Received: 7 July 2021 | Revised: 2 October 2023 | Accepted: 21 December 2023

DOI: 10.1002/smj.3575

RESEARCH ARTICLE

SMS | Strategic Management Journal WILEY

Status and consensus: Heterogeneity in audience evaluations of female- versus male-lead films

Bryan K. Stroube[1] | **David M. Waguespack**[2]

[1]Strategy and Entrepreneurship Area, London Business School, London, UK

[2]Management & Organization Department, Robert H. Smith School of Business, University of Maryland, College Park, Maryland, USA

Correspondence
Bryan K. Stroube, Strategy and Entrepreneurship Area, London Business School, Regent's Park, London NW1 4SA, UK.
Email: bstroube@london.edu

Abstract

Research Summary: Extant research finds that status characteristics such as gender are frequently related to average quality evaluations by external audiences, but little is known about whether such characteristics are also related to consensus in quality evaluations. We examine 383 million film ratings by consumers to document that female-lead movies elicit less consensus in quality evaluations than male-lead movies. In split-sample analyses, we find that male raters are more negative than female raters about female-lead titles, and that the two audiences differ on dispersion and skew. A subsequent experiment helps distinguish between various mechanisms that might be driving these results, including actor sorting, audience sorting, and treatment effects on audience quality perceptions. Finally, we find that independent studios yield greater box office revenue from female-lead movies.

Managerial Summary: Consumers often lack consensus about product quality. Does product gender-typing influence perceived quality consensus? We examine this question in the film industry, where 28.5% of films from 1992 to 2018 had a female actor in the lead role. Using 383 million consumer ratings from a popular website, we find less consensus in ratings of female-

lead films compared to male-lead films. Some of this effect stems from male audiences who, compared to female audiences, rate female-lead films lower than male-lead films and disagree more on their quality. We use an experiment with fictional AI-generated movie plots and random lead-actor gender to better understand what drives this effect. Finally, we find independent studios have higher box office revenue from female-lead films.

KEYWORDS

audience evaluations, film industry, gender, ratings, status

1 | INTRODUCTION

Do consumers tend to agree more on the quality of products associated with men as opposed to women? There is substantial work documenting that status characteristics, such as gender, independently influence the *average* quality evaluations delivered by various external audiences (Abraham, 2017; Cattani et al., 2014; Lee & Huang, 2018; Waguespack & Sorenson, 2010). An observed shift in the relative mean quality rating between a low- and a high-status object, however, does not imply that all consumers perceived equivalent quality differentials, or even regarded the low-status object as inferior. As such, an understanding of how status influences commercial viability requires examining variations in reactions to producer identity.

In this paper, we propose that the same mechanisms that arguably drive mean gender bias, widely held descriptive and prescriptive beliefs (Correll et al., 2017; Heilman & Eagly, 2008), may simultaneously translate to greater evaluative dispersion for a number of related reasons. First, when status serves as a quality heuristic, the amount of ambiguity present may vary substantially among audience members. Second, when status invokes ideological preferences, extreme reactions are possible, in effect making producer identity a primary rather than a secondary screening attribute. If either of these mechanisms is not constant across consumers, then low-status producers will necessarily face higher variance in how they are evaluated. Further, we believe the idea of culturally dominant product types—for example, male-lead action movies—is consistent with audience sorting mechanisms that may also contribute to greater variance in evaluations of less "conventional" products (Bourdieu, 1984). In short, we propose that status characteristics may be associated with evaluative distributions that have different central tendencies, as the literature has established, as well as greater dispersion.

We explore these issues in the context of 4012 general audience movies theatrically released in the United States from 1992 to 2018, for which consumers provided 383 million review scores on a 1–10 scale. Our main finding is that there is greater rating variance for female-lead titles, even though, and consistent with extant research on status, films with female lead actors receive average rating scores that are slightly lower than those with male leads. Splitting individual ratings into subsets coming from women ($n = 74$ million) and men ($n = 308$ million), and comparing differential reactions to the same movie, we show that the female-lead penalty exists for both male and female audiences but is substantially larger for male audiences. Male

audiences, however, are also more likely to disagree on the quality of female-lead movies than male-lead movies. Moreover, we find evidence of evaluative extremism, in that a female lead actor changes the skew of ratings for a given film in opposite directions for each audience subsegment: the skew of ratings from male audiences becomes more negative, with the mass of the distribution shifting towards the left tail, and female audience rating skew becomes more positive. We found consistent results using a survey experiment where we asked participants to evaluate fictional plots with a random lead actor gender. Importantly, the experiment also allows us to rule in a number of potential mechanisms related to evaluator sorting and quality uncertainty.

Finally, in our observational data, we show that female-lead titles, as should be expected from the variance and skew effects, indeed tend to have more raters in the right tail (quality rating of 8–10) than male-lead movies with the same average rating. For the 81% of titles with "mediocre" average ratings in the 5–7.5 range, we then compare the correlation between the presence of a female lead and (1) logged box office revenue and (2) logged male and female rating counts (a crude proxy for differential revenue from male and female audiences). For major studios, which typically seek broad audience appeal, the presence of a female lead is associated with a moderate decrease in box office performance and a substantial dip in male attendance after accounting for other movie content attributes. Conversely, for independent studios, which typically seek a more targeted audience, female leads are associated with greater performance. These differences and trade-offs, female leads moving the rating mean downwards and the right tail upwards, may partially explain why the presence of female leads has increased over time at independent studios while remaining flat at the majors.

This study makes several contributions to research on status and gender. First, to our knowledge, it is among the first to establish a link between a status characteristic and formal measures of consensus. We explore a number of channels through which such a relationship can exist, including heterogeneous sub-audiences and heterogeneity within sub-audiences as well as audience- and actor-sorting mechanisms. Second, we highlight that the traditional focus on the mean effects of status may neglect the importance of dispersed tastes on performance outcomes, because depressing the mean does not necessarily decrease the size of the right tail. In particular, this finding underlines that smaller producers may find it fruitful to release atypical products even if there is a mean penalty for doing so.

In Section 2 of the manuscript, we outline the potential sources and consequences of evaluative heterogeneity. Sources of evaluative heterogeneity include variance in audience quality perceptions, audience sorting, and actor sorting. The consequences of evaluative heterogeneity largely relate to strategic positioning. In Section 3, we introduce the context and data. In Section 4, we turn to multivariate analysis of the relationship between film lead gender, consumer ratings dispersion, and box office performance. We also present evidence from a survey experiment that sheds light on the mechanisms observed in the observational data. We conclude with a discussion of the limits and generalizability of these findings.

2 | THE POTENTIAL SOURCES AND CONSEQUENCES OF EVALUATIVE HETEROGENEITY

The two main research questions in this paper are, first, whether gender influences consumer evaluative consensus, and, second, whether gender has differential effects on performance. In the following sections, we delve into these questions in turn. Before going further, we want to

first acknowledge that, even while our data analysis was guided by theoretical priors, we were aware of the empirical patterns observed and then utilized theory to explain those patterns. As such, more research and more definitive testing are clearly warranted.

Second, our core premise in the following discussion is that the film setting is a context traditionally dominated by men, and in which there is ample anecdotal evidence of anti-female biases. For example, the #MeToo movement is closely linked to the industry (Luo & Zhang, 2022). Male dominance is also apparent in the data. In our sample of theatrically released films, titles with female lead actors constitute a minority of the sample (28.5%), and those titles are on average placed on fewer opening screens (1704 vs. 2026) and on average have lower box office revenue ($42.3 M vs. $57.2 M).

The male–female differentiation in the film industry also shows up on the audience side, with titles featuring female lead actors receiving lower average consumer ratings (6.39 vs. 6.61 out of 10) on the Internet Movie Database (IMDb) than those with a male lead. As we will show in the analysis sections, ratings for female titles are also more dispersed. In the following subsections, we delve into complementary explanations of why we might observe greater consumer disagreement on female-typed products, before turning to a discussion of alternative notions of the commercial consequences of dispersion.

2.1 | Sources of evaluative heterogeneity

Why might *observed* ratings become more volatile—less prone to evaluative consensus—when an object is female-typed? In this section, we consider three sets of mechanisms drawn primarily from the literatures on gender and aesthetics. First, audience members may systematically vary in the extent to which gender is used as a heuristic for marginal quality or appropriateness. Second, female-typed products may attract a higher proportion of consumers with less rigid, and consequently less predictable, taste profiles. Third, female actors may tend to sort (or be sorted) into product types with low agreement among raters.

In a nutshell, our taxonomy of the relevant literature posits that differences between male- and female-typed *observed* ratings distributions will reflect a mix of "treatment effects" (audiences rate female-typed products differently) and "selection effects" (female-typed products attract different audiences; female-typed products are different on other attributes). Clearly, these explanations are not fully orthogonal. For instance, film producers and studios may conduct biased evaluations that influence hiring, and in turn make varied decisions about marketing. While we cannot fully address these alternatives in the archival data that comprise our main analysis, we will also present results from an experimental survey that attempts to more cleanly tease apart mechanisms.

2.1.1 | Audience quality perception

The conventional explanation for status advantage is that systematic prior beliefs about unobserved quality influence evaluation under uncertainty (Podolny, 1993; Ridgeway & Correll, 2004). As Correll and Ridgeway (2003, p. 32) state: "Status characteristics are attributes on which people differ (e.g., gender, computer expertise) and for which there are widely held beliefs in the culture associating greater social worthiness and competence with one category of the attribute (men, computer expert) than another (women, computer novice)." In essence, the

status theory contends that perceptions of identity and outputs/behavior are independent. Consequently, individuals and organizations endowed with a higher (or lower) social rank face reactions from other parties that are more positive (or negative) than their capabilities and accomplishments merit (Merton, 1968), a process sometimes described as "status-based discrimination" (Correll & Benard, 2006). Gender is a particularly salient attribute that influences cultural perceptions of worth (Ridgeway, 2011), to the extent that both male and female evaluators discount women compared with men (Heilman, 2012).

In line with this theory, a broad range of experimental and observational research has demonstrated an evaluative double standard for women compared with men (Foschi, 1996). For example, women receive smaller raises than men with equivalent average performance ratings (Castilla, 2008), female investors get less attention than their male counterparts (Botelho & Abraham, 2017), and female-managed funds receive fewer capital inflows than equivalent male-managed funds (Niessen-Ruenzi & Ruenzi, 2019). A parallel line of research has explored the scope conditions on gender status effects, looking at the congruency between identity and social role expectations associated with different behaviors and contexts (Eagly & Karau, 2002). For instance, business plans from female founders are seen as more viable if oriented towards social impact (Lee & Huang, 2018), female film directors are rated more poorly except when creating unconventional films (Parker et al., 2020), and the gap between business referrals offered to men and women increases for male-typed occupations (Abraham, 2020).

Finally, related work has extended differential gender identity treatment to social categories and objects that become associated with men or women. For example, occupations viewed as "women's work" are often devalued (Cohen & Huffman, 2003). Likewise, Tak et al. (2019, p. 556) offer evidence that "... many product have gendered associations, which offers empirical evidence that gender-typing is pervasive in product markets." Moreover, even presenting the same identical product as more or less masculine can change how it is evaluated (Worth et al., 1992), and thus it is not only people but occupations, products, and other social constructions that can become socially ascribed with specific gender traits that are at risk of altering quality evaluations.

With that preamble in mind, the topic of quality consensus is one on which we were unable to find either direct empirical work or direct theorizing. However, there is an emerging line of inquiry that abuts the relationship between status, audience characteristics, and consensus: Dimitriadis et al. (2017) find that the appropriateness of female involvement in social ventures varies as a function of local cultural standards; in a study of crowd-funding, Greenberg and Mollick (2017) show that female investors are more likely to support female candidates; in law firms, Carnahan and Greenwood (2018) establish that the political ideology of partners influences gendered hiring; and Koning et al. (2021) demonstrate that female inventors increase the supply of inventions that benefit and appeal to women.

Our particular interest is in whether "widely held" beliefs about gender-typing are, in theory, largely uniform or systematically variable. If variable, then consensus *must* decrease when an object is gender-typed regardless of what happens to the mean. We postulate two reasons why evaluator beliefs may be variable. First, there may be systematic variance within audiences regarding how status maps to quality. A variety of studies posit that person/object status is a secondary *descriptive* attribute (Heilman, 2012), deployed heuristically by assessors when quality is uncertain (Abraham, 2020; Benjamin & Podolny, 1999; Botelho & Abraham, 2017; Correll et al., 2017). Consequently, for instance, when the quality and cost of acquiring information improves, the use and utility of status-based heuristics will tend to diminish (Simcoe & Waguespack, 2011). Thus, there is little reason to believe that everyone within an audience will

infer exactly the same thing from a status characteristic; and if they do not, this must increase the variance in reactions. Even while evidence to date of differential bias between male and female evaluators is scant (Heilman, 2012), it seems possible that the utility of status heuristics varies systematically within an audience. Note here that "quality" is a fairly expansive term, in that it could refer to beliefs about how others would evaluate the same object (Correll et al., 2017) or to internally held beliefs about quality if the object under consideration was associated with the opposite gender type.

Second, and resonant with the Heilman *prescriptive* attributes concept, there may be systematic disagreement on the legitimacy of the status hierarchy (Frake, 2017; Greenberg & Mollick, 2017; Kuppuswamy & Younkin, 2020; Siegel et al., 2019). Social identity provides an explanation for why individuals personally care about and respond to status characteristics such as gender. In the context of product markets, "identity is defined as any category label to which a consumer self-associates that is amenable to a clear picture of what the person in the category looks like, thinks, feels and does" (Reed et al., 2012, p. 310). Thus, any variance in the nature or strength of identity within an audience may also translate into variance in how status characteristics are ultimately evaluated.

For example, some male audiences may have a greater propensity to regard the presence of females as a threat (Rudman et al., 2012; Willer et al., 2013). Conversely, some women may embrace and feel empowered by the presence of women. In either case, status transitions from a marginal heuristic for quality assessment to a primary and potentially substantial influence on evaluation. From this view, evaluation is largely decoupled from perceptions of the quality of the object or person under scrutiny and instead reflects preferences for a particular social order. Theories explicitly focused on gender identity also lend support to the idea that not everyone will respond at the same magnitude to gendered objects. For example, the "precarious manhood" thesis argues that "manhood, in contrast to womanhood, is seen as a precarious state requiring continual social proof and validation" (Vandello et al., 2008, p. 1325). Two implications of this theory are that men experience more anxiety and stress than women when their gender is challenged and that men will "eschew feminine behaviors, preferences, traits, and desires" (Vandello & Bosson, 2013, p. 105). This implies that men's reactions to gender-typed objects should vary according to how they perceive their own personal masculinity. Any variance across men in perceived masculinity should lead to more variance in how men as a group—compared with women as a group—respond to female-typed products versus male-typed products.

More generally, there may be reason to believe that such heterogeneity in the prescriptive employment of status attributes is not limited to comparisons across audiences (e.g., men/women) but also occurs within audiences. This is in part because the use of gender schema for making sense of the world varies across individuals (Bem, 1981). For example, Worth et al. (1992) find that experiment participants who were provided beer described in masculine versus feminine language rated it higher when they viewed themselves as more masculine; and Cooper (1997) finds that women who are "traditional" view other women in leadership roles differently than do women who are "nontraditional." Such variance may naturally lead to nonuniform and even extreme reactions. For example, if male chauvinism is limited to only some men, then one would expect average penalties in how women are evaluated to also be associated with higher variance in how men evaluate women simply because it is a minority of men applying a larger penalty. This kind of variance in ideology is consistent with survey evidence that measures related views: in the 2018 General Social Survey, 28% of male respondents—some but not all—agreed or strongly agreed with the statement that "It is much better for everyone involved if the

man is the achiever outside the home and the woman takes care of the home and family."[1] Prescriptive beliefs about the appropriateness of gender roles are not uniformly held within gender, meaning they provide an external mechanism by which consensus in the quality of female-lead movies might be lower than for male-lead movies, both across and within sub-audiences.

In closing this section, note that the descriptive and prescriptive mechanisms discussed above are grounded in beliefs and preferences that are distributed within an audience rather than directly encapsulated in product features. Therefore, what our empirical analysis of the observed relationship between gender typing and perceived quality dispersion will accomplish is to shed some light, in the context of movies, on the character of the heretofore unobserved "widely held beliefs" that are reasoned to drive differential treatment. Put more succinctly, if audience members vary on the strength of *descriptive* and/or *prescriptive* priors when comparing counterfactual male- and female-typed objects, then the variance of perceived quality *must* increase regardless of whether the mean changes. Ascertaining the extent to which differences result from descriptive mechanisms (gender type is a secondary marginal attribute) versus prescriptive mechanisms (gender type is a primary attribute) is more challenging. A change in evaluative skew is arguably consistent with prescriptive responses, but not dispositive.

2.1.2 | Audience sorting

In the preceding section, we posited that, among those who choose to consume and rate films, consumer perceptions of female-typed product quality are potentially more volatile than the male counterfactual. In this section, we propose a complementary line of reasoning: consumers who choose to consume and rate female-typed products may have systematically more varied evaluative procedures, on average, than those who choose to consume and rate male-typed products.

In his work on aesthetic judgments, Bourdieu (1984) argues that class-based differences in artistic consumption are inherently social, and partially reflect differences in cultural capital. Put succinctly, "cultural capital" represents an individual's underlying knowledge and sophistication with respect to cultural products. One manifestation of cultural capital is variance in artistic preferences across social strata. For instance, Bourdieu (1984, p. 28) reports results from a wide-ranging survey of French consumers. When asked to identify a favorite musical piece, the modal favorite of lower-class respondents is the relatively low-brow "Blue Danube," while among the most elite respondents, the modal favorite is the higher-brow "Well-Tempered Clavier."

At a base level, these class differences may simply represent identity-based homophily: expressing taste preferences is an act of group identification. In the movie context, for example, it is possible that titles with female leads attract a greater proportion of female viewers who prefer female leads, and it is this shifting of the male/female audience ratio that produces greater observed variance.

More importantly, for the purposes of this paper, is Bourdieu's argument that those class differences also represent differences in methods for evaluating cultural objects. Evaluators with lower cultural capital will tend to have narrow conventional definitions of the constituents of quality when compared to those with higher cultural capital. For instance, in a particularly

[1]The corresponding number for female respondents was 22%, indicating variance both across and within male and female respondents. See "fefam" variable at https://gssdataexplorer.norc.org/.

vivid illustration of what Bourdieu (1984, p. 58) labels "aesthetic distancing," survey participants were shown a detailed photo of the gnarled hands of an old woman. When asked to remark on whether the photo was beautiful, lower-class respondents tended to focus on the photo's object and the pain it obviously implied. Higher-class respondents were more likely to respond positively and abstractly, invoking connections to related art and history. The upshot, we argue, is that greater cultural capital implies less predictability in tastes. Indeed, returning to the issue of favorite compositions: among the lower class 65% name "Blue Danube" and 1% "Well-Tempered Clavier," while among elites the same proportions are 11.5% and 29.5%.

A variety of scholars have utilized the notion of audience heterogeneity on evaluative schemas. Shrum (1996) argues that professional critics approach popular and fringe theatrical work with different expectations. Rindova and Petkova (2007) propose that customer perception of the value of technological innovation, which is inherently incongruous, is shifted by product design choices that better align with consumer expectations. Kovács and Sharkey (2014) note that when literary books win prestigious awards, consumer ratings tend to decline, in part because the reading audience expands to those with more conventional tastes. Kim and DellaPosta (2021) argue that some participants on a beer-rating platform are motivated to signal their evaluative skill and discernment. Finally, Cancellieri et al. (2022) posit that consumer reaction to innovative (as opposed to traditional) operatic productions is mediated by the prior expertise of the consumer.

How then do Bourdieu's ideas, based on observation of consumption of low/high culture across French social classes, map onto consumer ratings in the US mainstream film industry? In ethnographic work applying Bourdieu's theories to the US market, Holt (1997) advises scholars to focus on consumption practices and mass culture. Accordingly, we posit that in the mainstream film industry, the conventional dominant title, the US equivalent of "Blue Danube," is an action-oriented big-budget project with a male lead. The conventional film will tend to attract a high proportion of consumers with circumscribed and conventional consumption practices. In contrast, titles with female leads are less-conventional fringe objects, and will tend to attract more audience members with more heterogeneous tastes and evaluation criteria. In short, female-typed titles may attract raters with more varied tastes and more varied schemas, which in turns lead to observed ratings distributions that are more dispersed.

As a final consideration here, while we have discussed treatment and selection effects as complementary mechanisms, there are plausible scenarios in which homophilous selection entirely drives increases in dispersion of observed aggregated ratings. Supporting Information Appendix section A.1 details via simulations how the mechanisms of status deference (on average male leads are more highly regarded by all) and preference (on average each rater prefers leads that match their gender) might play out. The upshot of that discussion is that disentangling male and female audiences' reactions is critical for theoretical clarity.

2.1.3 | Actor sorting

The final potential explanation for differences in evaluative heterogeneity between female-lead and male-lead films relates to differential actor sorting. There is ample evidence in labor market research that women have unequal access to the same jobs as men as a result of discrimination in the hiring pipeline (Fernandez-Mateo & Fernandez, 2016; Fernandez-Mateo & King, 2011). If, in our setting, the labor market pushes female actors into different types of films than male actors—for example, less-conventional and niche-oriented plots—then it may not be female

lead actors per se that face less consensus. Rather, these are objects that tend to have highly variable appeal, and the gender association is coincidental.[2] Note we view this alternative explanation primarily as an empirical challenge. In our observational data, we will include controls to attempt to account for underlying differences of films, such as genre, but these are obviously imperfect. However, our experiment will be able to more robustly rule out this explanation by randomly assigning lead-actor gender to movie plots.

2.2 | Consequences of evaluative heterogeneity

The preceding section argues that male- and female-lead films are likely to face different levels of consensus in their quality evaluations. But what are the potential consequences of such disagreement? It is possible that consensus is linked to overall performance via a right-tail effect. Holding the mean evaluation constant and flattening the distribution pushes more people into both tails of the distribution. However, in product markets, overall product performance is strongly driven by the right tail—how many people love the product. How many people dislike the product is clearly less relevant.[3]

Therefore, shifts in mean quality assessment—as studied by prior research—may mask qualitatively different types of response heterogeneity that, in turn, matter for commercial viability. For instance, for producers targeting narrower audiences, the right tail of the distribution that "loves" a product is arguably much more important than the average reaction to the product. This insight is a basic tenet of strategic positioning (Porter, 1991), so we do not elaborate on it further in the main text. However, section A.2 in the Supporting Information Appendix presents a number of simulations that demonstrate how critical it is to investigate the distributional properties of evaluations and not simply the mean. In short, different underlying processes can generate the same mean outcomes yet very different outcomes with regard to consensus measures and performance. As a consequence, in Section 4.2, we will examine the differential box office performance of female-lead movies produced by independent studios compared to the major studios.

3 | CONTEXT AND DATA

The context for the study is the American film industry. Our empirical sample consists of 4012 theatrically released general-audience titles released from 1992 through 2018. The starting point for this sample was Kids-in-Mind—an independent organization that produces detailed reports of potentially objectionable content in films—which began reviewing general audience theatrically released movies in 1992 (Waguespack & Sorenson, 2010). We removed movies with MPAA ratings of "G" and "NC-17," as they are deliberately targeted to specific audiences. We then further limited the sample to movies that were rated by at least 1000 men and 1000 women. Table 1 reports descriptive statistics for our analysis data, and correlations are reported in

[2]We expect that perceived quality is partly driven by the fit between variable observable "horizontal" product attributes and variable consumer tastes. For instance, assuming all else is equal in terms of materials and manufacture, left-handed guitars and right-handed guitars are quite different in terms of mass-market appeal. We thank an anonymous reviewer for suggesting this analogy.

[3]Note this does not apply to evaluations conducted by groups, such as hiring committees, where a negative evaluation of a candidate by one evaluator may directly offset a positive evaluation by another evaluator.

TABLE 1 Descriptive statistics.

Statistic	Mean	St. dev.	Min	Max
Female lead actor	0.29	0.45	0	1
Amount of violence	5.01	2.20	0	10
Amount of sex/nudity	4.10	2.29	0	10
Amount of profane language	5.05	2.58	0	10
Major studio (0/1)	0.53	0.50	0	1
Opening theaters (log)	6.42	2.51	0.00	8.42
Total genres	2.59	0.63	1	3
Pooled audience ratings				
Rating mean	6.55	0.90	2.48	9.15
Rating Stdev	1.85	0.32	1.22	4.18
Rating skew	−0.51	0.48	−3.32	1.91
Male audience ratings				
Rating mean	6.46	0.94	2.36	9.16
Rating Stdev	1.82	0.31	1.21	4.31
Rating skew	−0.51	0.49	−3.36	2.04
Rating count (log)	10.52	1.25	7.13	14.13
Female audience ratings				
Rating mean	6.76	0.84	2.78	9.24
Rating Stdev	1.95	0.32	1.23	3.89
Rating skew	−0.53	0.45	−3.14	1.53
Rating count (log)	9.15	1.17	6.91	12.50
Box office gross revenue (log)	16.92	1.64	5.58	20.66

Tables A.1 and A.2 in Supporting Information. The raw data and code required to reproduce the main analyses can be accessed at the link provided in the Open Research section at the end of the paper.

3.1 | Independent variables

Our primary independent variable of interest is whether the principal actor on the title is female *(Female lead actor)*. The gender of the lead actor in each film was collected from IMDb, where each film in the sample has either a male or a female actor listed in the principal role.[4] Only 29% of films in the sample have a female lead actor (we report mean differences between male-lead and female-lead movies in Table A.3 in Supporting Information). This binary variable

[4]Actor data come from the official cast and crew dataset released by IMDb (imdb.com/interfaces/). This file typically includes the top four acting roles for each film, listed as either "actor" or "actress." The ordering of these principals appears consistent with the "Stars" list for each film that is managed by IMDb staff and is displayed on each film's main public web page.

construction has the advantage of simple comparisons across gender type, but it may introduce noise given that movies typically feature multiple significant leads, and relative star power may outweigh relative contribution in determining credit order. In robustness checks following our main analysis below, we show that we obtain consistent results with alternative variable constructions, such as the ratio of female actors in the first one to four credit positions.

We also employ control variables from three different sources. We use the three separate 0–10 numerical scores produced by Kids-in-Mind that reflect a film's level of *Violence*, *Sex/nudity*, and *Profane language* (Waguespack & Sorenson, 2010). Marketing aspects of the film are accounted for using data from Box Office Mojo, a film-industry data aggregator. These data include the number of opening theater screens the movie was shown on (*Opening theaters*), the month the film was released (to account for seasonality), and the year of release. We also created an indicator variable for whether a film was released by one of the *Major studios* (Warner Bros., Buena Vista, Universal Pictures, Twentieth Century Fox, Sony Pictures Entertainment, Paramount Pictures, and Metro-Goldwyn-Mayer). Genre data come from IMDb, which documents up to three different genres for each film. We created a count of the *Total genres* to which the film belongs as well as separate indicators for each genre to which a film could belong (documentary films were excluded).

3.2 | Dependent variables

For each film in our sample, we calculated ratings distribution attributes based on individual audience quality evaluations collected from IMDb, the largest online source of audience movie ratings. IMDb allows the general public to rate films using a 1–10 numerical scale, with 1 representing the lowest quality and 10 the highest quality. We gathered every rating for the movies in our sample that was provided by IMDb users who had a gender associated with their IMDb account.[5]

To illustrate the nature of the underlying ratings data as well as the cross-gender comparisons we make, Figure 1 reports the pooled histograms of 1–10 ratings for two average titles. In the left panel is a male-lead title, *From Paris with Love*, an action film starring John Travolta. In the right panel is a female-lead title, *The Congress*, a sci-fi drama starring Robin Wright. These two titles have pooled mean ratings of 6.55 that are identical to our sample mean. The male-title distribution is clearly less dispersed, however, and the relative standard deviations are 1.77 and 2.04. Neither distribution is symmetrical, and the relative asymmetry is quite similar, with skews of −0.58 and −0.61. Note that these ratings distribution "moments," the mean, standard deviation, and skew, constitute our main movie-level dependent variables.

[5]IMDb reports counts at each rating level by audience type. We collated this data manually, by clicking on the ratings link for each of the 4012 movies in the sample. Some reviewers also provide narrative reviews that accompany their score. While this potentially allows us to track some individual raters across film titles, the gender of those accounts is not revealed, nor are ratings where a narrative was not provided. Further, IMDb's terms of service explicitly prohibit bulk automated scraping of data. Finally, in our full sample 80% of IMDb ratings were submitted by (self-identified) men. We do not have a clear explanation for this imbalance, and it is possible that male and female IMDb raters systematically differ on other attributes. For example, the average female IMDb rater might be more "serious" about movies than the average male IMDb rater. However, as we will explore in the Discussion section in relation to professional critics, it is not immediately clear what, if any, effect this will have on ratings consensus, particularly in the split samples.

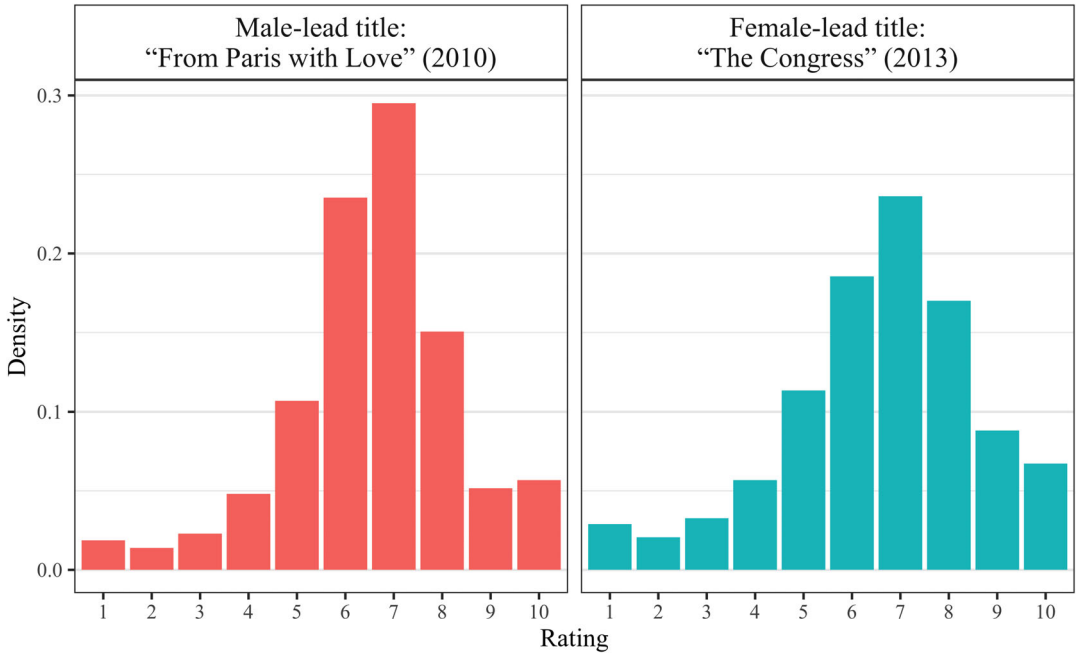

FIGURE 1 IMDb ratings histograms for a representative male-lead and a representative female-lead title. IMDb, Internet Movie Database.

Figure 2 further breaks out the ratings by scores from self-identified male and female raters, items that form the basis of the audience-specific distribution measures. While in these two examples, male and female audience ratings distribution moments are quite similar, one item not observable is rating counts. For *From Paris with Love*, female raters constitute 9.5% (8978 female vs. 85,383 male) of the total. For *The Congress*, the equivalent ratio is 18.6% (2583 vs. 11,232). Again, the distributional moments by audience are calculated at the movie level, as are rating counts by audience type.

The figures also clarify important issues with the multivariate analysis reported below. *From Paris with Love* and *The Congress* differ on more than just actors. We control for multiple observable film characteristics, but such a comparison is inherently sensitive to unobservable attributes related to content type or quality. By contrast, comparisons between male and female audiences are based on the same titles; thus, any differences capture divergent responses to the same stimulus.

To reiterate, the movies in our dataset were rated 74,056,616 times by women and 308,471,644 times by men. These ratings were used to calculate our main film-level dependent variables, which consist of the *Mean*, *Standard deviation (StDev)*, and *Skew* of the ratings distributions for the pooled audience (ratings from men and women combined), the male audience, and the female audience.[6] For the performance analyses, we use the logged *Box office gross* in

[6]*Skewness* is known technically as the third moment of a distribution, after the mean and standard deviation. A simplified version of the skew formula, suitable for calculation by hand, is (Mean − Median)/Standard Deviation. The intuition here is that skew captures symmetry: skew equals 0 when the mean and median are identical, is positive (skewed right) when the median is greater than the mean, and is negative (skewed left) when the median is less than the mean.

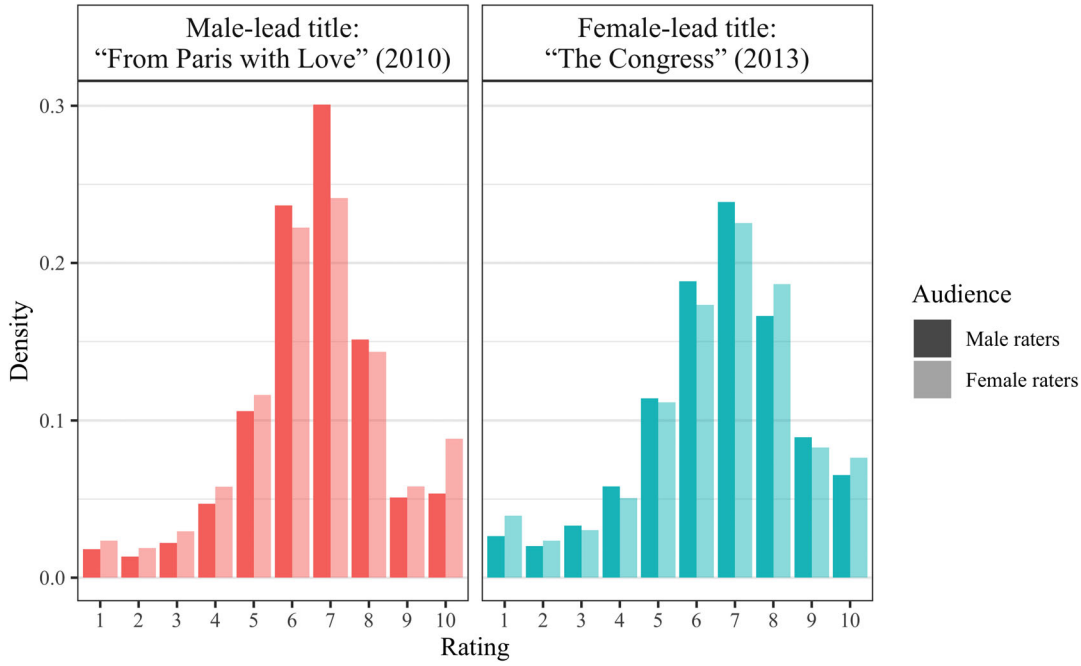

FIGURE 2 IMDb ratings histograms by audience gender for a representative male-lead and a representative female-lead title. IMDb, Internet Movie Database.

the United States (from Box Office Mojo) as well as the logged *Count* of reviews submitted by audience type.

3.3 | Data advantages and limitations

The film industry has a number of benefits for testing theories about audience perceptions of quality and the influence of status characteristics. First, movies are cultural products where "true" or unobserved quality does not exist in the same way that it does in settings where uncertainty about objective (yet unobserved) future quality is often a key motive for status-based discrimination (e.g., hiring decisions in labor markets). This is advantageous because post-consumption ratings are less likely to reflect beliefs about unobserved quality. Second, the setting allows us to study literal fictional "roles" that could have been theoretically occupied by men or women. In theory, the number of female-lead movies being produced could be altered at any time if producers in the industry chose to do so, and the percentage of women in lead roles does appear to be increasing. In Section 4.2 on performance, we explore which producers appear to be most likely to make female-lead films and how this is related to audience evaluations.

There are likewise some substantial limitations with the IMDb data that bear explaining before we move to the results. First is that all we know about a given rating is the movie, the 1–10 score, and whether the rater is male or female. We cannot attach ratings to individuals, and thus we cannot explore within-rater changes; nor do we know when the rating was posted, and thus we cannot explore within-title changes over time (nor how a rater might be influenced by prior ratings). Second, ratings are integers censored at 1 and 10. Consequently, the mean of a distribution has a direct effect on the possible range of the standard deviation and the skewness. For instance, for a

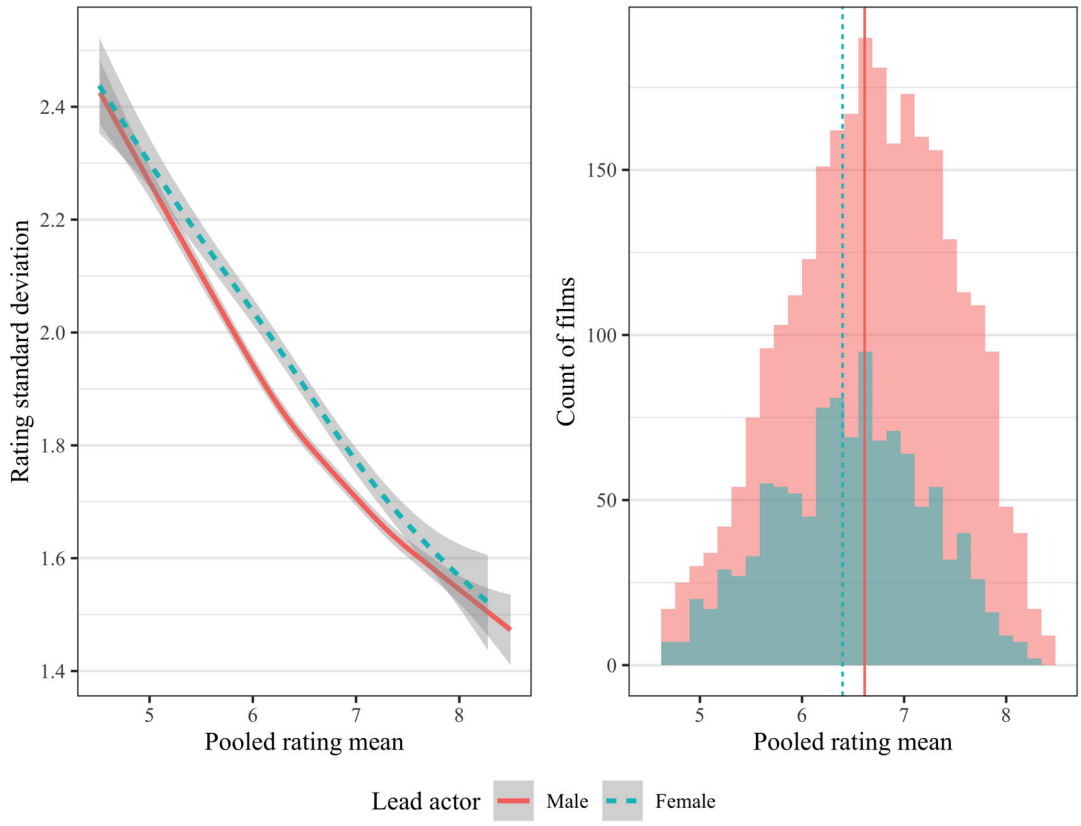

FIGURE 3 Mean standard deviation by mean rating and lead-actor gender (left), and overlaid mean rating histograms divided by male- and female-lead movies (right); the vertical lines are the mean ratings for each subsample. This figure excludes 88 titles with a mean rating below 4.5 and 24 titles with a mean above 8.5. The remaining sample size is 3942.

distribution with a mean of 8, the variance on the right side of the distribution, where most of the mass is located, is reduced because no rater can give a score greater than 10. For these reasons, our estimates of distribution standard deviation include the mean as a control variable, and the estimates of skewness include both the mean and the standard deviation as controls.

4 | RESULTS

Figure 3 reports a smoothed scatterplot of pooled rating *Standard Deviation* by rating *Mean* and *Lead-Actor Gender* on the left panel, and a histogram of rating *Mean* for reference on the right panel. One thing that is clear in the figure is that there is a negative relationship between mean rating and rating standard deviation. On the one hand, this relationship is likely somewhat mechanical in that scores are censored at 1 and 10, which limits the range on one side of a distribution. On the other hand, there is probably some behavioral element here, in that individual deviancy from the mean appears more likely for movies that are generally loathed than for those that are generally loved. In any event, and following on the discussion on data limitations above, this figure provides visual evidence of why controlling for the mean is important when comparing other aspects of dispersion in this setting.

The substantive takeaway from Figure 3 is that, at any given mean rating level, titles with female leads have perceived quality distributions that are more dispersed, and that this gap is larger for titles with mean ratings near the sample average. For example, 42.4% of titles in the sample have mean ratings between 6 and 7, and for these movies, the female-lead title standard deviation is roughly 0.1 points greater. While the figure is intriguing, we next turn to multivariate regression analysis of the relationship between gender type and rating dispersion for two reasons. First, is the relationship between actor gender and rating dispersion driven by differences in movie marketing and content? Second, can we further decompose the drivers of increased dispersion?

Table 2 presents regression results for pooled ratings distributions using OLS.[7] As expected, in model 1, *Female lead actor* is associated with a decrease in mean rating, all else held equal ($-0.171; p<0.01$). While this pattern is consistent with other work on evaluations, it is not possible to rule out the alternative interpretation that movies with female leads are of slightly worse quality. The omitted quality story, however, does not have clear implications for the shape of the distribution, and we find that a female lead also increases the *standard deviation* (model 2: $0.047; p<0.01$) and generates positive *skew* (model 3: $0.016; p<0.01$). These models include the full suite of controls for the amount of violence, sex/nudity, profane language, major studio, opening theaters, total count of genres, and indicator variables for each genre, release year, and release month. Note that we control for the mean in model 2, and for both mean and standard deviation in model 3, to deal with artificial constraints on the range of possible values due to censoring (i.e., the lowest possible rating is 1 and the highest is 10).

Table 3 repeats the above analyses, splitting results for the male and female audiences. These models are run using seemingly unrelated regressions so that the coefficients for male and female audiences can be statistically compared. It is clear in the table that male and female audiences respond differently to the same content. The first two models report the effects of a female lead on the mean rating provided by men (model 1) and women (model 2). Consistent with extant research on status characteristics, the female-lead penalty exists for both audiences. However, it is much larger for male audiences ($-0.273; p<0.01$) than for female audiences ($-0.060; p=0.027$), a difference of 0.213 ($p<0.01$). By implication, this difference extends to audience differences within the same movie (the same coefficients are obtained if we stack the data into male and female audience records for each title and then run a single model with audience × actor interactions and fixed effects for the title), meaning that movie-level omitted variables cannot explain why the audiences vary in their reception. The larger penalty for the non-identifying audience (men) suggests that some sort of identity mechanism is at play.

The third and fourth models report the effects of female lead using the standard deviation as the dependent variable. We again see differences across male and female audiences. Male audiences disagree more about quality when the lead is female ($0.032; p<0.01$). However, there is little evidence that women disagree more as a function of lead-actor gender ($0.009; p=0.223$). Therefore, a female lead results in less consensus for male audiences compared with female audiences ($0.023; p<0.01$).

Finally, the final two models report the results for skewness. Female leads appear to be associated with extreme reactions in opposite directions: male audience ratings become more negatively skewed ($-0.024; p<0.01$) and female audience ratings more positively skewed ($0.052; p<0.01$), a difference of 0.076 ($p<0.01$). Thus, the presence of a female lead moves

[7]Note that while the individual raw ratings are bounded integers, the dependent variables are all continuous measures at the film level.

TABLE 2 Effect of lead gender on pooled ratings.

	Dependent variable		
	Mean	**Stdev**	**Skewness**
	(1)	**(2)**	**(3)**
Female lead actor	−0.171	0.047	0.016
	(0.029)	(0.008)	(0.005)
Rating mean		−0.234	−0.504
		(0.004)	(0.004)
Rating Stdev			−0.127
			(0.010)
Violence	0.044	0.012	−0.008
	(0.009)	(0.003)	(0.002)
Sex/nudity	−0.052	0.012	−0.004
	(0.007)	(0.002)	(0.001)
Language	0.044	−0.015	−0.008
	(0.007)	(0.002)	(0.001)
Log(Opening theaters)	−0.080	0.006	0.018
	(0.006)	(0.002)	(0.001)
Major studio (0/1)	0.066	−0.040	−0.001
	(0.029)	(0.008)	(0.005)
Total genres	−0.008	−0.057	−0.055
	(0.197)	(0.054)	(0.034)
Constant	6.886	3.379	3.104
	(0.136)	(0.048)	(0.045)
Release year indicators	Yes	Yes	Yes
Release month indicators	Yes	Yes	Yes
Genre indicators	Yes	Yes	Yes
Observations	4012	4012	4012
R^2	0.279	0.562	0.925
Adjusted R^2	0.267	0.555	0.924

Note: The unit of analysis is a movie. Standard errors are in parentheses.

distributional symmetry in different directions for each audience type. Consistent with the idea that identification promotes divergent identity-based responses, female leads are associated with male audience ratings distributions becoming more left-skewed, and female audience ratings becoming more right-skewed.

Finally, section A.3 in the Supporting Information Appendix reproduces versions of the left panel of Figure 3 using a number of alternative definitions of "female-typed" movies, by the typicality of a film's lead actor's gender with respect to that film's genres, and by time periods. We find similar results using count of female actors rather than just lead actors (section A.3.1). The

TABLE 3 Seemingly unrelated regressions predicting rating outcomes from female and male audiences.

	Means		Stdev		Skew	
	Male Audience	Female Audience	Male Audience	Female Audience	Male Audience	Female Audience
	(1)	(2)	(3)	(4)	(5)	(6)
Female lead actor	−0.273	−0.060	0.032	0.009	−0.024	0.052
	(0.030)	(0.027)	(0.008)	(0.008)	(0.005)	(0.006)
Rating mean (M)			−0.192		−0.468	
			(0.003)		(0.003)	
Rating mean (F)				−0.253		−0.500
				(0.003)		(0.004)
Rating Stdev (M)					−0.085	
					(0.010)	
Rating Stdev (F)						−0.201
						(0.011)
Violence	0.049	0.034	0.011	0.019	−0.005	−0.015
	(0.009)	(0.009)	(0.002)	(0.002)	(0.002)	(0.002)
Sex/nudity	−0.052	−0.059	0.012	0.011	−0.002	−0.003
	(0.008)	(0.007)	(0.002)	(0.002)	(0.001)	(0.001)
Language	0.056	0.017	−0.015	−0.010	−0.005	−0.013
	(0.007)	(0.006)	(0.002)	(0.002)	(0.001)	(0.001)
Log(Opening theaters)	−0.092	−0.060	0.005	0.008	0.017	0.024
	(0.006)	(0.006)	(0.002)	(0.002)	(0.001)	(0.001)
Major studio (0/1)	0.079	0.037	−0.041	−0.035	−0.005	0.003
	(0.030)	(0.027)	(0.008)	(0.008)	(0.005)	(0.006)
Total genres	0.020	−0.051	−0.066	0.020	−0.044	−0.068
	(0.203)	(0.186)	(0.053)	(0.052)	(0.034)	(0.039)
Constant	6.838	7.143	3.078	3.640	2.763	3.280
	(0.141)	(0.129)	(0.043)	(0.044)	(0.043)	(0.053)
Year FEs	Yes	Yes	Yes	Yes	Yes	Yes
Month FEs	Yes	Yes	Yes	Yes	Yes	Yes
Genre FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,012	4,012	4,012	4,012	4,012	4,012
R^2	0.295	0.263	0.561	0.596	0.928	0.889
Adjusted R^2	0.284	0.252	0.554	0.589	0.927	0.887

Note: The unit of analysis is a movie. Standard errors are in parentheses.

male–female gap in consensus is stronger for films in relatively more "male" genres, and more "female" genres face less consensus (section A.3.2). We did not find strong evidence that the consensus gap has changed over time (section A.3.3).

4.1 | Experimental confirmation

The observational IMDb data have a number of strengths, including that they reflect hundreds of millions of real-world public evaluations of products in one of the largest cultural product industries across multiple decades. However, the data present some natural limitations related to identifying the relative extent of actor sorting, audience sorting, and treatment effects on audience quality perceptions. This is because lead actors are not randomly assigned to films, so viewers may be reacting to aspects of films that are correlated with actor gender rather than actor gender itself. Although we control for genre and other film characteristics in our regressions, there are likely other dimensions that we cannot observe. Evaluators may also choose to watch a movie based on actor gender, meaning our results reflect selection processes that we are unable to directly observe with our data. These issues do not invalidate our findings about how evaluators respond to male-lead and female-lead films, but unpacking them would better help explain what processes are driving our results. This was the motivation for the experiment described in this section, where we randomly assign actor gender and an independent quality signal (production crew experience) to movie content in a controlled setting.

4.1.1 | Experiment design

The main component for the experiment was a randomly constructed fictional "movie proposal" that was evaluated by experiment participants. Each movie proposal consisted of three elements: (1) a short description of a movie from 1 of the 20 different genres, (2) the level of experience of the production crew: experienced/inexperienced, and (3) whether the director planned to cast a male-lead actor or a female lead actor in the lead role. These three elements were randomly combined to create each movie proposal.

To create the fictional movie descriptions, we used ChatGPT-4 to generate a non-gendered pitch for a new movie in each of the 20 genres that were reflected in our main IMDb data.[8] We then reviewed each movie description to ensure it was suitable for the experiment (Table A.4 in the Supporting Information Appendix lists the 20 movie plots). These were then randomly combined with an experienced/inexperienced production team and a male/female actor in the lead role to create a complete movie proposal. For example, participants who were randomly assigned a mystery movie with an experienced production team and a female lead actor were asked to evaluate the following movie proposal:

> "The sudden disappearance of a renowned author leaves the literary world baffled. An avid fan, tasked with uncovering the truth, discovers cryptic clues in the author's latest work. As each enigma is unraveled, a dark conspiracy comes to light. In the end, the truth behind the author's disappearance proves more shocking than anyone could have imagined. The project has an experienced production team

[8]We used a version of the following GPT prompt to generate movie descriptions: "Generate a list of pitches for new movies. The pitches should not include pronouns nor proper names. Each summary should be four sentences long. Please do this for each of the following 20 genres: horror, romance, musical, mystery, drama, music, biography, comedy, history, fantasy, family, thriller, war, sci-fi, crime, animation, adventure, action, sport, and western."

(director, writer, producer). The director of the movie plans to cast a female actor in the lead role."

Each participant evaluated proposals from four different genres, each with a random lead-actor gender and production crew experience level. After participants read each movie proposal, they were asked to make three assessments: (1) Their anticipated 1–10 rating (how do you think you would rate this movie?), (2) their expected propensity to consume on a 1–10 scale (how likely is it that you would watch this movie?), and (3) their likelihood of sharing their opinion of the movie if they did watch it (assuming this movie gets made and you watch it, how likely are you to anonymously post your review score (good or bad) to a rating website like IMDb (the Internet Movie Database)?).

We recruited 804 participants from Prolific for the study. Participants were required to be in the United States, be fluent in English, and were roughly balanced by gender (48% of respondents self-reported female, 50% male, 1% non-binary, and the rest preferred not to say). Qualtrics was used for treatment randomization and data collection.

4.1.2 | Experiment results

Consistent with the main analyses, our primary variable of interest was the expected rating (1–10) of each movie proposal. We converted the data to rater dyads on the same genre (i.e., each of the 20 different movie descriptions), and in line with the research question in the paper, measured the amount of absolute pairwise disagreement on anticipated quality as a function of lead-actor gender. We present two main sets of findings from the experiment.

First, audience selection into consumption influences levels of observed disagreement. Figure 4 plots the level of ratings disagreement based on whether participants indicated they would likely watch the movie. For the full sample of respondents (threshold = 1), there is more disagreement about the quality of male-lead films compared to female-lead films. However, this result flips as the sample is limited to those that indicated they would likely watch the movie that they evaluated. Note that for participants who indicated they would likely watch the movie (watch score >5), there is more disagreement about the quality of female-lead films than male-lead films. This directly mirrors our findings from the observational data, and points to the importance of selection into consumption by audiences. Observed ratings agreement depends on people selecting into consumption.

Second, male audiences appear to be driving the disagreement about female-lead films, and this is only partly mitigated by providing information about quality (production team). Figure 5 plots ratings disagreement as a function of production team experience level (experienced team vs. inexperienced team) and gender of the evaluators (female vs. male) for respondents that indicated they were likely to watch the movie. Mirroring the observational data findings, female evaluators do not agree more or less on the quality of a movie as a function of lead-actor gender. This is true regardless of production team experience, as there is nearly no agreement gap between male and female lead actors for female evaluators. Male evaluators, on the other hand, agree less on female-lead movies. Although the male–female lead gap is largest for inexperienced teams, it remains for experienced teams. This is, we believe, indicative of men holding a descriptive bias against female-lead films that is lessened when male audiences are told the production team is high quality. However, some prescriptive elements seemingly remain, as the male–female actor difference does not fully dissipate.

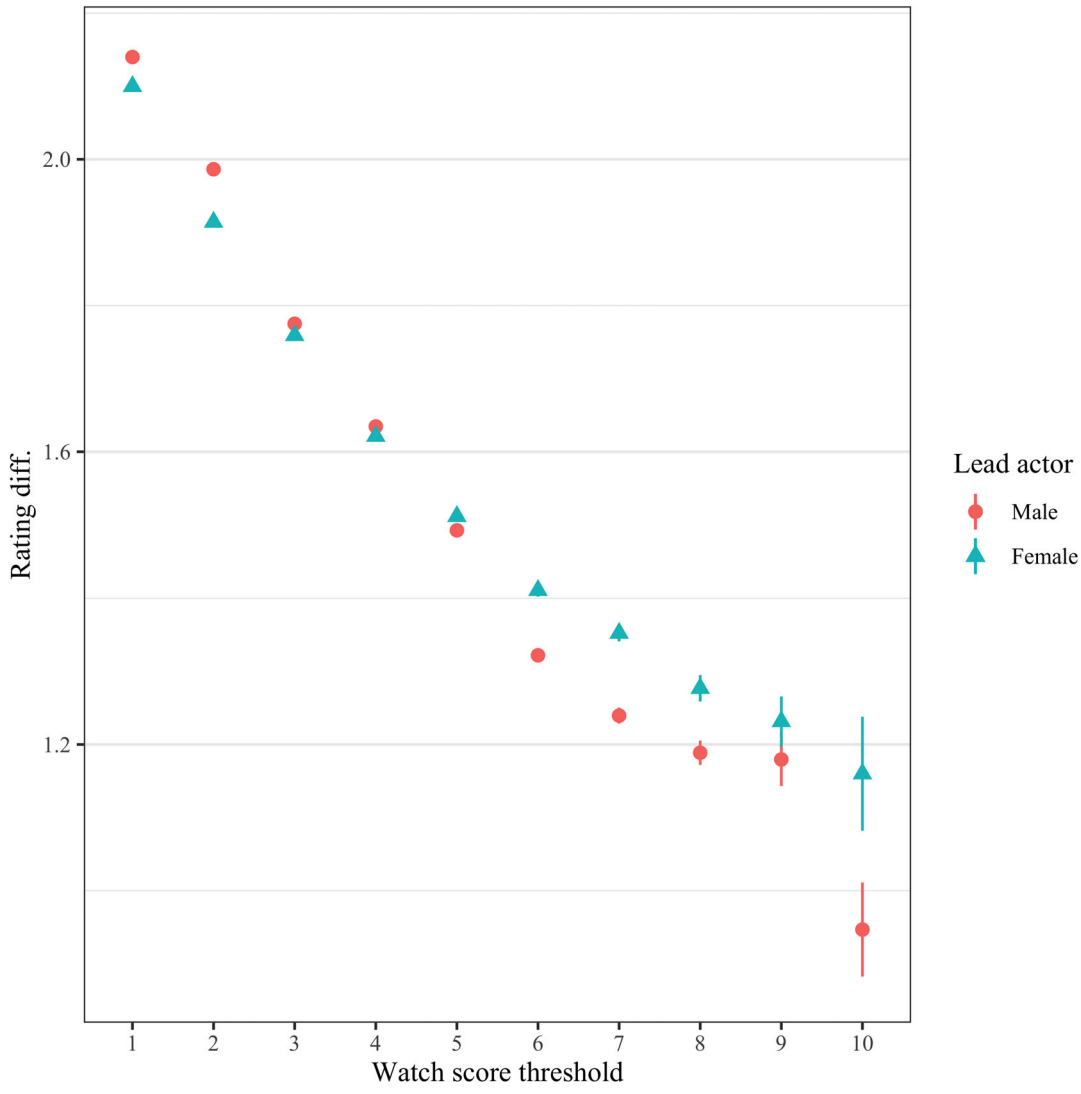

FIGURE 4 Pairwise rating disagreement (*y*-axis) as a function of whether participants indicated they were likely to actually watch the movie. The *x*-axis subsamples the data for respondent pairs that are more likely to indicate they would watch the movie they are evaluating. A watch score threshold of ≥1 comprises the full sample of respondents regardless of whether they said they would watch the movie. A watch score threshold =10 is only participants who indicated with highest confidence (10) they would watch the movie.

4.2 | Consensus and performance

We now turn to the potential performance consequences of evaluative heterogeneity. There is an interesting and consequential first-order supply–demand anomaly in the US theatrical film industry if one assumes demographics determine demand: women and men are equally represented among moviegoers, but female-lead titles are only 28.5% of the 1992–2018 market.[9] Our results on ratings consensus may partially explain why female-lead movies do not reach market

[9]MPAA 2016 Theatrical Market Statistics (p. 15), women represent 52% of moviegoers and 50% of tickets sold: https://www.motionpictures.org/wp-content/uploads/2017/03/MPAA-Theatrical-Market-Statistics-2016_Final.pdf.

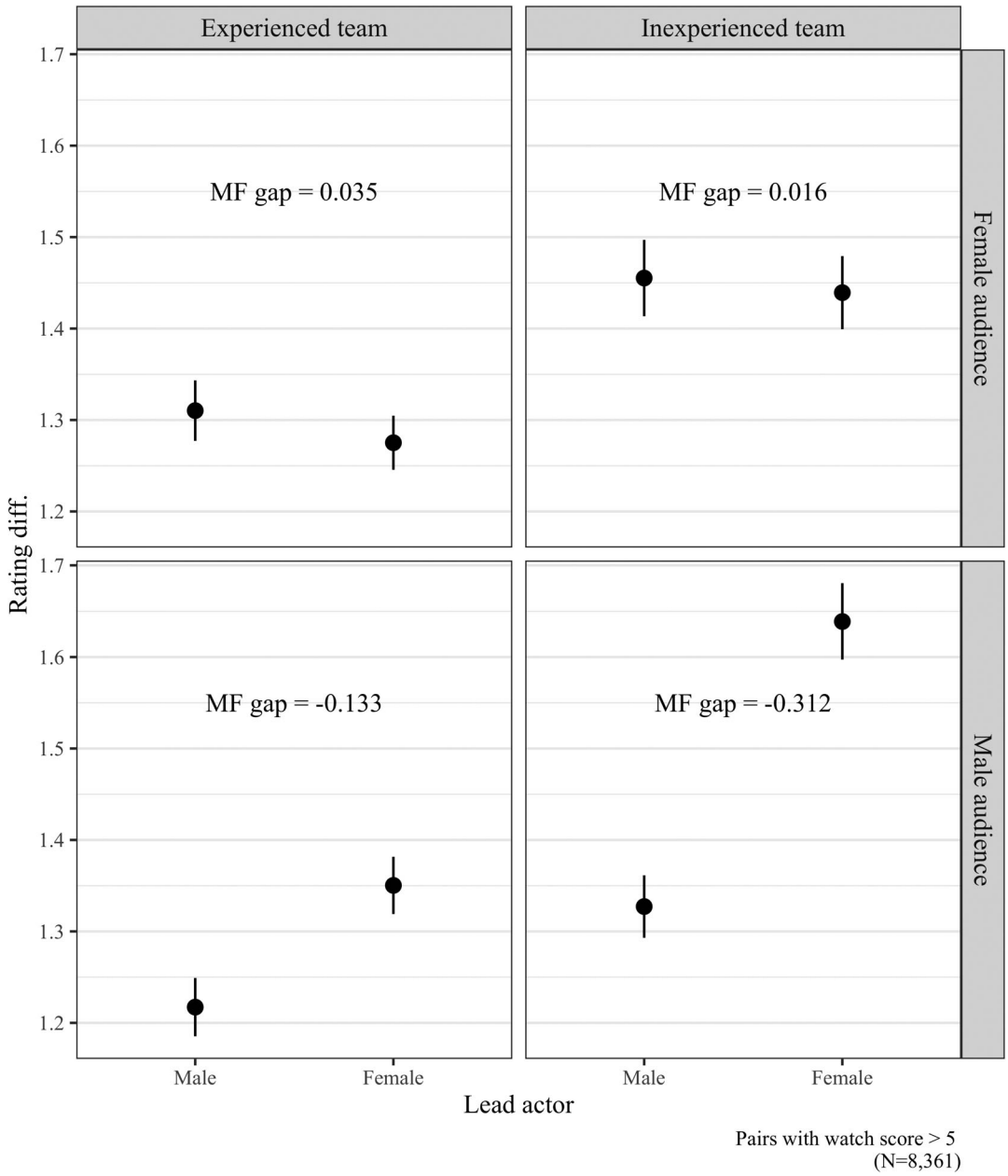

FIGURE 5 Pairwise rating disagreement for participants who said they were likely to watch the movie, split by (1) experience level of the production crew team and (2) gender of the evaluators.

parity. Consistent with a status-based theory, we do find that both male and female audiences rate female-lead movies lower on average. Despite this, we also find pockets of enthusiasm, as indicated by the higher dispersion among male audiences and positive skew among female audiences. These dispersion results hint at possible differential performance implications for gender-typing depending on the scope of market addressed.

We explore differential performance outcomes in Table 4, which restricts analysis to film titles with pooled mean ratings in the central part of the distribution, the 5–7.5 range (80.7% of all films).[10] Confirming the results in Table 2, models 1 and 2 in Table 4 show that a female lead is correlated with a larger left and right tail (measured as the percentage, 1–100, of ratings that are either 1, 2, or 3 for the *Left tail* variable or 8, 9, or 10 for the *Right tail* variable).

The remaining models assess the joint relationship of gender-typing and studio type for performance metrics. The key assumption here is that studio type is a crude proxy for market scope in that *Major* studios typically attempt to appeal to wide audiences, while *Independent* distributors have more modest commercial ambitions. In model 3 of Table 4, female lead actors employed by independent studios are associated with 1.13 times ($=exp(0.118); p=0.043$) the revenue compared with male lead actors. By contrast, for major studios, female lead actors are associated with 0.94 times ($=exp(0.118-0.183); p=0.286$) the revenue compared with male lead actors. In other words, for producers with lower market scope producing female-lead films appears to produce more positive results compared with similar titles that have the same mean ratings.

Finally, while data availability precludes us from directly testing revenue across male and female audiences, if we assume that the number of ratings submitted by different audiences for a film roughly correlates with the revenue from those audiences, then we can use the total (logged) number of reviews submitted by men and women as proxies for box office performance with each audience.[11] Models 4 and 5 of Table 4 report these results. In model 4, we see that a female lead results in more ratings from women, regardless of whether the title is produced by a major or an independent studio: female leads for independent studios are associated with 1.29 ($=exp(0.255); p<0.01$) times the number of ratings from women, and for major studios 1.28 times ($=exp(0.255-0.006); p<0.01$) the number of ratings from women. Model 5 then estimates the count of ratings from men. Compared with male-lead movies, female leads for independent studios are associated with 0.815 ($=exp(-0.205); p<0.01$) times the number of ratings from men, and for major studios 0.722 times ($=exp(-0.205-0.121); p<0.01$) the number of ratings from men. The negative female × major interaction ($-0.121; p<0.01$) indicates that female leads have a relatively smaller effect on the size of male audiences for independent studios compared with major studios.

Our interpretation of the results in Table 4 is that both deference and preference mechanisms are likely at play in this industry. On the one hand, the "widely held beliefs" hypothesized to underlie status dynamics are consistent with an equilibrium where violating norms is punished overall. On the other hand, strategy scholars have long theorized that firms will exploit heterogeneous demand to beneficially target a niche audience (Barney, 1991; Porter, 1991), and we see independent studios yielding some benefits from casting women. Consistent with this supposition, in our sample, not only do independent studios produce more female-lead films than major studios, at 34.4% versus 23.3% of their respective releases, but their propensity to do so is increasing over time while staying relatively flat at major studios. As

[10]We exclude titles with low/high scores on the theoretical assumption that universal loathing/acclaim has a mechanical association with performance. In terms of measurement, titles in the mean rating tails are also problematic in that rating variance is censored due to the 1–10 scale, and in that there are low cell counts for female-lead titles. Both problems are visible in Figure 3. In sensitivity analysis, we find that broadening the sample results in consistent point estimates, 95% confidence intervals that begin to cross 0, and decreasing model explanatory power.

[11]Not readily apparent in Table 4 is the statistic that female raters provide 31.7% of ratings for female-lead movies, versus 18.5% for male-lead movies.

TABLE 4 Effect of lead gender and studio type for "mediocre" movies (rated between 5 and 7.5 stars).

	Dependent variable				
	Left tail	Right tail	Revenue (log)	Female rating Count (log)	Male rating Count (log)
	(1)	(2)	(3)	(4)	(5)
Female lead actor	0.299	0.474	0.118	0.255	−0.205
	(0.112)	(0.180)	(0.058)	(0.043)	(0.042)
Major studio (0/1)	−0.480	−0.557	0.483	0.241	0.292
	(0.114)	(0.184)	(0.051)	(0.037)	(0.036)
Female × Major			−0.183	−0.006	−0.121
			(0.082)	(0.060)	(0.059)
Rating mean	−7.656	16.306	0.540	0.817	0.803
	(0.087)	(0.139)	(0.034)	(0.025)	(0.024)
Violence	0.172	0.256	−0.026	0.018	0.077
	(0.037)	(0.060)	(0.014)	(0.011)	(0.010)
Sex/nudity	0.108	0.166	−0.039	0.035	0.016
	(0.029)	(0.047)	(0.011)	(0.008)	(0.008)
Language	−0.152	−0.240	−0.012	−0.028	0.030
	(0.027)	(0.043)	(0.010)	(0.008)	(0.007)
Log(Opening theaters)	−0.024	−0.223	0.428	0.149	0.129
	(0.026)	(0.042)	(0.010)	(0.007)	(0.007)
Total genres	−0.791	−1.485	−0.577	−0.718	−0.516
	(0.764)	(1.227)	(0.296)	(0.217)	(0.212)
Constant	56.819	−76.674	10.690	1.452	3.121
	(0.810)	(1.302)	(0.315)	(0.231)	(0.226)
Release year indicators	Yes	Yes	Yes	Yes	Yes
Release month indicators	Yes	Yes	Yes	Yes	Yes
Genre indicators	Yes	Yes	Yes	Yes	Yes
Observations	3239	3239	3239	3239	3239
R^2	0.760	0.853	0.587	0.488	0.584
Adjusted R^2	0.755	0.850	0.578	0.478	0.576

Note: These movies represent 80.7% of the sample. The unit of analysis is a movie. Standard errors are in parentheses.

illustration, Figure 6 plots the percentage of female-lead movies released by major studios versus independent studios in each year of our sample. The difference in trend lines is clear.

In closing this section, we caution that more research is needed on the nature of the link between evaluative dispersion and niche/specialized producer performance. One issue is that while cultural product producers address niches based on content type (Barroso et al., 2016), our results suggest addressing the "female niche" is not a straightforward matter of identity matching: all audiences discount female-lead titles on average, male audiences are left-skewed

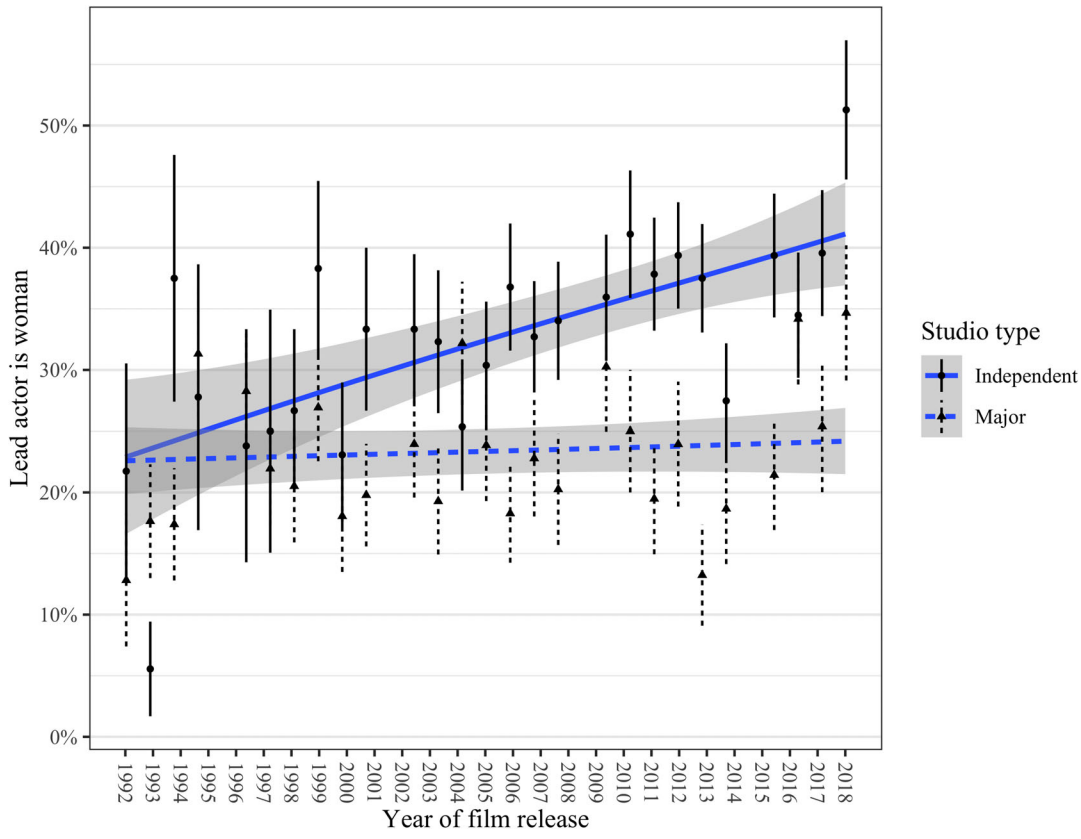

FIGURE 6 The share of female-lead films in the full sample is 23% for major studios and 34% for independent studios. However, the proportion of female-lead films released by major studios has stayed relatively consistent over time, yet the percentage released by independent studios has increased. Major studios include Warner Bros., Buena Vista, Universal Pictures, Twentieth Century Fox, Sony Pictures Entertainment, Paramount Pictures, and Metro-Goldwyn-Mayer (MGM).

and dispersed, and female audiences are right-skewed. Finally, even while we find selected instances where gender-typing is correlated with better performance metrics, it is possible that producing appealing feminine content is itself a variable capability.

5 | DISCUSSION

In this manuscript, we provide evidence that gender—one of the most studied status characteristics—is associated with the dispersion of audience quality evaluations. "Female-typed" films, primarily measured here as those with female lead actors, not only face mean rating penalties but also have more audience disagreement on quality. Moreover, gender status may be associated with some degree of polarization by its very nature. Although both male and female audiences rated female-lead films lower on average, some members of the male audience reacted particularly negatively, and some members of the female audience (but not the majority) experienced the opposite and reacted particularly positively. These ratings dispersion effects, while theoretically and empirically interesting in their own right, have further

implications for commercial performance. For instance, one can imagine scenarios where a ratings distribution with a lower mean and higher dispersion is actually preferred to the counterfactual of a higher mean and lower dispersion: in the former there may be more consumers in the right perceived quality tail.

These findings also help connect this paper to current work in strategy that is directly focused on the effects of audience polarization on firm performance. For example, our findings relate to the emerging work that seeks to understand how investors or customers respond to firms' often deliberate engagement with potentially contentious social issues: Mohliver and Hawn (2019) find that both firms that rank high and those that rank low in their provision of LGBTQ-friendly policies receive positive stock market reactions; Hou and Poliquin (2023) find evidence that firms whose CEOs voiced support for gun control faced temporarily lower sales in conservative geographies but not liberal locations; McDonnell and Darnell (2021) find that the consequences of boycotts are contingent on the ideological alignment between the firm and those conducting the boycott. Thus, the consequences of firm engagement with polarizing issues become largely a function of what type of audience a firm caters to. We highlight that product status characteristics in general may have the potential to decrease consensus, even in contexts where explicit polarization at the issue level does not exist.

Even while this work is the first, to our knowledge, to document a relationship between gender-typing and evaluative consensus, it is not immediately clear whether the dynamics we observe in consumer ratings would be mirrored in ratings by professional critics. For example, Shrum (1996, p. 161) concludes that professional critics reviewing performances at the Edinburgh Festival Fringe are less likely to agree on the quality of new shows compared with shows that have already been extensively performed in the market. If female-typed movies are considered relatively "new" in the market, then it is possible that what we find extends to professional critics.

That said, this paper more clearly relates to existing research conducted in the context of the film industry, which has taken a number of approaches. First, similar to our paper, research has studied film-level outcomes as a function of film-level attributes (e.g., Hsu, 2006; Parker et al., 2020). Future work might revisit existing findings to understand whether findings of mean effects are masking other processes related to consensus. Second, other studies have attempted to explain individual-level success within the industry by tracing the careers of actors (e.g., Rossman et al., 2010). Although we do not examine individual actor careers in this study, it seems plausible that there are labor market implications of acting in movies that feature less consensus. The most obvious might be that women may be more successful finding work in movies produced by independent rather than major studios, or that some actors may be gender-typecast more strongly than others. Thus, future work might consider the implications of evaluative consensus on gender inequality in the labor market by directly examining the ultimate career consequences for actors appearing in films with higher or lower levels of consensus in their ratings. Further, recent work indicates that firms may strategically differentiate through "counterpositioning" on potentially contentious issues such as CSR (Mohliver et al., 2023). For example, it is possible that independent studios are strategically reacting to the production decisions of the majors. Third, other studies have developed novel computational methods that allow more granular measurement of women's and men's roles within movies (e.g., Kagan et al., 2020; Luo & Zhang, 2021). These methods might allow for finer-grained metrics of gender-typed content for future attempts to explain the sources of consensus.

In line with these measurement issues, for our main analyses, we used the lead actor's gender to classify whether a film was more or less female-typed or male-typed. In our supplemental

analyses, we found that operationalizations using other cast members (section A.3.1 in the Supporting Information Appendix) and that an actor-gender-based classification of genres (section A.3.2 in the Supporting Information Appendix) produced consistent patterns related to consensus. It is of course possible that other aspects of a movie, such as the story or plot, might be gender-typed, and the congruency literature suggests that there could be important interactions between lead-actor gender and other film elements, further complicating measurement. In an ideal world, one could imagine creating a measure for each movie that directly captures audiences' first-order perceptions of the film's gender-typing. For example, studying variance in gender-typing across product markets, Tak et al. (2019, p. 555) asked a sample of respondents to rate on a 1–7 scale the degree to which they believed that "most people" found a product category masculine (1) versus feminine (7); they then analyzed the variance in this measure across 60 different product categories. Although doing so is not feasible in practice, asking audiences to provide such an assessment for each movie in our sample would sidestep the indirect measurement challenges. Future research might explore ways to approximate such a measure.

In the meantime, a practical question remains whether producers would be better or worse off from adding more masculine content to female-lead movies. The literature on gender congruence indicates that men are rewarded for behaving more masculine, while women are penalized for the same behavior. One possible archetypal gendered action concerns violence. In our sample, greater amounts of violence and gore are associated with higher mean ratings from both men and women (see the "Violence" coefficient in Table 2 model 1 and Table 3 models 1 and 2). Films featuring female leads have lower levels of violence, at a mean of 4.57 on the Kids-in-Mind 0–10 point scale, when compared to male-lead titles, which have an average of 5.19. The congruency literature suggests, however, that simply increasing violence in female-lead titles is problematic. Figure 7 and Figure 8 plot the ratings mean and standard deviation, respectively, by audience type and level of violence. The figures show a clear pattern: for both audiences, but males more strongly, the mean gap between male- and female-lead titles increases as violence increases (note that only 3.6% of titles have violence levels less than two, and therefore the divergence on the left tail is based on sparse data). Furthermore, greater violence also increases dispersion more strongly for female-lead titles.

These comparisons are intriguing, but a limitation of our large-scale quantitative dataset is that it does not directly capture the motivation of raters. The review narratives selectively posted to IMDb, however, suggest some possibilities. For instance, the 2016 film *Ghostbusters* is a remake that replaced male leads with female leads. It generated significant controversy on social media, as evidenced by high rating variance (standard deviation of 2.75 vs. the sample mean of 1.85) and a dramatic difference between male and female audience reactions (4.77 mean vs. 7.05 mean). Equally interesting are the narrative reviews of the 2016 title that used prescriptive language and that directly referenced gender (Heilman, 2012; Rudman et al., 2012). Comments from users who provided 1-star reviews included: "This is why the feminists shouldn't get their way"; "This is horrible feminist garbage"; and "a five out of ten movie that gets one star for promoting a pointless gender competition." Conversely, comments from users who provided 10-star reviews included: "An awesome female cast depicting smart, funny women"; "What's so great about this movie? Women characters who are not sexualized and who get the right to be badass"; "I'm giving it the highest note because i tend to think that this movie is panned by misogynistic and racist guys..." In settings where more textual review data is available, future research could attempt to systematically measure variance in the use of language and the co-occurrence of gendered terms.

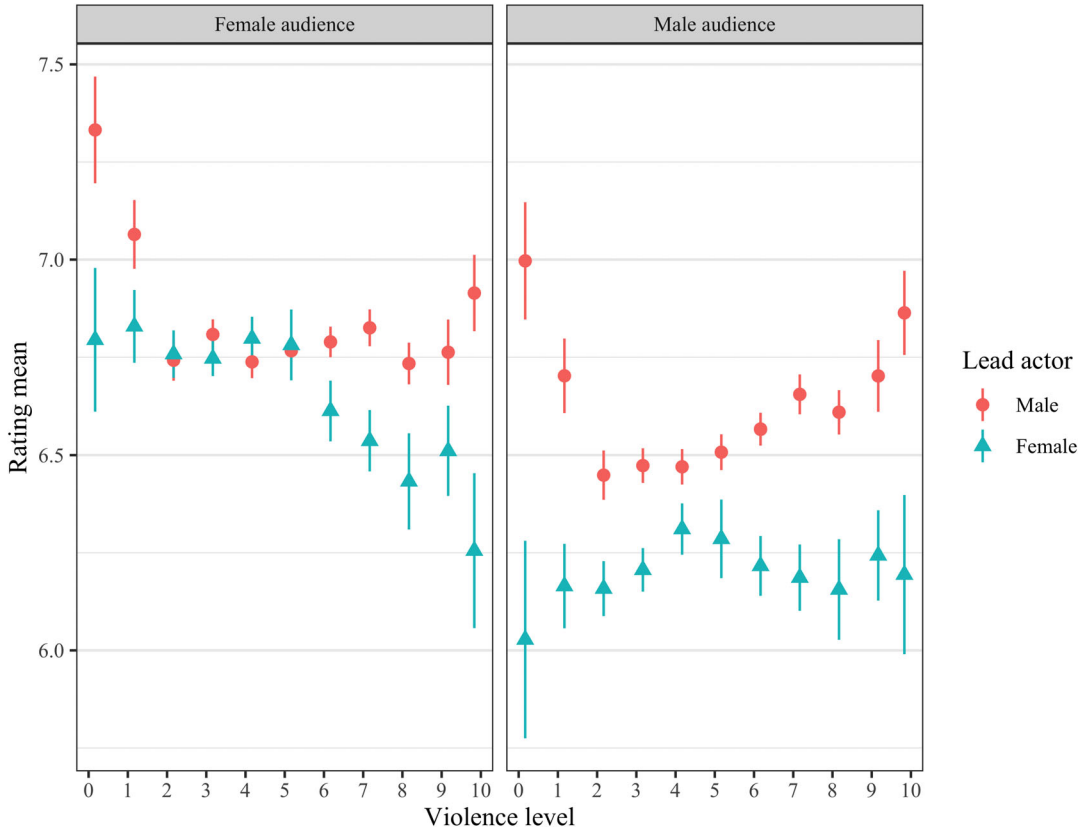

FIGURE 7 Mean rating by violence level and audience type.

Finally, there is the issue of whether these patterns generalize to other constructs and contexts. On the construct angle, there are a variety of other social status markers, such as race, religion, or politics, where the same consensus processes may be at play. For instance, Verkuyten (2005) finds that respondents in the Netherlands who strongly identify as Turkish are more likely to embrace multiculturalism than those who identify as Dutch. A related question is how imbalances on the supply side and demand side of group identification play out. In the US film market, female-typed products are clearly "outsiders" when considered from the supply side of the films made. On the audience side, however, the picture is less clear: there is no majority/minority audience in the industry as a whole, as ticket sales are roughly equal between men and women. If, however, men are considered the "majority" cultural force (an implicit argument of status characteristics theory), our findings may mirror those framed in more traditional insider/outsider terms.

6 | CONCLUSION

Work on status characteristics has found that female-typed objects are often evaluated less favorably—by both men and women—than equivalent male-typed products. This work has implicitly theorized a universal penalty and tested for a *mean* discount. In this paper, we asked

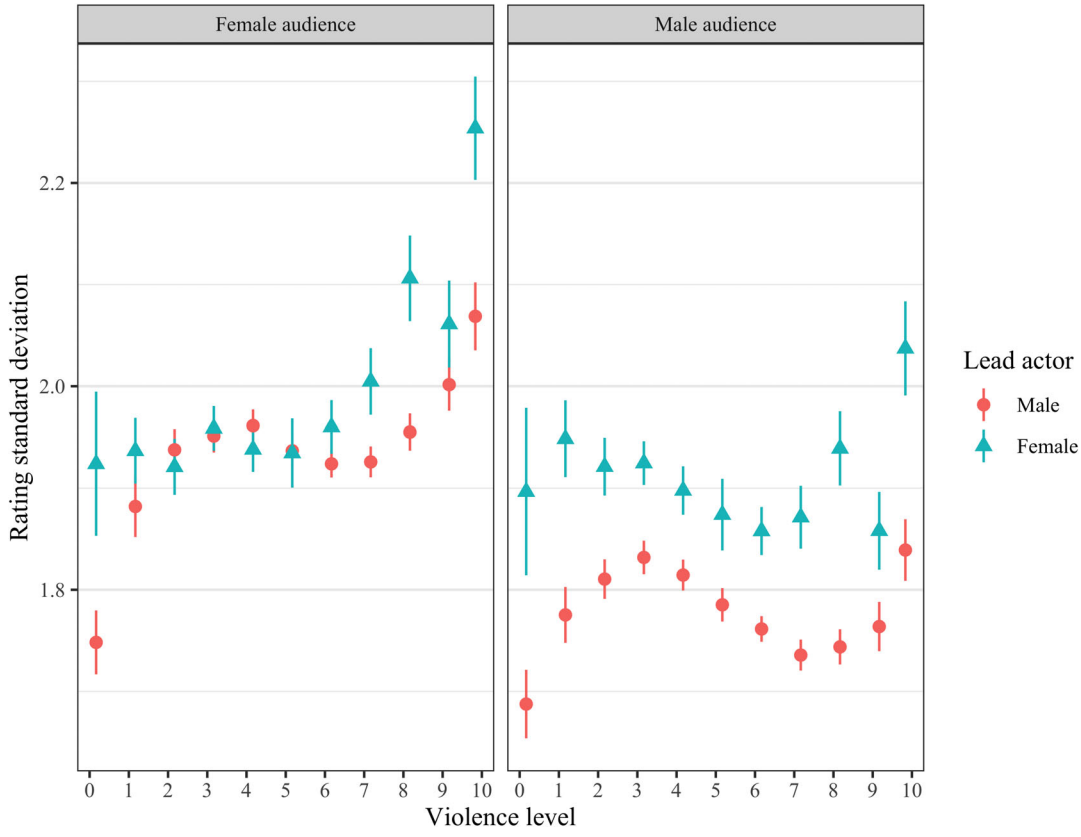

FIGURE 8 Standard deviation of rating by violence level and audience type.

whether status characteristics such as gender are also related to higher variance in evaluations. We introduced theory that demonstrates that although status characteristics may be widely held, it is unlikely—even implausible—that they are held *universally*. This means that for a given level of quality, a status trait should be associated with less consensus: some people employ such traits more strongly than others in their evaluations. We test for this in the context of 383 million consumer film ratings of 4012 widely released movies. We find that the gender-typed content of a film, operationalized as lead-actor gender, is associated with less consensus in ratings for any given mean quality level. The survey experiment sheds light on how audience sorting and treatment effects contribute to this relationship. The performance implications of this lack of consensus are reflected in how independent studios have better outcomes at the box office when releasing movies with female, rather than male, leads. This in turn may partially explain observed differences in trends for female-lead casting when comparing broadly focused and narrowly focused film studios.

ACKNOWLEDGMENTS

We would like to thank Violina Rindova, Ernesto Calvo, Olav Sorenson, Isabel Fernandez-Mateo, Seth Carnahan, Wilbur Chung, and Jóhanna Birnir, colleagues at London Business School and the University of Maryland, along with seminar participants at Purdue University, BI Norwegian Business School, the Oxford Reputation Symposium, the Academy of

Management Annual Meeting, and the Junior Faculty Organization Theory Conference for their helpful comments on earlier versions of this paper.

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://osf.io/tfj2k/

DATA AVAILABILITY STATEMENT

The data and code that support the findings of this study are available at the: https://osf.io/tfj2k/

ORCID

Bryan K. Stroube https://orcid.org/0000-0002-1785-3267
David M. Waguespack https://orcid.org/0000-0002-3723-6258

REFERENCES

Abraham, M. (2017). Pay formalization revisited: Considering the effects of manager gender and discretion on closing the gender wage gap. *Academy of Management Journal*, *60*(1), 29–54.

Abraham, M. (2020). Gender-role incongruity and audience-based gender bias: An examination of networking among entrepreneurs. *Administrative Science Quarterly*, *65*(1), 151–180.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99–120.

Barroso, A., Giarratana, M. S., Reis, S., & Sorenson, O. (2016). Crowding, satiation, and saturation: The days of television series' lives. *Strategic Management Journal*, *37*(3), 565–585.

Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, *88*(4), 354–364.

Benjamin, B. A., & Podolny, J. M. (1999). Status, quality, and social order in the California wine industry. *Administrative Science Quarterly*, *44*(3), 563–589.

Botelho, T. L., & Abraham, M. (2017). Pursuing quality: How search costs and uncertainty magnify gender-based double standards in a multistage evaluation process. *Administrative Science Quarterly*, *62*(4), 698–730.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste* (1st ed.). Routledge.

Cancellieri, G., Cattani, G., & Ferriani, S. (2022). Tradition as a resource: Robust and radical interpretations of operatic tradition in the Italian opera industry, 1989–2011. *Strategic Management Journal*, *43*(13), 2703–2741. https://doi.org/10.1002/smj.3436

Carnahan, S., & Greenwood, B. N. (2018). Managers' political beliefs and gender inequality among subordinates: Does his ideology matter more than hers? *Administrative Science Quarterly*, *63*(2), 287–322.

Castilla, E. J. (2008). Gender, race, and meritocracy in organizational careers. *American Journal of Sociology*, *113*(6), 1479–1526.

Cattani, G., Ferriani, S., & Allison, P. D. (2014). Insiders, outsiders, and the struggle for consecration in cultural fields: A core-periphery perspective. *American Sociological Review*, *79*(2), 258–281.

Cohen, P. N., & Huffman, M. L. (2003). Individuals, jobs, and labor markets: The devaluation of women's work. *American Sociological Review*, *68*(3), 443–463.

Cooper, V. W. (1997). Homophily or the queen bee syndrome: Female evaluation of female leadership. *Small Group Research*, *28*(4), 483–499.

Correll, S. J., & Benard, S. (2006). Biased estimators? Comparing status and statistical theories of gender discrimination. *Advances in Group Processes*, *23*, 89–116.

Correll, S. J., & Ridgeway, C. L. (2003). Expectation states theory. In J. Delamater (Ed.), *Handbook of social psychology* (pp. 29–51). Kluwer Academic/Plenum.

Correll, S. J., Ridgeway, C. L., Zuckerman, E. W., Jank, S., Jordan-Bloch, S., & Nakagawa, S. (2017). It's the conventional thought that counts: How third-order inference produces status advantage. *American Sociological Review*, *82*(2), 297–327.

Dimitriadis, S., Lee, M., Ramarajan, L., & Battilana, J. (2017). Blurring the boundaries: The interplay of gender and local communities in the commercialization of social ventures. *Organization Science*, *28*(5), 819–839.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598.

Fernandez-Mateo, I., & Fernandez, R. M. (2016). Bending the pipeline? Executive search and gender inequality in hiring for top management jobs. *Management Science*, *62*(12), 3636–3655.

Fernandez-Mateo, I., & King, Z. (2011). Anticipatory sorting and gender segregation in temporary employment. *Management Science*, *57*(6), 989–1008.

Foschi, M. (1996). Double standards in the evaluation of men and women. *Social Psychology Quarterly*, *59*(3), 237–254.

Frake, J. (2017). Selling out: The inauthenticity discount in the craft beer industry. *Management Science*, *63*(11), 3930–3943.

Greenberg, J., & Mollick, E. (2017). Activist choice homophily and the crowdfunding of female founders. *Administrative Science Quarterly*, *62*(2), 341–374.

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, *32*, 113–135.

Heilman, M. E., & Eagly, A. H. (2008). Gender stereotypes are alive, well, and busy producing workplace discrimination. *Industrial and Organizational Psychology*, *1*(4), 393–398.

Holt, D. B. (1997). Distinction in America? Recovering Bourdieu's theory of tastes from its critics. *Poetics*, *25*(2), 93–120.

Hou, Y., & Poliquin, C. W. (2023). The effects of CEO activism: Partisan consumer behavior and its duration. *Strategic Management Journal*, *44*(3), 672–703.

Hsu, G. (2006). Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. *Administrative Science Quarterly*, *51*(3), 420–450.

Kagan, D., Chesney, T., & Fire, M. (2020). Using data science to understand the film industry's gender gap. *Palgrave Communications*, *6*(1), 1–16.

Kim, M., & DellaPosta, D. (2021). The fickle crowd: Reinforcement and contradiction of quality evaluations in cultural markets. *Organization Science*, *33*(6), 2496–2518.

Koning, R., Samila, S., & Ferguson, J.-P. (2021). Who do we invent for? Patents by women focus more on women's health, but few women get to invent. *Science*, *372*(6548), 1345–1348.

Kovács, B., & Sharkey, A. J. (2014). The paradox of publicity: How awards can negatively affect the evaluation of quality. *Administrative Science Quarterly*, *59*(1), 1–33.

Kuppuswamy, V., & Younkin, P. (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.

Lee, M., & Huang, L. (2018). Gender bias, social impact framing, and evaluation of entrepreneurial ventures. *Organization Science*, *29*(1), 1–16.

Luo, H., & Zhang, L. (2021). *Gender inequality and the direction of ideas: Evidence from #MeToo*. SSRN Scholarly Paper 3817029, Social Science Research Network.

Luo, H., & Zhang, L. (2022). Scandal, social movement, and change: Evidence from #MeToo in Hollywood. *Management Science*, *68*(2), 1278–1296.

McDonnell, M.-H., & Darnell, S. (2021). Profiting from protest: A contingency model of the disruptive capacity of anti-corporate activism. *Working paper*.

Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.

Mohliver, A., Crilly, D., & Kaul, A. (2023). Corporate social counterpositioning: How attributes of social issues influence competitive response. *Strategic Management Journal*, *44*(5), 1199–1217.

Mohliver, A., & Hawn, O. (2019). *Rewarding the extremes: Market reaction to U.S. corporations' LGBTQ positions*. SSRN Scholarly Paper ID 3477837, Social Science Research Network.

Niessen-Ruenzi, A., & Ruenzi, S. (2019). Sex matters: Gender bias in the mutual fund industry. *Management Science*, *65*(7), 3001–3025.

Parker, O., Mui, R., & Titus, V. (2020). Unwelcome voices: The gender bias-mitigating potential of unconventionality. *Strategic Management Journal*, *41*(4), 738–757.
Podolny, J. (1993). A status-based model of market competition. *American Journal of Sociology*, *98*(4), 829–872.
Porter, M. E. (1991). Towards a dynamic theory of strategy. *Strategic Management Journal*, *12*, 95–117.
Reed, A., Forehand, M. R., Puntoni, S., & Warlop, L. (2012). Identity-based consumer behavior. *International Journal of Research in Marketing*, *29*(4), 310–321.
Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world*. Oxford University Press.
Ridgeway, C. L., & Correll, S. J. (2004). Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender and Society*, *18*(4), 510–531.
Rindova, V. P., & Petkova, A. P. (2007). When is a new thing a good thing? Technological change, product form design, and perceptions of value for product innovations. *Organization Science*, *18*(2), 217–232.
Rossman, G., Esparza, N., & Bonacich, P. (2010). I'd like to thank the academy, team spillovers, and network centrality. *American Sociological Review*, *75*(1), 31–51.
Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology*, *48*(1), 165–179.
Shrum, W. (1996). *Fringe and fortune: The role of critics in high and popular art*. Princeton University Press.
Siegel, J., Pyun, L., & Cheon, B. Y. (2019). Multinational firms, labor market discrimination, and the capture of outsider's advantage by exploiting the social divide. *Administrative Science Quarterly*, *64*(2), 370–397.
Simcoe, T. S., & Waguespack, D. M. (2011). Status, quality, and attention: What's in a (missing) name? *Management Science*, *57*(2), 274–290.
Tak, E., Correll, S. J., & Soule, S. A. (2019). Gender inequality in product markets: When and how status beliefs transfer to products. *Social Forces*, *98*(2), 548–577.
Vandello, J. A., & Bosson, J. K. (2013). Hard won and easily lost: A review and synthesis of theory and research on precarious manhood. *Psychology of Men & Masculinity*, *14*(2), 101–113.
Vandello, J. A., Bosson, J. K., Cohen, D., Burnaford, R. M., & Weaver, J. R. (2008). Precarious manhood. *Journal of Personality and Social Psychology*, *95*(6), 1325–1339.
Verkuyten, M. (2005). Ethnic group identification and group evaluation among minority and majority groups: Testing the multiculturalism hypothesis. *Journal of Personality and Social Psychology*, *88*(1), 121–138.
Waguespack, D. M., & Sorenson, O. (2010). The ratings game: Asymmetry in classification. *Organization Science*, *22*(3), 541–553.
Willer, R., Rogalin, C. L., Conlon, B., & Wojnowicz, M. T. (2013). Overdoing gender: A test of the masculine overcompensation thesis. *American Journal of Sociology*, *118*(4), 980–1022.
Worth, L. T., Smith, J., & Mackie, D. M. (1992). Gender schematicity and preference for gender-typed products. *Psychology & Marketing (1986-1998)*, *9*(1), 17.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Stroube, B. K., & Waguespack, D. M. (2024). Status and consensus: Heterogeneity in audience evaluations of female- versus male-lead films. *Strategic Management Journal*, 1–31. https://doi.org/10.1002/smj.3575