inference: What If*, Chapman & Hall/CRC. Available at: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017), Inherent trade-offs in the fair determination of risk scores, *in* '8th Innovations in Theoretical Computer Science Conference (ITCS 2017)', Vol. 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 43:1–43:23.

Kline, P., Rose, E. K. and Walters, C. R. (2022), 'Systemic discrimination among large us employers', *The Quarterly Journal of Economics* **137**(4), 1963–2036.

Kline, P., Rose, E. K. and Walters, C. R. (2023), 'A discrimination report card', *arXiv preprint arXiv:2306.13005* .

Kline, P. and Walters, C. (2021), 'Reasonable doubt: Experimental detection of job-level employment discrimination', *Econometrica* **89**(2), 765–792.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011), 'Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion'.

Koren, Y., Bell, R. and Volinsky, C. (2009), 'Matrix factorization techniques for recommender systems', *Computer* **42**(8), 30–37.

Mazumder, R., Hastie, T. and Tibshirani, R. (2010), 'Spectral regularization algorithms for learning large incomplete matrices', *Journal of Machine Learning Research* **11**, 2287–2322.

Moon, H. R. and Weidner, M. (2015), 'Linear regression for panel with unknown number of factors as interactive fixed effects', *Econometrica* **83**(4), 1543–1579.

Pesaran, M. H. (2006), 'Estimation and inference in large heterogeneous panels with a multifactor error structure', *Econometrica* **74**(4), 967–1012.

Phelps, E. S. (1972), 'The statistical theory of racism and sexism', *American Economic Review* **62**(4), 659–661.

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