NBER WORKING PAPER SERIES

RACIAL SCREENING ON THE BIG SCREEN? EVIDENCE FROM THE MOTION PICTURE INDUSTRY

Liang Zhong ⓡ
Angela Crema ⓡ
M. Daniele Paserman ⓡ

Working Paper 33186
http://www.nber.org/papers/w33186

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2024

We thank Matthew Gudgeon, Patrick Kline, Erzo F.P. Luttmer, Elena Manresa, Anna Weber and seminar participants at EEA-ESEM 2022, SOLE 2022, Georgia Tech, the United States Military Academy, the Boston University Labor Reading Group, and the New York University Econometrics Student Lunch for helpful suggestions. We also thank Venkata Andra, Tabitha Fortner, Srihaasa Kompella, Antonia Lehnert, Haoyu Liu, Jacob Park, Richard Pilleul, Jacob Pothen, Rishab Sudhir, Jason Wang, and Hongyi Yu for outstanding research assistance. All errors are our own. ⓡ Author names are in random order following Ray ⓡ Robson (2018). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Racial Screening on the Big Screen? Evidence from the Motion Picture Industry
Liang Zhong ⓡ Angela Crema ⓡ M. Daniele Paserman ⓡ
NBER Working Paper No. 33186
November 2024
JEL No. J15, J70, L82

ABSTRACT

We develop a model of discrimination that allows us to interpret observed differences in outcomes across groups, conditional on passing a screening test, as taste-based (employer,) statistical, or customer discrimination. We apply this framework to investigate the nature of non-white underrepresentation in the US motion picture industry. Leveraging a novel data set with racial identifiers for the cast of 7,000 motion pictures, we show that, conditional on production, non-white movies exhibit higher average revenues and a smaller variance. Our findings can be rationalized in the context of our model if non-white movies are held to higher standards for production.

Liang Zhong ⓡ
Boston University
samzl@bu.edu

Angela Crema ⓡ
Yale School of Management
165 Whitney Ave
New Haven, CT 06511
angela.crema@yale.edu

M. Daniele Paserman ⓡ
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and CEPR, IZA
and also NBER
paserman@bu.edu

1 Introduction

An employer must decide whether to hire a job applicant. An admission committee must decide whether to admit a candidate to its entering freshman class. A journal editor must decide whether to accept an article for publication. All these settings are characterized by a *decision maker* who must make an in-or-out decision about an *applicant*, having only imperfect information about the applicant's quality. The decision maker may use information about the applicant's race or gender to guide their decision, which may result in discrimination, i.e., the unequal treatment of applicants with otherwise identical characteristics. The econometrician, however, can typically observe only the ex-post outcomes of these decisions: the worker's productivity, the student's grades, or the number of citations received by an article. If we observe differences by race or gender in outcomes, what can we infer about the extent and nature of discrimination by the decision maker?

In this paper, we address this question in the context of the U.S. motion picture industry, where the producer is the decision maker. There are two main advantages to studying discrimination in the motion picture industry. First, this setting is of intrinsic interest because of the widespread perception of bias in the industry. For example, in the 2010s, only 7% of the nominees for the Academy Awards were African Americans, which is approximately half of their proportion in the population.[1] Does this underrepresentation reflect racial bias? Second, we can accurately measure productivity using box office revenue. This is an essential requirement to understand the nature of discrimination. The existence of discrimination in this industry can also have wider ranging implications, because actors can also serve as role models and impact students' educational attainment (Riley, 2024). Therefore, racial discrimination in movie production may differentially affect young viewers of different backgrounds. Understanding whether and to what extent discrimination can be reduced (e.g., via information; Chan, 2024) may guide the design of corrective policies.[2]

[1] https://www.washingtonpost.com/news/arts-and-entertainment/wp/2016/02/26/these-charts-explain-how-oscars-diversity-is-way-more-complicated-than-you-think/, accessed on October 26, 2021.

[2] Recently, the Academy of Motion Picture Arts and Sciences has announced a multitude of diversity-oriented changes, including diversity requirements for movies that wish to be nominated for the Academy Award in the Best Picture category. In this paper, we analyze a time period that precedes the inclusion of

We develop a model of discrimination that allows us to interpret differences in box-office revenue, conditional on production. In the model, a producer[3] receives an offer to produce a movie (a "script," similar to the applicant in the examples above). They observe the expected racial composition of the cast based on the script and receive a noisy signal of the movie's expected box-office revenue. Based on the information, they must choose whether to produce the movie and release it to the public or not. We define a "white" movie as a movie in which the leading roles are solely played by whites and a "non-white" movie as a movie in which the leading roles include non-whites. Our model nests different forms of discrimination within it and delivers a rich set of predictions regarding the extent and nature of discrimination. We distinguish between three types of discrimination: a) *customer discrimination*, whereby moviegoers have a preference for white movies over non-white movies; b) *employer* or *taste-based discrimination*, where the producer suffers a negative utility from producing a non-white movie (Becker, 1957); and c) *statistical discrimination*, where the signal conveyed by non-white movies is less informative about the movie's true quality (Phelps, 1972; Arrow, 1973). We show that the moments of the distribution of box-office revenue of movies *that are produced* allow one to distinguish between the three types of discrimination.

To test the model's predictions, we construct a novel data set with racial identifiers for the cast of more than 7,000 motion pictures released in the United States between 1997 and 2017. We obtained the data by scraping the popular website IMDB,[4] and combined it with extensive information from OpusData, a private company specialized in providing data and information on the movie industry.[5] The racial identifiers are constructed by combining human raters' classifications and a machine learning architecture that integrates a convolutional neural network (CNN) and support vector machine (SVM; Anwar and Islam, 2017).[6]

such standards.

[3]Throughout the paper we refer for simplicity to the agent deciding on whether to produce the movie as the "producer." This could be a studio executive or other decision maker, and does not necessarily have to coincide with the producer listed in the movie's credits.

[4]http://www.imdb.com

[5]www.opusdata.com

[6]We rely on the machine learning algorithm to classify the 8% of actors in our data whose racial classification was an object of disagreement for more than two of our eight (sometimes nine) human raters. We find that the algorithm obtains a classification accuracy of more than 95% in our validation data set, which

In our main analysis, we define a movie as "non-white" if two of the four top-billed performers are classified as non-white.[7] We document the following findings. First, the average box-office revenue of non-white movies is substantially *higher* than that of white movies. The raw non-white/white revenue gap is about 91 log points (150%). The inclusion of a standard set of control variables for other movie characteristics and the cast reduces the gap to between 43 and 34 log points (between 54% and 40%), still large and highly statistically significant. Second, the box office premium of non-white movies is driven primarily by movies in the bottom half of the distribution. Quantile regressions show that the adjusted gap is around 54 log points (about 72%) at the bottom quantiles of the distribution, but the gap at the upper end of the distribution shrinks to about 28 log points (about 33%). These results are robust to different definitions of non-white movies or different dependent variables (e.g., profit margins or profits). Third, we create a measure of the extent to which a movie's box-office revenue overperforms relative to expectations. Following Moretti (2011), we calculate this as the residual in a regression of opening weekend box-office revenue on the number of opening-weekend theaters. We find that relative to white movies, non-white movies substantially overperform relative to expectations.

These results are not consistent with either customer discrimination or statistical discrimination. Instead, we argue that the results are consistent with taste-based discrimination:[8] Non-white movies are held to a higher standard, i.e., they are produced only if the expected revenue surpasses a threshold that is higher than the one set for white movies. This pattern may result from either pure producer taste or a systematic underestimation of the box-office potential of non-white movies.

This paper is situated within a broad and interdisciplinary literature documenting and exploring discrimination in a variety of settings. While our goal in the next few paragraphs is to focus on the streams of this work to which we directly contribute, we refer the reader to

is considered excellent in the image classification literature. See section 4 for further details.

[7] The non-white category includes mostly African-Americans but may also include Asians, Hispanics, and other ethnicities.

[8] Our identification argument relies on the ordering of the means of the (observed) white and non-white box-office revenue distributions, as well as on the ordering of the variances. Therefore, our results may be interpreted as taste-based discrimination quantitatively dominating any other forms of discrimination that may be at play.

several excellent surveys in economics, including Fang and Moro (2011), Lang and Lehmann (2012), Bertrand and Duflo (2017), Lang and Spitzer (2020) and Onuchic (2022) for a broader overview.

We see our work contribute to the stream of the literature that aims to understand the nature of unequal treatments by either distinguishing between statistical and taste-based discrimination in the data,[9] or testing for the presence of one of the two in a specific market or context. These goals have been pursued experimentally (List, 2004; Zussman, 2013; Doleac and Stein, 2013; Agan and Starr, 2018; Cui et al., 2020; Bohren et al., 2023; Gallen and Wasserman, 2023; Chan, 2024)[10], as well as by testing theoretical predictions on secondary data (Altonji and Pierret, 2001; Knowles et al., 2001; Charles and Guryan, 2008). We contribute to this literature by focusing on a market where some salient interactions exist between employees and final customers, and customer demand drives profit maximization. We propose a simple theoretical framework that nests not only employer taste-based and statistical discrimination but also customer racial animus, and delivers testable predictions for each source of unequal treatment.

This paper is also related to the literature comparing outcomes between groups to detect the presence of taste-based discrimination. The overarching problem at the heart of average comparisons (or average-based outcome tests; Becker 1957) is that of *infra-marginality*, i.e., in the racial setting, differences in averages might mask both unequal treatment for candidates that are identical but for their race, as well as racial differences in the distributions of unobserved characteristics. Canay et al. (2023) present an extensive discussion on the conditions required for such tests to be valid. The existing literature has dealt with this problem via either random assignments of candidates to decision makers (Arnold et al., 2022); exploiting the timing of release decisions made by parole boards (Anwar and Fang, 2015); specifying equilibrium models (Knowles et al., 2001); or adding distributional assumptions (Simoiu et al., 2017; Pierson et al., 2018; Pierson, 2020). We contribute to the third stream by proposing a parametric approach that is suitable for describing a relatively broad

[9] For a review of the literature on the topic, see Guryan and Charles (2013) and Lippens et al. (2020).

[10] Chan (2024)'s field evidence and framework are particularly broad as they expand the focus beyond taste-based and statistical discrimination to include behavioral mechanisms such as biased beliefs and deniable prejudice.

class of screening problems and only relies on the first and second moments of the observed outcome distribution (in our case, box office revenue) for identification. Although we do rely on distributional assumptions to separately identify the different sources of discrimination, we argue that the identification results extend to a set of alternative parameterizations with which empirical researchers might feel comfortable in a variety of settings.

In using higher order moments of the outcome distribution, our test has a similar flavor to those recently proposed by Bharadwaj et al. (2024) and Benson et al. (2024). Bharadwaj et al.'s test is based on the comparison (in the sense of first-order stochastic dominance) between the entire wage distributions of different groups under the implicit assumption that all workers are employed, and there is no screening of workers based on expected productivity. While our approach nests within Bharadwaj et al.'s insight that studying a non-binary outcome (over a binary outcome, e.g., callback) adds margin to separately identify different sources of discrimination, our model explicitly considers the effect of different forms of discrimination on the (continuous) distribution of outcomes *conditional* on production. Moving away from exploiting the entire outcome distributions, in parallel work developed independently, Benson et al. propose a model of racial bias in hiring that nests taste-based discrimination, screening discrimination, and complementary production. They achieve separate identification through testable implications that rely on the mean and variance of workers' productivity under managers of different pairs of races, which they test within the retail context. While the modeling and identification approaches in the two papers are similar, our work departs from Benson et al. in that we do not require the race of the decision maker to be observable. We argue that this is an important contribution to study discrimination in contexts where decisions are likely made by groups rather than single individuals (e.g., admission committees, parole boards, lending organizations, grant review panels); or the decision maker in charge might be influenced by other layers of the organization or industry actors (e.g., media and artistic production, health care treatment approvals, charging decisions, regulatory or legal compliance decisions); or, as is probably quite common, the identity of the decision maker is not observed.

Through its empirical application, the paper also adds to the line of research that documents the presence of racial discrimination in the motion picture industry (Weaver, 2011; Fowdur et al., 2012). Closest to our work is the paper by Kuppuswamy and Younkin (2020), who find that movies with multiple African-American actors enjoy a box office premium. They rule out customer racial tastes as a discrimination mechanism through an experimental approach. We confirm their conclusion in a more comprehensive data set and provide an analytical framework that can be used to interpret racial differences in the mean and variance of the observed revenue distributions as a function of different forms of discrimination. Our application is also related to a broad empirical literature on labor market and recruiting discrimination, which among the most recent contributions include Åslund et al. (2014); Dustmann et al. (2016); Hedegaard and Tyran (2018); Kline et al. (2022)[11], as well as customer discrimination more broadly (Neumark et al. 1996; Bar and Zussman 2017; Combes et al. 2016; Leonard et al. 2010 in traditional labor market and service settings; Kahn and Sherer 1988; Nardinelli and Simon 1990; Stone and Warren 1999; Burdekin and Idson 1991 in sport contexts.)

The rest of the paper proceeds as follows. Section 2 describes the institutional background of the motion picture industry. Section 3 presents our theoretical model and discusses its empirical implications. Section 4 describes the data and the process used to classify performers by race. Section 5 presents the main empirical findings and assesses the robustness of the results to different definitions of race or dependent variables. Section 6 presents suggestive evidence of incorrect beliefs on the revenue potential of non-white movies within the industry. Section 7 discusses and concludes.

2 Institutional background of film production

Filmmaking is a complex industry that involves a multiplicity of skills, targets, and decision makers. Each movie displayed on the screen has been through three articulated macro-phases: script writing, production, and distribution. This paper studies racial discrimination at the production stage.

[11]See Benson et al. (2024) for a more comprehensive list.

A key decision maker in the production phase is the producer.[12] They decide whether a script is worth being turned into a movie and, if so, raise the money (sometimes supported by one or more executive producers.) The producer is then responsible for the financial and logistic aspects of the movie.[13] The producer oversees the hiring of the director, who is the creative soul of the movie, the cast, and the crew, and decides on the budget allocation.[14]

In our conceptual framework, we assume that the movie script itself determines the racial composition of the leading characters in a movie. Although the producer and the casting team[15] may have some latitude in choosing the supporting characters, we think it is plausible that the race of the main characters can be inferred directly from the script. In fact, casting notices for actors typically specify features such as race and ethnicity (and other aspects of physical appearance) for specific roles.

Our model, presented below, describes the producer's decision about whether to produce the movie after they have seen the movie's script and observed the racial composition of the cast and a signal of the movie's quality.

3 A model of the screening process

We present here a theoretical framework that helps us understand how observed box-office revenue can inform us about the extent and nature of discrimination in the industry. We assume that the movie production process has the following timeline.

Step 1: Script arrival

[12]See for reference Crimson Engine (2018).

[13]The Producers Guild of America (P.G.A.) has established that the producer's name in the film credits can be followed by the *p.g.a.* certification mark only if the producer has performed a significant portion of the producing duties, which includes being physically present on set for a substantial fraction of the production time (P.G.A., n.d.).

[14]In describing our model in Section 3, we will therefore refer to the decision maker as the "producer." It is likely more accurate, however, to think of the decision as made by several agents along a more complex chain of command, as illustrated in the following quote by popular American filmmaker Ed Zwick (Zwick, 2024): "When the creative executive says ''we're going to make this movie", it means she'll try to get the VP to read it. When the VP says he'll make it, it means he's read positive coverage. When the EVP says it, it means she'll take credit for finding it if the president of production likes it. When the president of production says it, it means he needs to tell the CEO which actor is starring in it. And at last, when the CEO says we're going to make this movie, it means it'll get made if he still has his job in six months."

[15]While the producer can be correctly thought of as the primary decision maker in the production process, casting decisions are typically shared among multiple roles.

There are two types of movies: white movies, denoted by w, and non-white movies, denoted by b. A risk-averse producer wishes to maximize *log* revenue, denoted by π. The producer receives a script and perfectly observes its type t. However, box office revenues are not observed. We assume that ex-ante box-office revenues of a movie of type t follow a log-normal distribution[16] with type-specific parameters μ_t and $\sigma^2_{\pi t}$:

$$\pi \mid t \sim N(\mu_t, \sigma^2_{\pi t}), \qquad \forall t \in \{w, b\}$$

.

Step 2: Signal and prior updating

Based on the script, the producer updates her prior about the movie's success. Formally, we can think of the producer observing a signal (y) of the movie's box-office revenue. The signal is normally distributed and is well-calibrated, meaning that in expectation, it is equal to the movie's actual (log) box-office revenue, but it is noisy. Critically, we assume that the precision of the signal may differ by movie racial type. Therefore:

$$y \mid \pi, t \sim N(\pi_t, \sigma^2_{yt}).$$

Given this setup, it is straightforward to calculate the posterior mean of log box-office revenue, conditional on the signal and the movie's type:

$$E(\pi \mid y, t) = \frac{\sigma^2_{\pi t}}{\sigma^2_{\pi t} + \sigma^2_{yt}} y + \frac{\sigma^2_{yt}}{\sigma^2_{\pi t} + \sigma^2_{yt}} \mu_t. \tag{1}$$

Step 3: Production decision

Producers produce a movie and release it to the public if the expected log box-office revenue, conditional on the movie's type and signal, exceeds a given threshold. This threshold (the *revenue threshold*) is exogenously given. We can think of it as the reservation revenue from a sequential search model, i.e., the value of the revenue that makes the producer

[16]The log-normal assumption is made for analytical convenience. In Appendix B, we explore alternative distributional assumptions. Most of our results are not sensitive to the specific distributional assumptions. Later in the paper, we highlight which results depend on the log-normal distribution.

indifferent between producing the movie or waiting for a better script.[17] We denote this revenue threshold π_{0t}, making the critical assumption that the threshold is type-specific. For example, this could result from the producer having a taste for producing movies of a given type.

The movie is produced if

$$E(\pi|y,t) > \pi_{0t}, \tag{2}$$

This is equivalent to saying that the movie is produced only if the signal y exceeds a given threshold (the *signal threshold*). Based on equation (1) and condition (2), it is easy to show that the signal threshold is

$$\bar{y}_t = \pi_{0t} + (\pi_{0t} - \mu_t)\frac{\sigma^2_{yt}}{\sigma^2_{\pi t}}. \tag{3}$$

In other words, the signal threshold is type-specific and depends on the revenue threshold, the parameters of the prior distribution, and the precision of the signal.

This threshold, together with the statistical features of the ex-ante distribution of box-office revenue and the distribution of revenue conditional on the signal, determines the ex-post distribution of box-office revenue. The following proposition establishes the comparative statics of the signal threshold with respect to the parameters of the model.

Proposition 1 *The following comparative statics results hold:*

(a) $\bar{y}_t$ *decreases in* μ_t.

(b) $\bar{y}_t$ *increases in* π_{0t}.

(c) If $\pi_{0t} > \mu_t$, $\bar{y}_t$ *increases in* σ^2_{yt}.

(d) If $\pi_{0t} < \mu_t$, $\bar{y}_t$ *decreases in* σ^2_{yt}.

[17] We think of the race-specific revenue threshold as capturing the disutility cost associated with producing a movie of a given racial type. In our model, the producer does not explicitly internalize the production cost. See Section 5.4 for a more extensive discussion of this assumption.

Proof. See Appendix A ■

The first two statements in Proposition 1 are straightforward and intuitive. If the ex-ante expected (log) revenue is higher (a high μ_t), the movie is produced even if the signal is not very good. Similarly, when the revenue threshold (π_{0t}) is high, the signal must be excellent to produce the movie. The third and fourth items in the Proposition are more involved but are familiar from the literature on statistical discrimination (Aigner and Cain, 1977; Lundberg and Startz, 1983; Neumark, 2012). Intuitively, if the signal is less precise (a high value of σ^2_{yt},) and the producer wants to produce only high-revenue movies, she will have to set a high signal threshold to make sure she only picks the right tail of the revenue distribution (item (c) in Proposition 1); on the other hand, if the producer only wants to cull out very low revenue movies and the signal is uninformative, the threshold must be set at a low value to ensure that only the very worst (i.e., lowest-revenue) movies are weeded out (item (d) in the proposition).[18]

3.1 Predictions for empirical work

Proposition 1 characterizes the properties of the signal threshold that determines whether a movie is produced. In practice, we do not observe the signal threshold, so the results are not useful for empirical analysis. However, we do observe the box office revenue of movies that are actually produced and released to the public. The mean and variance of log box-office revenue, conditional on production, are:[19]

[18]In Appendix B, we explore two departures from the normal-normal model. First, we consider a case where producers care only about the binary outcome "whether a movie is a hit" and decide to produce the script only if the posterior probability exceeds a certain threshold (we dub this the Beta-Binomial model). The comparative statics for the signal threshold in this model match exactly those of the normal-normal model, and so do the testable predictions. Second, we consider the case where the prior distribution of revenue is Pareto rather than log-normal (the Pareto model). The comparative statics for the signal threshold in the Pareto model match exactly those of the normal-normal model for the cases of customer and taste-based discrimination. The predictions are somewhat different, as in the Pareto model a) under taste-based discrimination, both the mean and the variance of log revenue conditional on production are predicted to be higher for non-white movies; and b) under statistical discrimination, both the mean and the variance appear to have a U-shaped relationship with the noise of the signal. Importantly, in the Pareto model, none of the three forms of discrimination can match the observed patterns that the mean log revenue is *higher* for non-white movies, while the variance of log-revenue is *lower* for non-white movies (see Section 5.5).

[19]Rosenbaum (1961).

$$E(\pi \mid y > \bar{y}_t) = \mu_t + \sigma \frac{\phi(\frac{\pi_0 - \mu_t}{\sigma})}{1 - \Phi(\frac{\pi_0 - \mu_t}{\sigma})} \tag{4}$$

$$Var(\pi \mid y > \bar{y}_t) = \sigma^2 \left(1 + \sigma_{yt}^2 + \lambda(\frac{\pi_0 - \mu_t}{\sigma}) \left(\frac{\pi_0 - \mu_t}{\sigma} - \lambda(\frac{\pi_0 - \mu_t}{\sigma})\right)\right), \tag{5}$$

where $\sigma = \frac{\sigma_{\pi t}^2}{\sqrt{\sigma_{\pi t}^2 + \sigma_{yt}^2}}$ and $\lambda(x) = \frac{\phi(x)}{(1-\Phi(x))}$.

We can then formulate our central proposition, which enables us to predict how different types of discrimination affect box-office revenues of white and non-white movies produced.

Proposition 2 *Let $E_t \equiv E(\pi|y > \bar{y}_t)$ and $Var_t \equiv Var(\pi|y > \bar{y}_t)$ be the mean and variance of log box-office revenue conditional on production, as defined in equations (4) and (5). Then, the following comparative statics results hold:*

(a) E_t and Var_t increase in μ_t.

(b) E_t increases in π_0, Var_t decreases in π_0.

(c) E_t decreases in σ_{yt}^2, Var_t increases in σ_{yt}^2.

Proof. See Appendix A ■

We first focus on the intuition behind the comparative statics of E_t with respect to the parameters. The intuition for the first two results is straightforward: Expected revenue conditional on production is higher, the more to the right lies the prior distribution of revenue (result (a)), and the higher is the revenue threshold (result (b)). The third result implies that expected revenue conditional on production increases with signal precision. This result may seem counter-intuitive, as the signal threshold can either increase or decrease with σ_{yt}^2 (Proposition 1, results from (c) and (d)). To gain intuition, it is useful to consider the extreme cases of a perfectly informative ($\sigma_{yt}^2 = 0$) vis-à-vis perfectly uninformative signal ($\sigma_{yt}^2 \to \infty$). If the signal is perfectly informative, the movie is produced only if the signal (which is exactly equal to box-office revenue) is above the revenue threshold. This implies that expected revenue conditional on production is strictly greater than μ_t because some movies will be below the threshold and are not produced. On the other hand, if the signal is perfectly uninformative, whether a movie exceeds the signal threshold conveys no information about its revenue – the expected revenue conditional on production is, therefore, μ_t.

For the variance results, it is helpful to consider the case of a perfectly informative signal. The distribution of revenue conditional on production is a truncated normal distribution, with the truncation point equal to the revenue threshold π_0. If the whole distribution is shifted to the right and the threshold remains the same, it is easy to see that the variance also increases (result (a)). If the revenue threshold π_0 increases, the truncation point shifts to the right and the distribution variance decreases (result (b)). As for the third result, it is again useful to consider the two polar cases of a perfectly informative vs. a perfectly uninformative signal: With a perfectly informative signal, the distribution of box-office revenue is a truncated normal distribution, which necessarily has a smaller variance than the untruncated distribution that results from a perfectly uninformative signal.

We can now use Proposition 2 to characterize the mean and variance of observed box-office revenues for white and non-white movies under different types of discrimination.

Case 1: Customer discrimination. Customer discrimination implies that the viewing public has a preference for white movies over non-white ones. In terms of our model, this means that the entire distribution of log box-office revenue for white movies is shifted to the right relative to the distribution for non-white movies, or $\mu_b < \mu_w$.

Then, by result 1, it follows that $E_b < E_w$, and $V_b < V_w$. We can, therefore, state the following prediction:

Prediction 1 *Under customer discrimination, the mean log box-office revenue for non-white movies is lower than for white movies, and the variance of log box-office revenue for non-white movies is lower than for white movies.*

Case 2: Taste-based discrimination. We can think of taste-based discrimination as the producer suffering a utility loss from producing non-white movies. Holding everything else constant, the producer will produce a non-white movie only if the expected log revenue exceeds a higher threshold than the one she sets for white movies to compensate her for the disutility of producing a non-white movie. In this case, $\pi_{0b} > \pi_{0w}$. By result 2, we have that $E_b > E_w$ and $V_b < V_w$. We can, therefore, state Prediction 2:

Prediction 2 *Under taste-based discrimination, the mean log box-office revenue for non-white movies is higher than for white movies. The variance of log box-office revenue for non-white movies is lower than for white movies.*

Case 3: Statistical discrimination. We classify under statistical discrimination the case where the informativeness of the signal for non-white movies is smaller than the one for white movies. We believe this assumption is plausible as historically there have been fewer movies with non-white characters, and the (mostly white) producers may find it more difficult to evaluate how successful a movie with non-white characters will be. In this case, $\sigma^2_{yb} > \sigma^2_{yw}$. By result 3, we have that $E_b < E_w$ and $V_b > V_w$. We can, therefore, state prediction 3:

Prediction 3 *Under statistical discrimination, the mean log box-office revenue for non-white movies is lower than for white movies, and the variance of log box-office revenue for non-white movies is higher than for white movies.*

Table 1 summarizes our model predictions. In the remainder of the paper, we use the above predictions to assess the extent and nature of discrimination in the motion picture industry.[20]

4 Data

4.1 Facial classification

A key ingredient of our paper is creating a data set with racial identifiers for movie casts. Some recent papers have used machine learning tools to classify images based on skin tone (Adukia et al., 2023; Colella, 2021). We note that these methods are only partially adequate for our purposes. First, we are interested in classifying images of all non-white actors,

[20]Throughout, we have assumed that a movie's script is not race-neutral. In fact, our empirical results are driven by genres in which the assumption of non-race-neutral scripts is more likely to hold (see Table 7). If scripts are race-neutral, the producer may decide both whether to produce the movie and the racial composition of the cast. However, if hiring a non-white cast is (at least on average) cheaper than hiring a white cast (Appendix Figure D.2), then a mere comparison of our race-specific revenue thresholds will likely *understate* the extent of taste-based discrimination in the market.

including those of Asian, Native American, and other ethnicities that are hard to classify based on skin tone alone. Second, even the most accurate machine-learning algorithm will yield some error rate and, most importantly, may not be able to fully capture all the shades of human perceptions, which is likely the most important dimension for classification in an entertainment context. Therefore, we relied on a team of ten human raters to assign racial identifiers to more than 7000 performer images downloaded from the popular website IMDB.[21]

Each rater was assigned 8 blocks of about 800 performers[22] and was asked to assess whether they thought the person in the image was *White/Caucasian*, *Black/African-American*, *Hispanic*, *Asian*, *Native American/Pacific Islander*, *South Asian*, or *Other*. The option *Unable to Tell* was also made available to the respondents. Raters were specifically instructed not to consult the internet for any information about the performer and to classify the image based on their perception alone. This procedure resulted in between 8 and 9 human ratings for each of the performers in our data set. We assigned to each image the modal classification as long as no more than two raters disagreed on that image's classification.[23] This allowed us to classify about 92% of the performers in our sample as either White (79.8%), Black (9.1%) or Asian (2.6%). For the remaining performers in the sample, we used the machine learning algorithm proposed by Anwar and Islam (2017)[24], described in more detail in Appendix C.

4.2 Additional variables

Our analysis is based on a sample of more than 7000 motion pictures released in the United States between 1997 and 2017. We obtain this information from Opus Data,[25] a private company that collects information on the industry, and rely on IMDb for the approximately 5% of observations in our sample for which OPUS revenues are unavailable. We gather aggregate financial data (box office revenue, production budget, opening weekend revenue,

[21] www.imdb.com

[22] One rater completed only four blocks.

[23] We grouped together the White/Caucasian and Hispanic categories, as we realized that it was difficult to accurately distinguish between the two. None of the substantive results in the paper are meaningfully affected if we do not impose this grouping.

[24] Link: https://arxiv.org/ftp/arxiv/papers/1709/1709.07429.pdf.

[25] www.opusdata.com

etc.) and metadata (genre, production method) for all movies in our sample.

The main variables of interest in our data set include the gross domestic box-office revenue,[26] production costs,[27] movie run time, Metacritic score, release date, MPAA rating, number of theaters in which the movie was released, and number of weeks in which the movie was in theaters. We also collect information on the gender and age of the four top-billed performers. We create a variable called "star power," equal to the cumulative box-office revenue of all movies in which each performer appeared up to the release date of the current movie.

4.3 Summary statistics

Summary statistics are shown in Table 2. The top panel shows that about 12 percent of the top-billed performers in our sample are non-white. About three-quarters of the movies have zero non-white performers, and about 18 percent have only one non-white performer. Our baseline analysis defines a movie as non-white if at least two of the four top-billed performers are non-white. Based on this definition, about eight percent of the movies in our sample are non-white. We also assess the robustness of the results to different definitions of non-white movies.

As for the other variables, the distribution of box office revenue is heavily skewed to the right. Therefore, we use its logarithm as the main dependent variable in our baseline analysis. We collapse the "niche" genres into broader categories so that all movies fall into one of five broad genres. For some of the variables, we only have incomplete data: For

[26]Our baseline definition of a movie's revenue includes domestic box-office revenues and excludes international box-office sales as well as DVD and Blu-ray revenues. While the information for revenues other than from domestic theaters is available in OPUS, it is not in IMDB, which is the data source we use for revenues whenever the information in OPUS is missing. Reassuringly, we note that, as we restrict the analysis to non-missing OPUS data, the progressive inclusion of DVD, Blu-ray, and international sales does not qualitatively alter our main results. The results are available upon request. Our definition of box-office revenues also excludes streaming revenues, which are instead not available in our data. In 2021, the digital market (which includes video streaming) accounted for 72% of the industry revenue composition, with online video subscription becoming the second largest subscription revenue market as a result of a 26% surge (Motion Picture Association, 2022). While we cannot directly test whether non-white movies account for a similar share of revenues across the streaming vs. non-streaming sectors, we note that our main result is robust (and even larger in magnitude; see Table 7) as we restrict our sample to the years before 2007, when streaming accounted for a negligible share of spending on entertainment (see Appendix Figure D.1). Also, no geographic breakdown of revenues is available in our data.

[27]All monetary values are expressed in 2005 dollars.

example, production costs are available for only about 56 percent of the sample,[28] while the Metacritic score is available only for 71 percent of the sample. To maximize sample size, in the empirical analysis, we replace missing values with zeros and add a dummy variable indicating that the variable is missing if the missing value is not central to the analysis.

5 Results

5.1 Non-parametric analysis

Figure 1 presents a box-whisker plot of box-office revenue by the number of non-white performers in the movie (out of the four top-billed actors.) The mean box-office revenue increases markedly with the number of non-white performers, while the dispersion of the distribution decreases. Also, the 25$^{\text{th}}$ percentile (and lower adjacent value) visibly increases with the number of non-white performers. On the other hand, the 75$^{\text{th}}$ percentile is quite stable across cast racial compositions, and the upper adjacent value reduces slightly. We interpret these patterns as the left tail of the non-white movie distribution being missing, which is consistent with the notion that non-white movies are held to a higher standard for production.

Of course, this analysis does not take into account other observable differences that may exist between white and non-white movies. In the following sections, we assess whether the non-white premium in box-office revenue is robust to the inclusion of a broad set of other movie and cast characteristics.

5.2 OLS regressions

The main regression model is the following:

$$\ln y_{it} = \beta_0 + \beta_1 Nonwhite_{it} + \beta_2 X_{it} + \delta_t + \varepsilon_{it}, \tag{6}$$

where y_{it} denotes domestic box-office revenue, in 2005 U.S. dollars, of movie i released in year t; $Nonwhite_{it}$, the key explanatory variable of interest, is a dummy variable indicating whether at least two of the four top-billed performers are non-white; X_{it} is a vector

[28] Probit and logit regressions suggest that movies with higher revenue have a significantly (in the statistical sense) higher probability of non-missing cost information, while the conditional difference between white and non-white movies is not statistically distinguishable from zero. The main result is robust to restricting the sample to observations with non-missing cost information: see columns 5 and 6 in Table 3.

of additional control variables, including both cast (average age, gender composition, the "star power" variable described previously) and movie (production budget, MPAA rating, Metacritic score, run time, genre dummies) characteristics; δ_t is a year-of-release fixed effect, and ε_{it} is the robust standard error clustered by distributor.[29]

The results are presented in Table 3. The first column of the table shows the unadjusted difference in mean log revenue between white and non-white movies without any controls. The mean box-office revenue of non-white movies is almost 2.5 times as high as that of white movies ($\exp(0.910) \approx 2.5$). In column 2, we include controls for other characteristics of the cast (average age, gender composition, and star power), and the coefficient remains almost unchanged. In column 3, we add controls for the production budget, a dummy for whether the production budget is missing, and all other movie characteristics, including genre and year-of-release fixed effects. The coefficient on the non-white indicator drops to 0.433, implying that non-white movies earn about 54 percent more than white movies at the box office. In column 4, we further add controls for the distributor-level fixed effects, and the coefficient drops to 0.336 (40% revenue gap) while remaining highly significant. For both column 5 and column 6, we restrict the analysis to movies with non-missing data on production costs. In column 5, we replicate column 2, and the coefficient drops from 0.926 to 0.488, which explains what drives the decrease of the coefficient from column 2 to column 3.[30] Finally, column 6 replicates column 3, and the results in this restricted sample are mostly unchanged – the coefficient on the non-white indicator rises to 0.522, implying that non-white movies earn on average about 69 percent more than white movies.[31]

These initial results on the differences between white and non-white movies are not con-

[29] We collapse all distributors with only one movie in our data set into one single distributor category (Other/Unknown).

[30] Conditional on being available in our data, the Metacritic score does not differ significantly on average across white and non-white movies, and a Kolmogorov-Smirnov test fails to reject that the distributions are the same. The Metacritic score is missing for 30% of white movies and 18% of non-white movies in our sample. The correlation coefficient between log revenues and the Metacritic score is equal to .11 and statistically significant at the 1% level.

[31] We have data on the script languages for approximately 70% of the working sample. Within this sample, approximately 86% percent of the movies in our sample have English as the only language on file, and this is a subset of the 95% that have English among the languages to which they are associated. The main result is robust to restricting the sample to English-language movies, indicating that our findings are not driven by foreign-language movies.

sistent with either a model of customer discrimination, where audiences prefer white movies to non-white movies nor a model of statistical discrimination, where the signal conveyed by non-white scripts is less informative about future box-office revenue. Both models predict that white movies should have, on average, higher box-office revenue than non-white movies, in contrast to our findings. Instead, the results are consistent with a model of taste-based discrimination, where non-white movies are held to a higher standard, i.e., they are only produced if the revenue exceeds a higher threshold than the one required of white movies. In what follows, we look at how other features of the distribution differ between white and non-white movies.

5.3 Quantile regressions

The model described in Section 3 derives predictions for not only the mean but also the variance of box-office revenues. In this subsection, we analyze other measures of dispersion, namely the white-nonwhite gap at different percentiles of the revenue distribution. Specifically, we estimate a series of quantile regressions of the following type:

$$Q_\tau(\ln y_{it}|Nonwhite, X) = \gamma_{0\tau} + \gamma_{1\tau} Nonwhite_{it} + \gamma_{2\tau} X_{it} + \delta_t,$$

where $Q_\tau(\ln y_{it}|Nonwhite, X)$ denotes the τth conditional quantile of the distribution of log box-office revenue, and $\tau \in \{0.05, 0.10, ...0.95\}$. The main coefficients of interest are the $\gamma_{1\tau}$'s, which measure the gap in conditional quantiles across the white and non-white box-office revenue distributions..

Figure 2 plots the quantile regression coefficients against the quantiles. As was already apparent from the box-whisker plots in Figure 1, from the 20$^{\text{th}}$ quantile onwards there is a clear downward trend in the quantile coefficients: The white-nonwhite gap at the lower quantiles is around 60 log points, while it is only about 20 log points at the upper quantiles. This finding reinforces the interpretation that there is "missing mass" in the left-tail of the non-white revenue distribution, or in other words, that non-white movies at the low end of the distribution of box-office revenue are not produced, while comparable white movies are.

5.4 Robustness

We next investigate the robustness of our results to different definitions of movie type and different dependent variables.

Classification of non-white movies. In Table 4, we consider additional definitions of "non-white" movies. The first column in the table reproduces the results using our baseline classification of non-white movies as those in which at least two of the four top-billed performers are non-white. The first row in the table shows the OLS. results from Table 3, while the remaining rows present the quantile regression coefficients at selected quantiles. All specifications include the full set of control variables.

In column 2, we change the definition of non-white movies to include all movies in which at least *one* of the four top-billed performers is non-white. We view this as a noisier indicator of the movie type, as a non-white actor may be cast in a supporting role in a movie that is mainly about white characters and storylines (a form of *tokenism*). Using this definition, the OLS coefficient is substantially reduced (about 23 log points) but still large and highly statistically significant. The pattern of quantile regression coefficients is also clearly downward sloping, with the gap going from about 26 log points at the 25^{th} to about 19 log points at the 90^{th} percentile. In column 3, we replace the dummy indicator for non-white movies with the share of non-whites among the four top-billed performers. The results are quantitatively and qualitatively similar to those of the baseline specification. Finally, in column 4, we classify a movie as non-white only if the top-billed performer is non-white. According to this definition, the average white-nonwhite premium is slightly smaller than in the baseline (46 log points), and the pattern of the quantile regression coefficients is also downward sloping.[32]

On the whole, Table 4 shows that the main conclusions regarding the white-nonwhite premium and the nature of discrimination in the industry are not sensitive to the exact definition of non-white movies.

Choice of the dependent variable. In all the analyses so far, we have looked at the logarithm of box-office revenue as the primary dependent variable of interest. The main

[32] Our main result is robust to restricting the sample to movies with cast popularity (see section 4 for a definition of "star power") below the median, ruling out that the non-white premium that we find is driven by "superstar" non-white movies exclusively.

reason for this choice is that box-office revenue is readily available for almost all movies, and it has been traditionally used as the primary metric for assessing the commercial success of a movie. However, producers also consider the expected cost of a movie when making production decisions. While we have addressed this in part by including production costs as an explanatory variable in Table 3, one may also want to work with profits directly. In the Opus data set, we observe a movie's production budget for about 56% of all movies so that we can calculate various measures of profit.[33] We report the results of this analysis in Table 5. The sample includes only those movies for which we observe the production budget. All specifications include the full set of control variables.

In column 1, we use the logarithm of the gross profit margin as a dependent variable, defined as the ratio of domestic box-office revenue to the production budget. The results are broadly consistent with those in the previous sections: Non-white movies have on average a substantially higher profit margin, and the white-nonwhite gap becomes smaller as we move from the low to the high end of the distribution.

In column 2, we focus on the total profit, calculated simply as the difference between box-office revenue and the production budget. It is still the case that the average non-white movie earns a higher profit than the average white movie (by about $8.7 million), holding other characteristics fixed. However, we no longer observe a clear declining pattern in the white-nonwhite gap as we move from lower to upper quantiles in the profit distribution. In fact, the gap appears to be fairly stable (at least in the statistical sense) at all quantiles of the distribution. This could be partly due to the shape of the profit distribution, which tends to be quite right-skewed. We confirm this in column 3, where we use the level of box-office revenue (rather than the logarithm) as the dependent variable. We find a positive premium favoring white movies, but now the pattern of quantile regression coefficients shows that the gap becomes larger as we move from the low to the high end of the distribution. We note that, given the substantial right skewness in the revenue distribution, the predictions

[33] Our measure of profits should only be viewed as a coarse estimate. First, the production budget does not represent the entirety of a movie's production costs, which typically also include marketing costs. Marketing costs are rarely disclosed. Second, the producer typically does not collect all of the box-office revenue, as theaters also receive a cut depending on bilateral negotiations as well as other factors such as the length of time that the movie has been in theaters. Third, cost sharing – a common practice in the movie industry (Weinstein, 1998) – is likely to reduce the extent to which producers internalize costs in their decision making.

regarding the variance of box-office revenues *in levels* conditional on production derived from a model that assumes normal distributions no longer hold necessarily.

5.5 The white-nonwhite gap in residual variance

An alternative approach to verify our dispersion predictions is to explore how the residual variance differs across white and non-white movies. Borrowing from the heteroskedasticity literature, we posit that the squared residuals from the OLS regression in equation 6 have the form:

$$u_{it}^2 = \exp(Z_{it}'\alpha),$$

where the vector Z_{it} contains a subset of the variables included in the main regression (potentially, all of them); We then estimate regressions of $\ln \hat{u}_{it}^2$ on the racial indicator and additional control variables. The results are reported in Table 6.

In column 1, the residual variance is assumed to depend only on the racial indicator. Consistent with the results of the box-whisker plot and quantile regressions, we find that non-white movies have a substantially lower residual variance than white movies. In columns 2 and 3, we progressively add additional controls to the variance regression. The results are essentially unchanged – the residual variance of non-white movies is lower than that of white movies.

In the remaining three columns, we experiment with different definitions of non-white movies. The coefficients on the racial variable in the residual regressions are somewhat smaller in absolute values, but still highly statistically significant.[34]

Overall, our results are consistent with what our theoretical model defines as taste-based discrimination, i.e., non-white movies being held to a higher standard, which results in a higher mean and lower variance of box-office revenue for the produced non-white movies.[35]

[34] The coefficient on the racial variable remains negative and statistically significant when the outcome variable is the log of the profit margin (column 1 of Table 5,) but it is imprecisely estimated for profits and revenues in levels (columns 2 and 3 of Table 5.) Results are available upon request.

[35] These observed patterns stand in stark contrast to the concept of mean-variance trade-off in the rational asset pricing literature, which traditionally assumes perfectly informed mean-variance utility-maximizing agents (Cochrane, 2005).

5.6 Heterogeneity Analysis

In Table 7, we explore the heterogeneity of our results along a number of different dimensions. First, we look at whether our results are driven by movies produced and distributed by specific segments of the industry. One concern is that our results may capture differences between movies produced by the major studios (the so-called "Big-Six")[36] vs. those produced by smaller studios. It could be that the smaller box-office revenue of white movies reflects the fact that these are often produced by small independent studios, while non-white movies are passed over by these studios altogether. Columns 1 and 2 of the table, however, show that this is not the case: the non-white revenue premium is present among movies distributed by both types of studios.

We next look at differences across genres (columns 3-5 of the table). The non-white premium is more pronounced among comedies and dramas, where the script is more likely to convey information about the racial composition of the cast. By contrast, the non-white premium is small and not statistically significant in action/adventure movies.

Columns 6 and 7 examine heterogeneity by time period. We look separately at movies produced before and after 2007, the median year in our sample. If taste-based discrimination declines over time, either because of a change in attitudes or because of a change in the competitive landscape, we would expect the non-white premium to shrink. There is some evidence in support of these hypotheses: The non-white premium is 57 log points in the pre-2007 period, but falls to 26 log points in the the post-2008 period.

Finally, in columns 8 and 9, we look at whether the results differ by the gender composition of the cast. We define "female" movies as those in which (strictly) more than 50% of the leading actors are women. The non-white premium is considerably larger among female movies, suggesting that non-white movies must pass an even higher threshold if the cast is predominantly female.

[36]These Big-Six studios are: Warner Bros., Paramount Pictures, Walt Disney, Sony / Columbia Pictures, Universal Studios, and 20th Century Fox. These six studios accounted for almost 90% of the US/Canadian market as of 2007. In 2020, Disney acquired 20$^{\text{th}}$ Century Fox, and the group is now commonly referred to as the "Big Five". The Big-Six control is included in our regression analysis as a control.

5.7 Producer analysis

In this section, we explore the role of the producer's race in explaining the non-white box-office premium. Neither IMDB nor OpusData contains demographic information on movie producers. We, therefore, rely on a human rater to code the producers' race for a subsample of our films. The first step is matching films to producer names, which are available in the *Credits* section of the Opus data set. We have information on producers for 3,878 out of the 6,943 movies in our sample: 9,842 distinct names are associated with those movies in the capacity of *Producer* or *Executive Producer.* Of these producers, fewer than 1% can be racially categorized via Wikipedia. For the remainder, we then randomly drew approximately 8% of the remaining producers associated with either white or non-white movies[37] and asked a human rater to racially classify these producers based on photos and text resources available online. We then matched the producer's racial information to our main data set. We end up with a working sample of 1,955 movies with racial information for at least one producer. Of these, 261 (13%) display more than two non-white actors, while 403 (21% – 257 of the white movies and 146 of the non-white movies) are associated with at least one non-white producer.

Table 8 shows the results of our producer analysis. Column 1 reports the estimated non-white premium in the sub-sample of interest. The coefficient is positive and statistically significant like the one obtained in the full sample (Table 3, column 3) but approximately half in size. Adding the producer's race to the controls (Column 2) does not change the coefficient of interest in any significant way, and the producer control itself is statistically insignificant.[38]

Column 3 reports the results obtained from interacting the racial indicator for the cast with the racial indicator for the producer. Our findings reveal that the non-white revenue

[37] In our sample, the average number of producers and executive producers (and co-producers) associated with a film is 8 (9), and the median is 7 (8). Therefore, to guarantee a large enough working sample for our producer analysis, we randomize at the producer level and not at the movie level. This implies that in our exercise, we are comparing movies with at least one non-white producer to movies that may or may not have any non-white producers. We stratified our randomization by the movie racial type to end up with a reasonably balanced data set.

[38] The findings are robust to the inclusion of studio fixed effects. Standard errors are clustered by the studio. Results are available upon request.

premium is driven by movies with at least one non-white producer, while on average, films associated with white producers do not display a non-white revenue premium. Taken at face value, these findings suggest that taste-based discrimination may be more concentrated among non-white producers. This evidence should be interpreted with caution, however, given the limited scope of our analysis. This pattern can be rationalized through the observation, discussed in Section 2, that producers may not be the pivotal decision makers in the film production decision and may be held to different standards themselves, depending on their racial group. An alternative interpretation is that non-white producers have a comparative advantage in producing non-white movies, and, in particular, may obtain a more precise signal of revenue when evaluating scripts. In the context of our model, such an informational advantage would indeed translate into higher revenues for non-white movies produced by non-white producers.

6 Alternative explanation: is the industry surprised?

The empirical results so far suggest that non-white movies are held to higher production standards than white movies. A candidate interpretation of these patterns is that producers dislike producing non-white movies and face a disutility cost every time they produce one. As a result, the expected revenue for producing non-white movies needs to be higher than the expected revenue for producing white movies (in the context of the model, $\pi_{0b} > \pi_{0w}$).

An alternative, non-mutually exclusive, interpretation is that the industry systematically underestimates the revenue potential of non-white movies relative to white movies.[39,40] In other words, actual box-office revenue for non-white movies is π_b, but producers perceive it to be $\hat{\pi}_b = \pi_b - e_b$, with $e_b > 0$. This explanation would yield similar predictions to the ones derived from taste-based discrimination, even if the nature of discrimination in the industry is quite different.

[39] The film industry is known for having a hard time forecasting movies' success, as well as analyzing past results: "Why was ''The Hunger Games" such a big hit? Because it had a built-in audience? Because it starred Jennifer Lawrence? Because it was released around spring break? The business is filled with analysts who claim to have predictive powers, but the fact that a vast majority of films fail to break even proves that nobody knows anything for sure" (Davidson, 2012.)

[40] See Chan (2024); Bohren et al. (2023); Esponda et al. (2022); Bordalo et al. (2016); Fong and Luttmer (2011) for evidence of inaccurate beliefs in other contexts.

Our model intrinsically cannot identify taste-based disutility costs and biased beliefs separately. Nevertheless, we can make some progress on this front by exploiting the decision that distributors make on the number of theaters at which the movie is displayed on the opening weekend. We argue that this is a proxy of the market's rational expectation of the movie's potential after production, as distributors' decision-making is less likely to be affected by taste-based or statistical discrimination: While producers "sign" a movie as a creation of theirs and create a permanent bond with the film, studios, and theater owners are more likely to make distribution choices based on purely profit-maximizing considerations once the movie has been produced. Moreover, statistical discrimination should also be of relatively less importance at the distribution stage, because distributors also observe the ex-post quality of the movie rather than just the script.

We conjecture that distributors choose the number of theaters based on expected customer demand. If non-white movies are displayed in fewer theaters than white movies, this indicates that distributors expect relatively smaller revenue from the non-white movies. Therefore, if non-white movies have the same level of customer demand but are displayed in fewer theaters, we conclude that distributors underestimate their revenue potential.

Using data on the number of screens in which a movie is shown, we test the hypothesis that the industry systematically underestimates the revenue potential of non-white movies. Specifically, we first regress first-weekend box office revenues on the number of theaters in which movies are projected over the first weekend upon their release. Following Moretti (2011), we interpret this as a proxy for the industry expectation of a movie's box-office revenue. The residuals from this regression can then be viewed as a measure of the industry's underestimation or overestimation of a movie's revenue potential. If the non-white mean residual is significantly larger than the white mean residual, this suggests that the industry systematically underestimates non-white movies' revenue potential relative to white movies.

We start by running a simple bivariate regression of log first-weekend revenues on the log number of theaters. The R-squared of this regression is 0.89 (column 1 of Table 9), and it remains relatively stable as further controls are added (columns 2 and 3). The number of theaters is, hence, a good predictor of first-weekend revenues.

We then test whether the residuals obtained from the regressions in Table are on average between non-white and white movies. The results are presented in the bottom panel of the table. We find that the mean residual for non-white movies is positive across specifications, while the mean residual for white movies is close to zero. That is, the industry underestimates the first-weekend success of non-white movies relative to white movies. The difference between the white and the non-white residual is always statistically significant. We conclude that our results might be at least partly explained by a systematic underestimation of non-white movies' box-office potential within the industry.

7 Conclusion

This paper presents a framework for detecting the extent and nature of discrimination in contexts in which decision-makers screen applicants. The econometrician can only observe the outcomes of applicants who successfully pass the screening process. The framework nests several leading theories of discrimination and derives a rich set of testable empirical predictions.

We apply these tests in the context of racial representation in the U.S. motion picture industry. We show that non-white movies earn a box-office premium. The gap is particularly pronounced at low quantiles of the distribution, suggesting that non-white movies with low box-office potential are never produced; in other words, non-white movies are held to a higher standard in the production decision. In the context of our model, this evidence is consistent with taste-based discrimination, i.e., producers suffering a utility loss from producing non-white movies. The evidence is also consistent with producers and distributors having inaccurate beliefs and systematically underestimating the revenue potential of non-white movies. On the other hand, the evidence is not consistent with simple customer discrimination against non-white movies, nor with a statistical discrimination story in which the signal sent by non-white movies is less precise.

These results may appear puzzling to the extent that they hint at lost profits and relatively slow learning in the industry for non-white movies' potential. While our results indicate that the non-white revenue premium has more than halved between 1997-2007 and

2008-2017 (and become harder to distinguish from zero in a statistical sense, despite the larger sample size,) the point estimate for the latter period is far from zero in an economically significant way. Some of the *a priori* plausible explanations do not seem applicable to our setting: it is unlikely that learning is hindered by the non-white premium being too small to be detected or consequential, or that too few non-white movies are produced. The fact that – conditional on production – non-white movies are relatively more successful rules out customers' attitudes and pre-market discrimination as leading explanations (Becker, 1957). We argue that other industry-specific forces might be at play, including the high concentration of the motion picture industry, with the "Big Six" studios typically accounting for more than 80% of the industry's total market share. In non-competitive industries, firms may have more latitude to indulge their discriminatory taste. The introduction and growth of streaming services and smartphone applications in the 2010s appear to have increased the amount of competition in the industry (Kuehn and Lampe, 2023), which is consistent with the declining non-white premium in the latter part of our sample. There are also documented challenges and uncertainty that industry actors face in predicting movies' revenues and profitability, even when data are available (Lash and Zhao, 2016).[41] We leave the investigation of this puzzle to future research, along with the analysis of the consequences of the recent diversity-promoting rules that the Academy of Motion Picture Arts and Sciences set for those aspiring to best-picture qualifications (Sperling, 2020a,b).

While the specific application in this paper looked at the motion picture industry, our model can be readily applied to other contexts in which decision makers can use group identifiers to screen applicants, and one can observe the outcome or productivity of successful applicants: the output of workers hired for a particular job, the academic performance of students admitted to a freshman class, or the number of citations accumulated by a published journal article. These other contexts are promising avenues for future research.

[41]For discussions and examples in the public press, see Yahr (2016); Gladwell (2006); Thompson (2013); Snee (2016).

References

Adukia, Anjali, Eble, Alex, Harrison, Elizabeth, Runesha, Haidong B., and Szasz, Tamas (2023). “What We Teach About Race and Gender: Representation in Images and Text of Children’s Books,” *Quarterly Journal of Economics*, 138 (4), 2225–2285.

Agan, Amanda and Starr, Sonja (2018). “Ban the box, criminal records, and racial discrimination: A field experiment,” *The Quarterly Journal of Economics*, 133 (1), 191–235.

Altonji, Joseph G. and Pierret, Charles R. (2001). “Employer learning and statistical discrimination,” *The Quarterly Journal of Economics*, 116 (1), 313–350.

Anwar, Imran and Islam, Nazrul Ul (2017). “Learned features are better for ethnicity classification,” *Cybernetics and Information Technologies*, 17 (3), 152–164.

Anwar, Shamena and Fang, Hanming (2015). “Testing for racial prejudice in the parole board release process: Theory and evidence,” *The Journal of Legal Studies*, 44 (1), 1–37.

Arnold, David, Dobbie, Will, and Hull, Peter (2022). “Measuring racial discrimination in bail decisions,” *American Economic Review*, 112 (9), 2992–3038.

Arrow, Kenneth et al. (1973). “The Theory of Discrimination, Discrimination in Labor Markets,” in Achenfelter, A. and Ress, R. eds. *Discrimination in Labor Markets*, Princeton, New Jersey: Princeton University Press.

Åslund, Olof, Hensvik, Lena, and Skans, Oskar Nordström (2014). “Seeking Similarity: How Immigrants and Natives Manage in the Labor Market,” *Journal of Labor Economics*, 32, 405–441.

Bar, Revital and Zussman, Asaf (2017). “Customer Discrimination: Evidence from Israel,” *Journal of Labor Economics*, 35 (4), 1031–1059.

Becker, Gary S. (1957). *The economics of discrimination*: University of Chicago Press.

Benson, Alan, Board, Simon, and ter Vehn, Moritz Meyer (2024). “Discrimination in hiring: Evidence from retail sales,” *Review of Economic Studies*, 91 (4), 1956–1987.

Bertrand, Marianne and Duflo, Esther (2017). "Field experiments on discrimination," in *Handbook of economic field experiments*, 1, 309–393.

Bharadwaj, Prashant, Deb, Rajat, and Renou, Ludovic (2024). "Statistical discrimination and the distribution of wages," Unpublished manuscript.

Bohren, J. Aislinn, Haggag, Kareem, Imas, Alex, and Pope, Devin G. (2023). "Inaccurate statistical discrimination: An identification problem," *Review of Economics and Statistics*, 1–45.

Bordalo, Pedro, Coffman, Katherine, Gennaioli, Nicola, and Shleifer, Andrei (2016). "Stereotypes," *The Quarterly Journal of Economics*, 131 (4), 1753–1794, 10.1093/qje/qjw029.

Burdekin, Richard CK and Idson, Todd L (1991). "Customer Preferences, Attendance and the Racial Structure of Professional Basketball Teams," *Applied Economics*, 23 (1), 179–186.

Canay, Ivan A., Mogstad, Magne, and Mountjoy, Jack (2023). "On the use of outcome tests for detecting bias in decision making," *The Review of Economic Studies*, Accepted for publication.

Chan, Alex (2024). "Customer Discrimination and Quality Signals."

Charles, Kerwin Kofi and Guryan, Jonathan (2008). "Prejudice and wages: an empirical assessment of Becker's The Economics of Discrimination," *Journal of Political Economy*, 116 (5), 773–809.

Cochrane, John H. (2005). *Asset Pricing*: Princeton University Press.

Colella, Francesco (2021). "Who Benefits from Support? The Heterogeneous Effects of Supporters on Athletes' Performance by Skin Colour," Mimeo., Université de Lausanne.

Combes, Pierre-Philippe, Decreuse, Bruno, Laouenan, Morgane, and Trannoy, Alain (2016). "Customer Discrimination and Employment Outcomes: Theory and Evidence from the French Labor Market," *Journal of Labor Economics*, 34 (1), 107–160.

Crimson Engine (2018). "What does a Producer ACTUALLY do?," https://www.youtube.com/watch?v=71Oh4gQ-1jM, February 13.

Cui, Ruomeng, Li, Jun, and Zhang, Dennis J (2020). "Reducing Discrimination with Reviews in the Sharing Economy: Evidence from Field Experiments on Airbnb," *Management Science*, 66 (3), 1071–1094, 10.1287/mnsc.2018.3253.

Doleac, Jennifer L. and Stein, Luke C. (2013). "The visible hand: Race and online market outcomes," *The Economic Journal*, 123 (572), F469–F492.

Dustmann, Christian, Glitz, Albrecht, and Schönberg, Uta (2016). "Referral-Based Job Search Networks," *Review of Economic Studies*, 83, 514–546.

Esponda, Ignacio, Oprea, Ryan, and Yuksel, Sevgi (2022). "Discrimination Without Reason: Biases in Statistical Discrimination."

Fang, Hanming and Moro, Andrea (2011). "Theories of statistical discrimination and affirmative action: A survey," in *Handbook of social economics*, 1, 133–200.

Fong, Christina M and Luttmer, Erzo FP (2011). "Do Fairness and Race Matter in Generosity? Evidence from a Nationally Representative Charity Experiment," *Journal of Public Economics*, 95 (5-6), 372–394.

Fowdur, Linley, Kadiyali, Vrinda, and Prince, Jeffrey (2012). "Racial bias in expert quality assessment: A study of newspaper movie reviews," *Journal of Economic Behavior & Organization*, 84 (1), 292–307.

Gallen, Yana and Wasserman, Melanie (2023). "Does Information Affect Homophily?" *Journal of Public Economics*, 222, 104876, 10.1016/j.jpubeco.2023.104876.

Gladwell, Malcolm (2006). "The Formula," October 16, https://www.newyorker.com/magazine/2006/10/16/the-formula, Accessed: 2024-11-10.

Guryan, Jonathan and Charles, Kerwin K. (2013). "Taste-based or statistical discrimination: the economics of discrimination returns to its roots," *The Economic Journal*, 123 (572), F417–F432.

Hedegaard, Morten S. and Tyran, Jean-Robert (2018). “The Price of Prejudice,” *American Economic Journal: Applied Economics*, 10, 40–63.

Kahn, Lawrence M and Sherer, Peter D (1988). “Racial Differences in Professional Basketball Players’ Compensation,” *Journal of Labor Economics*, 6 (1), 40–61.

Kline, Patrick, Rose, Evan K., and Walters, Christopher R. (2022). “Systemic Discrimination Among Large US Employers,” *The Quarterly Journal of Economics*, 137 (4), 1963–2036.

Knowles, John, Persico, Nicola, and Todd, Petra (2001). “Racial bias in motor vehicle searches: Theory and evidence,” *Journal of Political Economy*, 109 (1), 203–229.

Kuehn, Jörg and Lampe, Ryan (2023). “Competition and Product Composition: Evidence from Hollywood,” *International Journal of Industrial Organization*, 91, 102981, 10.1016/j.ijindorg.2023.102981.

Kuppuswamy, Venkat and Younkin, Peter (2020). “Testing the Theory of Consumer Discrimination as an Explanation for the Lack of Minority Hiring in Hollywood Films,” *Management Science*, 66 (3), 1227–1247.

Lang, Kevin and Lehmann, Jee-Yeon K. (2012). “Racial discrimination in the labor market: Theory and empirics,” *Journal of Economic Literature*, 50 (4), 959–1006.

Lang, Kevin and Spitzer, Ariella K. (2020). “Race discrimination: An economic perspective,” *Journal of Economic Perspectives*, 34 (2), 68–89.

Lash, Michael T. and Zhao, Kang (2016). “Early Predictions of Movie Success,” *Journal of Management Information Systems*, 33 (3), 874–903, https://www.jstor.org/stable/10.2307/26614064.

Leonard, Jonathan S, Levine, David I, and Giuliano, Laura (2010). “Customer Discrimination,” *The Review of Economics and Statistics*, 92 (3), 670–678.

Lippens, Louis, Baert, Stijn, Ghekiere, Astrid, Verhaeghe, Pieter-Paul, and Derous, Eva (2020). “Is labour market discrimination against ethnic minorities better explained by taste or statistics? A systematic review of the empirical evidence,"Technical report, IZA Discussion Paper No. 13523.

List, John A. (2004). "The nature and extent of discrimination in the marketplace: Evidence from the field," *The Quarterly Journal of Economics*, 119 (1), 49–89.

Moretti, Enrico (2011). "Social learning and peer effects in consumption: Evidence from movie sales," *The Review of Economic Studies*, 78 (1), 356–393.

Motion Picture Association (2022). "2021 THEME Report," https://www.motionpictures.org/wp-content/uploads/2022/03/MPA-2021-THEME-Report-FINAL.pdf, Accessed: 2024-11-10.

Nardinelli, Clark and Simon, Curtis (1990). "Customer Racial Discrimination in the Market for Memorabilia: The Case of Baseball," *The Quarterly Journal of Economics*, 105 (3), 575–595.

Neumark, David, Bank, Roy J, and Van Nort, Kyle D (1996). "Sex Discrimination in Restaurant Hiring: An Audit Study," *The Quarterly Journal of Economics*, 111 (3), 915–941.

Onuchic, Paulo (2022). "Recent contributions to theories of discrimination," *arXiv preprint*.

Phelps, Edmund S (1972). "The Statistical Theory of Racism and Sexism," *The American Economic Review*, 62 (4), 659–661.

Pierson, Emma (2020). "Assessing racial inequality in COVID-19 testing with Bayesian threshold tests," *arXiv preprint*.

Pierson, Emma, Corbett-Davies, Sam, and Goel, Sharad (2018). "Fast threshold tests for detecting discrimination," in *International conference on artificial intelligence and statistics*, 96–105: PMLR.

Riley, Emma (2024). "Role Models in Movies: The Impact of Queen of Katwe on Students' Educational Attainment," *The Review of Economics and Statistics*, 106 (2), 334–351.

Rosenbaum, S. (1961). "Moments of a Truncated Bivariate Normal Distribution," *Journal of the Royal Statistical Society. Series B (Methodological)*, 23 (2), 405–408.

Simoiu, Claudia, Corbett-Davies, Sam, and Goel, Sharad (2017). "The problem of infra-marginality in outcome tests for discrimination," *arXiv preprint*.

Snee, Tom (2016). "Predicting Box Office: Boffo or Bomb?," February, https://now.uiowa.edu/news/2016/02/predicting-box-office-boffo-or-bomb, Accessed: 2024-11-10.

Sperling, Nicole (2020a). "The Oscars Will Add a Diversity Requirement for Eligibility," *The New York Times*, https://www.nytimes.com/2020/09/08/movies/oscars-diversity-rules-best-picture.html, Accessed: 2024-11-10.

——— (2020b). "Academy Sets New Diversity Requirements for Oscars Best Picture Eligibility," *The New York Times*, https://www.nytimes.com/2020/09/09/movies/oscars-best-picture-diversity.html, Accessed: 2024-11-10.

Stone, Eric W and Warren, Ronald S (1999). "Customer Discrimination in Professional Basketball: Evidence from the Trading-Card Market," *Applied Economics*, 31 (6), 679–685.

Thompson, Ben (2013). "Solving Equation of a Hit Film Script, With Data," May 6, https://www.nytimes.com/2013/05/06/business/media/solving-equation-of-a-hit-film-script-with-data.html, Accessed: 2024-11-10.

Weaver, Andrew J. (2011). "The role of actors' race in white audiences' selective exposure to movies," *Journal of Communication*, 61 (2), 369–385.

Weinstein, Mark (1998). "Profit-sharing contracts in Hollywood: Evolution and analysis," *The Journal of Legal Studies*, 27 (1), 67–112.

Yahr, Emily (2016). "It's hard to predict a movie's profitability, but you learn some lessons along the way," https://www.

washingtonpost.com/news/arts-and-entertainment/wp/2016/05/16/its-hard-to-predict-a-movies-profitability-but-you-learn-some-lessons-along-the-way/, Accessed: 2024-11-10.

Zussman, Asaf (2013). "Ethnic discrimination: Lessons from the Israeli online market for used cars," *The Economic Journal*, 123 (572), F433–F468.

Zwick, Ed (2024). *Hits, flops, and other illusions: My Fortysomething Years in Hollywood*: Gallery Books.

Tables and Figures

Table 1: Model Predictions

Discrimination Source	**Mathematical Definition**	**Comparative Statics: Expected Value**	**Comparative Statics: Variance**
Taste-based The producer bears a utility loss producing non-white movies.	$\pi_{0b} > \pi_{0w}$ The production threshold is relatively higher for non-white movies.	$E_b > E_w$	$Var_b < Var_w$
Customer The viewing public has a preference for white movies over non-white movies.	$\mu_b < \mu_w$ The distribution of box-office revenues for white movies is shifted to the right, relative to that of non-white movies.	$E_b < E_w$	$Var_b < Var_w$
Statistical The producer has "less" or "worse" information on non-white movies' potential.	$\sigma^2_{yb} > \sigma^2_{yw}$ The signal for non-white movies is less informative.	$E_b < E_w$	$Var_b > Var_w$

Note: Summary of the model predictions. In our notation, w (b) denotes white (non-white) movies; π_{0b}, π_{0w} are the type-specific production thresholds; μ_b, μ_w denote the type-specific means of the box-office revenue distributions; $\sigma^2_{yb}, \sigma^2_{yw}$ stand for the type-specific signal variances. See Section 3 for the derivations.

Table 2: Summary Statistics

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
PANEL A: Classification of movies by type					
Share of non-white performers	7,840	0.12	0.24	0	1
At least one non-white	7,840	0.26	0.44	0	1
At least two non-whites	7,840	0.08	0.27	0	1
Distribution of the number of non-white performers (percentages):					
0	74.3				
1	17.9				
2	4.5				
3	2.1				
4	1.3				
PANEL B: Other variables					
Gross revenue(in Millions of 2005 Dollars)	7,205	25.9	54.1	$1.94 * 10^{-5}$	804
Ln (Gross revenue)	7,205	14.14	3.39	2.97	20.50
Cost(in Millions of 2005 Dollars)	3,955	37.4	44.7	$1.1 * 10^{-3}$	907
Ln(Cost)	3,955	16.72	1.45	7.00	20.63
Run time(minutes)	6,804	103.51	18.49	38	600
IMDB score	4,915	6.25	0.97	1.50	9
Metacritic score	4,915	51.58	17.10	1	100
Average age of billed performers	7,715	41.92	10.42	10	99
Star power(in Millions)	7,840	262	303	0	2,350
Ln(Star power)	7,840	17.50	4.54	0	21.58
Number of weeks	6,491	11.62	14.66	1	476
Ln(Number of screens)	6,078	4.88	2.95	0.69	8.43
Distribution of movies by genre (percentages):					
Action	16.62				
Animation	0.17				
Comedy	26.27				
Drama	36.57				
Other	20.37				

Note: Source: authors' calculations. Data sources are described in Section 4.

Table 3: The non-white revenue premium

	(1)	(2)	(3)	(4)	(5)	(6)
Sample:	Full	Full	Full	Full	Non-missing cost	Non-missing cost
	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)
Race: at least	0.914***	0.926***	0.433***	0.336***	0.488***	0.522***
two non-white	(0.200)	(0.183)	(0.093)	(0.076)	(0.157)	(0.096)
Share of		-1.059***	0.180**	0.096	-0.732***	0.130
female		(0.194)	(0.084)	(0.074)	(0.203)	(0.111)
ln(Star Power)		0.235***	0.023**	0.006	0.176***	-0.033***
		(0.022)	(0.010)	(0.008)	(0.020)	(0.011)
Average age		-0.065***	-0.004	0.003	-0.036***	-0.012***
		(0.006)	(0.004)	(0.003)	(0.007)	(0.004)
ln(Cost)			0.554***	0.403***		0.728***
			(0.051)	(0.034)		(0.044)
= 1 if ln(Cost)			6.249***	4.804***		
			(0.736)	(0.541)		
Movie controls			Y	Y		Y
Distributor FEs				Y		
N	6943	6943	6943	6943	3856	3856
R^2	0.006	0.125	0.698	0.341	0.071	0.596

Note: Data sources and specification are described in Sections 4 and 5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Robustness with respect to different definitions of "non-white" movies

	(1)	(2)	(3)	(4)
Race:	At least two non-white	At least one non-white	Share of non-white	Leading role is non-white
	Ln(Gross Revenue)	Ln(Gross Revenue)	Ln(Gross revenue)	Ln(Gross revenue)
Race	0.433***	0.227***	0.628***	0.464***
	(0.093)	(0.061)	(0.121)	(0.081)
Q10	0.542***	0.237**	0.664***	0.502***
	(0.104)	(0.113)	(0.229)	(0.100)
Q25	0.517***	0.263***	0.813***	0.520***
	(0.109)	(0.059)	(0.171)	(0.122)
Q50	0.375***	0.215***	0.572***	0.334***
	(0.078)	(0.060)	(0.120)	(0.089)
Q75	0.281**	0.189***	0.469***	0.327***
	(0.112)	(0.049)	(0.122)	(0.099)
Q90	0.283***	0.188***	0.442***	0.302***
	(0.058)	(0.061)	(0.120)	(0.062)
Cast controls	Y	Y	Y	Y
Movie controls	Y	Y	Y	Y
N	6943	6943	6943	6943

Note: Data sources and specification are described in Sections 4 and 5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Robustness to different dependent variables

	(1)	(2)	(3)
Sample:	Non-missing cost variable Ln(Profit Margin+1)	Non-missing cost variable Profit(in million)	Non-missing cost variable Revenue(in million)
Race: At least two non-white	0.553***	8.722***	1.746
	(0.095)	(2.156)	(3.004)
Q10	0.469***	5.849***	3.233***
	(0.114)	(1.569)	(0.919)
Q25	0.339***	6.934***	5.372***
	(0.095)	(1.250)	(1.192)
Q50	0.460***	8.352***	5.323***
	(0.104)	(1.287)	(1.654)
Q75	0.352***	9.494***	6.907**
	(0.085)	(3.679)	(3.368)
Q90	0.314***	11.883***	6.278
	(0.072)	(3.707)	(6.384)
Cast controls	Y	Y	Y
Movie controls	Y	Y	Y
N	3856	3856	3856

Note: Data sources and specification are described in Sections 4 and 5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in logs) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Conditional residual variance regressions:
robustness with respect to different definitions of non-white movies

	(1)	(2)	(3)	(4)	(5)	(6)
Race definition	At least two	At least two	At least two	At least one	Share	Leading role
	Dependent variable: Ln(residual square)					
Race	-0.463***	-0.471***	-0.324***	-0.117**	-0.379***	-0.212**
	(0.099)	(0.098)	(0.094)	(0.059)	(0.109)	(0.083)
Cast Controls		Y	Y	Y	Y	Y
Movie Controls			Y	Y	Y	Y
N	6943	6943	6943	6943	6943	6943
R^2	0.003	0.012	0.121	0.123	0.122	0.117

Note: Data sources and specification are described in Sections 4 and 5.5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors are clustered by distributor in the main regressions only. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Heterogeneity Analysis

	(1) Distributor Not Big-6	(2) Distributor Big-6	(3) Genre: Action/Adventure	(4) Genre: Comedy	(5) Genre: Drama
Race: At least two non-white	0.508***	0.355***	0.068	0.823***	0.395***
	(0.138)	(0.089)	(0.201)	(0.178)	(0.128)
Cast Controls	Y	Y	Y	Y	Y
Movie Controls	Y	Y	Y	Y	Y
P-Value of the difference	0.397			0.014	
N	4766	2177	1135	1880	2590

	(6) Period: Pre-2007	(7) Period: Post-2008	(8) Gender: ≤50% female	(9) Gender: >50% female
Race: At least two non-white	0.567***	0.262*	0.408***	0.674**
	(0.092)	(0.133)	(0.087)	(0.307)
Cast Controls	Y	Y	Y	Y
Movie Controls	Y	Y	Y	Y
P-Value of the difference	0.041		0.365	
N	2774	4169	5933	1010

Note: Data sources and specification are described in Sections 4 and 5.5. In all specifications, the sample is restricted to observations with non-missing data on production costs. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Producer Analysis

	(1) Producer Sub-Sample	(2) Producer Sub-Sample	(3) Producer Sub-Sample
	Dependent variable: Ln(gross revenue)		
Cast: more than two non-white	0.271**	0.253*	-0.004
	(0.132)	(0.132)	(0.138)
Producer: more than one non-white		0.045	-0.081
		(0.121)	(0.126)
Cast x Producer			0.556**
			(0.228)
Observations	1,955	1,955	1,955
R-squared	0.733	0.733	0.734
Baseline controls	Y	Y	Y

Note: Data sources and specification are described in Sections 4 and 5.7. In all specifications, the sample is restricted to observations with some information on the producer race. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, movie budget cost (in the log), year fixed effects, and indicators for missing run time, Metacritic score, MPAA rating, or budget cost. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Regressions of first-weekend theaters on number of theaters

VARIABLES	(1) First-weekend revenues, log	(2) First-weekend revenues, log	(3) First-weekend revenues, log
First-weekend	0.990***	0.980***	0.839***
# theaters, log	(0.016)	(0.017)	(0.018)
Cast controls	N	Y	Y
Movie controls	N	N	Y
Residuals: white vs non-white			
Average white	-0.014	-0.014	-0.016
Average non-white	0.151	0.152	0.178
Average difference	-0.165	-0.166	-0.194
p-value of t-test (two-sided)	0.001	0.001	0.000
N	6,276	6,276	6,276
R^2	0.889	0.890	0.927

Note: Data sources and specification are described in Sections 4 and 6. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, movie budget cost (in the log), and indicators for missing run time, Metacritic score, MPAA rating, and movie budget cost. Relative to the baseline sample used in Table 3, 633 observations are excluded due to missing data on first-weekend revenues or theaters, while 55 are lost due do taking logs (zero values.) Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

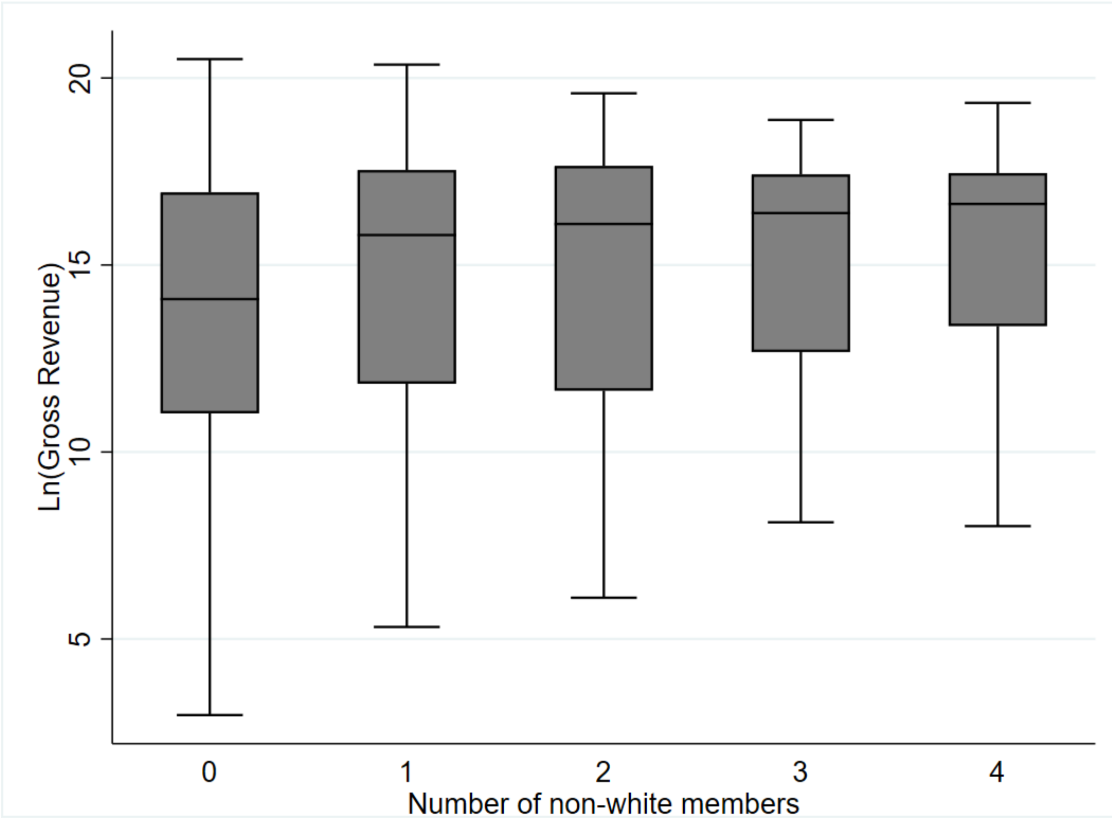

Figure 1: Revenue distribution by number of non-white members

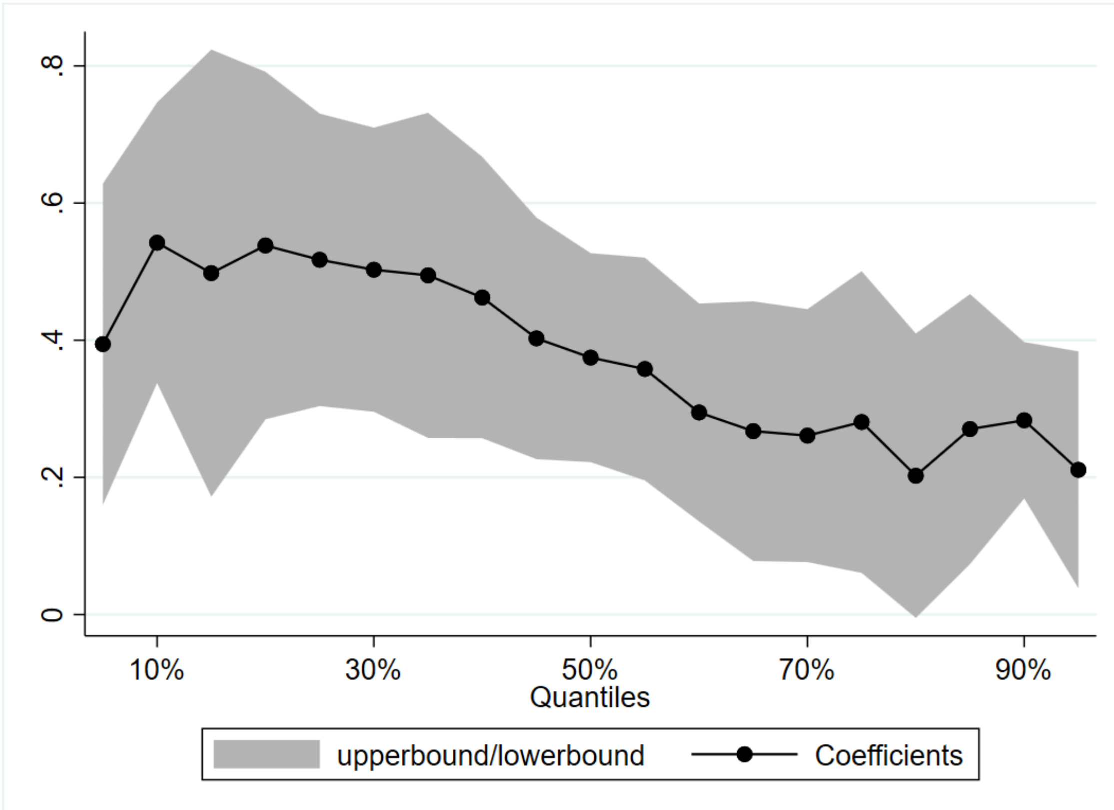

Figure 2: Coefficients are decreasing over quantiles

Appendix A Proofs

Lemma 1: (Inverse Mills ratio). If X is a normally distributed random variable with mean μ and variance σ^2, then

$$E(X \mid X > \alpha) = \mu + \sigma \frac{\phi(\frac{\alpha-\mu}{\sigma})}{1-\Phi(\frac{\alpha-\mu}{\sigma})}$$

where ϕ and Φ are the p.d.f. and c.d.f. of the Normal, respectively.

Proof of Proposition 2:

Part 1: mean and variance of box-office revenue conditional on production

(i) Given two normal distributions $\pi \mid t$ and $y|\pi, t$, $f(\pi|y,t) \propto f(y \mid \pi, t)f(\pi \mid t)$. Hence

$$\pi|y,t \sim N(E(\pi|y,t), Var(\pi|y,t)), \quad y|t \sim N(\mu_t, \sigma_{\pi t}^2 + \sigma_{yt}^2)$$

where:

$$E(\pi|y,t) = \frac{\sigma_{\pi t}^2}{\sigma_{\pi t}^2 + \sigma_{yt}^2} y + \frac{\sigma_{yt}^2}{\sigma_{\pi t}^2 + \sigma_{yt}^2} \mu_t \sim N(\mu_t, \frac{\sigma_{\pi t}^4}{\sigma_{\pi t}^2 + \sigma_{yt}^2})$$

$$Var(\pi|y,t) = \frac{\sigma_{\pi t}^2 \sigma_{yt}^2}{\sigma_{\pi t}^2 + \sigma_{yt}^2}$$

With an abuse of notation, denote π_t as $\pi \mid t$, and y_t as $y \mid t$, then by Lemma 1 and the law of total expectation:

$$\begin{aligned} E(\pi_t|y_t > \bar{y}_t) &= E(\pi_t \mid E(\pi_t \mid y_t) > \pi_0) \\ &= E(E(\pi_t \mid y_t) \mid E(\pi_t \mid y_t) > \pi_0) \\ &= \mu_t + \sigma \frac{\phi(\frac{\pi_0-\mu_t}{\sigma})}{1-\Phi(\frac{\pi_0-\mu_t}{\sigma})} (3) \end{aligned}$$

where $\sigma^2 = \frac{\sigma_{\pi_t}^4}{\sigma_{\pi_t}^2 + \sigma_{y_t}^2}$

END.

(ii) Now for variance:

$$
\begin{aligned}
Var(\pi_t|y_t > \bar{y}_t) &= Var(\pi_t|E(\pi_t|y_t) > \pi_0) \\
&= E(\pi_t^2|E(\pi_t|y_t) > \pi_0) - E^2(\pi_t|E(\pi_t|y_t) > \pi_0) \\
&= E(E(\pi_t^2|y_t)|E(\pi_t|y_t) > \pi_0) - E^2(\pi_t|E(\pi_t|y_t) > \pi_0) \\
&= E([Var(\pi_t|y_t) + E^2(\pi_t|y_t)]|E(\pi_t|y_t) > \pi_0) - E^2(E(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0) \\
&= \frac{\sigma_{\pi_t}^2\sigma_{y_t}^2}{\sigma_{\pi_t}^2 + \sigma_{y_t}^2} + E\Big(E^2(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0\Big) - E^2\Big(E(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0\Big) \quad (4)
\end{aligned}
$$

For a standard normal distribution, $z \sim N(0,1)$.

$$
\begin{aligned}
E(z^2|z > c) &= \frac{1}{1-\Phi(c)}\int_c^\infty \frac{z^2}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big)dz \\
&= \frac{1}{1-\Phi(c)}\int_c^\infty \Bigg(\frac{1}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big) - \Big(\frac{z}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg)dz \\
&= \frac{1}{1-\Phi(c)}\int_c^\infty \Bigg(\frac{1}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big) - \Big(\frac{z}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg)dz \\
&= \frac{1}{1-\Phi(c)}\int_c^\infty \Bigg(\frac{1}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big) - \Big(\frac{z}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg)dz \\
&= 1 + \frac{c\phi(c)}{1-\Phi(c)}
\end{aligned}
$$

So, for $x \sim N(\mu, \sigma^2)$

$$
\begin{aligned}
1 + \frac{\frac{c-\mu}{\sigma}\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})} &= E\Bigg(\Big(\frac{x-\mu}{\sigma}\Big)^2 | \frac{x-\mu}{\sigma} > \frac{c-\mu}{\sigma}\Bigg) \\
&= \frac{1}{\sigma^2}\Big(E(x^2|x>c) - 2\mu E(x|x>c) + \mu^2\Big)
\end{aligned}
$$

Combining with

$$
E(x|x>c) = \mu + \sigma\frac{\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})}
$$

we obtain

$$
E(x^2|x>c) = \sigma^2 + \sigma^2\frac{\frac{c-\mu}{\sigma}\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})} + \mu^2 + 2\mu\sigma\frac{\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})}
$$

Plugging in (4) yields

$$
Var(\pi_t|E(\pi_t|y_t) > \pi_0) = \frac{\sigma_{\pi_t}^2\sigma_{y_t}^2}{\sigma_{\pi_t}^2 + \sigma_{y_t}^2} + \sigma^2 + \sigma^2\frac{\frac{\pi_0-\mu_t}{\sigma}\phi(\frac{\pi_0-\mu_t}{\sigma})}{1-\Phi(\frac{\pi_0-\mu_t}{\sigma})} - \sigma^2\Bigg(\frac{\phi(\frac{\pi_0-\mu_t}{\sigma})}{1-\Phi(\frac{\pi_0-\mu_t}{\sigma})}\Bigg)^2 \qquad (I)
$$

Then, $(I) = \sigma_{\pi_t}^2 + \sigma^2(x\lambda(x) - \lambda^2(x))$, by $\sigma^2 = \frac{\sigma_{\pi_t}^4}{\sigma_{\pi_t}^2 + \sigma_{y_t}^2}$, $x = \frac{\pi_0 - \mu_t}{\sigma}$, $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$
END.

Part 2: comparative statics

Building blocks

Lemma 2: For $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$, $\frac{3x+\sqrt{x^2+8}}{4} < \lambda(x) < \frac{x+\sqrt{x^2+4}}{2}$ for $x \in R$.

Proof: Normally, a computer can confirm this lemma. However, when $x > 7$, both the numerator and the denominator of λ are so close to 0 that the value for λ is heavily biased. Hence, this proof will only target the case where $x > 7$.

First, taking the first derivative of $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ yields $\phi'(x) = -x\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = -x\phi(x)$. It follows that

$$\begin{aligned} 1 - \Phi(x) &= \int_x^\infty \phi(u)du \\ &= -\int_x^\infty \frac{\phi'(u)}{u}du \\ &= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} + \frac{3\phi(x)}{x^5} - \frac{15\phi(x)}{x^7} + \int_x^\infty \frac{105\phi(u)}{u^8}du \\ &= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} + \frac{3\phi(x)}{x^5} - \frac{15\phi(x)}{x^7} + \frac{105\phi(x)}{x^9} - \int_x^\infty \frac{945\phi(u)}{u^{10}}du \end{aligned}$$

Then

$$\frac{1}{\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7} + \frac{105}{x^9}} < \frac{\phi(x)}{1-\Phi(x)} < \frac{1}{\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7}} \qquad when \quad x > 7$$

Let the left (right) term of the inequality be denoted as LHS (RHS). We first prove that when $x > 7$, $LHS > \frac{3x+\sqrt{x^2+8}}{4}$. Assume that this is true. Then

$$LHS > \frac{3x+\sqrt{x^2+8}}{4}$$

$$\iff (x^9+3x^7-9x^5+45x^3-315x)^2 > (x^2+8)(x^8-x^6+3x^4-15x^2+105)^2$$

$$\iff x^{18}+6x^{16}-9x^{14}+36x^{12}-279x^{10}-27000x^8+7695x^6-28350x^4+99225x^2 >$$

$$x^{18}+6x^{16}-9x^{14}+20x^{12}-39x^{10}+1692x^8-1545x^6+3690x^4-14175x^2+88200$$

$$\iff 16x^{12}-240x^{10}-4392x^8+9240x^6-32040x^4+113400x^2-88200 > 0$$

Then for $x > 7$, $16x^{12} > 16 * 7^2 x^{10}$, i.e. $16x^{12} > 784x^{10}$, which is true.

We now prove that when $x > 7$, $RHS < \frac{x+\sqrt{x^2+4}}{2}$

$\iff (x^7+x^5-3x^3+15x)^2 < (x^2+4)(x^6-x^4+3x^2-15)^2$

$\iff x^{14}+2x^{12}-5x^{10}+24x^8+39x^6-90x^4+225x^2 <$

$x^{14}+2x^{12}-x^{10}-8x^8-105x^6+66x^4-135x^2+900$

$\iff x^{10}-8x^8-36x^6+39x^4-90x^2+225 > 0$

Then for $x > 7$, $x^{10} > 7^2x^8$, i.e. $x^{10} > 49x^8$, which is also true.

A computer can easily confirm that the lemma holds also for $x < 7$, which completes the proof. In addition, for $x > 0$, we can show that $x < \frac{3x+\sqrt{x^2+8}}{4} < \lambda(x) < \frac{x+\sqrt{x^2+4}}{2} < x + \frac{1}{x}$.

END.

Comparative statics 2(a):

(i) Let $x = \frac{\pi_0 - \mu}{\sigma}$. Then

$$\frac{\mathrm{d}(3)}{\mathrm{d}\mu} = 1+\sigma\lambda'(x) = 1+\sigma(-x\lambda(x)+\lambda^2(x))\Big(-\frac{1}{\sigma}\Big) = -\left(\lambda(x)-\frac{x+\sqrt{x^2+4}}{2}\right)\left(\lambda(x)-\frac{x-\sqrt{x^2+4}}{2}\right) \qquad (i)$$

By Lemma 2, $\frac{\mathrm{d}(3)}{\mathrm{d}\mu} > 0 \; \forall x \in R$.

END.

(ii) Again let $x = \frac{\pi_0 - \mu}{\sigma}$. Then

$$\begin{aligned}\frac{\mathrm{d}(I)}{\mathrm{d}\mu} &= \sigma^2(\lambda(x) + x\lambda'(x) - 2\lambda(x)\lambda'(x))\Big(-\frac{1}{\sigma}\Big) \\ &= -\sigma\Big(\lambda(x) - x^2\lambda(x) + 3x\lambda^2(x) - 2\lambda^3(x)\Big) \\ &= \sigma\lambda(x)\left(\lambda(x) - \frac{3x + \sqrt{x^2+8}}{4}\right)\left(\lambda(x) - \frac{3x - \sqrt{x^2+8}}{4}\right)\end{aligned}$$

By Lemma 2, $\frac{\mathrm{d}(I)}{\mathrm{d}\mu} > 0\ \forall x \in R$.

END.

Comparative statics 2(b):

(i)

$$\frac{\mathrm{d}(3)}{\mathrm{d}\pi_0} = \sigma\lambda'(x) = \sigma(-x\lambda(x) + \lambda^2(x))(\frac{1}{\sigma}) = (-x + \lambda(x))\lambda(x) > 0$$

.

END.

(ii) See the proof for 2(a), (ii).

END.

Comparative statics 2(c):

(i)

$$\frac{\mathrm{d}(3)}{\mathrm{d}\sigma} = \sigma\lambda'(x) + \lambda(x) = \Big(-x\lambda(x) + \lambda^2(x)\Big)\left(-\frac{y_0 - \mu_t}{\sigma}\right) + \lambda(x) = \lambda(x)\Big(1 + x^2 - x\lambda(x)\Big) \quad (4)$$

where, again, $\sigma = \frac{\sigma^2_{\pi t}}{\sqrt{\sigma^2_{\pi t} + \sigma^2_{yt}}}$, $x = \frac{\pi_0 - \mu_t}{\sigma}$, $x > 0$, and $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$.

When $x > 0$, we can write $(4) = x\lambda(x)(\frac{1}{x} + x - \lambda(x))$. By Lemma 2 and $\frac{x+\sqrt{x^2+4}}{2} < x + \frac{1}{x}$, $(4) > 0$ holds.

When $x < 0$, $(4) > 0$ clearly holds.

Hence, if σ^2_{yt} increases, then σ^2 decreases and (3), i.e. E_t decreases too.

END.

(ii)

$$\begin{aligned}
\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} &= \sigma_{yt}^4 \frac{\Big(x\lambda(x)-\lambda^2(x)\Big)'(\sigma_{\pi t}^2+\sigma_{yt}^2)-\Big(x\lambda(x)-\lambda^2(x)\Big)}{(\sigma_{\pi t}^2+\sigma_{yt}^2)^2} \\
&= \sigma_{\pi t}^4\left(\frac{\lambda(x)\Big(1-x^2+3x\lambda(x)-2\lambda^2(x)\Big)}{\sigma_{\pi t}^2+\sigma_{yt}^2}\Big(-\frac{\pi_0-\mu}{\sigma^2}\Big)\Big(-\frac{\sigma_{\pi t}^2}{2(\sigma_{\pi t}^2+\sigma_{yt}^2)^{\frac{3}{2}}}\Big)-\frac{x\lambda(x)-\lambda^2(x)}{(\sigma_{\pi t}^2+\sigma_{yt}^2)^2}\right) \\
&= -2x\frac{\lambda(x)\sigma_{\pi t}^4}{2(\sigma_{\pi t}^2+\sigma_{yt}^2)^2}\left(\lambda(x)-\frac{x}{2}\right)\left(\lambda(x)-x-\frac{1}{x}\right)
\end{aligned}$$

When $x > 0$, $x < \lambda(x) < x + 1/x$, which implies that $\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} > 0$.
When $x < 0$, it is easy to see that $\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} > 0$.
Hence, if σ_{yt}^2 increases, (I), i.e. Var_t increases too.
END.

Appendix B Alternative distributional assumptions

In this appendix, we explore the robustness of our results to alternative distributional assumptions.

B.1 Beta-Binomial distribution

B.1.1 Setup

We first consider the case where the producer cares only about a binary criterion, whether the movie will be a "hit" or not.

Assume the object of interest is p, the probability that the movie is a hit. The prior distribution of p is Beta with parameters α and β:

$$p \sim Beta(\alpha, \beta)$$

Therefore:

$$\begin{aligned}
f(p) &\propto p^{\alpha-1}(1-p)^{\beta-1} \\
E(p) &= \frac{\alpha}{\alpha+\beta} \\
V(p) &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}
\end{aligned}$$

Producers observe a signal y, which, conditional on the true p, is distributed binomial with parameters n and p. We can think of this as the producer consulting with n critics, and each one independently assessing whether the movie will be a hit or not, with probability p.

$$f(y|p) = \binom{n}{y} p^y (1-p)^{(n-y)}$$

It follows that the posterior density of p given y is

$$\begin{aligned}
f(p|y) &\propto p^{\alpha-1}(1-p)^{\beta-1} p^y (1-p)^{n-y} \\
&\Rightarrow p|y \sim Beta(\alpha+y, \beta+n-y)
\end{aligned}$$

Therefore,

$$E(p|y) = \frac{\alpha+y}{\alpha+y+\beta+n-y} = \frac{\alpha+y}{\alpha+\beta+n}$$

The producer will produce the movie if $E(p|y) > p_0$, for some predetermined p_0. Therefore, the signal threshold for production $\bar{y}$ is:

$$\bar{y} \equiv p_0(\alpha + \beta + n) - \alpha.$$

It is convenient to use a reparametrization, letting $\kappa = \alpha + \beta$. If $\alpha, \beta > 1$, then κ captures the spread of the distribution: for a given α, a higher value of κ means that the distribution is more concentrated, i.e., the prior is more informative.[1]

Let $s = y/n$ be the success rate of the signal. Expressing the signal threshold in terms of s, the movie is produced if and only if

$$s > \bar{s} \equiv p_0 + \frac{\kappa}{n}(p_0 - \frac{\alpha}{\kappa})$$

B.1.2 Comparative statics for production

(1) **Customer discrimination**. We interpret customer discrimination against non-white movies as $\alpha_b < \alpha_w$. That is, non white movies have a lower prior probability of being a hit. The signal threshold is decreasing in α:

$$\frac{\partial \bar{s}}{\partial \alpha} < 0.$$

Therefore, under customer discrimination, the signal threshold for non-white movies is higher than the signal threshold for white movies.

(2) **Taste-based discrimination**. We interpret taste-based discrimination against non-white movies as $p_{0b} > p_{0w}$. That is, non-white movies are held to a higher standard, and are produced only if the posterior probability of the movie being a hit exceeds a higher threshold. The signal threshold is increasing in p_0:

$$\frac{\partial \bar{s}}{\partial p_0} > 0.$$

Therefore, under taste-based discrimination, the signal threshold for non-white movies is higher than the signal threshold for white movies.

[1] Under this parameterization, the mean of the prior distribution is $E(p) = \frac{\alpha}{\kappa}$ and the variance is $V(p) = \frac{\alpha(\kappa-\alpha)}{\kappa^2(\kappa-1)}$. For $\alpha, \beta > 1$, the variance is strictly decreasing in κ.

(3) **Statistical discrimination**. Finally, we interpret statistical discrimination as $n_b < n_w$. That is, the signal for non-white movies being is less informative than that for white movies. The derivative of the signal threshold with respect to n is:

$$\frac{\partial \bar{s}}{\partial n} = -\frac{\kappa}{n^2}(p_0 - \frac{\alpha}{\kappa})$$

The sign of this derivative depends on $p_0 - \alpha/\kappa$. Under this parametrization, α/κ is the prior mean of p. In other words, we have the same qualitative result as in the log-normal model presented in the main text.

(i) If $p_0 > \alpha/\kappa$, (i.e., the producer wants to produce only movies with a very high probability of being a hit),

$$\frac{\partial \bar{s}}{\partial n} < 0;$$

that is, a less precise signal (lower n) raises the signal threshold. The signal threshold for non-white movies is higher.

(ii) If $p_0 < \alpha/\kappa$, (i.e. the producer wants to weed out the very low quality movies),

$$\frac{\partial \bar{s}}{\partial n} > 0;$$

now, a less precise signal (lower n) *lowers* the signal threshold. One can be a bit more tolerant of a bad signal for non-white movies, because it is difficult to say, based on the signal alone, whether the movie is really bad.

It is easy to see that the comparative statics with respect to the precision of the signal mirrors exactly what we had in the normal-normal case.

B.1.3 Comparative statics for the observed success rate, conditional on production: simulations

We only observe whether a movie is a hit, conditional on production. Therefore, as in the analysis in the main text, we need to characterize the the posterior distribution of p conditional on $s > \bar{s}$, and derive its comparative statics with respect to p_0, α and n. While it is not possible to derive an analytical solution for the comparative statics, we can proceed by simulation. Specifically, for each set of parameter values, we draw a sample of L movies, apply the production decision rule, and report the mean and standard deviation of the posterior distribution p conditional on production. The results are presented in Table B.1.

Table B.1: Simulation results: Beta-binomial distribution

A: Taste based discrimination: $p_0 \uparrow$ for Non-white movies Fixed $\alpha = 4$, $\kappa = 8$, $n = 5$											Trend
p_0	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	
mean	0.500	0.500	0.500	0.500	0.514	0.545	0.545	0.587	0.638	0.638	$\uparrow$
std	0.167	0.167	0.167	0.167	0.160	0.151	0.151	0.142	0.133	0.133	$\downarrow$
B: Customer discrimination: $\alpha \downarrow$ for Non-white movies Fixed $p_0 = 0.5$, $\kappa = 8$, $n = 5$											
α	6	5.5	5	4.5	4	3.5	3	2.5	2	1.5	
mean	0.752	0.692	0.650	0.597	0.587	0.542	0.555	0.515	0.538	0.497	$\downarrow$
std	0.142	0.151	0.148	0.151	0.142	0.143	0.136	0.137	0.135	0.135	$\downarrow$
C1: Statistical discrimination, $p_0 > \alpha/\kappa$: $n \downarrow$ for Non-white movies Fixed $p_0 = 0.6$, $\kappa = 8$, $\alpha = 4$											
n	11	10	9	8	7	6	5	4	3	2	
mean	0.673	0.656	0.679	0.659	0.638	0.662	0.638	0.666	0.637	0.600	$\downarrow$
std	0.115	0.119	0.117	0.122	0.128	0.126	0.133	0.131	0.139	0.148	$\uparrow$
C2: Statistical discrimination, $p_0 < \alpha/\kappa$: $n \downarrow$ for Non-white movies Fixed $p_0 = 0.4$, $\kappa = 8$, $\alpha = 4$											
n	11	10	9	8	7	6	5	4	3	2	
mean	0.549	0.557	0.540	0.548	0.528	0.535	0.545	0.520	0.527	0.500	$\downarrow$
std	0.145	0.143	0.149	0.148	0.154	0.153	0.151	0.159	0.158	0.167	$\uparrow$

Legend: simulated data with sample size $L = 10^6$, using R with seed 123. Mean, Std: sample average and standard deviation of the posterior distribution of $p|s, s > \bar{s}$ from the simulation.

Each panel in the table presents a different comparative statics exercise. For example, in Panel A, to examine the role of taste-based discrimination, we fix the values of α, κ and n, and study what happens to the mean and standard deviation of the posterior distribution of p as we increase the value of p_0. The results in Panel A show that as taste-based discrimination increases, the posterior expected value of p increases, and the posterior standard deviation decreases. These results match the predictions in the normal-normal model, derived analytically.

In Panel B we look at the effect of increasing customer discrimination increases. As in the normal model, both the expected value and the standard deviation increase as the value of α decreases.[2]

In Panels C1 and C2 we study the effect of statistical discrimination, distinguishing between the case in which the producer only wants to produce very high quality movies ($p_0 > \alpha/\kappa$) so that the signal threshold decreases in n (case 3.i in Section B.1.2); and the one in which the producer wants to weed out very low quality movies ($p_0 < \alpha/\kappa$) so that the signal threshold increases in n. We see that in both cases the mean of p decreases and the standard deviation increases as the extent of statistical discrimination increases (the signal becomes less prescise, or n decreases). Again, the pattern of comparative statics results mirrors exactly what we obtained in the normal-normal case (Section 3 in the main text).

We conclude that all of the main predictions of the theoretical model based on the normal-normal case in the main text remain identical under the beta-binomial model.

B.2 Pareto-Normal distribution

B.2.1 Setup

We now return to the case considered in the main text, where the producer cares about (log) revenue, but we now depart from the normal-normal model. Specifically, we assume that ex-ante revenue $\tilde{\pi}$ (in dollars) follows a Pareto distribution:

$$\tilde{\pi} \sim Pareto(x_m, a)$$

where x_m is the minimum, and a is the shape parameter. The CDF is:

$$F(\tilde{\pi}) = \begin{cases} 1 - (\frac{x_m}{\tilde{\pi}})^a, \; if\; \tilde{\pi} \geq x_m \\ 0, \; if\; \tilde{\pi} < x_m \end{cases}$$

Then, log revenue $\pi \equiv \log(\tilde{\pi})$ has a shifted exponential distribution: $\pi \sim Exp(a) + log(x_m)$, or, equivalently, $\log(\frac{\tilde{\pi}}{x_m}) \sim Exp(a)$.

[2] In each panel, as we move in the table from left to right, we *increase* the extent of discrimination. In the case of customer discrimination (Panel B) and statistical discrimination (Panel C), an increase in discrimination implies a decrease in the parameter of interest.

Therefore, the pdf of π is:

$$f(\pi) = \begin{cases} a * exp(-a(\pi - log(x_m))), & if\ \pi \geq log(x_m) \\ 0, & if\ \pi < log(x_m) \end{cases}$$

Producers observe a signal y, which, conditional on the true π is distributed $N(\pi, \sigma_y^2)$.

$$f(y|\pi) = \frac{1}{\sigma_y\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{y-\pi}{\sigma_y})^2)$$

It follows that the posterior distribution of π given y is

$$f(\pi|y) \propto exp(-a(\pi - \log(x_m)) - \frac{1}{2}(\frac{y-\pi}{\sigma_y})^2), \text{ if } \pi \geq \log(x_m)$$

When $\pi \geq log(x_m)$:

$$\begin{aligned} f(\pi|y) &\propto exp(-a(\pi - log(x_m)) - \frac{1}{2}(\frac{y-\pi}{\sigma_y})^2) \\ &= exp(-\frac{1}{2\sigma_y^2}(\pi^2 - 2y\pi + y^2 + 2\sigma_y^2 a\pi - 2\sigma_y^2 a \log(x_m))) \\ &= exp(-\frac{1}{2\sigma_y^2}((\pi - (y - a\sigma_y^2))^2 - a^2\sigma_y^4 + 2ya\sigma_y^2 - 2\sigma_y^2 a \log(x_m))) \\ &= exp(-\frac{1}{2\sigma_y^2}((\pi - (y - a\sigma_y^2))^2)) \times exp(-a(y - \frac{a\sigma_y^2}{2} - \log(x_m))) \end{aligned}$$

Given y, the second term is constant. Therefore, putting everything together, we have that

$$f(\pi|y) \propto exp(-\frac{1}{2\sigma_y^2}((\pi - (y - a\sigma_y^2))^2)), \text{ for } \pi > \log(x_m).$$

This implies that the posterior distribution of π given the signal y is a truncated normal derived from a normal distribution with mean $y - a\sigma_y^2$, variance σ_y^2 and lower truncation point $\log(x_m)$. The posterior mean is therefore

$$E(\pi|y) = (y - a\sigma_y^2) + \sigma_y \frac{\phi(\log(x_m))}{1 - \Phi(\log(x_m))}.$$

The producer will produce the movie if $E(\pi|y) > \pi_0$, for some predetermined π_0. Therefore, the signal threshold for production $\bar{y}$ is:

$$\bar{y} \equiv \pi_0 + a\sigma_y^2 - \sigma_y \frac{\phi(\log(x_m))}{1 - \Phi(\log(x_m))}.$$

B.2.2 Comparative statics for production

(1) **Customer discrimination**: The expectation of an exponential distribution with parameter a is $1/a$. Therefore, we interpret customer discrimination against non-white movies as $a_b > a_w$. The signal threshold is increasing in a:

$$\frac{\partial \bar{y}}{\partial a} > 0$$

As in the normal-normal case, the signal threshold for non-white movies is higher than the signal threshold for white movies.

(2) **Taste-based discrimination**: We interpret taste-based discrimination against non-white movies as $\pi_{0b} > \pi_{0w}$ – non-white movies are held to a higher standard and are produced only if the posterior mean exceeds a threshold that is higher than that set for white movies. The signal threshold increases in π_0:

$$\frac{\partial \bar{y}}{\partial \pi_0} > 0.$$

Therefore, under taste-based discrimination, the signal threshold for non-white movies is higher than that for white movies. This result mirrors that of the normal-normal case.

(3) **Statistical discrimination**: We interpret statistical discrimination as $\sigma_{yb} > \sigma_{yw}$, i.e., the signal for non-white movies is less precise. The derivative of the signal threshold with respect to σ_y is:

$$\frac{\partial \bar{y}}{\partial \sigma_y} = 2a\sigma_y - \frac{\phi(\log(x_m))}{1 - \Phi(\log(x_m))}$$

The sign of this derivative depends on the magnitude of σ_y.

(i) If $\sigma_y > \frac{\phi(\log(x_m))}{2a(1-\Phi(\log(x_m)))}$, (i.e., the movie has a high variance on the potential outcome),

$$\frac{\partial \bar{y}}{\partial \sigma_y} > 0;$$

that is, a less precise signal (lager σ_y) raises the signal threshold. The signal threshold for non-white movies is higher.

(ii) If $\sigma_y < \frac{\phi(\log(x_m))}{2a(1-\Phi(\log(x_m)))}$, (i.e. the signal has good information on the movie revenue),

$$\frac{\partial \bar{y}}{\partial \sigma_y} < 0;$$

now, a less precise signal (larger σ_y) *lowers* the signal threshold. One can be a bit more tolerant of a bad signal for non-white movies, because it is difficult to say, based on the signal alone, whether the movie is really bad.

Although the signal threshold for non-white movies could still have been either higher or lower than that for white movies, the result does not depend on whether producers only want to produce very high quality movies, or they just want to weed out very low quality movies (i.e., it does not depend on whether the revenue threshold π_0 is above or below the prior mean of π,w which is in contrast to the normal-normal model.

B.2.3 Comparative statics for observed revenue, conditional on production: simulations

As in Section B.1.3 we use simulations to characterize the posterior distribution of observed revenue, conditional on production. We choose the parameters of the Pareto distribution to roughly mimic the observed distribution of revenue in our sample. Therefore, in all simulations, we set $x_m = 20$ (roughly equal to the minimum observed revenue in our sample) and set the baseline value of a at 0.2.[3] In this case, $\frac{\phi(\log(x_m))}{2a(1-\Phi(\log(x_m)))} \approx 8.2$, so we choose $\sigma_y = 8$ as the baseline. The prior mean is $1/a + log(x_m) \approx 8$, so we choose $\pi_0 = 8$ as the baseline. The results of the simulations are presented in Table B.2.

[3]The maximum likelihood estimate of a in our full sample is 0.09; 0.18 if one excludes the bottom 10% of the distribution; and 0.22 if one excludes the bottom 25%. We chose a slightly higher value of a as the baseline in our simulations because lower values of a will result in an implausibly large fraction of movies with explosive revenues (the mean of a Pareto distribution with $a < 1$ is infinite).

Table B.2: Simulation results: Pareto distribution

A: Taste based discrimination: $\pi_0 \uparrow$ for Non-white movies											
	Fixed $a = 0.2$, $\sigma_y = 8$, $x_m = 20$										Trend
π_0	5	7	9	11	13	15	17	19	21	23	
mean	8.091	8.164	8.252	8.374	8.548	8.767	9.041	9.381	9.800	10.309	$\uparrow$
std	5.041	5.081	5.108	5.173	5.270	5.362	5.521	5.686	5.888	6.137	$\uparrow$
	Fixed $a = 0.5$, $\sigma_y = 8$, $x_m = 20$										Trend
π_0	5	7	9	11	13	15	17	19	21	23	
mean	5.697	5.858	6.020	6.215	6.474	6.747	7.105	7.373	7.844	8.518	$\uparrow$
std	2.545	2.670	2.781	2.951	3.141	3.327	3.621	3.731	4.094	4.364	$\uparrow$

B: Customer discrimination: $a \uparrow$ for Non-white movies											
	Fixed $\pi_0 = 8$, $\sigma_y = 8$, $x_m = 20$										
a	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	
mean	13.056	9.786	8.198	7.315	6.769	6.428	6.201	6.040	5.944	5.837	$\downarrow$
std	10.012	6.711	5.083	4.168	3.587	3.242	3.008	2.833	2.742	2.628	$\downarrow$

C: Statistical discrimination: $\sigma_y \uparrow$ for Non-white movies											
	Fixed $a = 0.2$, $\pi_0 = 8$, $x_m = 20$										
σ_y	1	2	3	4	5	6	7	8	9	10	
mean	9.736	8.344	8.156	8.112	8.114	8.139	8.156	8.203	8.254	8.324	?
std	5.082	5.072	5.045	5.026	5.027	5.061	5.078	5.087	5.128	5.168	?
	Fixed $a = 0.5$, $\pi_0 = 8$, $x_m = 20$										
σ_y	1	2	3	4	5	6	7	8	9	10	
mean	6.749	5.469	5.327	5.361	5.447	5.585	5.750	5.930	6.151	6.360	?
std	2.216	2.167	2.152	2.203	2.278	2.412	2.565	2.725	2.933	3.118	?

Legend: simulated data with sample size $L = 10^6$, using R with seed 123. Mean, Std: sample average and standard deviation of the posterior distribution of $\pi|y, y > \bar{y}$ from the simulation.

In Panel A we examine the role of taste-based discrimination. We fix the values of a and σ_y and study what happens to the posterior mean and standard deviation of (log) revenue conditional on production as we increase π_0. The posterior mean increases (as in the normal-normal case), while the standard deviation also increases which is different from the normal-normal case.

In Panel B we look at the effect of increasing customer discrimination by letting a increase. Both the posterior mean and standard deviation decrease as the extent of customer discrimination increases, matching the predictions of the normal-normal model.

Finally, in panel C we vary the extent of statistical discrimination by letting σ_y increase, i.e., making the signal less precise. Here the results stand in contrast with those of the normal-normal model: as the signal becomes less precise, both the posterior mean of log revenue and the posterior standard deviation have a U-shaped pattern, first decreasing and then increasing in the extent of noise in the signal.

The comparative statics in the Pareto model are not identical to those in the normal-normal model presented in the main text. However, the simulations show that both the mean and the variance of log box-office revenue always move in the same direction as we change the discrimination parameter, under all three forms of discrimination. This is in contrast with the observed patterns in the data, where the mean of log revenue is higher for non-white movies, but the variance of log revenue is smaller (see Tables 6 and **??** in the text).

Appendix C Machine learning algorithm for facial classification

For performers that were not unambiguously classified by the human raters, we applied the facial classification algorithm proposed by Anwar and Islam (2017)[1]. The algorithm is based on a machine learning architecture that combines a convolutional neural network (CNN) and support vector machine (SVM), described below.

Step 1. We started with a sample of more than 7000 motion pictures released in the United States between 1997 and 2017, taken from Opus Data,[2] a private company that collects data on the industry. For each movie, we took the names of the four top-billed performers. We then scraped and cropped the image appearing on each performer's page on the popular website IMDB.[3]

Step 2. We used the Visual Geometry Group[4] (V.G.G.) technique to locate the actor's face on each picture. The output of this step is a vector of information extracted from each image, or a "feature vector."

Step 3. We repeated step 2 on our training data set, the Chicago Face Database (CFD).[5] This database is intended for use in scientific research. It is useful as it contains images of 597 unique individuals (both male and female) who self-identify as White, Black, Asian, or Latino/a.

Step 4. We used CFD to train our algorithm using the Support Vector Machine (SVM) approach.[6] Intuitively, the purpose of SVM is to find the "best separation line," meaning the hyper-plane that correctly separates white from non-white performers when such performers are located in a multi-dimensional space through their feature vectors.

Step 5. We applied our trained algorithm to the pictures obtained from Steps 1 and 2. We validated our algorithm on a subsample of actors for which we manually coded the racial groups and obtained a success rate of 95%. A few examples of the outcomes of our

[1] Link: https://arxiv.org/ftp/arxiv/papers/1709/1709.07429.pdf.
[2] www.opusdata.com
[3] www.imdb.com
[4] See for reference https://www.robots.ox.ac.uk/~vgg/.
[5] The CFD is available at https://www.chicagofaces.org/.
[6] See for reference https://scikit-learn.org/stable/modules/svm.html.

classification algorithm are presented in Figure C.1.

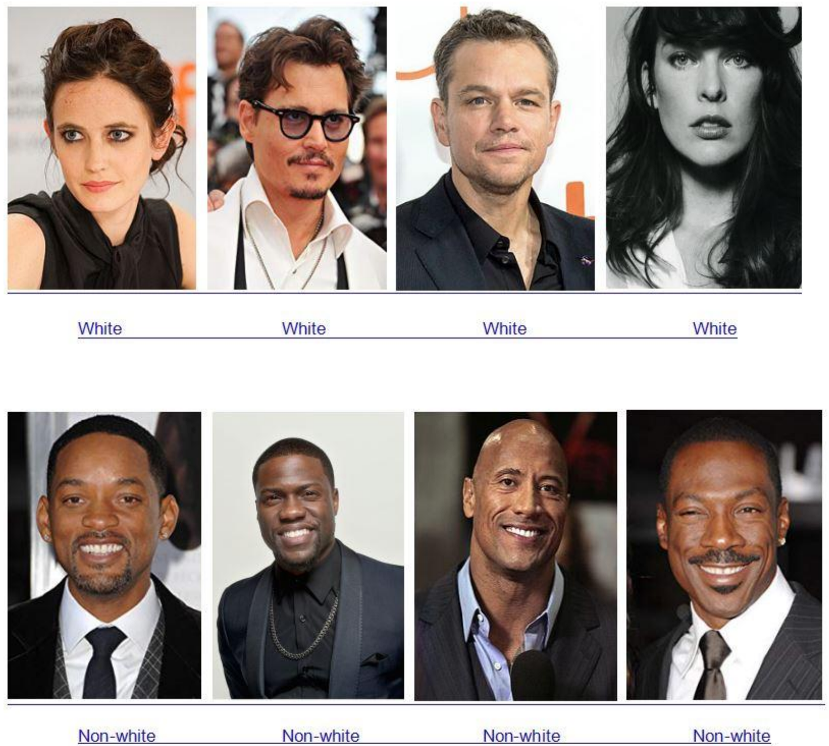

Figure C.1: Output of facial classification

Appendix D Other Figures

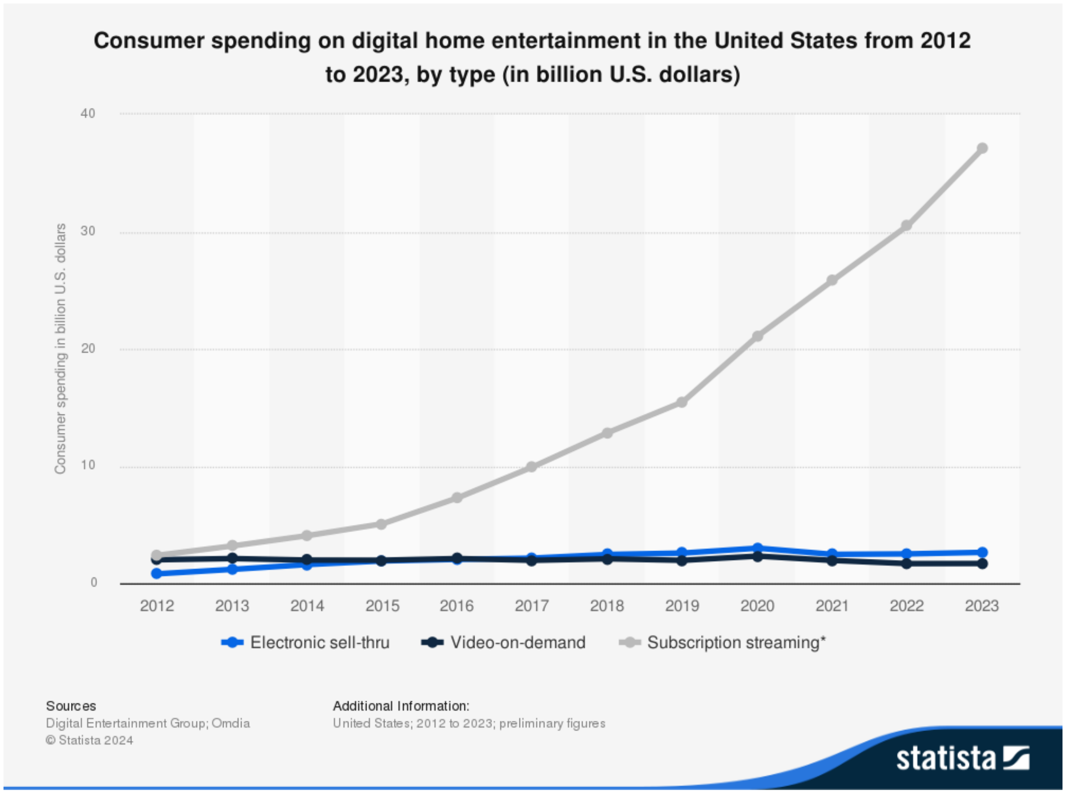

Figure D.1: Trend in consumer spending on digital home entertainment, by category
Source: Statista (link.)

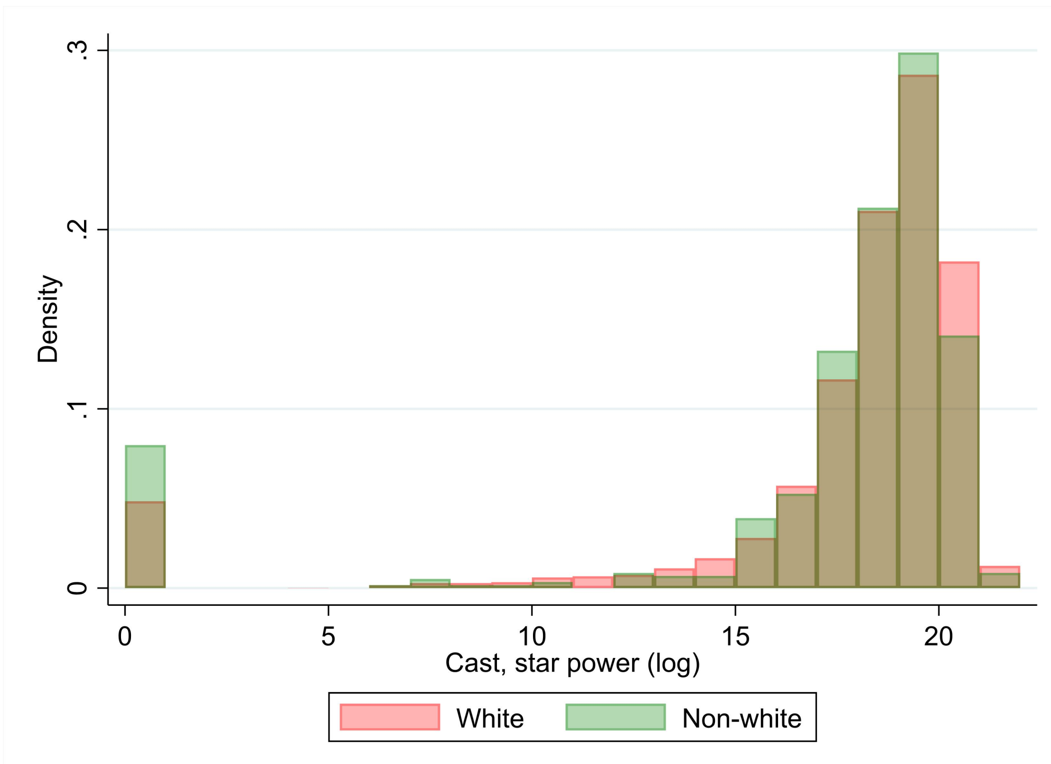

Figure D.2: Distribution of log cast star power across racial movie types
Note: A one-sided t-test on the means calculated off the full distributions fails to reject the null hypothesis that the white average is larger than the non-white average. Excluding the left tail of the distributions (i.e., truncating the distributions from below at 5) makes the means non significantly different.