Bias in Context: Small Biases in Hiring Evaluations Have Big Consequences

Jay H. Hardy III
Oregon State University

Kian Siong Tey
INSEAD

Wilson Cyrus-Lai
INSEAD

Richard F. Martell
Oregon State University

Andy Olstad
Oregon State University

Eric Luis Uhlmann
INSEAD

November 27th, 2020

Author notes: Correspondence concerning this manuscript should be sent to Jay Hardy, Oregon State University, College of Business, 370 Austin Hall, Corvallis, Oregon, 97331; E-mail: jay.hardy@oregonstate.edu; Phone: (720) 841-2167. Supplemental material for this article is available at http://xxx.sagepub.com/supplemental”

ABSTRACT

It is widely acknowledged that subgroup bias can influence hiring evaluations. However, the notion that bias continues to threaten equitable hiring outcomes in modern employment contexts continues to be debated, even among organizational scholars. In this study, we sought to contextualize this debate by estimating the practical impact of bias on real-world outcomes (a) across a wide range of hiring scenarios and (b) in the presence of diversity-oriented staffing practices. Toward this end, we conducted a targeted meta-analysis of recent hiring experiments that manipulated both candidate gender and qualifications to couch our investigation within ongoing debates surrounding the impact of small amounts of bias in otherwise meritocratic hiring contexts. Consistent with prior research, we found evidence of small gender bias effects (d = -0.30) and large qualification effects (d = 1.61) on hiring manager evaluations of candidate hireability. We then used these values to inform the starting parameters of a large-scale computer simulation designed to model conventional processes by which candidates are recruited, evaluated, and selected for open positions. Collectively, our simulation findings empirically substantiate assertions that even seemingly trivial amounts of subgroup bias can produce practically significant rates of hiring discrimination and productivity loss. Furthermore, we found that contextual factors alter, but cannot obviate the consequences of biased evaluations, even within apparently optimal hiring scenarios (e.g., when extremely valid assessments are used). Finally, our results demonstrate residual amounts of subgroup bias can undermine the effectiveness of otherwise successful targeted recruitment efforts. Implications for future research and practice are discussed.

Keywords: gender and diversity; bias; selection; hiring decisions; computer simulation

BIAS IN CONTEXT: SMALL BIASES IN HIRING EVALUATIONS HAVE BIG CONSEQUENCES

Researchers have spent over a century accumulating convincing evidence that well-designed hiring systems can enhance the measurement and prediction of human potential at work (Schmidt & Hunter, 1998; Schmidt, Oh, & Shaffer, 2016). However, human behavior is inherently complex. As such, it is unrealistic to expect that even the best predictors of applicant potential can eliminate all sources of systematic and unsystematic error from evaluation scores. Concerns over the possible consequences of these errors are further exacerbated by the prominent role of human judgment in nearly all hiring decisions, which as a form of social categorization, makes them particularly susceptible to the insidious influence of prejudice (Duckitt, 1992). Unfortunately, one risk inherent to this reality is the possibility that subgroup bias, defined by the Uniform Guidelines on Employee Selection Procedures as consistent, nonzero errors of prediction made for members of a subgroup based on group membership (Ledvinka, 1979), can influence the decision-making process. In cases where undetected or unaddressed biases impact a hiring manager's evaluations of job candidates to depart from purely merit-based assessment, discriminatory hiring outcomes can occur.

Given the seemingly inevitable influence of subgroup bias on hiring evaluations, it is somewhat surprising to find relatively little research within the recruitment and selection literature has focused on advancing practical strategies or techniques managers can use to directly identify and reduce the influence of bias on hiring outcomes. Instead, research has focused more on indirect tactics such as proactively seeking out qualified minorities to fill out applicant pools using targeted recruitment (Avery & McKay, 2006; Avery, McKay, & Volpone, 2012; Newman & Lyon, 2009) or reducing the impact of subgroup differences in assessment scores through inclusivity-oriented adjustments to the system itself (e.g., altering the choice of

predictors or changing how they are scored; Ployhart & Holtz, 2008; Sackett & Ellingson, 1997).

Although useful in their own right, these and many other established diversity-oriented staffing interventions are not designed to help identify and eliminate the influence of subgroup bias when it occurs. One possible explanation for the lack of interest in a more direct focus on bias in selection research is that effect sizes reported within the broader literature on subgroup bias in subjective evaluations tend to be relatively small, typically ranging between 1% and 7% of the total variance (see Greenwald, Banaji, & Nosek, 2015; Madera, Hebl, & Martin, 2009; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Olian, Schwab, & Haberfeld, 1988). Given these small effects, it seems reasonable to conclude that alternative avenues for combating hiring discrimination would provide more fruitful areas of inquiry. Some have even gone as far as to say, "exceedingly small effect sizes" found in these studies rule out the plausibility of powerful stereotypes in real-world settings altogether (Landy, 2008a).

However, it is not the process of assessment but rather the outcomes associated with the *use* of assessment phase predictions in decision making that are of primary importance to individuals, organizations, and society as a whole (Cascio & Boudreau, 2010). As such, it is essential that we collectively consider the potential downstream impact of factors that influence selection decisions before their importance or relevance can be established. Indeed, we argue that it is this context that is often missing from the conversation surrounding the impact of bias. Without a fuller consideration of how, where, and when bias can impact and subsequently shape hiring outcomes, it can be difficult for researchers, practitioners, the courts, and organizations alike to determine what (if anything) should be done about it.

The purpose of the present study is to better integrate these considerations into the conversation surrounding the impact of subgroup bias in modern hiring contexts. Specifically,

our primary research goals in the present study were twofold. First, we sought to generate estimates of the practical impact of small amounts of subgroup bias on hiring outcomes as they are likely to be felt across a range of hiring contexts. In doing so, we sought to determine whether the practical impact of discriminatory subgroup bias during the assessment phase is theoretically limited to fringe cases with particularly suboptimal conditions or is cause for concern across a broader range of plausible real-world hiring scenarios. Second, we sought to explore whether small amounts of unresolved subgroup bias can undermine otherwise successful targeted recruitment strategies. To that end, we hope to provide greater clarity regarding the extent to which directly addressing residual amounts of discriminatory subgroup bias is necessary when inclusivity-oriented hiring interventions are already in place.

To accomplish these goals, we conducted a meta-analysis of recent hiring experiments that manipulated both candidate gender and strength of qualifications, effectively updating an earlier meta-analytic investigation by Olian et al. (1988). This allowed us to ground our investigation within an ongoing debate on the impact of small amounts of gender bias in otherwise meritocratic hiring contexts. We then used these meta-analytic estimates to inform the starting parameters of a large-scale computer simulation designed to model conventional processes by which candidates are recruited, evaluated, and selected for open positions.

INTRODUCING CONTEXT: GENDER BIAS AND THE ASSESSMENT PHASE

To what extent does gender bias still matter in modern organizations? The empirical answer to this question is not as straightforward as it may initially seem. On one hand, there is widespread agreement that gender bias is morally and ethically wrong. Furthermore, few would dispute the fact that in the past, women faced significant, systematic barriers to entry in the workplace resulting from the insidious influence of gender bias (EEOC, 2010). However, there is

notably less consensus about whether the destructive influence of such biases remains relevant in modern organizations, even among organizational scholars (for competing perspectives on this topic, see Greenwald, 2008; Heilman & Eagly, 2008; Landy, 2008a, 2008b; Martell, Emrich, & Robison-Cox, 2012; Martell, Lane, & Emrich, 1996; Rudolph & Baltes, 2008).

In the following sections, we start by offering a formal definition of gender bias in the assessment phase. We then review the extant empirical literature on the prevalence of gender bias effects in hiring contexts and report the results of an updated meta-analysis of hiring experiments that manipulated both the strength of qualifications and candidate gender, which we use to inform the starting parameters of our computer simulation.

Defining Gender Bias

In the present study, we use the term *gender bias* to refer to a systematic preference or prejudice toward one of the two major genders (female or male) over the other when employee gender is not meaningfully relevant to the job.[1] These biases can be either explicit or implicit (Banaji & Greenwald, 1995) and can go either direction (i.e., a preference for males over females or for females over males). When a hiring manager's gender bias causes their evaluation of one or more candidates in the applicant pool to depart from purely merit-based assessment, sex-based discrimination can occur, which involves the differential treatment (either intentional or unintentional) of job-seekers on the sole basis of their sex. One challenge economists often note when attempting to estimate the magnitude of bias's impact on career outcomes (typically operationalized as gender pay gaps) is that differential outcomes for males versus females attributed to gender bias are often conflated with underlying subgroup differences in labor market experience (Blau & Kahn, 2003). A similar conflation may also exist in the formation of hiring recommendations. Thus, it is important to clarify here that the term gender bias, as used in

this study, speaks to specific preferences or prejudices independent of an applicant's experience, qualifications, and any actual underlying subgroup differences.

Defining the Assessment Phase

Our focus in the present effort was on the influence of bias during the assessment phase, which refers to the stage of the selection process following recruitment and preceding final hiring decisions. During this phase, hiring managers gather information using a range of objective, subjective, formal, and informal tests to develop definitive (and hopefully accurate) opinions regarding the qualifications and performance potential of various individuals within the applicant pool. In particular, we were interested in the influence of gender bias on the cumulative evaluations of applicant performance that hiring managers generate during the assessment phase, which collectively reflects not only the objective and subjective scores from the tools and tests hiring managers use, but also how they interpret and digest information derived from those tests when developing impressions of candidate hireability. Through the lens of this conceptualization, it is possible to conceive of a situation where hiring decisions based on entirely unbiased assessments could still be influenced by small amounts of bias when hiring managers apply their human judgment to combine, interpret, and evaluate assessment scores. Thus, to the extent that hiring manager evaluations of applicant qualifications are subject to any amount of subgroup bias, bias can be reasonably expected to influence hiring outcomes as well. For this reason, our model is not constrained to any specific type of assessment but is designed to capture the potential influence of bias on the hiring decision-making process as a whole.

Research on Gender Bias in Hiring Contexts

Numerous quantitative investigations attest to gender gaps in pay and representation in leadership positions in organizations (Aud et al., 2011; Blau & Kahn, 2017). Studies of

employment evaluations in field settings provide no evidence of biases in job performance and promotability ratings but are potentially confounded by unobserved differences in actual performance (Bowen, Swim, & Jacobs, 2000; Roth, Purvis, & Bobko, 2012). Indeed, if there are unjustified selection biases against a specific demographic group in employment settings, the survivors of such a process should, on average, outperform members of other groups at the same job rank (see Card, DellaVigna, Funk, & Iriberri, 2020, for evidence). An experimental approach is well suited to strong inferences (Platt, 1964) regarding the contribution of discrimination to unequal outcomes because of the ability to draw causal connections between candidate gender and selection decisions. It is further possible to cross target gender with other experimental manipulations (e.g., the strength of qualifications, or interventions against discrimination) to compare effect sizes and capture potential causal interactions.

Meta-analyses of experimental laboratory studies consistently find a small but statistically significant preference for male candidates for traditionally male-typed jobs such as managerial positions (Davison & Burke, 2000; Eagly, Makhijani, & Klonsky, 1992; Koch, D'Mello, & Sackett, 2015). This preference is much more pronounced among male evaluators, who hold a disproportionate number of decision-making roles in organizations. Audit studies, in which fake resumes are sent to real businesses and invitations for interviews serve as the outcome, corroborate the overall pattern of discrimination against women observed in controlled laboratory settings (Ayres, 2003; Neumark, Bank, & Van Nort, 1996; see also Moss-Racusin et al. 2012). The underlying "think manager, think male" cognitive prototype is common across nations and generations (Eagly & Karau, 2002; Heilman, 2001; Koenig, Eagly, Mitchell, & Ristikari, 2011; Schein, Mueller, Lituchy, & Liu, 1996). Discrimination against female job candidates can also result from perceptions—correct or incorrect—about the sexist biases of

superiors, clients, and customers (Trentham & Larwood, 1998; Vial, Brescoll, & Dovidio, 2019). Although experimental studies also find that the causal contribution of a strong vs. weak resume is substantial (Olian et al., 1988), gender biases may co-exist with and even co-opt seemingly meritocratic selection criteria (Norton, Vandello, & Darley, 2004; Uhlmann & Cohen, 2005).

In our view, a primary limitation of the experimental approach, whether in a laboratory or real employment setting, is the difficulty of examining cumulative processes across numerous similarly influenced decisions. A small causal effect could compound over time in a real-world setting, helping explain group-based inequities (Blank, 2005; Greenwald et al., 2015). Contrarily, small group-based biases could be progressively overwhelmed by a comparatively greater focus on candidate qualifications and meritocratic selection processes. Although they cannot answer such questions, experimental paradigms can be used to derive empirically informed estimates for key variables of interest (e.g., such as candidate gender and strength of qualifications). These estimates provide the empirical starting points for simulations of iterative organizational decision-making processes in the present research.

A META-ANALYSIS OF EXPERIMENTS MANIPULATING CANDIDATE QUALIFICATIONS AND GENDER

To provide updated empirical benchmarks for our simulations, we conducted a targeted meta-analysis of experiments manipulating the strength of qualifications and candidate gender on evaluations of candidate hireability. A list of search keywords was used to conduct systematic searches of the databases PsycINFO, Business Source Complete, and ProQuest. Our search terms included the word stems gender, sex*, qualification, ability, anti-women, bias, stereotyp*, prejudic*, discriminat*, job, employ*, personnel, hir*, perform*, manage*, résumé, appl*, recruit*, apprais*, select*, rating*, evaluat*, randomized, and experiment. Two members of the research team reviewed the abstracts and method sections of the initial set of 27 articles for

relevant experiments. The purpose of the meta-analysis was to provide empirical estimates for use in simulating hiring decisions and their long-term consequences for individuals and organizations that are influenced by gender bias as well as evaluations of applicant qualifications. Therefore, we included only directly relevant studies and statistical comparisons, focusing on quality over quantity, as our goal was to provide accurate estimates to use as a basis for the simulations rather than to summarize and aggregate the broader research literature. First and foremost, an experiment was included in the meta-analysis only if both qualifications and gender were directly manipulated within the same design. Second, to simplify our comparisons, we focused on hiring decisions involving either gender-neutral or stereotypically male jobs, excluding stereotypically female jobs (e.g., daycare center workers). Third, we excluded tests of intersectionality effects, for instance, comparing hiring preferences for lesbian women and gay men. In the present research, we are interested in the main effects of gender and modal job candidates rather than interactions with other social identities (we return to the intersectionality question in the General Discussion). Fourth, when studies used a new intervention to reduce discrimination, we included only the baseline, no-intervention condition, as this best captures real-world conditions. Fifth, we selected the ten most recent articles from the set to avoid biasing simulations of contemporary discrimination based on very old studies. Finally, for studies reporting incomplete statistics, we emailed the authors to request additional analyses. In such cases, we excluded articles when authors could not be contacted or did not provide the necessary data for the computation of effect sizes. Appendix A of the online supplement summarizes the final set of six articles containing seven experiments that fit these specific, targeted criteria.

We used a random-effects model, a more conservative test than a fixed-effects model, to estimate the average weighted effect size. A random-effects model assumes that the computed

effect sizes vary across studies and thus allows us to generalize beyond the current set of studies (Lipsey & Wilson, 2001). Consistent with prior research, we weighted each effect size by the inverse of its variance. The common effect size metric used was Cohen's d. We also computed the heterogeneity of effect sizes (Lipsey & Wilson, 2001).

The results of the meta-analysis showed an average effect size of d = -.30 favoring male over female candidates, p = .05, 95% CI [-.60, .00] (see Figure 1a). There is relatively little heterogeneity in effect sizes for gender discrimination (Q = 2.38, p = .80). Within this set of experiments, we observe a dramatically larger effect of qualifications, d = 1.61, p < .01, 95% CI [.86, 2.35] (see Figure 1b), with more qualified candidates strongly preferred over less qualified candidates. However, there is significant heterogeneity in this second set of estimates (Q = 29.99, p < .01), which is likely attributable to the tendency for some studies to compare strong vs. weak profiles and others to compare strong vs. moderate profiles.

Insert Figure 1 about here

The overall effect sizes for candidate gender and strength of qualifications were used to inform our simulation's starting parameters. Before moving on, it is important to note that studies included in our meta-analysis were mostly laboratory experiments. This methodology allows for stronger causality inferences due to the accurate control of the independent and extraneous variables and the use of random assignment. However, one downside of relying on laboratory experiments to estimate bias effects is that they are often based on student samples with little work and managerial experience, which can weaken the generalizability of the results. Although subgroup bias effects observed in field samples are often as big or even bigger than those observed in laboratory settings (Kraiger & Ford, 1985, Colella, Hebl, & King, 2017), future

research should estimate bias effect sizes with more diverse approaches, such as correlational studies and quasi-experiments using field data. In the present study, we simulate a wide range of gender bias estimates, including effects even smaller than our observed meta-analytic effects, to address the possibility that laboratory experiments that inform our meta-analytic results are overestimating the magnitude of bias effects in real-world hiring contexts.

SIMULATING THE IMPACT OF GENDER BIAS ON HIRING OUTCOMES

Computer simulation is a useful method for modeling the operation of abstract theoretical phenomena within the dynamic and complex domain of "real-world" processes, systems, or events (Law, Kelton, & Kelton, 1991). In particular, simulations facilitate a richer consideration of context and actors that complement the careful behavioral control provided by experiments. Using experimental results and simulations in concert can help circumvent intractable dilemmas inherent to using one methodology in isolation (McGrath, 1981). The simulations used in this study can be categorized as stochastic process models (Davis, Eisenhardt, & Bingham, 2007) in that their primary purpose is to understand the relative impact of gender bias on hiring outcomes for individuals and organizations in hiring settings containing parameters with partially random elements (e.g., the ratio of females to males in the applicant pool, the qualification levels of various applicants, error due to assessment unreliability, etc.). After specifying model parameters, we run a series of simulation-based experiments to estimate how the impact of gender bias can be expected to change based on underlying characteristics of the hiring context and the implementation of diversity-focused staffing initiatives. Model experimentation using this approach involves allowing certain stochastic elements to vary while constraining others to be constant. This methodology provides a systematic way to understand and estimate the likely impact of gender bias across various plausible hiring scenarios.

Model Development

The first step of model specification is developing a framework for simulating job applicants' typical progression through each of four stages in the hiring process. In the following sections, we describe how we developed each part of this four-stage model from which we derive the underlying parameters, which provides the foundation upon which subsequent simulation experiments are developed. Table 1 summarizes how each stage is represented in the model and lists key model assumptions made at each step.

Insert Table 1 about here

Stage 1: Generating the applicant pool. As in the real world, our model's selection process begins with assembling a pool of applicants from which to select. Once generated, the model assigns applicants within the simulated sample a gender identifier (female = 0; male = 1) and a qualifications rating. The probability of gender assignment in the model conforms to a Bernoulli distribution, where p represents the likelihood that any given applicant is male. Altering the p parameter within the model allows one to vary the proportion of males relative to females in the applicant pool ($p < .5$ = more female applicants, $p > .5$ = more male applicants).

In the language of Binning and Barrett (1989), the qualifications rating represents the targeted performance domain. That is, the qualifications rating embodies an idealized measure of performance potential in the form of a “true score” value reflecting how capable each applicant is at performing the job relative to other applicants in the pool. In our model, we assume this value is normally distributed (Vancouver, Li, Weinhardt, Steel, & Purl, 2016) and independent of gender. Qualification ratings represent a composite of applicant job-relevant qualifications related to job success. Thus, these ratings define how applicants would perform on the job should

they be hired. As such, the predictive validity of assessment evaluations is defined in our model as the extent to which variance in evaluation scores attributed to applicant qualification ratings is maximized and covers relevant performance domains while the variance in scores attributed to systematic and unsystematic random error is minimized.

Stage 2: Assessment phase. Once the applicant pool is generated, the next step is assigning evaluation ratings to each applicant to be used to make hiring decisions. In real-world hiring contexts, evaluating applicant qualifications for a position is an imperfect process which results in evaluation scores that comprise a combination of both construct-relevant variance (i.e., variance associated with the criterion of interest—in this case, the qualifications rating true-score) and construct-irrelevant variance (i.e., variance not associated with the criterion of interest). In our model, we further subdivide construct-irrelevant variance into (1) systematic error due to gender bias and (2) unsystematic random error due to the influence of construct irrelevant variance on hireability ratings, assessment unreliability, and random noise. This latter source of error is represented in our model using a value that conforms to a random, normal distribution with a mean of 0 and a standard deviation of 1 that is independent of both qualification ratings and the gender identifier. In other words, in our model, qualifications and gender are uncorrelated, which allows gender bias effects to be applied uniformly to all applicants, regardless of their qualifications ratings. Collectively, the assessment phase evaluation process is represented in our model using Equation 1 below.[2]

$$a_i = q_i(\sqrt{q_{\%}}) + 2g_i(\sqrt{b_{\%}}) + e_i(\sqrt{1 - q_{\%} - b_{\%}}) \quad (1)$$

This function works by assigning each applicant an individual assessment score (a_i), of which a certain percentage of the variance ($q_{\%}$) is attributable to applicant qualifications (q_i), a certain percentage of the variance ($b_{\%}$) is attributable to bias associated with applicant gender

(g_i), and the remaining variance ($1 - q_{\%} - b_{\%}$) is due to random error (e_i). Collectively, these evaluation ratings represent an accumulation of all formal and informal evaluative judgments attributed to each applicant used in determining their overall qualifications for the position.

Stage 3: Selection. Next, hiring decisions are made for each applicant in the simulation based on their overall assessment scores relative to other applicants in the sample. Consistent with prior simulation work on this topic (e.g., Murphy, 1986; Tam, Murphy, & Lyall, 2004), our model implements a top-down selection protocol in which the applicant with the highest evaluation score is selected first, followed by the applicant with the second-highest evaluation score second, and so on until all available positions are filled. The proportion of the applicant pool hired is determined by the selection ratio, reflecting the number of job openings relative to the total number of applicants in the pool (Taylor & Russell, 1939). Altering the selection ratio allows one to vary the competitiveness of the simulated selection context.

Stage 4: Evaluation. In the final stage of our model, we evaluate the outcomes of the selection process using five practical significance metrics relevant to the functioning of real-world organizations—namely, the impact ratio, the odds ratio, risk of disparate treatment, new hire failure rates, and system utility loss due to bias. The first three metrics estimate the impact of selection decisions on subgroup hiring outcomes. The latter two metrics define the expected impact of bias in hiring evaluations on organizational performance.

The impact ratio (*IR*) is a commonly used metric for determining the presence of adverse impact in personnel decisions. As shown in Equation 2 below, it is calculated as

$$impact\ ratio = \frac{p_1}{p_2} \tag{2}$$

where p_1 and p_2 are the selection rates (i.e., number of hired group members/number of applicants from that group) for the disadvantaged and advantaged groups, respectively. Impact

ratios deviating from 1 indicate discrepancies in hiring outcomes across the groups. In many cases, these calculations are coupled with significance testing, although the usefulness of this practice has recently been called into question due to the strong influence of sample size on test significance (Dunleavy, 2010; Morris, 2016; Murphy & Jacobs, 2012). The Uniform Guidelines on Employee Selection Procedures initially set the legal threshold to be used as prima facie evidence of discrimination (somewhat arbitrarily) at ratios less than .80. This means that when protected groups are hired at a rate of 20% or lower than the higher scoring group, adverse impact can be said to exist, thus compelling the organization to prove the validity and/or business necessity of their selection procedures. This cutoff is often referred to as the "4/5ths rule".[3]

The odds ratio (*OR*) is another effect size that has been proposed as a useful metric for determining the presence of adverse impact in personnel decisions (Oswald, Dunleavy, & Shaw, 2016). Like the impact ratio, the odds ratio is calculated using selection rates for each group. However, the odds ratio expresses these differences in terms of the relative probability of a particular hiring outcome for one group relative to another rather than as raw ratios. As shown in Equation 3 below, the odds ratio is calculated as:

$$odds\ ratio = \frac{[p_1/\,(1-p_1)]}{[p_2/\,(1-p_2)]} \tag{3}$$

where p_1 and p_2 are the selection rates for disadvantaged and advantaged groups, respectively. When selection rates are equal, the odds ratio equals 1. The primary advantage of the odds ratio is that it is sensitive to rejection rates as well as selection rates, which enables the odds ratio to detect the presence of adverse impact in hiring contexts where the overall proportion of the applicant pool receiving offers is relatively high. Gastwirth (1988) suggested that odds ratios greater than 1.4 or less than .71 could be considered meaningful disparities in the eyes of the law.

Although adverse impact metrics such as the *IR* and *OR* are commonly used as indicators

of hiring outcome fairness and legal risk, the types of discrimination lawsuits organizations are more likely to face in the real world pertain not to claims of adverse impact but to allegations of *disparate treatment*, which refers to assertions that an applicant was treated differently than other similarly-situated applicants based on group membership. For instance, in a hiring context, if (a) a qualified female applicant is passed over for the position for a less qualified male and (b) it can be demonstrated that the hiring decision was made on the basis of the applicant's gender, the applicant may have legal grounds to file a claim against the organization that disparate treatment has occurred. Thus, to provide a more complete picture of the legal risk to organizations associated with the influence of gender bias on hiring decisions, we sought to include a metric representing the risk of disparate treatment associated with the presence of bias in our model. To create this metric, we started by identifying each female applicant in our simulation samples that possessed true-score qualification ratings exceeding hiring cutoff thresholds. Then we assigned these applicants the "highly qualified" label, which indicates that if a genuinely meritocratic process were in place, they would have been hired for an open position. Then using this distinction, we defined observed rates of disparate treatment as:

$$\textit{rate of disparate treatment} = \frac{n_{q-nh}}{n_{q-tot}} \tag{4}$$

where $n_{q\text{-}nh}$ represents the number of highly qualified female applicants that were (a) not hired for a position and (b) were passed over by a less-qualified male and $n_{q\text{-}tot}$ represents the total number of highly qualified female applicants available in the applicant pool. We also report the percentage increase in disparate treatment in biased models relative to rates observed in a bias-free model to isolate the contribution of bias to this risk. Although the vast number of disparate treatment cases in the real world likely go uncontested, the disparate treatment metric helps supplement the adverse impact metrics by providing a more complete picture of the potential

legal risk to organizations associated with the presence of bias.

New hire failure rates (*NHFR*) is a selection system efficiency metric that specifies the proportion of hired employees that fail to meet the minimum standards required to succeed in the position (i.e., false positives). The *NHFR* is an important metric for organizations given estimates that the cost of replacing a failed hire can be up to three times the amount of that employee's annual base salary (Ruyle, 2012). In our model, each applicant that is selected for hire is marked as either a success or failure depending on whether the chosen applicant's underlying qualifications rating true-score exceeded a base rate cutoff value defining the position's minimum standards. *NHFR*s are then reported as a percentage of the total population of hired applicants marked as "failed hires." The probability of newcomer success in each position is heavily influenced by its base rate, represented in our model as the cutoff point in the distribution of applicant qualifications that defines the proportion of applicants who could succeed on the job if given the opportunity. To estimate the relative contribution of gender bias to new hire failure rates, we report the percentage increase in *NHFR*s in biased models relative to baseline failure rates observed in a bias-free model.

The final organizational performance metric we consider is system utility, which refers to the degree to which a selection system improves the average quality of new hires. Although interest in utility analysis as a decision aid has diminished notably in recent years, Sturman (2012) argued that it can still be useful as a theoretical tool for demonstrating how strategic human resource management concepts affect organizational value. In the present study, we use utility analysis to isolate the economic costs to organizations coinciding with the disruptive influence of bias on hiring decisions. This information is useful because it communicates the value of an HR system in financial terms organizational decision-makers can understand, rather

than in abstract psychometric terms. In our model, the value for system utility per hire is derived from the utility formula initially proposed by Brogden (1946), shown in Equation 5 below.

$$system\ utility\ per\ hire = \underline{z}_y(SD_y) \quad (5)$$

As explained by Schmidt, Hunter, McKenzie, and Muldrow (1979), Brogden's equation specifies that the average gain in system utility per hire can be estimated by multiplying the average criterion scores of those selected for the position ($\underline{z}_y$) times the standard deviation of the criterion in dollars (SD_y). This value can then be plugged into the Brodgen-Cronbach-Gleser utility formula along with contextual information on the number of new hires, the average tenure of each new hire, and the total cost of the assessment to calculate the total cost/savings to the organization. In our model, the value for $\underline{z}_y$ is calculated by averaging the qualifications z-scores of all applicants selected for the position.[4] The value of SD_y is calculated by multiplying the median salary of employees in a given position by .40. This approach is consistent with a conservative implementation of the average-salary method described by Schmidt and Hunter (1983) and can thus be viewed as a low-end estimate of system utility.

SIMULATION 1: THE PRACTICAL IMPACT OF GENDER BIAS IN TYPICAL SELECTION CONTEXTS

With the model in hand, we next turned our attention to the task of estimating the impact of bias on hiring outcomes using a series of simulation-based experiments. In Simulation 1, we sought to determine whether the practical impact of gender bias is limited to a specific set of conditions (e.g., for competitive or challenging jobs or when low-validity assessments are used) or is instead likely to be felt across a broader range of hiring scenarios. Toward this end, in Simulation 1a, we set model parameters to align with characteristics of a "statistically typical" selection context using four simulations that differed only in the relative amount of assessment score variance explained by gender bias and true-score qualification ratings. Then, in Simulation

1b, we further altered key characteristics of the hiring context (i.e., selection ratios, base rates, assessment phase validity, and SD_y) to examine how contextual factors might mitigate (or magnify) the felt impact of bias on individuals and organizations.

Simulation 1: Methods

All simulations presented in this paper were implemented using a macro developed for SAS version 9.4 (SAS Institute Inc., 2015). For each simulation, one million applicant data points were randomly generated using commonly seeded values.[5] The specific parameters specified for each simulation can be found in the online supplement in Appendix B.

To accurately represent a typical selection context as closely as possible, we sought to align the model's starting parameters in Simulation 1a to reflect real-world values. To simplify this process, we focused primarily on statistics derived from selection settings in the United States. The baseline *p* parameter for applicant pool gender representation was set at .56, reflecting statistics derived from the World Bank, suggesting there are slightly more males than females in the typical workforce. For selection ratios, we relied on values derived from recent hiring benchmark surveys, which report that applicant pools in U.S. companies typically range between 20 to 100 applicants per hire (corresponding with selection ratios between .05 and .01) depending on factors such as organizational size, industry, and occupation (ERE, 2016; iCIMS, 2016; Jobvite, 2017; Lever, 2016). Prior research has shown that low selection ratios increase the relative risk of adverse impact associated with subgroup differences in assessment scores (Sackett & Ellingson, 1997). Therefore, in Simulation 1a, we set the value for the selection ratio (*SR*) conservatively at the upward end of this range (i.e., 20 applicants per hire, *SR* = .05). To allow for an examination of findings in hiring contexts that deviate from this norm, in Simulation 1b, we report results in more competitive (i.e., 100 applicants per hire, *SR* = .01) and less

competitive (ten applicants per hire, *SR* =.10, four applicants per hire, *SR* = .25; two applicants per hire, *SR* = .50; and nine openings per ten applicants, *SR* = .90) hiring contexts as well. The base-rate parameter in Simulation 1a was set at .50, which produces new-hire failure rates in our baseline simulations that roughly correspond with the typical involuntary turnover rates reported in SHRM's 2016 Human Capital Benchmarking report (SHRM, 2016). In Simulation 1b, we present additional models representing job contexts with higher (*BR* = .80) and lower (*BR* = .20) base rates as well. To calculate SD_y in Simulation 1a, we multiplied the median salary of employees in the United States in 2020 ($49,348; Bureau of Labor Statistics, 2020) by .40. As noted above, the .40 standard is consistent with a conservative implementation of the average-salary method of utility analysis described by Schmidt and Hunter (1983). It can thus be viewed as a low-end estimate of system utility. Therefore, in Simulation 1b, we expanded our analyses to include SD_y values of .50 and .60 as well.

In all models examined in the present paper, gender bias effects were specified to cover a range of plausible effect sizes aligned with meta-analytic estimates of the impact of gender bias on personnel decisions. In the first model, gender bias and qualification effects were set at 4% (*d* = 0.41) and 35% (*d* = 1.47), respectively. These values are consistent with meta-analytic effects reported by Olian et al. (1988) and represent the upward bound of plausible gender bias estimates reported in the extant literature. In the second model, the gender-bias effect was set at 2.2% (*d* = 0.30), and the qualifications effect was set at 39.3% (*d* = 1.61). These values are consistent with estimates for gender and qualifications effects reported in our updated meta-analysis of experiments manipulating the strength of qualifications and candidate gender on evaluations of hireability. In the third model, the gender-bias effect was set at 1% of the total variance (*d* = 0.20). This third model provides an even more conservative estimated effect size of bias that

accounts for the possibility that meta-analytic estimates might still be overestimating the magnitude of gender bias effects. Finally, in the fourth model, the gender-bias effect was set at 0% (d = 0.00). This "no-bias" model provides a baseline standard of comparison for assessing the relative impact of gender bias on the various outcome criteria of interest.

Simulation 1a Results: The Impact of Bias in Typical Hiring Contexts

As shown in Table 2, the results of Simulation 1a showed that a substantial increase in the risk of discriminatory hiring outcomes in typical hiring contexts could be expected in any hiring process in which systematic gender bias is present. Furthermore, the magnitude of discriminatory hiring outcomes associated with even small amounts of gender bias generally proved to be quite substantial. For instance, in the 2.2% bias model, observed rates of disparate treatment were 13.5% higher than the incidental rates observed in the no bias model. Furthermore, the odds a female would receive a favorable hiring decision in models influenced by a 2.2% bias effect were 49% lower than the odds for comparable males (OR = .51), a rate that clearly violates established adverse impact thresholds. In the presence of a 4% bias effect, rates of disparate treatment associated with bias increased by 20.3%, and a female's overall odds of getting hired were 60% lower than the odds for comparable males (OR = .40). Notably, no evidence of adverse impact was observed in the absence of bias effects (i.e., the 0% model; IR = 0.99, OR = 0.99), indicating that the adverse impact violations reported in Simulation 1a were not the mere result of false positives associated with sampling error or assessment unreliability.

Insert Table 2 about here

A similar, albeit more equivocal, pattern of results emerged when examining the theoretical impact of gender bias on organizational performance metrics. As with adverse impact,

the results shown in Table 2 indicated that the influence of bias on hiring evaluations resulted in hiring inefficiencies (i.e., increases in new hire failure rates and system utility loss) in all models where bias was present. However, the severity of these costs was contingent on the overall amount of bias present in the formation of hiring evaluations, with the most substantial costs observed in the 4% bias model (Δ% in the rate of new hire failure due to bias = 50.2%, utility loss due to bias per hire = -$2,2125.64) and the smallest costs represented in the 1% model (Δ% in the rate of new hire failure due to bias = 7.7%, utility loss due to bias = -$355.36 per hire).

Simulation 1b Results: The Influence of Contextual Factors on the Impact of Bias

In Simulation 1b, we expanded our analysis to examine the influence of assessment validity, selection ratios, base rates, and various estimates of the dollar value of performance (SD_y) on the impact of bias on hiring outcomes. Consistent with prior research (cf. Sackett & Ellingson, 1997), the results of Simulation 1b shown in Table 3 indicate that selection ratios strongly influenced the risk of discriminatory hiring outcomes such that the risk of adverse impact due to bias and overall rates of disparate treatment were higher when lower selection ratios were modeled. However, our simulations also revealed that the range of selection ratios for which bias can be expected to yield practically significant levels of discriminatory hiring outcomes is surprisingly broad. As shown in Table 3, all three bias models produced impact ratios signaling adverse impact when selection ratios were less than .25 (i.e., when there were four or more applicants per position). Furthermore, the odds ratio, which considers rejection rates in addition to acceptance rates, signaled adverse impact in nearly *all* simulations in which bias was present, including those with selection ratios as large as .90 (i.e., when there were at least nine open positions for every ten applicants). In a similar vein, the results in Table 3 show that bias increased the risk of disparate treatment against highly qualified female applicants in all

models in which any amount of bias was present. However, the relative magnitude of this risk was particularly pronounced in models with higher selection ratios.

Insert Table 3 about here

In contrast, we found that variations in system validity had little to no influence on the impact of bias on discriminatory hiring outcomes. Although the results shown in Table 3 do support the notion that the use of more valid assessments *can* reduce the likelihood that highly qualified female applicants will experience disparate treatment, the magnitude of this reduction in risk was relatively small (a difference in rates of disparate treatment between 7-17% between the models simulating low vs. high validity). Furthermore, the proportion of disparate treatment cases directly attributable to biased assessments actually increased slightly (~3-7%) when moving from less valid to more valid assessments. This increase in proportional risk suggests that although improving assessment validity can reduce the risk of disparate treatment caused by unsystematic error, more valid assessments cannot, on their own, reduce the risk of disparate treatment when the underlying source of systematic bias remains unaddressed. Supporting this notion, the results of Simulation 1b showed that variations in system validity had no discernable mitigating effect on increased levels of adverse impact associated with biased hiring evaluations.

Turning next to the organizational performance metrics, we found, unsurprisingly, that lower selection ratios and higher base rates were associated with general reductions in the overall frequency of new hire failure, particularly in models where more valid assessments were used (see Table 4). On the other hand, the relative contribution of bias to new hire failure rates was actually *greater* in more competitive hiring contexts than in less competitive hiring contexts. Interestingly, the results of Simulation 1b shown in Table 4 support the notion that the use of

more valid assessments *can* partially mitigate the negative impact of bias on new hire failure rates. However, ancillary analysis providing a more in-depth examination of this effect across a broader range of validity estimates indicated that the mitigating effect of validity on bias's contribution to new hire failure was subject to diminishing returns at assessment validities >.25.

Insert Table 4 about here

A similar pattern of effects was observed for the impact of bias on system utility across a range of hiring contexts. Specifically, Table 5 shows that the negative impact of bias on system utility was greatest in models representing competitive hiring contexts characterized by low selection ratios and in high-stakes jobs where the financial implications of variations in employee performance are more pronounced. Furthermore, our results suggest that estimates of utility loss due to bias were substantially higher in models with lower validity assessments. This finding suggests that the results of Simulation 1a likely underestimate the negative consequences of bias in hiring contexts with less than optimal levels of validity. Furthermore, it should be noted that the negative impact of bias on organizational performance remained quite substantial, even when highly valid assessments were used. Collectively, these findings suggest that a diversity strategy focused exclusively on increasing assessment validity can reduce, but not eliminate the negative impact of underlying gender biases on organizational performance.

Insert Table 5 about here

Simulation 1: Discussion

The purpose of Simulation 1 was twofold. First, we sought to determine the extent to which small bias effects can meaningfully impact individual and organizational hiring outcomes in typical hiring contexts. Second, we sought to examine the extent to which characteristics of

the hiring context influence these findings. The results of these simulations highlight two key points that inform scholarly conversations surrounding the impact of bias in hiring contexts.

First, the results of Simulation 1a help establish a basic principle all hiring managers need to understand, which is that even seemingly small amounts of bias in the assessment phase can have a profound negative impact on a wide range of hiring outcomes for individuals and organizations alike. Supporting this notion, we found evidence of discriminatory hiring outcomes in every simulation in which gender bias was present, even when bias accounted for a mere 1% of the variance in overall assessment scores. Moreover, the risk of discrimination in these models was generally quite substantial, often far exceeding established cutoffs for practically significant subgroup differences in hiring outcomes, suggesting a substantial legal risk associated with failure to address even residual amounts of bias in candidate evaluations.

However, our results show the costs of gender bias to organizations go well beyond the risk of litigation. As a particularly problematic source of systematic construct irrelevant variance, gender bias in hiring evaluations contributes to suboptimal hiring decisions when less qualified members of a favored group are selected over more qualified members of a minority group. The consequences of these suboptimal decisions include inefficiencies in hiring decisions resulting from increases in new-hire failure rates and decreases in selection system utility. To put these costs into context, a typical Fortune 500 company that hires 8,000 new employees a year with a 1% gender-bias effect in the company's selection procedures can expect the botched hiring of an additional 32 new employees (8,000 hires × .04% increase in new hire failure rate) and a loss in productivity totaling approximately $2.8 million per year (8,000 hires × $355 utility loss per hire) resulting from suboptimal hiring decisions alone. A similar company with a 4% bias effect can expect an additional 192 failed hires (8,000 hires × 2.4% increase in new hire failure rate)

and a loss in productivity totaling $17 million per year (8,000 hires × $2,125 utility loss per hire). Of course, it is worth noting here that these estimates are assuming extremely valid assessments assist hiring decisions. Unfortunately, most organizations rely on assessment techniques with substantially lower amounts of predictive validity, and can thus expect much larger cost estimates ranging from 2x to 7x the magnitude of the figures provided above. Collectively, these results make a strong case that reducing or eliminating gender bias in hiring decisions is not just the ethically correct thing to do—it is also financially prudent.

Second, Simulation 1 findings suggest the negative impact of gender bias is not constrained to competitive or challenging jobs or hiring contexts where low-validity assessments are used but is likely to be felt in the vast majority of contexts in which hiring decisions are made. In particular, careful consideration of the role of assessment validity in mitigating the negative impact of bias raises questions about the extent to which highly valid assessments can offset suboptimal hiring outcomes for minority applicants when residual amounts of bias remain unaddressed. For instance, as noted above, Simulation 1a results likely underestimate the felt impact of bias on organizational performance when predictive validity coefficients are less than .50. Unfortunately, many HR professionals remain skeptical of assessments that have been shown to produce high validity coefficients (Rynes, Colbert, & Brown, 2002) and of the importance of predictive validity as a whole (Rynes, Giluk, & Brown, 2007). As such, we expect that in real-world hiring contexts, suboptimal validities are likely the norm, not the exception.

Furthermore, although the findings of Simulation 1 support the idea that a bias mitigation strategy focused on increasing validity can theoretically reduce the impact of bias on hiring outcomes, this mitigating effect only applies to the impact of bias on organizational performance, not disparate treatment of female and male job applicants. Further, our findings demonstrate that

a strategy built around addressing problems of bias through increases in assessment validity cannot address problems of inequitable hiring outcomes for females when even seemingly trivial amounts of bias remain in the evaluative process. A possible exception to this rule is when it can also be clearly demonstrated that increases in job-relevant information simultaneously root out *all* sources of systemic bias (however small) that exist in the making of hiring decisions. The extent to which this is possible remains an open question, subject to intense debate (Landy, 2008b), given that some subjective interpretation of performance metrics is unavoidable. Nevertheless, the findings of Simulation 1 raise the interesting, potentially provocative suggestion that validity-based legal defenses against claims of adverse impact should be required to speak explicitly to questions of how existing assessment strategies obviate the potential for bias to influence the decision-making process beyond improvements in predictive potential.

SIMULATION 2: THE INFLUENCE OF TARGETED RECRUITMENT ON THE IMPACT OF GENDER BIAS

Organizations interested in proactively increasing the representation of minorities and females in their workforce are, in some national contexts, constrained by legal rulings that disallow the use of formal quotas in selection decisions. More generally, many organizations turn to diversity-focused recruitment strategies as their primary means of diversifying their workforce through the hiring process. Along these lines, one common tacttic often advanced as a technique for combating adverse impact is targeted recruitment (Avery & McKay, 2006; Newman & Lyon, 2009), which refers to “practices and decisions that affect either the number or the types of targeted individuals who are willing to apply for, or to accept, a given vacancy” (Newman & Lyon, 2009, p. 299). Targeted recruitment works by using specialized recruitment strategies to increase the representation of qualified members from underrepresented groups in the applicant pool, thus increasing the likelihood that members from those groups will be selected. For

example, an organization in a traditionally male-dominated industry may actively seek out highly qualified female candidates and encourage them to apply for open positions hoping that such practices will increase the female representation of their workforce. Support for this strategy can be found in research showing that applicant pool characteristics can have a considerable influence on the risk of adverse impact (Murphy, Osten, & Myors, 1995; Ryan, Ployhart, & Friedel, 1998). However, it is less certain whether targeted recruitment can fully counteract the underlying issues of gender bias driving many instances of adverse impact in the first place.

Two distinct theoretical mechanisms can be argued to contribute to the success of targeted recruitment initiatives. The first is the overall increased representation of female and minority candidates in the applicant pool resulting from concerted recruitment efforts. The second is an overall increase in the underlying qualification levels of applicants resulting from the targeting of highly qualified members of protected groups. In Simulation 2, we modify our model to examine the relative contribution of each of these mechanisms and targeted recruitment as a whole on the impact of gender bias on hiring outcomes.

Simulation 2: Methods

We started by modeling a hiring scenario representing a traditionally male-dominated industry (90% male applicants, 10% female applicants). Then, in Simulation 2a, we examined outcomes in hiring scenarios in which female representation was increased by 10%, 20%, 50%, and 100% due to successful targeted recruitment initiatives. Simulation results for an extreme scenario in which females outnumber males in the applicant pool by 9 to 1 are also provided as an additional point of comparison. The purpose of Simulation 2a was to isolate the influence of increased female applicant pool representation on hiring outcomes. However, many targeted recruitment initiatives emphasize the need to target highly-qualified members of protected

groups in particular (Newman & Lyon, 2009). Therefore, in Simulation 2b, we increased the average qualifications rating for female applicants by d = .25 to model outcomes associated with successful attempts to attract higher-quality female applicants. As before, model parameters for the selection ratio, base rate, and SD_y were set at .05, .50, and .40, respectively.

Simulation 2a Results: Increasing Female Applicant Pool Representation

As shown in Table 6, the results of Simulation 2a showed that increasing female representation in the applicant pool had little impact on average hiring outcomes for female applicants as a whole. Even dramatic shifts in applicant pool characteristics (e.g., increasing female representation from 10% to 90%) increased female odds of hire by less than 2%. As a result, females in the biased models were hired at rates well below males, even when overall female representation in the applicant pool surpassed their male counterparts. The results of Simulation 2a further revealed that the negative impact of bias on organizational performance metrics actually *increased* slightly as a function of female applicant pool representation.

Insert Table 6 about here

Simulation 2b Results: Targeting Highly-qualified Female Applicants

Interestingly, the results of Simulation 2b shown in Table 7 supported the notion that directly targeting more qualified female applicants can improve overall hiring outcomes for women, even in the face of bias. Specifically, we found that even a mere 10% increase in the proportion of highly-qualified female applicants increased adverse impact ratios by 21% and 27% for female applicants when faced with 4% and 2.2% bias effects, respectively. However, reductions in the rate of disparate treatment against highly qualified females were substantially more modest (~Δ2%), and new hire failure rates and utility loss due to bias both *increased* as a

result of targeted recruitment efforts when sources of bias remained unresolved. Furthermore, it should be noted that in both the 1% and no bias models, targeted recruitment resulted in a substantial increase in the risk of adverse impact against male applicants.

Insert Table 7 about here

Simulation 2: Discussion

The rationale provided in support of targeted recruitment is that "adverse impact depends on the selection ratio in each group, and the selection ratio depends on the number of applicants." Thus, "the larger the pool of qualified applicants in the minority group, the higher the selection ratio and the lower the probability of adverse impact" (Cascio & Aguinis, 2011, p. 183). This proposition is attractive to many organizations because non-preferential strategies such as targeted recruitment do not carry the same legal barriers in some contexts as more direct affirmative action strategies in which clear preferences are given to minority group members in hiring decisions (Kravitz, 2008). In this regard, the results of Simulation 2 provide some evidence in support of the notion that efforts to increase the representation of females in the applicant pool can help organizations reduce the risk of adverse impact for organizations by increasing the likelihood that qualified members of protected groups will be selected.

However, our findings also suggest that targeted recruitment will not meaningfully reduce the risk of disparate treatment against highly qualified females. Indeed, we found that qualified female applicants continue to face difficulties in cases where hiring evaluations are influenced by bias, even as their representation in the applicant pool improves because they are still required to meet a higher standard of evaluation than their male counterparts. This holds true even when the performance standard is only slightly higher for female candidates than for male candidates (i.e., the gender bias in evaluations is quite small). Thus, although our results support

prior research suggesting that targeted recruitment can contribute to more equitable hiring outcomes (e.g., Newman & Lyon, 2009), we caution against overstating the potential of this strategy as a panacea for workplace discrimination.

GENERAL DISCUSSION

Taken together, the findings of our updated meta-analysis were somewhat enigmatic. Consistent with prior research (e.g., Davison & Burke, 2000; Olian et al., 1988), we found evidence that gender bias still systematically influences evaluations of candidate hireability. However, observed bias effects remained quite small, particularly compared to the far more robust influence of applicant qualifications. So what then is one to do with this information? In the absence of context, it can be challenging to say when residual amounts of bias matter or even if they matter at all. Selection researchers who aspire to have a discernible positive impact on employees and the organizations for whom they work cannot say their task is complete until the consequences of bias in the hiring process are well understood (Messick, 1995).

Toward this end, we used a series of simulation-based experiments to better integrate this context within the conversation surrounding the impact of bias on hiring evaluations. The results of these simulations (summarized in Table 8) proved to be quite informative and point to three fundamental principles that we believe can help guide future research and practice focused on solving problems of subgroup bias in hiring evaluations. In the following sections, we present and then elaborate on the evidence supporting each of these three principles, discuss the practical implications of our findings, and outline future opportunities for research on this topic.

Insert Table 8 about here

Principle 1: All Bias Matters, No Matter How Small

The first principle—all bias matters, no matter how small—is not an entirely new

observation. Scholars in the fields of economics (Blank, 2005), sociology (Reskin, 2011), and psychology (Greenwald et al., 2015) have each speculated on the possibility that significant societal consequences can result from small, seemingly insignificant causal effects. Our findings substantiate these speculations with empirical support. Indeed, across the various simulations examined in the present study, we consistently found that it was the mere *presence* of bias in a given model—not its amount—that defined whether bias presented a practically significant problem in hiring contexts. In fact, when expanding the results of Simulation 1a to examine the impact of bias effects less than 1%, we found that gender bias in hiring evaluations would need to be reduced to less than 0.3% of the overall variance to avoid practically significant inequities in hiring outcomes.[6] Given these results, we advocate for a "no tolerance" policy when it comes to the question of how much bias can be reasonably tolerated in the evaluative process, as our simulation findings show that a failure to address even small amounts of existing biases can have substantial negative consequences on hiring outcomes.

Furthermore, exploring the implications of this principle in hiring contexts suggest that confronting problems of bias in organizations can yield benefits not just at the societal level (as has previously been suggested) but at the organizational and individual level as well. Specifically, our findings indicated that small amounts of bias in hiring evaluations increased not only legal risk to organizations associated with adverse impact, but led to substantial increases in rates of disparate treatment, new hire failure rates, and productivity losses as well. By demonstrating the potential costs to organizations resulting from unaddressed biases, the present study shifts the impetus of action from societal policy to the organizations themselves. Securing the support of an organization's leadership is a critical catalyst for enacting organizational change (Yukl, 2008). Toward that end, the present study provides new financial justification for

organizations to invest the resources needed to develop innovative solutions for addressing problems of residual bias that continue to plague organizational functioning.

Before moving on, it is worth noting that in practice, detecting subtle gender bias effects in real-world hiring contexts using statistical significance tests can be quite difficult. For instance, our analyses indicate that a sample of at least 2704 applicants is needed to ensure significance tests are sufficiently powerful to confidently rule out practically meaningful (but statistically tiny) effects. As noted earlier, most hiring decisions involve between 20-100 total applicants, which can make it difficult to tell the difference between unbiased and biased organizations in the context of a single hiring decision. These findings underscore the point that in many hiring scenarios, significance tests alone may not be enough to indicate whether hiring disparities are large enough to be of practical concern (Morris, 2016). Such evaluations are better reserved for long-term analyses based on accumulations of a large sample of applicants (Baldus & Cole, 1980; Morris, 2016). However, waiting until bias effects are detectable before acting risks bringing substantial harm to individuals and organizations alike. Thus, in the short term, we argue it is in an organization's best interest to assume at least small amounts of bias exist in all hiring decisions and to work proactively to mitigate its effects.

Principle 2: Context Alters, but Does Not Obviate the Impact of Bias on Hiring Outcomes

No two hiring contexts are the same. Even within organizations, the likelihood of success for a hiring decision is contingent on variations in the validity of the assessments used to evaluate candidate qualifications, the overall number of candidates applying for each open position, characteristics of the applicant pool, department hiring policies, and the anticipated costs of making a wrong choice. This point is particularly relevant when attempting to estimate the practical impact of bias in selection contexts, as prior work on systematic subgroup

differences in assessment scores has shown that lower selection ratios can amplify the risk of adverse impact (Sackett & Ellingson, 1997). In the present study, we sought to determine how much the impact of bias is contingent on contextual variations across hiring scenarios.

Along these lines, our findings point to a second fundamental principle, which is that context alters, but does not obviate the need to address problems of bias in hiring evaluations. Specifically, our findings showed that the estimated impact of bias on key outcome metrics did change (in some cases substantially) as a function of variations in contextual parameters. For instance, consistent with research by Sackett and Ellingson (1997), we found that the risk of adverse impact directly attributable to systematic bias was generally higher when selection ratios were lower. Conversely, the relative contribution of bias to disparate treatment actually *increased* as selection ratios went up. Interestingly, increases in system validity partially mitigated the negative impact of bias on organizational performance. However, even in simulated hiring contexts utilizing extremely valid assessments (r = .50), a substantial impact of bias on organizational performance remained. Female representation in the applicant pool is another contextual factor thought to play a critical role in facilitating progress toward more equitable hiring outcomes. However, our simulations produced no evidence supporting the idea that efforts to increase female representation in the applicant pool alone can meaningfully reduce the risk of discriminatory hiring outcomes against female candidates in the face of bias.

Indeed, looking across the 100+ hiring scenarios examined in the present study, we could not find any combination of factors where the presence of bias produced outcomes that would not cause some degree of concern for organizational decision-makers. This principle is particularly true when examining the impact of bias on hiring discrimination. Across Simulations 1 and 2, the only scenarios in which bias did not unambiguously signal practically significant

levels of adverse impact was (a) in a single model where bias was 1%, and selection ratios were equal to .50 and (b) when females in the applicant pool possessed average qualification levels far surpassing those of their male counterparts. However, even in these isolated cases, bias increased the risk of disparate treatment by 2-27%, suggesting that the traditional overreliance on impact ratios as the primary indicator of discriminatory hiring outcomes may be obscuring the actual risk of discrimination associated with biased decision-making processes. In sum, our results suggest that although context can indeed shape the magnitude of bias's impact, the safest position is to assume that bias will cause significant problems when it is present.

Principle 3: Unaddressed Bias will Undermine Even Well-Intentioned Diversity Initiatives

In their 2008 paper on the diversity-validity dilemma, Ployhart and Holtz (2008) summarized the existing body of research on strategies for reducing subgroup differences and adverse impact in selection decisions. At the end of their review, they settled on recommending a combination of strategies focused on (a) using less cognitively loaded predictors to reduce the likelihood of subgroup differences (Hough et al., 2001; Schmidt et al. 1996) and (b) fostering favorable applicant reactions through approaches like targeted recruitment to increase the number of qualified subgroup members in the applicant pool (Ryan, Sacco, McFarland, and Kriska, 2000, Tam, Murphy, & Lyall, 2004). In their review of diversity staffing practices, Avery et al. (2012) further emphasized the importance of targeted recruitment (Newman & Lyon, 2009) as a primary means of facilitating organizational diversity.

Non-preferential interventions such as these are attractive to organizations that wish to both (a) hire high-quality candidates and (b) acquire a diverse workforce (Pyburn, Ployhart, & Kravitz, 2008). Unfortunately, the results of our simulations suggest that even established diversity initiatives will struggle to realize their full potential in cases where systematic subgroup

bias continues to influence the hiring process. As such, the final principle derived from our simulations is that unaddressed sources of systematic subgroup bias in the formation of hiring evaluations will undermine the effectiveness of even well-intentioned diversity initiatives. In support of this notion, our simulation results revealed that neither increases in assessment validity nor attempts to expand female representation in the applicant pool led to equitable hiring outcomes for qualified female applicants in the presence of bias. One possible exception to this is that recruitment strategies proactively targeting highly qualified female applicants could forestall signals of adverse impact against female applicants, even in models in which a pro-male bias in evaluations remained. However, a closer examination of other hiring metrics beyond impact ratios revealed that the use of this strategy might not always be the panacea to problems of discrimination that it first appears. For instance, our simulations showed that the practice of targeting highly qualified female applicants did not meaningfully reduce rates of disparate treatment against the same group of female applicants it purports to attract. Moreover, this approach paradoxically exacerbated the negative impact of bias on new hire failure rates and system utility loss due to bias, suggesting that gains in impact ratios may come at the cost of organizational performance when the influence of bias remains. Furthermore, a closer examination of impact ratios across the models revealed that once bias is successfully removed from the evaluative process, organizations will face potentially counterproductive side effects resulting from a corresponding increase in the paradoxical risk of adverse impact against male applicants. All this is not to say that we recommend organizations give up in their attempts to use more valid assessments or improve the representation of protected class members in their applicant pools. Rather we argue that organizational scholars need to start thinking about subgroup bias in hiring decisions as a problem that cannot be ignored or resolved indirectly.

Model Assumptions and Study Limitations

Any computer simulation is only as good as the assumptions upon which it is built. In this regard, we made several key assumptions in the model specification process that have implications for interpreting the present findings. For instance, one assumption inherent to our model is that staffing is a one-way process in which applicants are entirely subject to the organization's whims. We made this assumption to allow us to focus on the consequences of bias in the assessment and evaluation phase, which is typically under the purview of the organization rather than the applicant. Of course, this is an oversimplification, as researchers have long acknowledged the role of applicant decisions within the scope of the hiring decision-making process (Carlson & Connerley, 2003; Murphy, 1986). Nevertheless, we felt this over-simplification was justifiable in this case, as it is unlikely that male applicants differ dramatically from female applicants in their willingness to accept a job offer in general. Although failing to account for two-way decisions may contribute to overestimates of the overall utility of the selection system by between 30 to 80% (Murphy, 1986), it is unlikely to meaningfully impact the relative contribution of gender bias to these estimates. Fortunately, one of the advantages of using computer simulations is that the model can easily be expanded to incorporate other selection system characteristics (e.g., applicant decisions) to understand their influence on hiring outcomes. Toward this end, prior simulation work by Tam et al. (2004) may prove useful as a starting point for model expansion in that they provide several competing models of applicant withdrawal through which models of the impact of gender bias can be compared.

A second critical assumption of our model is that males and females in the applicant pool do not systematically differ in their group-level true-score qualification ratings. Although the gender similarities hypothesis supports the tenability of this assumption in regards to applicant

psychological characteristics most relevant to job performance (Hyde, 2005), systematic subgroup differences might still arise in some cases as a result of societal, educational, or opportunity-based advantages for one group over the other (Moughari, Gunn-Wright, & Gault, 2012). Furthermore, underlying subgroup differences in predictor domains might manifest as differences in true-score qualifications, which is potentially problematic because it raises the theoretical possibility that fair assessments may still contribute to inequitable outcomes. To consider the implications of this assumption's violations for our findings vis-à-vis gender bias, we conducted sensitivity analyses in which varying levels of underlying subgroup differences in qualifications were added to the model.[7] These tests' results reveal that underlying subgroup differences can, indeed, suppress the negative impact of bias on new hire failure rates and system efficiency, thus potentially weakening the business case presented in Simulation 1. However, the range of situations in which this mitigating effect is likely to be realized is limited to scenarios where (a) underlying subgroup qualification differences are substantial (i.e., differences in the magnitude of d = 0.5 or higher) *and* (b) the amount of bias present in the assessment phase is relatively small. Furthermore, these tests' results suggest that variation in subgroup differences had little impact on the contribution of bias to discriminatory hiring outcomes and can even amplify the risk of discrimination when the direction of advantages of subgroup differences and bias favors one group over another.

Another assumption of our model is that performance is normally distributed. Sensitivity analyses testing the implications of violations of this assumption again suggest that our findings were indeed somewhat sensitive to variations in performance distributions. For instance, when a lognormal distribution represents variation in applicant true-score qualifications rather than a normal distribution, the influence of bias on adverse impact was slightly less pronounced.

However, this alternative distribution also led to increases (in some cases, substantial increases) in the impact of bias on disparate treatment and financial performance metrics. Although interesting as a potential boundary condition, a specific focus on the implications of performance distributions for bias's effects was somewhat beyond the scope of the present investigation. Moreover, research by Vancouver et al. (2016) suggests that in typical selection contexts using composite evaluations, normal distributions in performance can be expected to emerge, even when highly skewed individual performance metrics used to form these evaluations (i.e., number of publications produced by a researcher throughout their career) are not, which substantiates the viability of the normality assumption for performance variance in most hiring contexts. Nevertheless, future research should explore the implications of performance distributions for the impact of bias in hiring contexts in greater depth, as our tests suggest departures from normality in performance distributions can influence the magnitude of biases effects on hiring outcomes.

Finally, to render our research questions tractable, we carried out simplified simulations of group-based discrimination in hiring that ignored complicating factors such as the potential organizational benefits of a more diverse workforce and which focused exclusively on biases involving one major demographic distinction (i.e., between women and men). As emphasized earlier, gender is not binary, involving further identities and categories (Hyde et al., 2019), and future work should seek to capture this when modeling selection decisions and organizational diversity and performance. There are also reasons to expect racial biases against negatively stereotyped ethnic minorities (e.g., Black Americans in the United States) to be as strong, if not stronger, than against women. For example, recent meta-analytic evidence indicates the stability of race-based discrimination in field audits over time (Quillian, Pager, Hexel, & Midtbøen, 2017), coupled with meaningful change in gender stereotypes regarding competence (Eagly,

Nater, Miller, Kaufmann, & Sczesny, in press). Further, there are theoretical and empirical grounds to expect discrimination in favor of female candidates in contexts in which feminine traits are prototypical for the job in question (Eagly & Karau, 2002; Glick et al., 1988; Kalin & Hodgins, 1984). Finally, research on intersectionality indicates that target demographic characteristics (e.g., gender and race) interact to predict discriminatory treatment from others, such that group-based inequalities cannot be examined solely in isolation (Browne & Misra, 2003; Sidanius & Pratto, 2001). More comprehensive simulations of inequalities in employment settings will need to grapple with these complex and interacting variables.

Practical Implications and Future Research Directions

Acknowledging the potency of small bias effects requires a shift in the goal of evidence-based diversity and inclusion initiatives away from a mere reduction of bias's role in hiring decisions and toward a new standard defined by complete elimination of bias in the hiring process. Adopting this more extreme standard can paradoxically make the goal of developing equitable hiring processes more attainable by allowing researchers and practitioners to focus on innovative approaches that do not require overcoming the persistent and deeply-wired human tendency to allow bias to influence decision-making processes.

For instance, one unique strategy that has been proposed for reducing the impact of bias on hiring decisions is limiting the extent to which unnecessary group membership information is available to decision-makers. This solution's potential power was famously demonstrated in orchestra settings, where the inclusion of a physical barrier to conceal candidate identity from a selection jury increased the probability that females would advance out of preliminary rounds by ~50% (Goldin & Rouse, 2000). Similarly, anonymous job application procedures were shown to improve the probability of job offers for female applicants across a broader range of jobs and

occupations (Åslund & Skans, 2012). However, such concealments are not always possible or practical in contexts where contact between applicants and decision-makers is unavoidable.

As such, there is a need to identify and apply practical interventions to reduce the extent to which distracting group-identity related information influences hiring decisions. More generally, accepting that some degree of bias in the hiring process is very likely where human judgment is involved highlights the need to mitigate adverse outcomes as soon as they start to emerge and to search for potential sources of bias where we may not be currently looking (Klein, Hill, Hammond, & Stice-Lusvardi, 2020). A few such interventions include eliminating unstructured aspects of the interview process as much as possible (Bragger, Kutcher, Morgan, & Firth, 2002), committing to both the hiring criteria and their relative weighting before knowing the demographic group memberships of the applicant (Uhlmann & Cohen, 2005), encouraging joint rather than separate evaluation of candidates (Bohnet, Van Geen, & Bazerman, 2016), and keeping hiring managers informed about the ratio at which they are selecting female and male candidates relative to the available pool (see Bohnet, 2016, for a review). As another example, many U.S. states have begun outlawing the extent to which organizations can demand applicants share salary history information until after the offer is extended to prevent gender-based pay disparities from following them throughout their career (Dive, 2019). Future laboratory and field research should focus on identifying additional ways that psychological biases' influence on hiring outcomes can be directly counteracted within the structure of the hiring process itself.

REFERENCES

Åslund, O., & Skans, O. N. 2012. Do anonymous job application procedures level the playing field? *ILR Review, 65*(1), 82-107.

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. 2011. The condition of education 2011. NCES 2011-033. *National Center for Education Statistics*.

Avery, D. R., & McKay, P. F. 2006. Target practice: An organizational impression management approach to attracting minority and female job applicants. *Personnel Psychology, 59*(1), 157-187.

Avery, D. R., McKay, P. F., & Volpone, S. D. 2012. Diversity staffing: Inclusive personnel recruitment and selection practices. *The Oxford handbook of diversity and work*, 282-296.

Ayres, I. 2003. *Pervasive prejudice?: Unconventional evidence of race and gender discrimination*: University of Chicago Press.

Baldus, D. C., & Cole, J. W. 1980. *Statistical proof of discrimination*. New York: McGraw-Hill.

Banaji, M. R., & Greenwald, A. G. 1995. Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology, 68*(2), 181-198.

Binning, J. F., & Barrett, G. V. 1989. Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*(3), 478-494.

Blank, R. M. 2005. Tracing the economic impact of cumulative discrimination. *American Economic Review, 95*(2), 99-103.

Blau, F. D., & Kahn, L. M. 2003. Understanding international differences in the gender pay gap. *Journal of Labor Economics, 21*(1), 106-144.

Blau, F. D., & Kahn, L. M. 2017. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature, 55*(3), 789-865.

Bohnet, I. 2016. *What works: Gender equality by design.* Cambridge, MA: Harvard University Press.

Bohnet, I., Van Geen, A., & Bazerman, M. 2016. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science, 62*(5), 1225-1234.

Bowen, C. C., Swim, J. K., & Jacobs, R. R. 2000. Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology, 30*(10), 2194-2215.

Bragger, J.D., Kutcher, E., Morgan, J., & Firth, P. 2002. The effects of the structured interview on reducing biases against pregnant job applicants. *Sex Roles, 46*, 215–226.

Brogden, H. E. 1946. On the interpretation of the correlation coefficient as a measure of

predictive efficiency. *Journal of Educational Psychology, 37*(2), 65-76.

Browne, I., & Misra, J. 2003. The intersection of gender and race in the labor market. *Annual Review of Sociology, 29,* 487-513.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. 2020. *Gender differences in peer recognition by economists*. Unpublished manuscript.

Carlson, K. D., & Connerley, M. L. 2003. The staffing cycles framework: Viewing staffing as a system of decision events. *Journal of Management, 29*(1), 51-78.

Cascio, W. F., & Aguinis, H. 2011. Fairness in employment decisions. In W. F. Cascio & H. Aguinis (Eds.), *Applied Psychology in Human Resource Management* (Seventh Edition ed., pp. 167-192). Upper Saddle Lake, NJ: Prentice Hall.

Cascio, W. F., & Boudreau, J. 2010. *Investing in people: Financial impact of human resource initiatives*. Upper Saddle River, New Jersey: Ft Press.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. 2007. Developing theory through simulation methods. *Academy of Management Review, 32*(2), 480-499.

Davison, H. K., & Burke, M. J. 2000. Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior, 56*(2), 225-248.

Duckitt, J. H. 1992. Psychology and prejudice: A historical analysis and integrative framework. *American Psychologist, 47*(10), 1182.

Dunleavy, E. M. 2010. A consideration of practical significance in adverse impact analysis. *Washington, DC: DCI Consulting Group, July.*

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

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. 1992. Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin, 111*(1), 3-22.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (in press). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist*.

Equal Employment Opportunity Commission (EEOC). 2010. *Women's work group report.* Retrieved from https://www.eeoc.gov/federal-sector/report/eeoc-womens-work-group-report

ERE. 2016. *ERE talent acquisition benchmarking survey.* Retrieved from https://info.eremedia.com/benchmarking/

Gastwirth, J. L. 1988. *Statistical reasoning in law and public policy: Tort law, evidence and*

health (Vol. 2): Elsevier.

Glick, P., Zion, C., & Nelson, C. 1988. What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology, 55*, 178–186.

Goldin, C., & Rouse, C. 2000. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American economic review, 90*(4), 715-741.

Greenwald, A. G. 2008. Landy is correct: Stereotyping can be moderated by individuating the out-group and by being accountable. *Industrial and Organizational Psychology, 1*(4), 430-435.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology, 108*(4), 553-561.

Heilman, M. E. 2001. Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57*(4), 657-674.

Heilman, M. E., & Eagly, A. H. 2008. Gender stereotypes are alive, well, and busy producing workplace discrimination. *Industrial and Organizational Psychology, 1*(04), 393-398.

Hyde, J. S. 2005. The gender similarities hypothesis. *American Psychologist, 60*(6), 581-591.

iCIMS. 2016. *U.S. Hiring Trends Q1 2016*. Retrieved from https://www.icims.com/monthly-hiring-indicator/

Jobvite. 2017. *10th Annual Recruiting Funnel Benchmark Results*. Retrieved from https://www.jobvite.com/wp-content/uploads/2017/05/Jobvite_2017_Recruiting_Funnel_Benchmark_Report.pdf

Kalin, R., & Hodgins, D.C. 1984. Sex bias in judgments of occupational suitability. *Canadian Journal of Behavioral Science, 16,* 311–325.

Klein, F. B., Hill, A. D., Hammond, R., & Stice-Lusvardi, R. 2020. The gender equity gap: A multistudy investigation of within-job inequality in equity-based awards. *Journal of Applied Psychology*, Advanced online publication.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. 2015. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology, 100*(1), 128.

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. 2011. Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137*(4), 616.

Kravitz, D. A. 2008. The diversity–validity dilemma: Beyond selection-the role of affirmative action. *Personnel Psychology, 61*(1), 173-193.

Landy, F. J. 2008a. Stereotypes, bias, and personnel decisions: Strange and stranger. *Industrial and Organizational Psychology, 1*(4), 379-392.

Landy, F. J. 2008b. Stereotyping, implicit association theory, and personnel decisions: I guess we will just have to agree to disagree. *Industrial and Organizational Psychology, 1*(4), 444-453.

Law, A. M., Kelton, W. D., & Kelton, W. D. 1991. *Simulation modeling and analysis* (Vol. 2): McGraw-Hill New York.

Ledvinka, J. 1979. The statistical definition of fairness in the federal selection guidelines and its implications for minority employment. *Personnel Psychology, 32*(3), 551-562.

Lever. 2016. *The Little Grey Book of Recruiting Benchmarks 2016*. Retrieved from https://www.lever.co/resources/little-grey-book-of-recruiting-benchmarks/

Lipsey, M. W., & Wilson, D. B. 2001. *Practical meta-analysis*: SAGE publications, Inc.

Madera, J. M., Hebl, M. R., & Martin, R. C. 2009. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology, 94*(6), 1591.

Martell, R. F., Emrich, C. G., & Robison-Cox, J. 2012. From bias to exclusion: A multilevel emergent theory of gender segregation in organizations. *Research in Organizational Behavior, 32*, 137-162.

Martell, R. F., Lane, D. M., & Emrich, C. 1996. Male-female differences: A computer simulation. *American Psychologist, 51*(2), 157-158.

McGrath, J. E. 1981. Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist, 25*(2), 179-210.

Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Morris, S. B. 2016. Statistical Significance Testing in Adverse Impact Analysis *Adverse Impact Analysis* (pp. 91-111): Routledge.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474-16479.

Moughari, L., Gunn-Wright, R., & Gault, B. (2012). Gender segregation in fields of study at community colleges and implications for future earnings. Fact Sheet# C395. *Institute for Women's Policy Research*.

Murphy, K. R. 1986. When your top choice turns you down: Effect of rejected offers on the

utility of selection tests. *Psychological Bulletin, 99*(1), 133-138.

Murphy, K. R., & Jacobs, R. R. 2012. Using effect size measures to reform the determination of adverse impact in equal employment litigation. *Psychology, Public Policy, and Law, 18*(3), 477-499.

Murphy, K. R., Osten, K., & Myors, B. 1995. Modeling the effects of banding in personnel selection. *Personnel Psychology, 48*(1), 61-84.

Neumark, D., Bank, R. J., & Van Nort, K. D. 1996. Sex discrimination in restaurant hiring: An audit study. *The Quarterly Journal of Economics, 111*(3), 915-941.

Newman, D. A., & Lyon, J. S. 2009. Recruitment efforts to reduce adverse impact: targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology, 94*(2), 298-317.

Norton, M. I., Vandello, J. A., & Darley, J. M. 2004. Casuistry and social category bias. *Journal of Personality and Social Psychology, 87*(6), 817-831.

Olian, J. D., Schwab, D. P., & Haberfeld, Y. 1988. The impact of applicant gender compared to qualifications on hiring recommendations: A meta-analysis of experimental studies. *Organizational Behavior and Human Decision Processes, 41*(2), 180-195.

Oswald, F. L., Dunleavy, E. M., & Shaw, A. 2016. Measuring practical significance in adverse impact analysis. In S. B. Morris & E. M. Dunleavy (Eds.), *Adverse Impact Analysis: Understanding Data, Statistics, and Risk* (pp. 92-112). New York: Routledge.

Platt, J. R. 1964. Strong inference. *Science, 146*(3642), 347-353.

Ployhart, R. E., & Holtz, B. C. 2008. The diversity–validity dilemma: strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*(1), 153-172.

Pyburn, J. K. M., Ployhart, R. E., & Kravitz, D. A. 2008. The diversity–validity dilemma: overview and legal context. *Personnel Psychology, 61*(1), 143-151.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences, 114*(41), 10870–10875.

Reskin, B. F. 2011. Rethinking employment discrimination and its remedies. *The inequality reader: Contemporary and foundational readings in race, class, and gender*, 378-388.

Roth, P. L., Bobko, P., & Switzer, F. S. 2006. Modeling the behavior of the 4/5ths rule for determining adverse impact: reasons for caution. *Journal of Applied Psychology, 91*(3), 507-522.

Roth, P. L., Purvis, K. L., & Bobko, P. 2012. A meta-analysis of gender group differences for

measures of job performance in field studies. *Journal of Management, 38*(2), 719-739.

Rudolph, C. W., & Baltes, B. B. 2008. Main effects do not discrimination make. *Industrial and Organizational Psychology, 1*(04), 415-416.

Ruyle, K. 2012. Measuring and mitigating cost of employee turnover. *Society of Human Resource Management*.

Ryan, A. M., Ployhart, R. E., & Friedel, L. A. 1998. Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*(2), 298-307.

Rynes, S. L., Colbert, A. E., & Brown, K. G. 2002. HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management, 41*(2), 149-174.

Rynes, S. L., Giluk, T. L., & Brown, K. G. 2007. The very separate worlds of academic and practitioner periodicals in human resource management: Implications for evidence-based management. *Academy of Management Journal, 50*(5), 987-1008.

Sackett, P. R., & Ellingson, J. E. 1997. The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*(3), 707-721.

Schein, V. E., Mueller, R., Lituchy, T., & Liu, J. 1996. Think manager—think male: A global phenomenon? *Journal of Organizational Behavior, 17*(1), 33-41.

Schmidt, F. L., & Hunter, J. E. 1983. Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68*(3), 407-414.

Schmidt, F. L., & Hunter, J. E. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. 1979. Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*(6), 609-626.

Schmidt, F. L., Oh, I. S., & Shaffer, J. A. 2016. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings. *Fox School of Business Research Paper*

SHRM. 2016. *Human Capital Benchmarking Report*. Retrieved from https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/pages/2016-human-capital-report.aspx

Sidanius, J. & Pratto, F. 2001. *Social dominance: an intergroup theory of social hierarchy and oppression.* New York: Cambridge University Press

Sturman, M. C. 2012. Employee value: Combining utility analysis with strategic human resource

management research to yield strong theory *The Oxford Handbook of Personnel Assessment and Selection*.

Tam, A. P., Murphy, K. R., & Lyall, J. T. 2004. Can changes in differential dropout rates reduce adverse impact? A computer simulation study of a multi-wave selection system. *Personnel Psychology, 57*(4), 905-934.

Taylor, H. C., & Russell, J. T. 1939. The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23*(5), 565-578.

Trentham, S., & Larwood, L. 1998. Gender discrimination and the workplace: An examination of rational bias theory. *Sex Roles, 38*(1-2), 1-28.

Uhlmann, E. L., & Cohen, G. L. 2005. Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16*(6), 474-480.

Vancouver, J. B., Li, X., Weinhardt, J. M., Steel, P., & Purl, J. D. 2016. Using a computational model to understand possible sources of skews in distributions of job performance. *Personnel Psychology, 69*(4), 931-974.

Vial, A. C., Brescoll, V. L., & Dovidio, J. F. 2019. Third-party prejudice accommodation increases gender discrimination. *Journal of Personality and Social Psychology, 117*(1), 73-98.

FOOTNOTES

[1] It is important to emphasize that gender is not binary (Hyde, Bigler, Joel, Tate, & van Anders, 2019). We model selection decisions between women and men because (a) binary choices are more straightforward to model, and (b) self-identified women and men are the numerical majority of the population and we are interested in explaining aggregate-level unequal outcomes. For similar reasons, we do not model intersectionality effects for between gender and ethnic minority status, or gender and sexual orientation.

[2] Because the gender identifier is coded 0 for females and 1 for males, the gender identifier is multiplied by 2 in this equation to scale the unweighted gender bias effect to have the same amount of baseline variance and standard deviation (i.e., 1 SD instead of .5 SD) as the corresponding values for qualifications and error.

[3] The psychometric properties of the 4/5ths rule have been questioned over the years due to the test's relatively high probability of producing false-positive readings in small samples or in cases where minorities have limited representation in the applicant pool (Roth, Bobko, & Switzer, 2006). This is a non-trivial issue given that the costs of a false-positive reading can be quite substantial. Nevertheless, the 4/5ths rule remains a relevant rule of thumb that continues to see use by organizations and regulatory agencies as "a warning light that signals potentially important and costly underlying problems" (Roth et al., 2006, p. 508). As such, we use the 4/5ths rule when evaluating our findings as a general rule of thumb for determining what comprises practically significant differences in hiring outcomes. As justification for this decision, we designed our simulations to circumvent the limitations of the 4/5ths rule by calculating the influence of gender bias on impact ratios using large samples of applicants aggregated across multiple selection protocols, thus increasing the accuracy of estimates provided and all but eliminating the possibility of false positive results in our analyses. This approach is consistent with the spirit of the majority of adverse impact investigations, which are concerned primarily with long-term hiring outcomes cumulated across a large sample of applicants (Baldus & Cole, 1980; Morris, 2016).

[4] In research contexts where average criterion scores of individuals are not known, $\bar{z}_y$ can be estimated by taking the product of the test's validity and the mean standard test score of hired applicants [for a table where the mathematics of these values are worked out for various validity coefficients and selection ratios, see Brown and Ghiselli, (1953)]. However, in our model, the use of validity coefficients in this calculation was not necessary because the true-score performance domain scores (i.e., the qualifications rating) of hired applicants were readily available.

[5] The decision to use large applicant sample sizes in our simulations was made to avoid problems pertaining to unreliability in impact ratios and odds ratios when smaller sample sizes are used. Nevertheless, ancillary analyses revealed that this decision was arbitrary, as all simulation results successfully replicated when averaging results over a large number of smaller, independent applicant pools instead. As such, our findings can be reasonably expected to generalize to smaller sample hiring decisions that are more commonly observed in real world hiring contexts.

[6] See Appendix C in the online supplement for the analyses

[7] Full results of these and other sensitivity analyses discussed are available upon request.

TABLES

Table 1

Summary of Simulation Model Steps, Parameters, and Assumptions

Stage	Model function	Model parameters	Model assumptions
Stage 1: Generating the applicant pool	A pool of applicants is created, and applicant characteristics (i.e., their gender and their true-score qualifications rating) are randomly assigned.	*p* – ratio of males to females in the applicant pool	Applicant true-score qualifications are normally distributed Applicant gender and true-score qualifications are independent
Stage 2: Assessment phase	Applicant qualifications are assessed, and an evaluation score is assigned. Evaluation scores reflect a combination of true score qualifications variance, systematic error variance due to bias, and unsystematic random error variance.	*q%* – variance in assessment scores reflecting applicant true-score qualifications *b%* – variance in assessment scores due to bias	Assessment phase error is normally distributed Assessment phase error is independent of applicant gender and applicant true-score qualification ratings
Stage 3: Selection	Applicants are ranked in order of their evaluation scores, and the candidates with the highest available evaluation scores are selected to fill open positions.	*selection ratio* – ratio of available job openings per applicant	Hiring decisions are made solely based on cumulative assessment evaluation scores using a top-down selection protocol
Stage 4: Evaluation	Selection system performance is assessed based on rates of adverse impact, risk of disparate treatment, new hire failure rate due to bias, and system utility loss per hire due to bias.	*base rate* – proportion of applicants that could succeed in the position if given the opportunity *SDy* – estimate of the financial impact of 1 SD difference in job performance	*SD*y is set to values representing 40%, 50%, and 60% of the annual salary of the typical employee, respectively (Schmidt and Hunter, 1983)

Table 2

Simulation Results for Simulation 1a: The Impact of Bias on Hiring Outcomes in a Typical Selection Context

	Adverse impact		Disparate treatment	Financial impact	
Bias models	Impact ratio	Odds ratio	(Δ% in rate due to bias)	New hire failure rate	Utility loss due to bias
4% bias	**.42**	**.40**	.790 (20.3%)	.072 (50.2%)	-$2,125.64
2.2% bias	**.53**	**.51**	.746 (13.5%)	.056 (16.1%)	-$710.54
1% bias	**.66**	**.64**	.714 (8.7%)	.052 (7.7%)	-$355.36
0% bias	.99	.99	.657 (0.0%)	.048 (0.0%)	$0.00

Note. See Table B1 in Appendix B of the online supplement for model parameter values used in Simulation 1a. Values in **bold** indicate adverse impact effect sizes that exceed traditional practical significance cutoffs by > .05. Values in *italics* indicate marginal adverse impact effect sizes that fall within +/- .05 of traditional practical significance cutoffs. Values in (parentheses) reflect rate changes observed in the bias models relative to the no bias model.

Table 3

Results for Simulation 1b: The Impact of Bias across a Range of Hiring Contexts on Discriminatory Hiring Outcomes

Bias models	Selection ratios	Assessment Battery Validity = .10			Assessment Battery Validity = .25			Assessment Battery Validity = .50		
		Adverse impact		Disparate treatment	Adverse impact		Disparate treatment	Adverse impact		Disparate treatment
		IR	OR	(Δ% in rate due to bias)	IR	OR	(Δ% in rate due to bias)	IR	OR	(Δ% in rate due to bias)
4% bias	1 in 100 hired *(SR = .01)*	**.34**	**.33**	.99 (3.1%)	**.34**	**.33**	.98 (3.7%)	**.34**	**.33**	.92 (7.9%)
2.2% bias		**.46**	**.45**	.98 (2.2%)	**.46**	**.45**	.97 (2.7%)	**.45**	**.45**	.91 (6.2%)
1% bias		**.60**	**.60**	.97 (1.3%)	**.61**	**.61**	.96 (1.7%)	**.59**	**.59**	.89 (4.1%)
0% bias		1.03	1.03	.96 (0.0%)	1.03	1.03	.94 (0.0%)	1.01	1.01	.86 (0.0%)
4% bias	1 in 20 hired *(SR = .05)*	**.42**	**.40**	.96 (7.6%)	**.42**	**.40**	.92 (8.5%)	**.42**	**.40**	.84 (14.4%)
2.2% bias		**.53**	**.51**	.94 (5.2%)	**.53**	**.51**	.90 (6.0%)	**.53**	**.52**	.81 (10.3%)
1% bias		**.65**	**.63**	.92 (3.2%)	**.65**	**.63**	.88 (3.8%)	**.65**	**.64**	.78 (6.8%)
0% bias		.98	.98	.89 (0.0%)	.99	.99	.85 (0.0%)	.99	.99	.73 (0.0%)
4% bias	1 in 10 hired *(SR = .10)*	**.48**	**.44**	.91 (11.3%)	**.48**	**.44**	.87 (11.9%)	**.48**	**.44**	.77 (17.8%)
2.2% bias		**.58**	**.55**	.88 (7.4%)	**.58**	**.55**	.84 (8.3%)	**.58**	**.55**	.73 (12.8%)
1% bias		**.70**	*.67*	.86 (4.5%)	**.70**	*.67*	.82 (5.4%)	**.70**	*.67*	.70 (8.2%)
0% bias		.99	.99	.82 (0.0%)	1.00	1.00	.78 (0.0%)	1.00	1.00	.65 (0.0%)
4% bias	1 in 4 hired *(SR = .25)*	**.59**	**.50**	.79 (19.8%)	**.59**	**.50**	.73 (19.3%)	**.59**	**.50**	.62 (24.5%)
2.2% bias		**.68**	**.60**	.74 (12.9%)	**.68**	**.60**	.69 (13.4%)	**.68**	**.60**	.58 (17.4%)
1% bias		*.77*	*.71*	.71 (7.9%)	*.77*	*.71*	.66 (8.4%)	*.77*	*.71*	.55 (11.2%)
0% bias		1.00	1.00	.66 (0.0%)	1.00	1.00	.61 (0.0%)	1.00	1.00	.50 (0.0%)
4% bias	1 of 2 hired *(SR = .50)*	**.72**	**.52**	.56 (30.3%)	**.72**	**.52**	.51 (28.3%)	**.72**	**.52**	.43 (33.1%)
2.2% bias		*.78*	**.62**	.51 (19.6%)	*.78*	**.62**	.48 (19.6%)	*.78*	**.62**	.39 (23.5%)
1% bias		*.85*	*.73*	.48 (12.0%)	*.85*	*.73*	.45 (12.4%)	*.85*	*.73*	.37 (15.1%)
0% bias		1.00	1.00	.43 (0.0%)	1.00	1.00	.40 (0.0%)	1.00	1.00	.32 (0.0%)
4% bias	9 of 10 hired *(SR = .90)*	.92	**.45**	.14 (49.1%)	.92	**.45**	.13 (47.7%)	.92	**.45**	.11 (50.5%)
2.2% bias		.94	**.56**	.12 (34.1%)	.94	**.56**	.12 (33.9%)	.94	**.55**	.10 (36.5%)
1% bias		.96	*.67*	.11 (22.0%)	.96	*.67*	.11 (22.1%)	.96	*.67*	.09 (23.8%)
0% bias		1.00	1.00	.09 (0.0%)	1.00	1.00	.09 (0.0%)	1.00	.99	.07 (0.0%)

Note. See Table B2 in Appendix B of the online supplement for model parameter values used in Simulation 1b. Assessment battery validity values are presented as validity coefficients. Values in **bold** indicate adverse impact effect sizes that exceed traditional practical significance cutoffs by > .05. Values in *italics* indicate marginal adverse impact effect sizes that fall within +/- .05 of traditional practical significance cutoffs. Values in parentheses represent the percent change in rates of disparate treatment to bias.

Table 4

Results for Simulation 1b: The Impact of Bias across a Range of Hiring Contexts on New Hire Failure Rates

Bias models	Selection ratios	Assessment Battery Validity = .10			Assessment Battery Validity = .25			Assessment Battery Validity = .50		
		BR = .20	BR = .50	BR = .80	BR = .20	BR = .50	BR = .80	BR = .20	BR = .50	BR = .80
4% bias	1 in 100 hired *(SR = .01)*	.72 (19.9%)	.40 (47.4%)	.14 (89.3%)	.57 (14.9%)	.25 (33.2%)	.06 (58.5%)	.29 (16.2%)	.07 (39.2%)	.01 (70.3%)
2.2% bias		.66 (9.0%)	.33 (20.6%)	.10 (37.7%)	.54 (8.1%)	.22 (16.8%)	.05 (27.6%)	.27 (8.1%)	.06 (22.2%)	.01 (54.1%)
1% bias		.62 (3.6%)	.30 (8.4%)	.08 (14.5%)	.52 (3.4%)	.20 (7.7%)	.04 (7.3%)	.26 (3.4%)	.05 (10.0%)	.00 (24.3%)
0% bias		.60 (0.0%)	.27 (0.0%)	.07 (0.0%)	.50 (0.0%)	.19 (0.0%)	.04 (0.0%)	.25 (0.0%)	.05 (0.0%)	.00 (0.0%)
4% bias	1 in 20 hired *(SR = .05)*	.74 (13.3%)	.42 (30.1%)	.15 (58.4%)	.63 (9.7%)	.30 (20.8%)	.08 (39.0%)	.42 (10.6%)	.12 (22.9%)	.02 (45.0%)
2.2% bias		.69 (6.4%)	.37 (13.4%)	.12 (25.0%)	.61 (5.3%)	.27 (11.0%)	.07 (20.2%)	.40 (6.3%)	.11 (12.3%)	.02 (25.1%)
1% bias		.67 (2.6%)	.34 (5.2%)	.10 (9.7%)	.59 (2.5%)	.26 (5.1%)	.07 (9.8%)	.39 (2.9%)	.11 (4.9%)	.01 (11.1%)
0% bias		.65 (0.0%)	.32 (0.0%)	.09 (0.0%)	.58 (0.0%)	.25 (0.0%)	.06 (0.0%)	.38 (0.0%)	.10 (0.0%)	.01 (0.0%)
4% bias	1 in 10 hired *(SR = .10)*	.75 (10.2%)	.43 (24.7%)	.15 (47.9%)	.66 (7.3%)	.33 (17.3%)	.09 (32.4%)	.49 (6.9%)	.16 (16.8%)	.03 (37.6%)
2.2% bias		.71 (4.7%)	.38 (11.3%)	.13 (20.8%)	.64 (3.8%)	.30 (8.9%)	.08 (16.3%)	.47 (3.9%)	.15 (9.3%)	.02 (20.9%)
1% bias		.69 (2.0%)	.36 (4.8%)	.11 (8.5%)	.63 (1.7%)	.29 (4.0%)	.08 (6.8%)	.46 (1.8%)	.15 (4.3%)	.02 (11.9%)
0% bias		.68 (0.0%)	.35 (0.0%)	.10 (0.0%)	.62 (0.0%)	.28 (0.0%)	.07 (0.0%)	.45 (0.0%)	.14 (0.0%)	.02 (0.0%)
4% bias	1 in 4 hired *(SR = .25)*	.76 (6.8%)	.45 (16.5%)	.17 (31.6%)	.70 (4.3%)	.37 (10.9%)	.12 (21.2%)	.59 (3.4%)	.24 (9.7%)	.05 (21.5%)
2.2% bias		.74 (3.1%)	.41 (7.5%)	.14 (14.1%)	.69 (2.3%)	.35 (5.6%)	.11 (10.9%)	.58 (1.8%)	.23 (5.3%)	.05 (11.6%)
1% bias		.72 (1.4%)	.40 (3.3%)	.13 (5.8%)	.68 (1.0%)	.34 (2.6%)	.10 (4.8%)	.58 (0.8%)	.22 (2.3%)	.04 (5.0%)
0% bias		.72 (0.0%)	.39 (0.0%)	.13 (0.0%)	.67 (0.0%)	.34 (0.0%)	.10 (0.0%)	.57 (0.0%)	.22 (0.0%)	.04 (0.0%)
4% bias	1 of 2 hired *(SR = .50)*	.78 (3.7%)	.47 (9.3%)	.18 (18.5%)	.74 (2.1%)	.42 (5.8%)	.14 (12.3%)	.69 (1.4%)	.33 (4.4%)	.09 (11.2%)
2.2% bias		.76 (1.7%)	.45 (4.2%)	.16 (8.5%)	.74 (1.1%)	.41 (3.0%)	.14 (6.4%)	.68 (0.7%)	.33 (2.4%)	.08 (6.1%)
1% bias		.76 (0.7%)	.44 (1.7%)	.16 (3.5%)	.73 (0.5%)	.40 (1.3%)	.13 (2.9%)	.68 (0.3%)	.32 (1.1%)	.08 (2.7%)
0% bias		.75 (0.0%)	.43 (0.0%)	.15 (0.0%)	.73 (0.0%)	.40 (0.0%)	.13 (0.0%)	.68 (0.0%)	.32 (0.0%)	.08 (0.0%)
4% bias	9 of 10 hired *(SR = .90)*	.79 (0.7%)	.49 (2.0%)	.19 (4.2%)	.79 (0.3%)	.48 (1.1%)	.18 (2.8%)	.78 (0.1%)	.46 (0.5%)	.17 (2.1%)
2.2% bias		.79 (0.3%)	.49 (0.9%)	.19 (2.0%)	.79 (0.2%)	.48 (0.6%)	.18 (1.5%)	.78 (0.0%)	.46 (0.3%)	.16 (1.1%)
1% bias		.79 (0.1%)	.48 (0.4%)	.19 (0.8%)	.79 (0.1%)	.48 (0.2%)	.18 (0.6%)	.78 (0.0%)	.46 (0.1%)	.16 (0.5%)
0% bias		.79 (0.0%)	.48 (0.0%)	.19 (0.0%)	.79 (0.0%)	.48 (0.0%)	.18 (0.0%)	.78 (0.0%)	.46 (0.0%)	.16 (0.0%)

Note. See Table B2 in Appendix B of the online supplement for model parameter values used in Simulation 1b. Assessment battery validity values are presented as validity coefficients. BR = Base rate or the percent of the applicant pool that possesses at least a minimal level of qualifications necessary for job success; Values in parentheses represent the percent change in new hire failure rates due to bias.

Table 5

Results for Simulation 1b: The Impact of Bias across a Range of Hiring Contexts on System Utility

Bias models	Selection ratios	Assessment Battery Validity = .10			Assessment Battery Validity = .25			Assessment Battery Validity = .50		
		SD_y = .40	SD_y = .50	SD_y = .60	SD_y = .40	SD_y = .50	SD_y = .60	SD_y = .40	SD_y = .50	SD_y = .60
4% bias	1 in 100 hired *(SR = .01)*	-$6,687.25	-$8,359.15	-$10,030.70	-$3,647.53	-$4,559.46	-$5,471.21	-$2,020.26	-$2,525.35	-$3,030.34
2.2% bias		-$2,998.72	-$3,748.43	-$4,498.00	-$1,918.44	-$2,398.07	-$2,877.60	-$1,166.68	-$1,458.36	-$1,749.98
1% bias		-$1,218.77	-$1,523.47	-$1,828.12	-$826.67	-$1,033.35	-$1,239.99	-$598.41	-$748.02	-$897.60
0% bias		$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00
4% bias	1 in 20 hired *(SR = .05)*	-$4,980.21	-$6,225.32	-$7,470.18	-$2,815.56	-$3,519.48	-$4,223.27	-$1,667.94	-$2,084.94	-$2,501.86
2.2% bias		-$2,281.71	-$2,852.17	-$3,422.51	-$1,502.33	-$1,877.93	-$2,253.46	-$949.19	-$1,186.50	-$1,423.76
1% bias		-$907.69	-$1,134.62	-$1,361.51	-$703.17	-$878.97	-$1,054.73	-$430.79	-$538.49	-$646.17
0% bias		$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00
4% bias	1 in 10 hired *(SR = .10)*	-$4,217.26	-$5,271.63	-$6,325.79	-$2,431.83	-$3,039.82	-$3,647.69	-$1,359.48	-$1,699.37	-$2,039.18
2.2% bias		-$1,920.41	-$2,400.54	-$2,880.56	-$1,253.30	-$1,566.64	-$1,879.92	-$748.96	-$936.21	-$1,123.42
1% bias		-$811.10	-$1,013.88	-$1,216.63	-$558.24	-$697.81	-$837.34	-$346.11	-$432.64	-$519.16
0% bias		$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00
4% bias	1 in 4 hired *(SR = .25)*	-$3,108.34	-$3,885.46	-$4,662.43	-$1,756.40	-$2,195.53	-$2,634.56	-$970.13	-$1,212.68	-$1,455.17
2.2% bias		-$1,411.51	-$1,764.41	-$2,117.23	-$910.09	-$1,137.62	-$1,365.11	-$525.10	-$656.39	-$787.64
1% bias		-$603.63	-$754.54	-$905.42	-$409.96	-$512.45	-$614.93	-$227.12	-$283.90	-$340.67
0% bias		$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00
4% bias	1 of 2 hired *(SR = .50)*	-$1,947.52	-$2,434.42	-$2,921.23	-$1,091.49	-$1,364.38	-$1,637.21	-$613.72	-$767.16	-$920.56
2.2% bias		-$880.61	-$1,100.78	-$1,320.90	-$566.18	-$707.73	-$849.25	-$335.96	-$419.96	-$503.93
1% bias		-$356.88	-$446.10	-$535.31	-$250.05	-$312.56	-$375.06	-$149.40	-$186.75	-$224.10
0% bias		$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00
4% bias	9 of 10 hired *(SR = .90)*	-$475.88	-$594.86	-$713.81	-$271.61	-$339.52	-$407.41	-$145.07	-$181.34	-$217.61
2.2% bias		-$217.01	-$271.26	-$325.50	-$139.88	-$174.85	-$209.81	-$76.86	-$96.08	-$115.29
1% bias		-$91.64	-$114.55	-$137.45	-$61.20	-$76.50	-$91.79	-$32.41	-$40.51	-$48.61
0% bias		$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00

Note. See Table B2 in Appendix B of the online supplement for model parameter values used in Simulation 1b. Values represent the change in system utility per applicant in the biased models relative to bias-free model.

Table 6

Simulation Results for Simulation 2a: Influence of Increasing Overall Rate of Female Representation in the Applicant Pool on the Impact of Gender Bias

Bias models	Applicant pool gender ratios (*p*)	Adverse impact		Disparate treatment	Financial impact	
		Impact ratio	Odds ratio	(Δ rate due to bias)	New hire failure rate	Utility loss due to bias
4% bias	*Male dominated industry (10% female)*	**0.41**	**0.39**	.834 (25.7%)	.068 (40.4%)	-$1,716.67
2.2% bias		**0.53**	**0.52**	.783 (18.1%)	.054 (11.6%)	-$515.15
1% bias		**0.65**	**0.64**	.746 (12.6%)	.051 (5.3%)	-$248.66
0% bias		1.00	1.00	.663 (0.0%)	.048 (0.0%)	$0.00
4% bias	*10% more female applicants*	**0.40**	**0.39**	.833 (25.2%)	.068 (40.4%)	-$1,730.32
2.2% bias		**0.53**	**0.51**	.782 (17.6%)	.054 (11.7%)	-$513.86
1% bias		**0.64**	**0.63**	.746 (12.1%)	.051 (5.6%)	-$249.69
0% bias		0.99	0.99	.665 (0.0%)	.048 (0.0%)	$0.00
4% bias	*50% more female applicants*	**0.41**	**0.39**	.827 (24.7%)	.068 (41.5%)	-$1,759.34
2.2% bias		**0.52**	**0.51**	.777 (17.2%)	.054 (12.4%)	-$527.13
1% bias		**0.64**	**0.63**	.742 (12.0%)	.051 (5.7%)	-$239.93
0% bias		0.99	0.99	.663 (0.0%)	.048 (0.0%)	$0.00
4% bias	*100% more female applicants*	**0.41**	**0.40**	.820 (24.0%)	.069 (42.9%)	-$1,808.81
2.2% bias		**0.53**	**0.51**	.772 (16.7%)	.055 (13.1%)	-$559.09
1% bias		**0.65**	**0.64**	.737 (11.4%)	.051 (6.3%)	-$265.92
0% bias		1.00	1.00	.661 (0.0%)	.048 (0.0%)	$0.00
4% bias	*Female dominated industry (90% female)*	**0.46**	**0.43**	.711 (8.1%)	.069 (43.3%)	-$1,838.94
2.2% bias		**0.56**	**0.54**	.683 (3.8%)	.054 (12.4%)	-$564.25
1% bias		**0.67**	*0.66*	.672 (2.2%)	.051 (4.8%)	-$251.85
0% bias		0.99	0.99	.658 (0.0%)	.048 (0.0%)	$0.00

Note. See Table B3 in Appendix B of the online supplement for model parameter values used in Simulation 2a. Values in **bold** indicate adverse impact effect sizes that exceed traditional practical significance cutoffs by > .05. Values in *italics* indicate marginal adverse impact effect sizes that fall within +/- .05 of traditional practical significance cutoffs.

Table 7

Simulation Results for Simulation 2b: The Influence of Targeted Recruitment of Highly-qualified Female Applicants on the Impact of Gender Bias

Bias models	Applicant pool gender ratios (*p*)	Adverse impact		Disparate treatment	Financial impact	
		Impact ratio	Odds ratio	(Δ rate due to bias)	New hire failure rate	Utility loss due to bias
4% bias	*Male*	**0.58**	**0.56**	.818 (27.2%)	.070 (43.7%)	-$1,981.44
2.2% bias	*dominated*	*0.75*	*0.74*	.769 (19.6%)	.055 (14.0%)	-$726.83
1% bias	*industry (10%*	0.92	0.91	.731 (13.6%)	.052 (7.2%)	-$396.06
0% bias	*female)*	1.39	1.42	.643 (0.0%)	.048 (0.0%)	$0.00
4% bias	*10% more female applicants*	*0.79*	*0.78*	.806 (27.2%)	.071 (51.1%)	-$2,415.97
2.2% bias		1.02	1.02	.755 (19.3%)	.056 (19.2%)	-$1,026.76
1% bias		1.26	1.28	.715 (12.9%)	.051 (9.5%)	-$574.36
0% bias		1.88	1.96	.633 (0.0%)	.047 (0.0%)	$0.00
4% bias	*50% more female applicants*	*0.79*	*0.79*	.801 (26.5%)	.072 (57.8%)	-$2,678.91
2.2% bias		1.03	1.03	.750 (18.5%)	.056 (23.1%)	-$1,197.49
1% bias		1.26	1.28	.712 (12.4%)	.052 (12.5%)	-$694.92
0% bias		1.89	1.97	.633 (0.0%)	.046 (0.0%)	$0.00
4% bias	*100% more female applicants*	*0.80*	*0.79*	.793 (24.7%)	.075 (62.9%)	-$2,903.06
2.2% bias		1.03	1.04	.743 (16.9%)	.058 (26.3%)	-$1,348.28
1% bias		1.27	1.29	.707 (11.2%)	.052 (14.4%)	-$774.19
0% bias		1.90	1.98	.636 (0.0%)	.046 (0.0%)	$0.00
4% bias	*Female*	*0.80*	*0.79*	.700 (6.8%)	.073 (55.6%)	-$2,156.89
2.2% bias	*dominated*	1.03	1.03	.674 (2.8%)	.056 (19.6%)	-$765.78
1% bias	*industry (90%*	1.27	1.29	.665 (1.5%)	.052 (9.8%)	-$370.33
0% bias	*female)*	1.98	2.04	.655 (0.0%)	.047 (0.0%)	$0.00

Note. See Table B4 in Appendix B of the online supplement for model parameter values used in Simulation 2b. To model the impact of successful targeted recruitment initiatives, all-female applicants in these simulations were set to have .25 higher average qualification ratings than equivalent male applicants. Values in **bold** indicate adverse impact effect sizes that exceed traditional practical significance cutoffs by > .05. Values in *italics* indicate marginal adverse impact effect sizes that fall within +/- .05 of traditional practical significance cutoffs.

Table 8

Summary of Simulation Findings

Purpose	Summary of simulation findings	Theoretical and practical implications
Simulation 1a: The Impact of Bias in Typical Hiring Contexts		
Estimating the impact of gender bias in typical selection contexts	Practically significant levels of hiring discrimination and substantial inefficiencies in the hiring process were reported in all models in which even small amounts of bias were present in the formation of hiring evaluation scores.	Even a seemingly trivial amount of gender bias in the candidate evaluation process can have a profound negative impact on a wide range of hiring outcomes for both applicant and organization alike.
Simulation 1b: The Influence of Contextual Factors on the Impact of Bias		
Examining the influence of variations in contextual factors (i.e., assessment validity, selection ratios, base rates, and estimates of SD_y) on the impact of bias	Selection ratios strongly influenced the risk of adverse impact associated with bias such that risk was most significant when lower selection ratios were modeled. However, the range of selection ratios for which bias's felt impact was meaningful was surprisingly broad (i.e., in all applicant pools with 4 or more applicants per opening) and other signals of hiring discrimination (i.e., low odds ratios and increased rates of disparate treatment) were observed in nearly all simulations in which bias was present. System validity had little to no influence on the impact of bias on discriminatory hiring outcomes but did partially mitigate the impact of bias on financial and performance metrics. Variations in base rate had a minimal influence on the impact of bias on new hire failure rates. The financial impact of bias increased as a function of variations in SD_y.	The negative impact of gender bias is not constrained to competitive or challenging jobs or hiring contexts where low-validity assessments are used but is likely to be felt in the vast majority of contexts in which hiring decisions are made.
Simulation 2a: Increasing Female Applicant Pool Representation		
Determining the influence of increasing overall female applicant pool representation on the impact of gender bias	Increasing female representation in the applicant pool had little impact on typical hiring outcomes for female applicants as a whole. Females in the biased models were hired at rates well below their similarly qualified males, even when overall female representation in the applicant pool surpassed that of their male counterparts.	Efforts to increase the representation of female candidates in the applicant pool is unlikely to improve hiring outcomes for female candidates when underlying biases remain unresolved.
Simulation 2b: Targeting Highly-qualified Female Applicants		
Determining the influence of targeted recruitment of highly qualified female applicants on the impact of gender bias	Directly targeting more qualified female applicants can reduce rates of adverse impact against women, even in the face of bias. However, reductions in the rate of disparate treatment against highly qualified females were minimal, and new hire failure rates and utility loss due to bias increased as a result of targeted recruitment efforts when sources of bias remained unresolved.	Although targeted recruitment can contribute to more equitable hiring outcomes, qualified female applicants will continue to face discrimination when hiring evaluations are influenced by bias.

FIGURES

Figure 1a

Distribution of Effect Sizes Comparing Evaluations of Male vs. Female Applicants

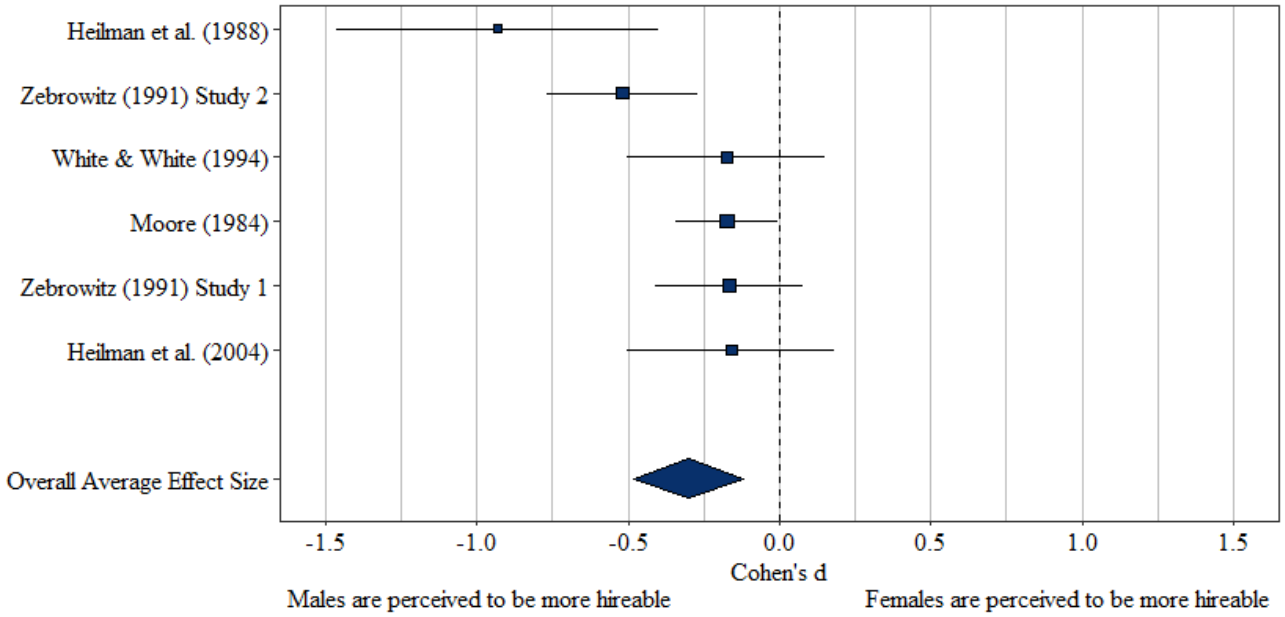

Figure 1b

Distribution of Effect Sizes Comparing Evaluations of Applicants with Relatively Higher (Vs. Lower) Qualifications