

Does Algorithmic Uncertainty Sway Human Experts? Evidence from a Field Experiment in Selective College Admissions

HANSOL LEE, Stanford University, USA

AJ ALVERO, Cornell University, USA

RENÉ F. KIZILCEC, Cornell University, USA

THORSTEN JOACHIMS, Cornell University, USA

Algorithmic predictions are inherently uncertain: even models with similar aggregate accuracy can produce different predictions for the same individual, raising concerns that high-stakes decisions may become sensitive to arbitrary modeling choices. In this paper, we define *algorithmic sensitivity* as the extent to which arbitrary modeling choices propagate into human decisions: how much a decision outcome shifts when a more favorable versus less favorable algorithmic prediction is presented to the decision-maker for the same individual. We estimate this in a randomized field experiment ($n = 19,545$) embedded in a selective U.S. college admissions cycle, in which admissions officers reviewed each application alongside an algorithmic score while we randomly varied whether the score came from one of two similarly accurate prediction models. Although the two models performed similarly in aggregate, they frequently assigned different scores to the same applicant, creating exogenous variation in the score shown. Surprisingly, we find little evidence of algorithmic sensitivity: presenting a more favorable score does not meaningfully increase an applicant's probability of admission on average, even when the models disagree substantially. These findings suggest that, in this expert, high-stakes setting, human decision-making is largely invariant to arbitrary variation in algorithmic predictions, underscoring the role of professional discretion and institutional context in mediating the downstream effects of algorithmic uncertainty.

CCS Concepts: • **Social and professional topics** → **Computing / technology policy**; • **Computing methodologies** → *Artificial intelligence*; • **Human-centered computing** → *Empirical studies in HCI*.

Additional Key Words and Phrases: algorithmic sensitivity, predictive multiplicity, human-algorithm interaction, algorithmic uncertainty, randomized field experiment, college admissions

ACM Reference Format:

Hansol Lee, AJ Alvero, René F. Kizilcec, and Thorsten Joachims. 2026. Does Algorithmic Uncertainty Sway Human Experts? Evidence from a Field Experiment in Selective College Admissions. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*, June 25–28, 2026, Montreal, QC, Canada. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3805689.3812380>

1 Introduction

Predictive models are increasingly used to inform high-stakes social decisions, from criminal justice and lending to medical diagnosis, hiring, and college admissions [28, 44]. The case for using them rests on the complementary strengths of machines and humans. Models can synthesize structured information at scale and with a level of consistency that human decision-makers, working under time pressure and subject to noisy judgment, often

Authors' Contact Information: Hansol Lee, Stanford University, Stanford, CA, USA, hansol@stanford.edu; AJ Alvero, Cornell University, Ithaca, NY, USA, ajalvero@cornell.edu; René F. Kizilcec, Cornell University, Ithaca, NY, USA, kizilcec@cornell.edu; Thorsten Joachims, Cornell University, Ithaca, NY, USA, tj@cs.cornell.edu.



This work is licensed under a Creative Commons Attribution 4.0 International License.

FAccT '26, Montreal, QC, Canada

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2596-8/2026/06

<https://doi.org/10.1145/3805689.3812380>

cannot achieve [27, 33]. Human decision-makers, in turn, bring contextual knowledge, professional judgment, and the capacity to integrate information that is not encoded in training data. The promise of human-algorithm collaboration is that the combination can perform better than either alone [2, 45].

At the same time, the algorithmic prediction assigned to an individual is not a fixed property of that individual. Similarly accurate models can produce different predictions for the same individual [5, 32], predictions can shift as models are retrained on new data [17, 37], and choices made in feature engineering, calibration, and model selection can each move a prediction up or down [12]. Encoded biases in training data introduce further variation [3]. The algorithmic prediction a human decision-maker sees is therefore shaped in part by aspects of the modeling process that often have little to do with the underlying individual. This raises a concern that has motivated a substantial literature on algorithmic fairness and accountability: whether decisions made downstream of such predictions end up depending on arbitrary modeling choices rather than substantively meaningful differences between individuals [11, 46].

In response to these concerns, a dominant decision-making pattern has emerged in which the algorithm is positioned not as a decision-maker but as a decision aid, with trained human experts retaining final authority over consequential decisions. College admissions is one such setting. Virginia Tech, for instance, recently announced a new undergraduate admissions process that pairs human essay readers with an AI-supported scoring system developed by university researchers [35], with final decisions made exclusively by trained admissions professionals. Such arrangements are widespread: industry surveys suggest roughly half of higher-education admissions offices were already using AI tools by 2023 [25]. The justifying logic is often that the human-in-the-loop will serve as a corrective force, absorbing the imperfections of the algorithm rather than passing them through to outcomes.

Whether this safeguard actually works depends on a specific empirical question. When a human decision-maker sees a more favorable rather than a less favorable score for the same individual, is their decision swayed by which particular score they happen to see? If decisions track the score, then arbitrary aspects of the modeling process propagate into outcomes even when humans nominally retain authority. If decisions are largely invariant to the particular score shown, then expert judgment is performing much of the corrective role that human-in-the-loop arrangements presuppose. We call this property **algorithmic sensitivity**: the extent to which arbitrary modeling choices propagate into consequential human decisions, irrespective of whether the underlying decisions are correct.

Existing evidence on this question is mixed and largely comes from laboratory experiments or observational field settings, where decision-makers and algorithmic input are not randomly paired. The most direct test would be a field experiment in which decision-makers are systematically shown a more favorable score for some individuals and a less favorable score for others, but such a design would be ethically impermissible: it would knowingly disadvantage some individuals by presenting decision-makers with information that is, by construction, less favorable than what those individuals would otherwise have received. Our approach engages this challenge by exploiting *predictive multiplicity*, the phenomenon that similarly accurate models can produce different predictions for the same individual [5, 32]. When two such models disagree, the score a decision-maker happens to see is more or less favorable purely as a function of which model was deployed. Randomizing which of two such models' scores is presented therefore induces exogenous variation in the favorability of the displayed score without assigning any individual a score from a model known to be inferior. Predictive multiplicity is one of several sources of model-contingent variation in algorithmic scores, and findings about sensitivity to this form of variation have implications for the broader concern.

We apply this design in a randomized field experiment embedded in a selective U.S. college admissions process ($n = 19,545$ applications) during the 2022 admissions cycle, in close collaboration with the university's admissions leadership [29, 30]. We developed two predictive models using the same features and modeling pipeline but trained on slightly different historical data, reflecting realistic variation in modeling choices. The two models achieve similar aggregate performance but assign different score deciles to over 70% of applicants. For each

application, we randomized which model’s score was presented, allowing us to identify the causal effect of presenting a more favorable versus less favorable algorithmic prediction for the same applicant.

Despite the substantial individual-level disagreement, we find little evidence of algorithmic sensitivity. Presenting a more favorable score does not meaningfully increase an applicant’s probability of admission on average, even when the models disagree substantially. Across specifications, admission outcomes are largely invariant to which of the two disagreeing algorithmic scores is shown. These findings suggest that, in this expert, high-stakes decision-making setting, model-contingent variation in algorithmic scores does not translate directly into variation in admission outcomes. Rather than responding to which particular algorithmic score they observe, admissions officers appear to integrate algorithmic scores as one input among many within a holistic evaluation process.

This paper makes three contributions. Conceptually, we propose *algorithmic sensitivity* as a framework for studying whether arbitrary variation in algorithmic input propagates into consequential human decisions, applicable to domains where no canonical ground truth exists against which to evaluate decisions. Methodologically, we show that predictive multiplicity can serve as a source of exogenous variation for identifying algorithmic sensitivity in real institutional settings, opening up causal analysis where direct manipulation of algorithmic input would be ethically impermissible. Empirically, we apply this design in a randomized field experiment in selective college admissions and find that reviewers’ decisions are largely invariant to model-contingent variation in algorithmic scores, suggesting that the downstream consequences of algorithmic uncertainty depend on the sociotechnical context in which the algorithm is embedded, not just on properties of the model.

2 Related Work

Our study sits at the intersection of three concerns that have been pursued largely in parallel: a literature on how human decision-makers respond to algorithmic input, a literature on predictive uncertainty and the arbitrariness of algorithmic outputs, and a smaller body of recent work that studies algorithm-assisted decision-making in real institutional contexts. We outline what each contributes and identify the gaps that motivate our framework and design.

2.1 Human Reliance on Algorithmic Advice

Early experimental work on human responses to algorithmic advice documented two opposing tendencies: *algorithm aversion*, in which decision-makers discount algorithmic forecasts after observing them err, even when the algorithm continues to outperform the available human alternative [14], and *algorithm appreciation*, in which lay decision-makers weight algorithmic advice more heavily than equivalent human advice [31]. A related literature on *automation bias* examines settings in which decision-makers over-defer to algorithmic outputs, failing to detect errors even when they have access to information that would allow them to do so [20, 36]. Even nominally human-in-the-loop designs can degrade decision quality when human oversight is mechanical rather than discretionary: Sele and Chugunova [42] show that human monitors are less likely to intervene on the least accurate recommendations, raising concerns about whether human oversight achieves the corrective function that often justifies its inclusion [21].

When the decision-maker is an expert working in a high-stakes domain, the question takes on a sharper form: does professional judgment work as the intended corrective to algorithmic limitations? Recent work in real institutional settings suggests that it can. Cheng et al. [10] analyzed call-screen workers at the Allegheny County Department of Human Services using an algorithmic family screening tool, and found that workers exercising holistic judgment substantially reduced racial disparities relative to the algorithm alone, narrowing the gap in screen-in rates between Black and white children from 20% to 9%. De-Arteaga et al. [13] document a related pattern in child welfare hotline screening: workers were less likely to adhere to the algorithm’s recommendation

when the displayed score was an incorrect estimate of risk, even when overriding required supervisory approval. Schechtman et al. [39] study an algorithm-assisted advising program at Georgia State University with a randomized controlled trial design, and estimate that roughly two-thirds of advisor interventions were plausibly targeted using non-algorithmic context, with advising style itself shaping student outcomes.

But expert discretion does not always serve a corrective role. Albright [1] examines a Kentucky bail reform that set a recommended default based on algorithmic risk scores, and finds that the recommended default was disproportionately overridden in favor of harsher bond conditions for Black defendants relative to comparable white defendants. The result was that judicial discretion amplified, rather than attenuated, racial disparities in pretrial detention. Green and Chen [22] provide complementary lab-based evidence that algorithmic risk assessments can systematically alter how decision-makers weight risk relative to competing considerations, in ways that can undermine policy goals even when prediction accuracy improves.

Read together, this literature establishes that the relationship between algorithmic input and expert decisions is real, consequential, and highly contextual. It does not, on its own, tell us whether human reviewers will buffer or transmit arbitrary variation in the algorithmic input itself. A second feature of this literature also limits how directly it bears on our question: much of it evaluates human reliance against a known ground truth, asking whether deference to the algorithm improves decision accuracy or fairness with respect to an observable outcome [15, 23, 40]. This is less applicable to college admissions and other social decisions that weigh multiple incommensurable criteria and lack a single objectively correct answer [8, 34, 43]. Our framework of algorithmic sensitivity takes up this gap: by asking only whether different scores lead to different decisions for the same individual, it decouples the question of how algorithmic input propagates through expert judgment from the question of whether the algorithm or the human is right.

2.2 Predictive Multiplicity and Algorithmic Arbitrariness

A second literature speaks to the property of algorithmic systems that motivates our methodological design. Breiman [7]’s early observation of the *Rashomon Effect*, that many distinct models can achieve similar aggregate performance on the same task, has been formalized in recent years as *predictive multiplicity* [5, 32] and *under-specification* [12]: equally accurate models can yield meaningfully different predictions for the same individual. A growing normative literature treats this contingency as a problem for the legitimacy of algorithmic decision-making. When outcomes depend on which of several equally defensible models is deployed, decisions become difficult to justify on substantive grounds [11], individuals bear the cost of these arbitrary choices [19], and these concerns extend into a broader argument against predictive optimization as a foundation for consequential decisions [46].

What is largely absent from this literature is empirical evidence on the question that determines whether the concern about model-level arbitrariness translates into a concern about decision-level arbitrariness in deployed systems. The literature documents that model outputs can differ for the same individual under defensible modeling choices; it generally does not test whether such differences propagate through the human decision-makers who actually issue consequential decisions in the institutional settings where these models are deployed. Our methodological move turns predictive multiplicity from a property to be characterized into a tool: by randomly assigning which of two equally accurate models’ scores a reviewer sees, we obtain exogenous variation in the favorability of the algorithmic input and can directly identify whether that variation propagates into outcomes.

2.3 Field-Based Studies of Algorithm-Assisted Decision-Making

A smaller body of work studies algorithm-assisted decision-making in its actual institutional context. This work is closest in spirit to ours, both in setting and in methodological ambition, and it shapes how we interpret what we find.

Stevenson and Doleac [44] study the introduction of algorithmic risk assessments for felony sentencing in Virginia. Their finding is nuanced: judges' decisions were influenced by the risk scores, with higher scores leading to longer sentences and lower scores to shorter ones, but judicial discretion mediated the tool's impact in systematic ways. Notably, judges granted leniency to young defendants despite their high risk scores, and over time used the scores less. The introduction of the tool did not produce the public safety gains its designers had anticipated, illustrating how expert discretion can substantially reshape the practical effect of algorithmic input even when input does enter decisions. Imai et al. [24] develop a statistical framework for experimentally evaluating algorithm-assisted human decision-making and apply it to a randomized controlled trial of the pretrial Public Safety Assessment, finding that providing the PSA to judges had little overall impact on judges' decisions or subsequent arrestee behavior. Brayne and Christin [6] provide ethnographic evidence from a large urban police department and a midsized criminal court, showing that algorithmic tools meet substantial professional resistance and that, where they are absorbed, they tend to relocate discretion to less visible and less accountable areas of organizational practice rather than displacing it entirely.

Across these studies, a common pattern emerges: the downstream consequences of algorithmic input in real institutional settings are often more muted, and more mediated by professional judgment, than analyses of model behavior in isolation would predict. Our empirical finding sits comfortably within this pattern, and the design we develop extends this body of work in two ways. Conceptually, prior field-based work largely retains the reliance framework's implicit appeal to ground truth, which constrains its portability to domains like admissions; our framework of algorithmic sensitivity is designed to be usable where that appeal is not available. Methodologically, prior studies vary the presence or salience of an algorithmic input, while we vary which of two equally defensible algorithmic inputs is shown, holding the institutional context, the reviewer, and the application constant. This lets us isolate the effect of the input's specific value from the effect of the input's presence, and to do so without imposing a manipulation that disadvantages applicants relative to what they would otherwise have received.

3 Context: Algorithm-Assisted College Admissions at a Selective U.S. University

Selective college admissions is a useful setting for studying how human experts respond to algorithmic input. Admissions decisions are not reducible to ranking applicants by a single notion of merit; they involve constructing a class that satisfies multiple, often competing institutional goals [43]. Admissions officers therefore engage in holistic review, synthesizing quantitative indicators with qualitative assessments of context, potential, and institutional fit [34]. As a result, admissions decisions are inherently judgment-based and normatively contested rather than objectively correct [8]. In this setting, the question of whether algorithmic input propagates into decisions can be asked directly, without recourse to a notion of decision correctness.

Our study is situated at a highly selective U.S. university during the 2021–2022 admissions cycle, following the widespread adoption of test-optional policies in response to the COVID-19 pandemic. These policies substantially reduced the availability of standardized test scores [4], historically a central quantitative input into admissions decisions. At the same time, application volumes rose, placing additional time and attention constraints on admissions officers. To support holistic review under these conditions, the institution introduced predictive algorithms trained on historical admissions data. The goal was not to automate admissions decisions but to provide a test-free signal that could help reviewers allocate limited attention and ensure that strong applicants were not inadvertently overlooked. Details on the design of these algorithms are reported in prior work [29, 30].

The algorithms produce predicted probabilities of admission based on information available at the time of application, including academic records and other observed characteristics. For operational use, these probabilities are discretized into deciles (1–10, with 10 indicating the highest predicted probability) and displayed to admissions officers alongside each applicant's materials. The scores are intended as a coarse, directional signal that reviewers can use to prioritize higher-scoring applications for earlier review or to flag applications that might otherwise

receive limited attention. They are not used as thresholds at any stage of the process: no applicant is automatically advanced, eliminated, or routed based on the score alone. Final admissions decisions remain with trained admissions professionals, who exercise holistic judgment based on essays, recommendation letters, contextual background, and institutional priorities not captured by the model.

Algorithmic scoring in this context raises ethical concerns specific to admissions that lie beyond the scope of the empirical question we study. Models trained on historical admissions data risk encoding patterns shaped by structural disadvantages [3, 18], and prior work at this institution shows that removing race data from applicant ranking algorithms reduces the diversity of the top-ranked pool without meaningfully improving academic merit, while increasing arbitrariness in outcomes for most applicants [30]. Our findings on algorithmic sensitivity do not resolve these concerns or speak to whether the deployed models are themselves fair. The Ethical Considerations Statement examines the design of this study in greater depth.

4 Methods

We define algorithmic sensitivity as a population-level causal estimand and develop the experimental design that permits its identification in a field setting.

4.1 Defining algorithmic sensitivity

We study whether decision outcomes for the same individual depend on which algorithmic prediction is presented when multiple plausible predictions exist. We define *algorithmic sensitivity* (AS) as the population-level causal effect of presenting a more favorable algorithmic prediction rather than a less favorable one on downstream decisions, holding fixed the individual.

Let A denote the space of algorithmic predictions that may be presented to a decision-maker. For any $a \in A$, let $Y_i(a) \in \{0, 1\}$ denote the potential decision outcome for individual i if prediction a were presented, where $Y_i(a) = 1$ indicates a favorable decision outcome (e.g., admission).

Let \succ denote an ordering over A such that $a^H \succ a^L$ means a^H is more favorable than a^L . For any ordered pair $(a^H, a^L) \in A \times A$ with $a^H \succ a^L$, define *algorithmic sensitivity* as

$$AS(a^H, a^L) := \mathbb{E}[Y_i(a^H) - Y_i(a^L)]. \quad (1)$$

Under this definition, algorithmic sensitivity has the following behavioral interpretation:

- $AS(a^H, a^L) \approx 0$: showing a^H rather than a^L does not materially change decisions.
- $AS(a^H, a^L) > 0$: showing a^H rather than a^L *increases* the likelihood of a favorable decision; larger values indicate stronger sensitivity to the algorithmic signal.
- $AS(a^H, a^L) < 0$: showing a^H rather than a^L *decreases* the likelihood of a favorable decision; larger negative values indicate stronger counteracting or aversive responses to the algorithmic signal.

This definition does not assume the existence of a ground-truth or objectively correct decision, nor does it require that either the human decision or the algorithmic prediction be accurate. Algorithmic sensitivity is defined purely in terms of how observed decisions respond to changes in the algorithmic input.

For example, if across individuals decisions are favorable 40% of the time when a^H is shown and 30% of the time when a^L is shown, then $AS(a^H, a^L) = 0.10$: presenting a^H rather than a^L increases the probability of a favorable decision by 10 percentage points on average. In settings characterized by predictive multiplicity, algorithmic sensitivity captures whether model-contingent differences in algorithmic scores translate into differences in real outcomes for the same individual.

4.2 Experiment Design

Our field experiment was conducted during the 2022 Regular Decision admissions cycle at a highly selective U.S. university with $n = 19,545$ applications, under a research protocol approved by an Institutional Review Board (IRB protocol number #2001009365) and in close collaboration with senior admissions leadership at the participating institution. The design exploits *predictive multiplicity*, the fact that multiple predictive models can achieve similar aggregate performance while producing different scores for the same individual [5, 32]. Rather than manipulating the favorability or quality of algorithmic advice, our design exploits naturally occurring, model-contingent variation among equally plausible predictions. This generates exogenous variation in the algorithmic score observed by decision-makers without degrading model accuracy or assigning any applicant a score from a model known to be inferior. Figure 4 in Appendix A illustrates the causal structure of the experimental design and highlights the specific edge under study.

4.2.1 Model development. We developed two predictive models, Model 1 and Model 2, using the same modeling pipeline but slightly different training data. Both models are gradient boosted decision tree classifiers trained to predict the probability of admission, following institutional practice and prior work in this setting [29, 30].¹

The data used to train the two models comes from the university’s 2019–20 and 2020–21 admissions cycles, including all information submitted via the Common Application: academic records (high school GPA, class rank, courses taken, AP/IB scores, and optional SAT subject scores) and personal information (essays, extracurricular activities, honors and awards, intended major, legacy status, and demographic indicators including gender, race and ethnicity, and first-generation status). Recommendation letters and school-specific counselor reports were not available. SAT/ACT and TOEFL/IELTS scores were explicitly excluded to simulate the test-optional environment of the experiment.

The two models differ only in their training data, reflecting realistic variation in modeling choices one would encounter in practice. Model 1 was trained on applicants from the 2020–21 Regular Decision cycle only; Model 2 was trained on both the 2019–20 and 2020–21 cycles. Each model outputs a predicted probability of admission, discretized into deciles (1–10) before being displayed to reviewers, consistent with institutional practice.

Table 1. Out-of-sample model validation performance.

	2020–21 RD		2021–22 ED	
	AUROC	Avg. Precision	AUROC	Avg. Precision
Model 1	86.9	32.5	80.9	52.1
Model 2	87.8	34.5	82.0	54.9

As reported in Table 1, the two models exhibit similar aggregate predictive performance on out-of-sample validation data. Model 2 performs slightly better on average in terms of AUROC and Average Precision, reflecting its larger training set. However, the inclusion of older data also increases the risk of distributional shift [37], making it uncertain whether this advantage would persist in the upcoming cycle—part of what motivates treating both models as equally defensible. Figure 1 also shows the two models are similarly calibrated with respect to observed decisions, supporting the comparability of their score scales. While aggregate performance is similar, the models show substantial individual-level disagreement: **the two models assign different score deciles**

¹Predicting admission probability grounds the signal in the institution’s own historical decisions, which is appropriate for a tool designed to support attention allocation within that same process. Alternative targets such as post-admission outcomes (e.g., GPA) were unavailable at application time and would not reflect the multi-criteria nature of admissions decisions, which involve constructing a class that satisfies multiple institutional goals beyond academic performance alone [43].

to **73.2% of applicants** in the experimental sample, providing the variation we exploit to study algorithmic sensitivity (Section 5.2).

4.2.2 Randomized Algorithmic Score Presentation. For each applicant i , both model scores are computed offline prior to human review, denoted $(S_{i1}, S_{i2}) \in \{1, \dots, 10\}^2$. Exactly one of these scores is shown to admissions officers, with assignment $W_i \sim \text{Bernoulli}(0.5)$, where $W_i = 0$ indicates that the Model 1 score is presented and $W_i = 1$ that the Model 2 score is presented. All other aspects of the applications and the decision-making processes are held fixed. Admissions officers were blind to which model produced the scores they were presented with.

Randomization yielded 9,765 applications assigned to the Model 1 condition and 9,780 to the Model 2 condition. The outcome of interest, $Y_i \in \{0, 1\}$, indicates whether applicant i was ultimately admitted; the overall admission rate in the analytic sample is 5.44%. Applicant characteristics are well balanced across experimental arms (Table 2), consistent with successful randomization.

Table 2. Proportion of applicants by demographic group across the two experimental conditions.

Condition	First-generation	International	URM	Female
Model 1 presented	0.171	0.269	0.144	0.298
Model 2 presented	0.165	0.262	0.147	0.310

4.3 Estimating algorithmic sensitivity

Our experimental design randomizes which model’s score is displayed, not whether reviewers see a more favorable versus a less favorable prediction. When the two models disagree for a given applicant, however, the randomized model assignment induces exogenous variation in the favorability of the displayed score. We show how this permits identification and unbiased estimation of algorithmic sensitivity for the subset of prediction pairs that arise from model disagreement.

4.3.1 Setup and estimand. We use the notation introduced in Section 4.2: $(S_{i1}, S_{i2}) \in \{1, \dots, 10\}^2$ are the two model scores for applicant i , $W_i \in \{0, 1\}$ indicates which is displayed, and $Y_i \in \{0, 1\}$ is the realized admission outcome.

Define the disagreement set $\mathcal{D} := \{i : S_{i1} \neq S_{i2}\}$. For each $i \in \mathcal{D}$, define the more and less favorable realized scores as

$$H_i := \max(S_{i1}, S_{i2}), \quad L_i := \min(S_{i1}, S_{i2}).$$

Let $Y_i(H_i)$ and $Y_i(L_i)$ denote the potential admission outcomes if the more favorable versus less favorable score were shown.² We define algorithmic sensitivity in our setting as

$$\text{AS} := \mathbb{E}[Y_i(H_i) - Y_i(L_i) \mid i \in \mathcal{D}]. \quad (2)$$

Equivalently, $\text{AS} = \mathbb{E}[\text{AS}(H_i, L_i) \mid i \in \mathcal{D}]$ in the notation of equation (1): the AS estimand is the expectation of the pair-level algorithmic sensitivity over the distribution of disagreement pairs in \mathcal{D} .

²This potential-outcomes notation invokes the Stable Unit Treatment Value Assumption (SUTVA): applicant i ’s outcome under a given displayed score depends only on the score shown for applicant i , not on scores shown for other applicants. In principle, admissions decisions could exhibit interference through class-size constraints. We expect any such spillover to be negligible in practice given the size of the applicant pool ($n = 19,545$) relative to the admitted class, the absence of within-cycle communication between reviewers about score assignments, and the balanced random assignment across applicants.

4.3.2 *Decomposition.* By the law of iterated expectations,

$$\begin{aligned} \text{AS} = & \underbrace{\Pr(S_{i1} > S_{i2} \mid i \in \mathcal{D})}_{\omega_{1>2}} \cdot \underbrace{\mathbb{E}[Y_i(H_i) - Y_i(L_i) \mid S_{i1} > S_{i2}]}_{\text{AS}_{1>2}} \\ & + \underbrace{\Pr(S_{i1} < S_{i2} \mid i \in \mathcal{D})}_{\omega_{2>1}} \cdot \underbrace{\mathbb{E}[Y_i(H_i) - Y_i(L_i) \mid S_{i1} < S_{i2}]}_{\text{AS}_{2>1}}. \end{aligned} \quad (3)$$

Here $\omega_{1>2}$ and $\omega_{2>1}$ are the frequencies of the two disagreement directions, and $\text{AS}_{1>2}$ and $\text{AS}_{2>1}$ are the corresponding stratum-specific components of algorithmic sensitivity.

4.3.3 *Identification.* Within \mathcal{D} , whether the higher score is shown is a deterministic function of (S_{i1}, S_{i2}) and W_i :

- If $S_{i1} > S_{i2}$, the higher score is shown if and only if $W_i = 0$.
- If $S_{i1} < S_{i2}$, the higher score is shown if and only if $W_i = 1$.

Since W_i is randomized independently of (S_{i1}, S_{i2}) and the potential outcomes $\{Y_i(a) : a \in A\}$, the stratum-specific components satisfy

$$\text{AS}_{1>2} = \mathbb{E}[Y_i \mid W_i = 0, S_{i1} > S_{i2}] - \mathbb{E}[Y_i \mid W_i = 1, S_{i1} > S_{i2}],$$

and

$$\text{AS}_{2>1} = \mathbb{E}[Y_i \mid W_i = 1, S_{i1} < S_{i2}] - \mathbb{E}[Y_i \mid W_i = 0, S_{i1} < S_{i2}].$$

In our analysis, we therefore estimate $\text{AS}_{1>2}$ and $\text{AS}_{2>1}$ by sample differences in means and combine them using the empirical weights $\widehat{\omega}_{1>2}$ and $\widehat{\omega}_{2>1}$ to obtain

$$\widehat{\text{AS}} = \widehat{\omega}_{1>2} \widehat{\text{AS}}_{1>2} + \widehat{\omega}_{2>1} \widehat{\text{AS}}_{2>1}.$$

Each stratum-specific component ($\widehat{\text{AS}}_{1>2}$ and $\widehat{\text{AS}}_{2>1}$) is computed as the difference in sample means of Y_i across the two values of W_i within the stratum, equivalent to the coefficient on W_i in an unadjusted linear probability model $Y_i = \alpha + \beta W_i + \epsilon_i$ fit within the stratum. Standard errors reported alongside our estimates are HC3 robust standard errors from this regression.

5 Results

Our analysis proceeds in three steps. First, we establish that the algorithmic scores shown to admissions officers are meaningful inputs to admissions decisions: higher scores are associated with higher admission probabilities, and this relationship is similar between the two models. Second, we characterize the nature of the algorithmic variation introduced by the experiment, documenting how often and in what ways the two models disagree in their predictions of the same applicant. Finally, leveraging this randomized variation in score presentation, we estimate algorithmic sensitivity: whether admission decisions are systematically influenced by which of two disagreeing but equally plausible algorithmic scores is shown. Together, these analyses allow us to assess whether individual-level algorithmic instability propagates into downstream admission decisions made by admissions officers.

5.1 How do algorithmic scores relate to admission decisions?

Before presenting our estimates of algorithmic sensitivity, we first examine how the algorithmic scores used in the experiment relate to downstream admission outcomes. Understanding this relationship is important for two reasons. Substantively, it establishes whether the scores are meaningful inputs into the decision process. Methodologically, it clarifies whether the two models operate on comparable score scales, which matters for interpreting the experimental manipulation.

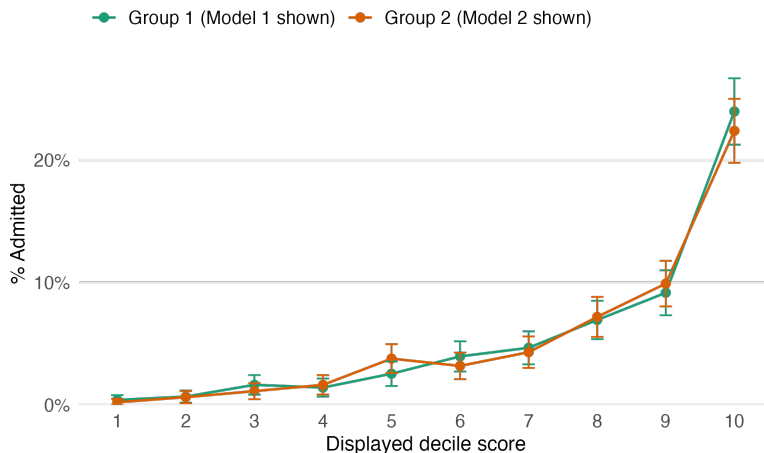


Fig. 1. Proportion of applicants admitted, by displayed algorithmic score decile, shown separately for applicants randomly assigned to have the Model 1 score (green) or Model 2 score (orange) displayed to admissions officers. Error bars indicate 95% confidence intervals.

Figure 1 plots observed admission rates by presented score, shown separately for applicants for whom the Model 1 or Model 2 score was presented. We observe two things: first, in both conditions, admission rates increase monotonically with the presented score. Second, we observe that across the full score distribution, the two curves closely overlap. This indicates that the score scales produced by the two models are similarly calibrated with respect to observed decisions, despite the two models differing in their assessments of individual applicants.

Algorithmic scores are therefore consequential inputs, and the two models operate on comparable score scales. We next characterize the disagreement that the experimental manipulation introduces.

5.2 What algorithmic disagreement does the experiment introduce?

5.2.1 How much do the models disagree at the individual level? We begin by quantifying the extent of individual-level disagreement between the two models. Although the models are similar in aggregate, they frequently diverge in their assessments of the same applicant. In our analytic sample, the two models assign different decile scores to 73.2% of applicants. The mean absolute difference in assigned scores is 1.49 deciles (SD = 1.38), and the correlation between the two scores is moderate ($\rho = 0.75$).

Figure 2 illustrates both the prevalence and structure of this disagreement. Panel (a) shows that disagreement is lowest at the extremes of the score distribution, corresponding to clearer admit and clearer deny decisions, and highest in the middle of the score distribution, which contains the majority of applicants for whom admission decisions are less certain. Panel (b) shows that while many applicants lie near the diagonal, indicating agreement or near-agreement, there is substantial mass off the diagonal, including cases where the two models differ by three or more deciles.

Together, these patterns indicate that the experimental manipulation introduces meaningful variation in the algorithmic scores shown to admissions officers, concentrated among applicants for whom decision-making is most difficult. Because applicants are randomly assigned to which model’s score is shown, this disagreement creates exogenous variation in the algorithmic information presented to reviewers, providing the key source of variation we exploit in subsequent analyses to study algorithmic sensitivity.

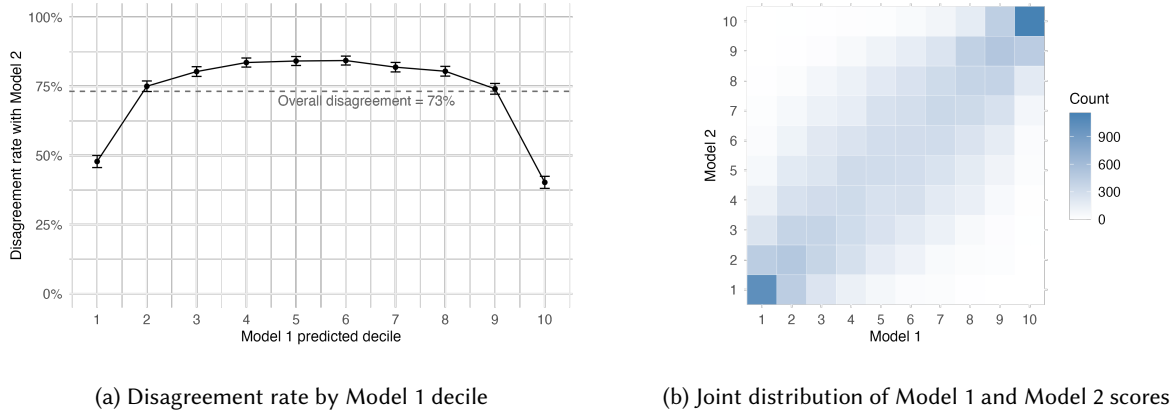


Fig. 2. Individual-level disagreement between Model 1 and Model 2. Panel (a) shows the probability that the two models assign different decile scores, conditional on the Model 1 decile; the solid line connects per-decile disagreement rates with 95% confidence interval error bars, and the dashed horizontal line denotes the overall disagreement rate (73%) across all applicants. Panel (b) shows the joint distribution of decile scores assigned by the two models, where darker shading indicates a higher number of applicants and diagonal cells correspond to exact agreement.

5.2.2 *Is one model systematically more favorable?* The two models are symmetric in the aggregate. Mean predicted scores are nearly identical (5.4822 under Model 1, 5.4821 under Model 2), and the distribution of signed score differences is approximately symmetric and centered near zero (Figure 3). Conditional on disagreement, the two directions occur with nearly equal frequency: $\omega_{1>2} = 0.501$ and $\omega_{2>1} = 0.499$ in the notation of Section 4.3. Neither model is systematically more lenient, so the AS estimand does not mechanically favor one model’s scores over the other; any estimated algorithmic sensitivity reflects responses to individual-level disagreement rather than global, systematic differences between the models. Aggregate symmetry masks modest subgroup-level variation (Model 2 assigns somewhat higher scores to URM and first-generation applicants), but as shown in Appendix B, these score-level differences do not translate into differential admission outcomes.

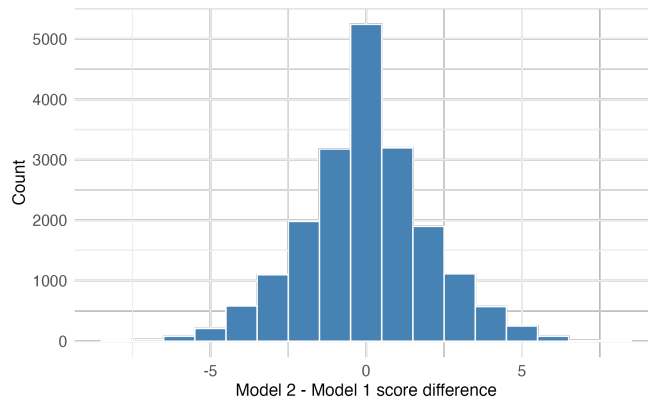


Fig. 3. Distribution of signed score differences (Model 2 minus Model 1) on the analytic sample.

Table 3. Estimates of algorithmic sensitivity (effect of showing the higher vs. lower score)

Case	N	$\mathbb{E}[Y W=0]$	$\mathbb{E}[Y W=1]$	\widehat{AS}	SE	p -value
$S_1 > S_2$ (Model 1 assigns the higher score)	7,166	0.0448	0.0420	+0.0029	0.0048	0.554
$S_1 < S_2$ (Model 2 assigns the higher score)	7,139	0.0371	0.0427	+0.0056	0.0046	0.225
$S_1 \neq S_2$ (All disagreement cases, \mathcal{D})	14,305	—	—	+0.0042	0.0033	0.206

Notes. Y indicates admission. $W = 0$ denotes that the Model 1 score is shown and $W = 1$ denotes that the Model 2 score is shown. Rows labeled $S_1 > S_2$ and $S_1 < S_2$ report estimates of the stratum-specific components $AS_{1>2}$ and $AS_{2>1}$ of algorithmic sensitivity as defined in Section 4.3. The pooled estimate \widehat{AS} is the weighted average $\omega_{1>2}AS_{1>2} + \omega_{2>1}AS_{2>1}$. Positive values indicate that showing the more favorable score increases the probability of admission. Standard errors are from linear probability models with HC3 robust standard errors.

5.3 Is there algorithmic sensitivity in admission decisions?

We now report estimates of the algorithmic sensitivity defined in Section 4.3: the average causal effect, among applicants for whom the models disagree, of showing the more favorable rather than the less favorable algorithmic score. Of the 19,545 applicants in the analytic sample, the models disagree for 14,305 applicants (73.2%). For the remaining 5,240 applicants (26.8%), the models assign identical scores. This agreement group provides no identifying variation for algorithmic sensitivity, because the favorability of the displayed score does not change with random assignment. Consistent with this, we observe no statistically detectable difference in admission outcomes across experimental arms (absolute difference = 0.34 pp, SE = 0.79 pp, $p = 0.67$).

Table 3 reports estimates of algorithmic sensitivity by disagreement direction. When Model 1 assigns the higher score ($S_1 > S_2$), showing the more favorable rather than the less favorable score increases admission probability by an estimated 0.29 percentage points (SE = 0.48 pp, $p = 0.55$). When Model 2 assigns the higher score ($S_1 < S_2$), showing the more favorable score increases admission probability by an estimated 0.56 percentage points (SE = 0.46 pp, $p = 0.23$). Pooling across disagreement directions yields an overall estimate of $\widehat{AS} = 0.42$ percentage points (SE = 0.33 pp, $p = 0.21$). None of these estimates are statistically distinguishable from zero at conventional significance levels. This null result holds among higher-scoring applicants, for whom admission rates are substantially higher. Restricting to those receiving a score of 9 or 10 from at least one model yields a weighted average AS of -0.005 , with only 2 strata reaching significance at the 5% level in opposite directions (see Appendix C). AS estimates stratified by demographic group are similarly small and statistically indistinguishable from zero across all subgroups examined, including URM and first-generation applicants (Appendix B).

5.4 Does algorithmic sensitivity vary with the magnitude of score disagreement?

Thus far, we have shown that algorithmic sensitivity is small and statistically indistinguishable from zero when aggregating across all disagreement cases. We next examine whether algorithmic sensitivity varies with the magnitude of disagreement between the two model scores. Specifically, we restrict attention to applicants in the disagreement set \mathcal{D} and stratify the sample by the absolute score gap, $|S_{i1} - S_{i2}|$. Within each stratum and disagreement direction, estimated effects correspond to conditional versions of algorithmic sensitivity that hold fixed both the direction and magnitude of model disagreement.

Table 4 reports the resulting estimates. Across most values of $|S_{i1} - S_{i2}|$, estimated effects are small and statistically indistinguishable from zero. A small number of strata reach significance at the 5% level, but the signs of these significant estimates are inconsistent across adjacent disagreement magnitudes within the same direction: $|S_{i1} - S_{i2}| = 3$ in the $S_1 > S_2$ case is negative and significant, while neighboring strata are positive; in the $S_1 < S_2$ case, $|S_{i1} - S_{i2}| = 4$ and 5 are significant in opposite directions.

Table 4. Estimates of algorithmic sensitivity (effect of showing the higher vs. lower score) by magnitude of model disagreement

$ S_1 - S_2 $	N	$\mathbb{E}[Y W=0]$	$\mathbb{E}[Y W=1]$	\widehat{AS}	SE	p -value
<i>Case 1: $S_1 > S_2$ (Model 1 assigns the higher score)</i>						
1	3,178	0.0442	0.0516	+0.0074	0.0076	0.327
2	1,981	0.0533	0.0426	-0.0108	0.0096	0.262
3	1,099	0.0531	0.0157	-0.0374	0.0111	0.001
4	583	0.0179	0.0461	+0.0281	0.0144	0.052
5	211	0.0094	0.0476	+0.0382	0.0230	0.099
6	82	0.0238	0.0000	-0.0238	0.0238	0.326
Overall	7,166	0.0448	0.0420	+0.0029	0.0048	0.554
<i>Case 2: $S_1 < S_2$ (Model 2 assigns the higher score)</i>						
1	3,197	0.0509	0.0480	-0.0029	0.0077	0.705
2	1,899	0.0307	0.0450	+0.0143	0.0088	0.102
3	1,114	0.0164	0.0265	+0.0102	0.0087	0.242
4	573	0.0204	0.0573	+0.0369	0.0162	0.023
5	251	0.0472	0.0000	-0.0472	0.0190	0.013
6	82	0.0244	0.0244	-0.0000	0.0345	1.000
Overall	7,139	0.0371	0.0427	+0.0056	0.0046	0.225

Notes. Y indicates admission. $W = 0$ denotes that the Model 1 score is shown and $W = 1$ denotes that the Model 2 score is shown. Within each panel, \widehat{AS} denotes the estimated component of algorithmic sensitivity, where positive values indicate a higher probability of admission when the more favorable score is shown. Rows labeled “Overall” correspond to $\widehat{AS}_{1>2}$ and $\widehat{AS}_{2>1}$ as reported in Table 3. Standard errors are from linear probability models with HC3 robust standard errors.

If decision-makers were increasingly influenced by score favorability as disagreement grew, we would expect effects to become larger in magnitude and more consistently signed at higher values of $|S_{i1} - S_{i2}|$. Instead, estimated effects fluctuate in both direction and size, suggesting that the isolated significant estimates are unlikely to reflect a consistent pattern of sensitivity.

6 Discussion

We find little evidence that admission decisions track which of two equally defensible algorithmic scores reviewers see. Across the 14,305 applications for which the two models disagreed, often by several deciles and most often in the middle of the score distribution where decisions are least certain, estimated sensitivity effects are small, statistically indistinguishable from zero in pooled specifications, and inconsistent in direction across strata and subgroups. In this setting, the model-contingent variation that predictive multiplicity introduces does not propagate into final admission outcomes.

A null estimate of algorithmic sensitivity is, by itself, compatible with two distinct decision-making patterns. Reviewers may place little weight on the algorithmic score, so that variation in the displayed score does not move their decisions much. Alternatively, reviewers may place substantial weight on the score while discounting modest variation across models, on the implicit understanding that small differences in displayed score need not reflect substantively meaningful differences between applicants. Our design cannot distinguish between these patterns: both yield small sensitivity estimates, but they imply different things about how the algorithm functions

in the decision process and about whether the absence of sensitivity here would generalize to settings in which the algorithm carries more weight in the decision.

This indeterminacy does, however, bound what mechanisms may be at work. If reviewers were strongly anchoring on the displayed score under time pressure, as participants do in laboratory settings [38], we would expect the more favorable score to produce systematically more favorable outcomes; we do not see this. If reviewers felt their decision-making authority was constrained by the algorithm and responded by deferring to whatever score was displayed, as decision-makers do in some algorithm-assisted settings [26], we would similarly expect the score to drive outcomes; again, we do not. Whatever decision-making pattern is operative in our setting, it is not one in which the displayed score is propagated mechanically into outcomes.

If a mechanism of active discounting is at work, several features of the decision-making context plausibly contribute to it. Admissions decisions are holistic, integrating essays, recommendation letters, contextual background, and institutional priorities alongside the algorithmic score [43]. The score is framed as a coarse, directional signal for attention allocation rather than as a recommendation, and is not used as a threshold at any stage of the process [29, 30]. Reviewers are experienced admissions professionals operating under institutional norms that legitimize discretion. The muted downstream influence we observe is consistent with prior field-based work showing that algorithmic input in real institutional settings is often more mediated by professional judgment than algorithm-centric accounts would predict [6, 24, 44]. Consistency with this broader pattern does not, however, establish that these specific features explain the attenuation we observe here, since our design varies the algorithmic input rather than the institutional context surrounding it.

Beyond the mechanism question, our finding has implications for how predictive multiplicity has been theorized as a problem. The literature has largely concentrated on properties of models: whether equally accurate models produce different predictions for the same individual, and what to do about it through ensembling, model selection, or calibration [5, 11, 32]. Our result addresses a different question, namely whether model-level multiplicity translates into outcome-level arbitrariness once the model is deployed in a human-mediated decision process. In our setting, it does not. Du et al. [16] make a complementary observation from the opposite direction: even when two predictive models agree on their individual predictions almost everywhere, they can lead downstream decision-makers to take substantially different actions. Taken together, these findings caution against assuming that what models do at the prediction level determines what decisions look like at the outcome level: the sociotechnical context in which a model is deployed can produce arbitrary outcomes from agreeing models or non-arbitrary outcomes from disagreeing ones. Which direction the context takes is itself an empirical question [41], and our results provide one data point on the attenuating side.

Identifying these patterns required a framework that does not presume a notion of decision correctness, since the reliance literature’s accuracy-based evaluation [23] is unavailable in many social decision settings, including admissions [8]. The algorithmic sensitivity framework, by defining the estimand purely in terms of how decisions respond to changes in the algorithmic input, makes the propagation question tractable in such settings.

A null effect on algorithmic sensitivity is consistent with the absence of one specific harm—the systematic determination of individual outcomes by arbitrary modeling choices—but does not address the broader set of concerns motivating recent normative work on algorithmic decision-making. Wang et al. [46] argue that predictive optimization in consequential domains is presumptively illegitimate because of a recurring set of flaws (target mismatch, distribution shift, limited contestability, susceptibility to gaming) that are not readily addressed by technical means; our results speak to one specific failure mode within this critique but do not address the broader argument. Models trained on historical admissions data may still encode patterns shaped by structural disadvantage, even when the choice between two such models does not produce differential outcomes for the groups we observe [3, 18]. Algorithmic scoring may still obscure the basis for consequential decisions from applicants and the public [9]. And the attenuation documented here should not be read as a general endorsement of human-in-the-loop arrangements: Green [21] shows that human oversight requirements have repeatedly been

used to legitimize the deployment of algorithmic systems whose underlying problems remained unaddressed. Our results establish only that, in this setting, decisions were not systematically swayed by which of two model scores was shown; they do not establish that reviewers were performing the oversight that human-in-the-loop policy typically assumes, nor that the underlying models are appropriate to the decisions they inform. Although the experiment did not produce detectable disparate impact across the demographic groups we observed, individual applicants whose displayed score was less favorable than the alternative model would have produced bore a real consequence of the design.

6.1 Limitations

Four limitations bear on the interpretation of our findings. First, the variation we exploit comes from two specific models trained on overlapping but non-identical historical cohorts, rather than from randomized splits of identical data. Some of the disagreement between the two models therefore reflects cohort-level distributional shift [37], not just the pure sampling variability that most of the predictive multiplicity literature has focused on. This reflects realistic deployment conditions, since models in practice are retrained as new data arrive and any institution choosing between two such models encounters both sources of variation at once. But it means our estimates cannot be cleanly interpreted as sensitivity to sampling noise alone. Other forms of multiplicity, across model families, feature sets, or hyperparameter choices, may produce different patterns of disagreement and different patterns of sensitivity; whether our null extends to them is an open question.

Second, the algorithmic sensitivity estimand is identified on the subset of applicants for whom the two models disagree, which is not a random sample of the full applicant pool. Disagreement is concentrated in the middle score deciles, where admission rates are low (Figure 2), so the pooled estimates speak most directly to applicants whose scores were not strongly signaling either admission or rejection. The robustness check in Appendix C addresses the case where sensitivity would matter most: applicants receiving high scores from at least one model, for whom admission rates are substantially higher and any sensitivity would have larger practical consequences. The null holds in this subgroup as well, but we cannot rule out that sensitivity exists in other subpopulations our analysis does not cover.

Third, our findings are specific to a setting with substantial procedural structure, experienced reviewers, and an algorithmic input framed as a coarse, directional signal rather than as a recommendation. We have argued that this configuration plausibly contributes to the attenuation we observe, but we have not varied it directly. Settings in which algorithmic scores are tightly coupled to decisions, in which reviewers lack expertise or discretion, or in which institutional pressure rewards mechanical compliance with algorithmic outputs may show very different patterns of sensitivity [22, 44, 47]. Our design cannot identify which specific features of the institutional process are responsible for the attenuation, or whether the attenuation would survive in settings with different features.

Finally, our null result is bounded by statistical power. The 95% confidence interval around the pooled estimate rules out effects larger than roughly one percentage point in either direction, but smaller effects are not detectable in our sample. Our data therefore rule out substantial algorithmic sensitivity in this setting, but they do not establish that decisions are exactly invariant to which score is shown.

Acknowledgments

This research was supported in part by NSF Awards IIS-2312865, OAC-2311521, and EDU-2237593. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

- [1] Alex Albright. 2019. If you give a judge a risk score: Evidence from Kentucky bail decisions. *Law, Economics, and Business Fellows' Discussion Paper Series* 85 (2019), 2019–1.

- [2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
- [3] Solon Barocas and Andrew D Selbst. 2016. Big data’s disparate impact. *Calif. L. Rev.* 104 (2016), 671.
- [4] Christopher T Bennett. 2022. Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies. *American Educational Research Journal* 59, 1 (2022), 180–216.
- [5] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 850–863.
- [6] Sarah Brayne and Angèle Christin. 2021. Technologies of crime prediction: The reception of algorithms in policing and criminal courts. *Social problems* 68, 3 (2021), 608–624.
- [7] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.
- [8] Elizabeth Bruch and Fred Feinberg. 2017. Decision-making processes in social contexts. *Annual review of sociology* 43, 1 (2017), 207–227.
- [9] Jenna Burrell. 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big data & society* 3, 1 (2016), 2053951715622512.
- [10] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghui Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How child welfare workers reduce racial disparities in algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [11] A Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelman, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. 2024. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22004–22012.
- [12] Alexander D’Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* 23, 226 (2022), 1–61.
- [13] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.
- [14] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.
- [15] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1639–1656.
- [16] Ally Du, Dung Daniel Ngo, and Steven Wu. 2025. Reconciling model multiplicity for downstream decision making. In *International Conference on Learning Representations*, Vol. 2025. 59133–59164.
- [17] João Gama, Indrè Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 1–37.
- [18] Denisa Gándara, Hadis Anahideh, Matthew P Ison, and Lorenzo Picchiarini. 2024. Inside the black box: Detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. *AERA open* 10 (2024), 23328584241258741.
- [19] Prakhar Ganesh, Ihsan Ibrahim Daldaban, Ignacio Cofone, and Golnoosh Farnadi. 2024. The cost of arbitrariness for individuals: Examining the legal and technical challenges of model multiplicity. *arXiv preprint arXiv:2407.13070* (2024).
- [20] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.
- [21] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681.
- [22] Ben Green and Yiling Chen. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.
- [23] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 221–236.
- [24] Kosuke Imai, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. 2023. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society* 186, 2 (2023), 167–189.
- [25] Intelligent.com. 2023. 8 in 10 Colleges Will Use AI in Admissions by 2024. <https://www.intelligent.com/8-in-10-colleges-will-use-ai-in-admissions-by-2024/>. Accessed: 2026-05-11.
- [26] Anupama A Jolly, Patrick D Dunlop, Sharon K Parker, and Lisette Kanse. 2026. It’s Not My Responsibility: Working with Autonomy-Restricting Algorithms Facilitates Unethical Behavior and Displacement of Responsibility: AA Jolly et al. *Journal of Business Ethics* 203, 2 (2026), 405–424.

- [27] Daniel Kahneman, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser. 2016. Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harvard Business Review* (Oct. 2016). <https://hbr.org/2016/10/noise> Section: Decision making and problem solving.
- [28] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* (Aug. 2017).
- [29] Hansol Lee, René F Kizilcec, and Thorsten Joachims. 2023. Evaluating a learned admission-prediction model as a replacement for standardized tests in college admissions. In *Proceedings of the tenth acm conference on learning@ scale*. 195–203.
- [30] Jinsook Lee, Emma Harvey, Joyce Zhou, Nikhil Garg, Thorsten Joachims, and René F Kizilcec. 2024. Algorithms for college admissions decision support: Impacts of policy change and inherent variability. *arXiv preprint arXiv:2407.11199* (2024).
- [31] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
- [32] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *International conference on machine learning*. PMLR, 6765–6774.
- [33] Paul E Meehl. 1954. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. (1954).
- [34] Rochelle S Michel, Vinetha Belur, Bobby Naemi, and Harrison J Kell. 2019. Graduate admissions practices: A targeted review of the literature. *ETS Research Report Series* 2019, 1 (2019), 1–18.
- [35] Virginia Tech News. 2025. Virginia Tech updates undergraduate admissions process. *Virginia Tech News* (2025). <https://news.vt.edu/articles/2025/07/admissions-changes-2025.html>
- [36] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI literature review. *Microsoft Research* 339 (2022), 340.
- [37] Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2022. *Dataset shift in machine learning*. MIT Press.
- [38] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–22.
- [39] Kara Schechtman, Benjamin Brandon, Jenise Stafford, Hannah Li, and Lydia T Liu. 2025. Discretion in the Loop: Human Expertise in Algorithm-Assisted College Advising. *arXiv preprint arXiv:2505.13325* (2025).
- [40] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.
- [41] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.
- [42] Daniela Sele and Marina Chugunova. 2024. Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *Plos one* 19, 2 (2024), e0298037.
- [43] Mitchell L Stevens. 2009. *Creating a class*. Harvard University Press.
- [44] Megan T Stevenson and Jennifer L Doleac. 2024. Algorithmic risk assessment in the hands of humans. *American Economic Journal: Economic Policy* 16, 4 (2024), 382–414.
- [45] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.
- [46] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing* 1, 1 (2024), 1–45.
- [47] Kyra Wilson, Mattea Sim, Anna-Maria Gueorguieva, and Aylin Caliskan. 2025. No Thoughts Just AI: Biased LLM Hiring Recommendations Alter Human Decision Making and Limit Human Autonomy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2692–2704.

Ethical Considerations Statement

This study was conducted under a research protocol approved by an Institutional Review Board (IRB) and in close collaboration with senior admissions leadership at the participating institution. The research team worked alongside the institution's admissions office, which had independently determined that algorithmic scoring tools were needed to support holistic review under test-optional conditions. The widespread adoption of test-optional policies during the COVID-19 pandemic had eliminated standardized test scores as a universally available quantitative signal, while application volumes continued to rise, placing substantial time and attention constraints on admissions officers tasked with conducting individualized review of each application. Algorithmic scoring at the institution predated this study: a prior heuristic-based tool, which relied heavily on standardized test scores, had become ineffective under test-optional policies. Rather than having the admissions office develop new heuristics on their own, the research collaboration meant that the successor tools were developed and evaluated rigorously, and their effects on admissions decisions studied empirically, providing stronger empirical grounding than would typically be available in operational deployments without research involvement. All applicants still underwent human review, and the tools were not used to automate or replace human judgment. The admissions office was also an active partner in the research design: institutional leadership were specifically interested in understanding how sensitive their staff's decisions would be to algorithmic scores, a question with direct operational relevance to how these tools were deployed and communicated. All data were de-identified prior to analysis.

Potential for harm. A central ethical concern in this design is whether presenting a lower score to some applicants constituted a disadvantage. We addressed this on two grounds prior to deployment. First, the two models were selected to be equally defensible: both achieved similar aggregate predictive performance and were trained on the same feature set and modeling pipeline. Although Model 2 performed slightly better on out-of-sample validation data, the inclusion of older training data increases the risk of distributional shift [17], making it uncertain whether this advantage would persist in the upcoming cycle. Neither model was therefore clearly preferable, and no applicant was assigned a score from a model known to be inferior. Second, no applicant was automatically advanced or eliminated based on the score alone at any stage of the process; scores served exclusively as coarse, directional signals for attention allocation, with all final decisions made through holistic review by trained admissions professionals.

Our empirical findings further support these assurances. Appendix B reports predicted score distributions for both models by demographic group. Model 2 assigns modestly higher scores to URM and first-generation applicants on average (mean differences of 0.48 and 0.51 deciles respectively), likely reflecting year-to-year variation in admission patterns between the two training cohorts, while differences for other groups are negligible. These score-level differences did not translate into differences in admission outcomes: the two models were similarly calibrated with respect to observed admission decisions (Figure 1), and AS estimates are small and statistically indistinguishable from zero across the main analysis (Table 3) and across demographic subgroups (Table 6). The randomization did not systematically disadvantage applicants in either condition. This does not establish that the models are free of bias in any broader sense, only that the randomization did not produce differential harm across the groups we observed.

Broader considerations. Ongoing debates concern whether algorithmic tools should be used in university admissions decisions at all. Models trained on historical admissions data risk encoding patterns shaped by structural disadvantage: if prior admissions decisions reflected disadvantages facing certain groups, models trained on those decisions may systematically undervalue applicants from those groups [3, 18]. Algorithmic scoring may also obscure the basis for consequential decisions from applicants and the public [9]. Algorithmic

input may subtly reshape holistic review by directing reviewer attention in ways that disadvantage already-marginalized applicants, even when no individual decision is explicitly automated. These concerns are not resolved by our findings. Our study addresses a specific, narrow question, namely whether human reviewers are sensitive to arbitrary variation in algorithmic scores, and a null result on this question is not an endorsement of algorithmic admissions tools broadly. In particular, our findings say nothing about whether the deployed models produce fair predictions, whether algorithmic scoring improves or undermines admissions decisions overall, or whether the use of such tools is appropriate in this context. The appropriate role of algorithms in high-stakes admissions decisions remains an open normative question for institutional leaders, affected communities, and researchers.

A Causal Graph of the Experimental Design

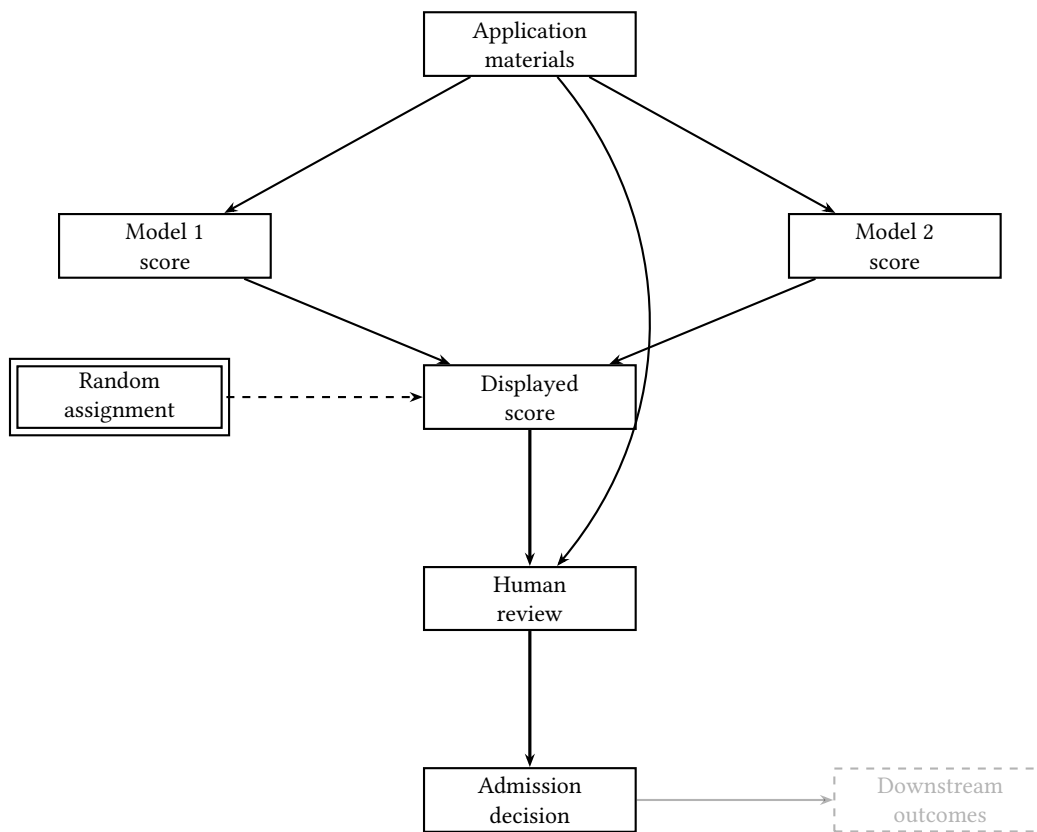


Fig. 4. Causal graph of the experimental design. The bold path from displayed score to human review to admission decision represents the AS estimand: the causal effect of the displayed algorithmic score on the admission decision, mediated through the human reviewer’s judgment. Random assignment (double border) generates exogenous variation in the displayed score. The dimmed node and edge indicate that downstream decision outcomes are outside the scope of this study.

B Model Score Distributions and Algorithmic Sensitivity by Demographic Group

Figure 5 shows mean predicted score deciles for Model 1 and Model 2 by demographic group, with 95% confidence intervals. The two models produce closely aligned score distributions for most groups. Model 2 assigns modestly higher scores to URM applicants (mean difference = 0.48 deciles) and first-generation applicants (mean difference = 0.51 deciles) relative to Model 1, while differences for other groups are negligible. Table 5 reports the corresponding summary statistics.

These subgroup-level differences likely reflect year-to-year variation in admission patterns across the two training cohorts: because URM and first-generation applicants are evaluated more holistically and their admission rates are more sensitive to institutional priorities that may shift across cycles, small differences in training data composition disproportionately affect their predicted scores relative to groups whose admission patterns are more stable [37].

Despite these score-level differences, Table 6 shows that AS estimates are small and statistically indistinguishable from zero across all demographic groups, indicating that the modest score differences between models did not propagate into differential admission outcomes for any group. Taken together, these results suggest that neither model systematically disadvantaged any demographic subgroup in terms of realized admission decisions.

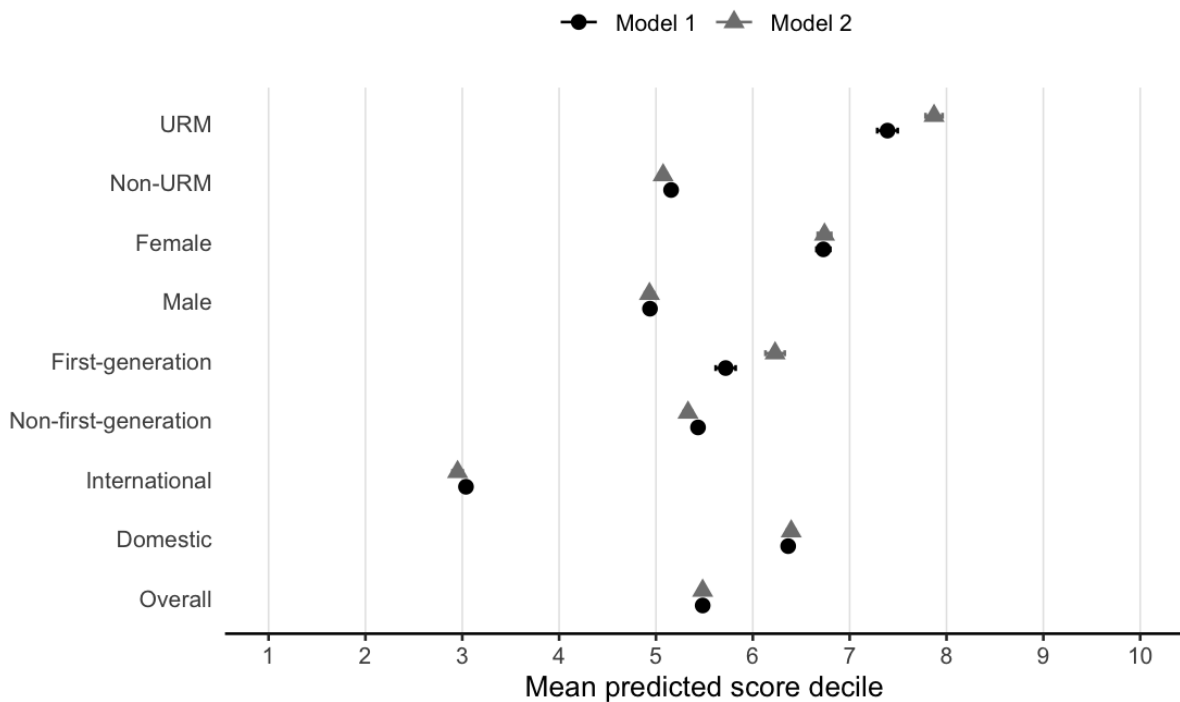


Fig. 5. Mean predicted score decile by demographic group for Model 1 and Model 2. Error bars indicate 95% confidence intervals. The two models are closely aligned for most groups; Model 2 assigns modestly higher scores to URM and first-generation applicants.

Table 5. Mean predicted score decile by demographic group.

Group	N	\bar{S}_1	\bar{S}_2	Mean diff.	SD diff.
Overall	19,545	5.482	5.482	+0.000	2.031
URM	2,852	7.391	7.871	+0.481	1.808
Non-URM	16,693	5.156	5.074	-0.082	2.055
Female	5,940	6.728	6.740	+0.012	2.004
Male	13,605	4.938	4.933	-0.005	2.042
First-generation	3,285	5.720	6.230	+0.511	1.968
Non-first-generation	16,260	5.434	5.331	-0.103	2.028
International	5,187	3.038	2.952	-0.087	1.907
Domestic	14,358	6.365	6.396	+0.031	2.073

Notes. \bar{S}_1 and \bar{S}_2 denote mean predicted score deciles for Model 1 and Model 2 respectively. Mean diff. is Model 2 minus Model 1. SD diff. is the standard deviation of the within-applicant score difference.

Table 6. Estimates of algorithmic sensitivity by demographic group.

Group	N	\widehat{AS}	SE	p -value
URM	1,740	+0.0156	0.0124	0.210
Non-URM	12,565	+0.0025	0.0034	0.467
Female	4,210	+0.0002	0.0072	0.983
Male	10,095	+0.0057	0.0036	0.118
First-generation	2,342	+0.0052	0.0088	0.552
Non-first-generation	11,963	+0.0040	0.0036	0.264
International	3,540	+0.0013	0.0032	0.680
Domestic	10,765	+0.0049	0.0043	0.254

Notes. Sample restricted to applicants for whom the two models disagree. \widehat{AS} is the estimated effect of being shown the more favorable score on admission probability, estimated via linear probability model with HC3 robust standard errors. No estimate is statistically significant at the 5% level.

C Algorithmic Sensitivity Among Higher-Scoring Applicants

A potential concern with the main analysis is that disagreement cases are concentrated in middle deciles where admission rates are low, such that near-zero estimates of algorithmic sensitivity could partly reflect a floor effect rather than genuine invariance. To address this, we restrict attention to applicants receiving a score of 9 or 10 from at least one model and for whom the two models disagree ($n = 3,561$), where admission rates are substantially higher and algorithmic sensitivity would be most detectable if present.

Table 7 reports estimates stratified by the magnitude of model disagreement within this subset. The weighted average AS is -0.005 . Only 2 strata reach significance at the 5% level, and these are in opposite directions, providing no evidence of a consistent pattern of sensitivity. The null result thus holds precisely where sensitivity would be most detectable, alleviating concerns that the main findings are an artifact of floor effects among lower-scoring applicants.

Table 7. Estimates of algorithmic sensitivity by magnitude of model disagreement (subset: applicants scored 9 or 10 by at least one model)

$ S_1 - S_2 $	N	$\mathbb{E}[Y W=0]$	$\mathbb{E}[Y W=1]$	\widehat{AS}	SE	p -value
<i>Case 1: $S_1 > S_2$ (Model 1 assigns the higher score)</i>						
1	838	0.1019	0.1202	+0.0183	0.0218	0.401
2	449	0.1132	0.0928	-0.0204	0.0289	0.481
3	244	0.0957	0.0310	-0.0646	0.0317	0.042
4	134	0.0345	0.0658	+0.0313	0.0377	0.408
5	75	0.0000	0.0227	+0.0227	0.0230	0.326
6	31	0.0000	0.0000	+0.0000	0.0000	—
<i>Case 2: $S_1 < S_2$ (Model 2 assigns the higher score)</i>						
1	841	0.1394	0.1181	-0.0213	0.0232	0.358
2	405	0.0686	0.1244	+0.0558	0.0294	0.059
3	242	0.0268	0.0308	+0.0040	0.0217	0.854
4	149	0.0405	0.1333	+0.0928	0.0461	0.046
5	75	0.0811	0.0000	-0.0811	0.0461	0.083
6	37	0.0000	0.0000	+0.0000	0.0000	—

Notes. Y indicates admission. $W = 0$ denotes that the Model 1 score is shown and $W = 1$ denotes that the Model 2 score is shown. Within each panel, \widehat{AS} denotes the estimated component of algorithmic sensitivity, where positive values indicate a higher probability of admission when the more favorable score is shown. Boldface rows indicate estimates significant at the 5% level. Standard errors are from linear probability models with HC3 robust standard errors.

Generative AI Usage Statement

Generative AI tools were used in a limited capacity during manuscript preparation to assist with grammar and fluency editing of author-written text, as well as with formatting tables. Generative AI tools were not used to generate substantive text, arguments, results, analyses, or interpretations. All intellectual contributions, empirical analyses, and conclusions are solely those of the authors, who take full responsibility for the originality, accuracy, and integrity of the manuscript.