Organizational Behavior and Human Decision Processes 161 (2020) 291–309

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Creative destruction in science[☆]

Warren Tierney[a,*], Jay H. Hardy III[b], Charles R. Ebersole[c], Keith Leavitt[d], Domenico Viganola[e], Elena Giulia Clemente[f], Michael Gordon[g], Anna Dreber[h], Magnus Johannesson[f], Thomas Pfeiffer[g], Hiring Decisions Forecasting Collaboration, Eric Luis Uhlmann[a,*]

[a] *INSEAD, Singapore*
[b] *Oregon State University, United States*
[c] *University of Virginia, United States*
[d] *Oregon State University, United States*
[e] *The World Bank*
[f] *Stockholm School of Economics, Sweden*
[g] *Massey University, New Zealand*
[h] *Stockholm School of Economics, Sweden, and University of Innsbruck, Austria*

ARTICLE INFO

Keywords:
Replication
Theory pruning
Theory testing
Direct replication
Conceptual replication
Falsification
Hiring decisions
Gender discrimination
Work-family conflict
Cultural differences
Work values
Protestant work ethic

ABSTRACT

Drawing on the concept of a gale of creative destruction in a capitalistic economy, we argue that initiatives to assess the robustness of findings in the organizational literature should aim to simultaneously test competing ideas operating in the same theoretical space. In other words, replication efforts should seek not just to support or question the original findings, but also to replace them with revised, stronger theories with greater explanatory power. Achieving this will typically require adding new measures, conditions, and subject populations to research designs, in order to carry out conceptual tests of multiple theories in addition to directly replicating the original findings. To illustrate the value of the creative destruction approach for theory pruning in organizational scholarship, we describe recent replication initiatives re-examining culture and work morality, working parents' reasoning about day care options, and gender discrimination in hiring decisions.

Significance statement: It is becoming increasingly clear that many, if not most, published research findings across scientific fields are not readily replicable when the same method is repeated. Although extremely valuable, failed replications risk leaving a theoretical void— reducing confidence the original theoretical prediction is true, but not replacing it with positive evidence in favor of an alternative theory. We introduce the creative destruction approach to replication, which combines theory pruning methods from the field of management with emerging best practices from the open science movement, with the aim of making replications as generative as possible. In effect, we advocate for a Replication 2.0 movement in which the goal shifts from checking on the reliability of past findings to actively engaging in competitive theory testing and theory building.

Scientific transparency statement: The materials, code, and data for this article are posted publicly on the Open Science Framework, with links provided in the article.

1. Introduction

As Meehl (1978, p. 817) writes, it is the job of scientists to "subject theories… to grave danger of refutation… A theory is corroborated to the extent that we have subjected it to such risky tests; the more dangerous tests it has survived, the better corroborated it is." We suggest that for too long, theories in the organizational and psychological literatures have been akin to domesticated animals—sheltered and nurtured by supporters, rather than subject to the fitness and survival pressures Meehl (1978), Popper (1963), and others envisioned.

Indeed, organizational scholars have long lamented the proliferation of new theories within management research (Hambrick, 2007), with meaningful attempts at theory reduction remaining largely absent from the literature (Aguinis, Pierce, Bosco, & Muslin, 2009; Leavitt, Mitchell, & Peterson, 2010). Platt (1964) used the term *strong inference* to describe at a high level how faster-moving sciences tend to pit theories against one another to accelerate progress (see also Albertini, 2017). To address this challenge, management scholars have slowly

[☆] Action Editors: Don Moore and Stefan Thau.

[*] Corresponding authors at: INSEAD, Organisational Behaviour Area, 1 Ayer Rajah Avenue, 138676 Singapore, Singapore. Tel.: +65 8468 5671.
E-mail addresses: warrentierney@hotmail.com (W. Tierney), eric.luis.uhlmann@gmail.com (E.L. Uhlmann).

https://doi.org/10.1016/j.obhdp.2020.07.002
Received 29 November 2019; Received in revised form 6 July 2020; Accepted 16 July 2020
0749-5978/ © 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

adopted a loosely described set of techniques known as "theory pruning," which are defined as theory testing techniques which "can move us in the direction of limiting, bounding, and perhaps reducing theory" (Leavitt et al., 2010).

Concerns about theory proliferation are compounded by the limited number of studies focusing on replication (Bergh, Sharp, Aguinis, & Li, 2017; Earp & Trafimow, 2015; Lykken, 1968; Tsang & Kwan, 1999; Brandt et al., 2014), and new findings regarding a general lack of replicability within organizational scholarship (Bergh et al., 2017; Bosco, Aguinis, Field, Pierce, & Dalton, 2016). Accordingly, commentators have recently described the risk of a crisis of confidence in organizational research (Gelman, 2015; Köhler & Cortina, in press). Thus, while scholars continue to generate new theory at an accelerated pace, their propositions typically enjoy preliminary rather than definitive support, and are rarely subjected to attempts at direct replication (Schmidt, 2009; Simons, 2014) or placed in competition against adjacent (and sometimes contradictory) theories.

The current paper introduces and applies the concept of *creative destruction* of management and psychological theory, wherein best practices for replication and transparency (Nosek, Spies, & Motyl, 2012; Open Science Collaboration, 2015) are combined with epistemological strategies of theory pruning. The goal is to draw strong inferences (Platt, 1964) by carrying out severe tests (Mayo, 2018) of two or more competing theories that occupy shared theoretical space. We begin by identifying the limits of traditional approaches to bounding theory, and define the optimal features of the creative destruction approach. To illustrate how the creative destruction paradigm provides information gain beyond either traditional replication or theory pruning methods, we describe the results of recent initiatives to revisit findings regarding the role of a Puritan-Protestant heritage in American work morality, as well as motivated reasoning on the part of would-be parents facing difficult child care choices. We also report a combined direct and conceptual replication (Crandall & Sherman, 2016; Schmidt, 2009; Simons, 2014) of past work on psychological rationalizations for gender discrimination. This original data collection is used as a vehicle to test four theories of hiring decisions involving female and male candidates, specifically motivated gender discrimination, assimilation to cognitive expectations, motivated liberal ideologies, and study savviness. Under the taxonomy of replications introduced by Köhler and Cortina (in press), these investigations constitute semi-independent replications rather than independent replications, since they include one member of the original research team.

In each case, high-powered and in some cases cross-national samples, combined with pre-registered (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) empirical predictions from each theoretical perspective, allow for strong inferences (Platt, 1964) in the absence of publication bias (Kvarven et al., in press). In addition to repeating the original design, we systematically include further measures, conditions, and populations, allowing for novel tests of competing theoretical accounts operating in the same domains. We suggest that the creative destruction paradigm can serve the long-sought goal of encouraging the development of new theories and insights for the study of management and organizations, while also rigorously pruning and bounding theories as they emerge (Porter, 1996).

2. The need for theory pruning in management scholarship

Scientific theories are like toothbrushes—no one wants to use anyone else's (Mischel, 2008). Editors and reviewers at journals, and selection and promotion committees at universities, reward the introduction of new theoretical ideas more so than adjudicating between existing theories. A study of prestigious medical journals found that the outlets with the highest impact factors preferred publishing novel research, not necessarily the most robust research (Evangelou, Siontis, Pfeiffer, & Ioannidis, 2012). The professional incentive to develop one's own distinctive intellectual brand leads to a proliferation of theories, frameworks, and models (Köhler & Cortina, in press; Hambrick, 2007; Mischel, 2008), many of these attracting relatively little attention from other scientists. As a result, theories in social and organizational psychology are rarely made vulnerable to disproof.

Pitting competing empirical predictions against one another in the same experimental paradigm provides the opportunity to bound, qualify, and reduce theory (Aguinis et al., 2009; Hambrick, 2007; Kluger & Tikochinsky, 2001; Van de Ven & Johnson, 2006; Vandenberg & Grelle, 2008). By directly considering and testing theories in tandem, scholars are able to determine the necessity of additional constructs introduced by a novel theory, or identify which of two theories provides predictive validity across a broader range of criteria (Leavitt et al., 2010). Such an approach may generate support for one theoretical explanation over another (Schlaegel & Koenig, 2014), reconcile apparent contradictions that are later explained by differences in assumptions underlying divergent theoretical orientations (Peteraf, Di Stefano, & Verona, 2013), or facilitate new discovery by identifying previously hidden moderators that emerge when one theory directly antagonizes another (Latham, Erez, & Locke, 1988).

To date, five general categories of theory pruning strategies have been identified, with definitiveness for identifying a champion between two theories increasing with the more sophisticated strategies (Leavitt et al., 2010). First, scholars may simply apply a basic parsimony test of the two theories, and demonstrate that the novel constructs from one theory add additional predictive variance beyond those constructs present in both theories (e.g., Barrick & Zimmerman, 2005). A second approach involves comparing two models (one more parsimonious than the other) which "nest" with regard to total terms and propositions required for an explanation (e.g., Barger & Grandey, 2006). The third approach involves testing the direction and magnitude of effect sizes predicted by the two theories, across a range of studies (e.g., Thau & Mitchell, 2010). Fourth, scholars may apply a comparison of the predictive robustness of two theories, favoring the theory which best describes stable relationships across a greater range of predictors and criteria (e.g., Reynolds, Dang, Yam, & Leavitt, 2014). Finally, the most definitive approach to theory pruning involves carefully constructing tests where two truly incompatible theories are introduced in the same space. Within this approach, a finding in support of propositions from one theory may seriously call into questions propositions from the second theory (Supplement 6).

These approaches to theory pruning are often limited by the constraints of existing data or under-powered studies which are unlikely to be definitive. We will describe how a creative destruction approach may build upon the existing paradigm of theory pruning by combining these methodologies with best practices gleaned from the open science movement.

3. The crisis of confidence in science

Replication is a cornerstone of scientific progress, and can take the form of a direct/literal replication (same method, new participants), or conceptual/constructive replication (different method, new participants) (Köhler & Cortina, in press; Schmidt, 2009; Simons, 2014). Replications of past findings increase confidence in a given phenomenon and can demonstrate the ability of theories to make successful predictions. Furthermore, previous studies become the inspiration for future studies and orient researchers toward new avenues for theory expansion. If prior work is not replicable, it is difficult to gain confidence in a finding or theory, and researchers will likely have a harder time finding productive avenues for new inquiry. Conducting conceptual replications, for example repeating a laboratory manipulation in a field setting, or testing the same idea using different experimental approaches within the same paper, is already commonplace and rightly treated as important in organizational scholarship. In contrast, direct replication is far less frequent across fields of inquiry (Köhler & Cortina, in press; Zwaan, Etz, Lucas, & Donnellan, 2017).

Unfortunately, recent attempts at directly replicating findings have raised concerns about the strength of this cornerstone. Across many disciplines, including medicine (Begley & Ellis, 2012; Prinz, Schlange, & Asadullah, 2011), economics (Camerer et al., 2016; Chang & Li, 2017; McCullough, McGeary, & Harrison, 2006), psychology (Ebersole et al., 2016; Klein et al., 2014, 2018; Open Science Collaboration, 2015), and the social sciences, broadly defined (Camerer et al., 2018), researchers have found that a concerning number of studies fail to replicate when the same methodology is repeated in new samples. At a minimum, these results pose challenges to our understanding of the phenomena tested in the replication studies. More broadly, the overall lack of replicability of prior findings poses a threat to scientific progress. The need to adopt more robust methodologies and achieve more reliable results is a common challenge for psychology, management, education, ecology, medicine, and other fields (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Bedeian, Taylor, & Miller, 2010; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; John, Loewenstein, & Prelec, 2012; Ramagopalan et al., 2014; Makel et al., 2019).

These concerns surrounding replication and research practices appear similarly relevant within myriad organizational literatures and across management research methodologies (Bamberger, 2019; Bergh et al., 2017; Pratt, Kaplan, & Whittington, 2019; Aguinis & Solarino, in press). While our search was unable to identify a systematic assessment of the direct replicability of organizational behavior or human resources research, a survey by Bedeian et al. (2010) found that the majority of organizational scholars had first-hand knowledge of questionable research practices, which are likely fueling poor replicability across methodologies and domains of inquiry (Byington & Felps, 2017). Other meta-scientific work identifies a “Chrysalis Effect” such that published articles in management are far more likely to report statistically significant effects than are unpublished dissertations on the same research (Cairo et al., in press; O’Boyle, Banks, & Gonzalez-Mulé, 2017). Such findings are especially alarming at a time when popular press books, TED talks, and podcasts allow for interesting or provocative management research findings to reach a broad practitioner audience and make their way into practice.

4. The informational value critique of replications

Researchers do update their beliefs about prior findings in light of replications. For instance, in prediction markets, researchers have less confidence in a finding in light of a failed replication (Dreber et al., 2015). Conversely, researchers report more confidence in a finding following a successful replication. From a Bayesian perspective, these adjustments seem sensible. Researchers should update their priors concerning research claims in response to new information about those claims.

However, the information provided by replications may be more ambiguous than is often appreciated. Critics have pointed out that there are many reasons why a replication study might fail to support the original predictions (Schwarz & Strack, 2014; Strack, 2016; Stroebe & Strack, 2014; Petty & Cacioppo, 2016; Schnall, 2014). The original study may have been a false positive, meaning that there was no “true” effect for the replication study to detect. Conversely, the replication may have been underpowered, making the observed null effect a false negative. It is also possible that the replication study used suboptimal methods for eliciting the effect (Luttrell, Petty, & Xu, 2017). Even when the same methodology from an original study is used, it is possible that those methods are not applicable to the setting or sample of the replication (Schwarz & Strack, 2014). Finally, it is possible that there are unknown moderators of the finding in question that systematically varied between the original study and replication contexts (Schweinsberg et al., 2016).

Despite these challenges, replication studies can be designed to reduce some of this ambiguity. For instance, some scholars have advocated for adding conditions and measures to replications to test new research questions in addition to those tested in the original study, such as an *a priori* individual differences moderator (Brainerd & Reyna, 2018). Although post-hoc appeals to “hidden moderators” are generally unpersuasive, especially in light of the low cross-site heterogeneity of effects that fail to replicate (Klein et al., 2018), contextual moderators that were predicted beforehand and then demonstrated empirically can be extremely informative. The creative destruction approach adopts and extends this mentality, arguing replications are the perfect ground for systematic theory pruning.

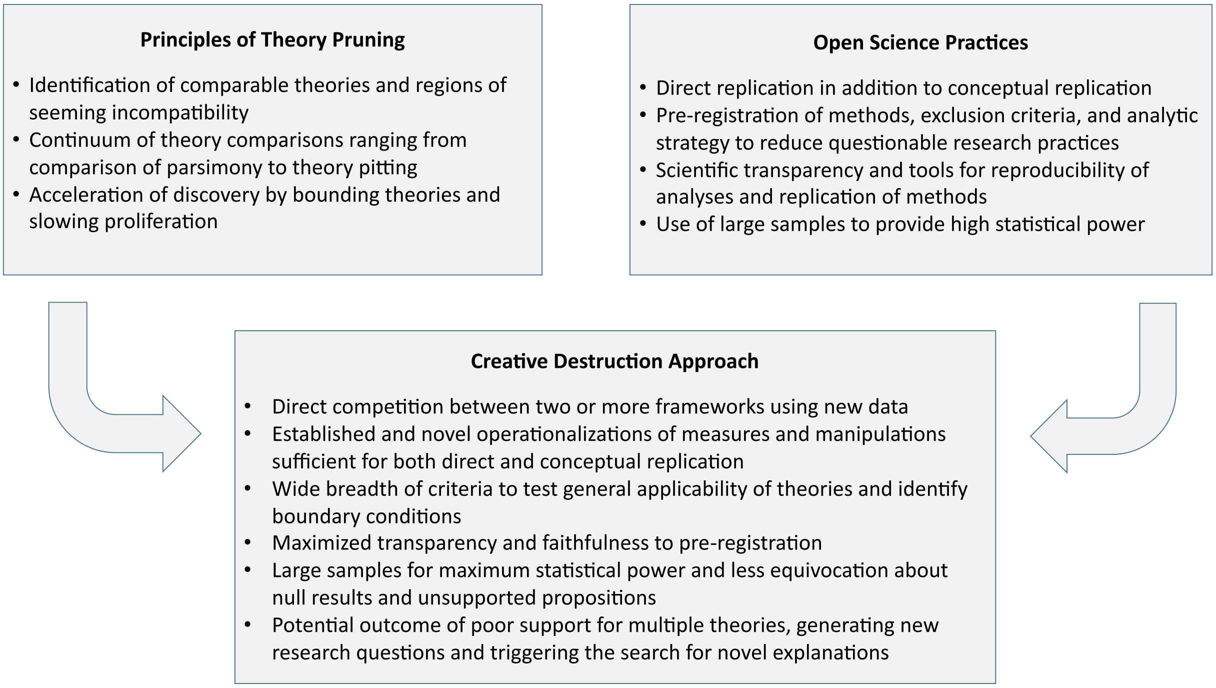
5. A creative destruction approach to organizational scholarship

Drawing on the concept of Schumpeter’s gale in a capitalistic economy (Schumpeter, 1942/1994), in which outmoded organizations and processes are continually replaced by newer, more effective ones, we argue that replication initiatives should regularly pit competing ideas against one another. Adding new conditions, measures, and subject populations to replication designs allows for accomplishing so much more than merely supporting the original findings or producing null results. It could prove an ineffective use of resources to conduct a large-scale replication assessing many moderators if the original finding, or context sensitivity of that finding, were the only theoretically interesting outcome. However, one of the goals of the creative destruction approach is to introduce further theories and expected findings, such that a completely different pattern of results can still be highly informative. Through this process, outmoded intellectual ideas can be replaced with revised, stronger theories with greater explanatory power (see Fig. 1).

The creative destruction approach is fully aligned with existing epistemological goals of theory pruning, but is distinct in leveraging open science innovations, such as direct replication and pre-registration of predictions, to achieve especially strong inferences (Platt, 1964). There are at least four key defining characteristics that enhance the effectiveness of a creative destruction approach. Specifically: 1. testing at least two competing theoretical frameworks using new data; 2. including sufficient measures and operationalizations of key constructs to carry out both direct and conceptual replications; 3. applying maximum transparency, including pre-registration of analyses; and 4. relying on large samples in order to maximize statistical power to detect a specified effect size.

First, traditional methods of theory pruning often rely on extant data to reconcile or compare theoretical predictions. For example, Schlaegel and Koenig (2014) used meta-analytic path analysis to examine two competing explanations for entrepreneurial intentions in predicting propensity to start a firm. Although such sophisticated analytic techniques are useful for combining studies testing different theoretical orientations into a single analysis, the full set of terms and propositions for both theories may not appear within any single existing study or dataset. Moreover, because research finding support for the proposed hypotheses is far more likely to lead to a publication (i.e., publication bias; Fanelli, 2010; Kepes, Banks, McDanel, & Whetzel, 2012), available reports using such an approach are unlikely to result in the conclusion that a third explanation may be superior (i.e., that neither of the pitted theories is supported). By contrast, creative destruction involves collecting novel data, explicitly including measures for all key constructs and propositions specified by both theories, and allowing for the possibility that an unexpected pattern of results will emerge and neither theory will find strong support.

Second, creative destruction leverages both direct (same method) and conceptual (different method) replication, including measurements and experimental operationalizations of as many key variables as possible within the competing theories. Although replication is not the only way to prune theory, it has distinct advantages in terms of the information it adds. In particular, direct replication is better positioned to cast doubt on the original findings that are the building-blocks for the original theory than are other replication approaches. This is because



Principles of Theory Pruning

- Identification of comparable theories and regions of seeming incompatibility
- Continuum of theory comparisons ranging from comparison of parsimony to theory pitting
- Acceleration of discovery by bounding theories and slowing proliferation

Open Science Practices

- Direct replication in addition to conceptual replication
- Pre-registration of methods, exclusion criteria, and analytic strategy to reduce questionable research practices
- Scientific transparency and tools for reproducibility of analyses and replication of methods
- Use of large samples to provide high statistical power

Creative Destruction Approach

- Direct competition between two or more frameworks using new data
- Established and novel operationalizations of measures and manipulations sufficient for both direct and conceptual replication
- Wide breadth of criteria to test general applicability of theories and identify boundary conditions
- Maximized transparency and faithfulness to pre-registration
- Large samples for maximum statistical power and less equivocation about null results and unsupported propositions
- Potential outcome of poor support for multiple theories, generating new research questions and triggering the search for novel explanations

Fig. 1. The creative destruction approach to replication, and its roots in theory pruning methods and open science practices.

null results from a conceptual replication can be readily attributed to deviations from the original method (Schmidt, 2009; Simons, 2014). Thus, direct replications are more suited to disconfirmation than are new conceptual tests. At the same time, conceptual tests have an important place, testing the generalizability and broader validity of the theoretical ideas. Notably, recent evidence indicates that prior successful (i.e., statistically significant) conceptual replications do not predict a higher likelihood of direct replication (Kunert, 2016), underscoring the importance of repeating the original method again.

Strong theories should produce evidence that both directly replicates and is conceptually robust to alternative approaches to testing the underlying ideas. As others have noted, it is possible that theories are true only within specific measurements of key terms; that is, they are highly sensitive to the approach to measurement or conceptualization (Baribault et al., 2018; Landy et al., 2020). A strong theory should show a stable relationship across a greater range of criteria and operationalizations of variables. Creative destruction aims to establish “neutral territory” with regard to how key constructs are operationalized when placing multiple theories into competition. One pragmatic means of achieving such fair tests is to directly and conceptually replicate a collection of past findings on the same narrowly defined topic (e.g., work morality, or gender discrimination), and applying multiple theories to them, often importing new measures from prior research within those theoretical traditions.

Third, the creative destruction approach seeks to maximize transparency in making critical decisions about how data is excluded and how hypotheses are tested. Scholars have increasingly discovered that theory-supporting findings may fail to replicate under scrutiny (Tsang & Kwan, 1999), in part because hypothesizing after the results are known (i.e., HARKing; Kerr, 1998) and publication bias may put forward only tests and patterns of control variables that support a conclusion (O’Boyle et al., 2017). Moreover, researchers often include multiple versions of a dependent variable or surrogate outcomes in their work, publishing only those relationships which demonstrate the largest effect sizes and best support their conclusions (Murphy & Aguinis, 2019). Possibly most troubling is the recent discovery that a large proportion of findings do not replicate, even when replication attempts simply involve subjecting the original data to reanalysis (Bergh et al., 2017). By contrast, novel creative destruction data collections create especially high transparency, such that all targeted relationships subject to testing are pre-identified, the statistical approach is registered in advance, and all variables measured within the study are visible and reported.

Fourth, creative destruction draws conclusions from especially large sample sizes, as per the lessons of recent replication initiatives (Alogna et al., 2014; Klein et al., 2018). The problem of under-powered studies is well-known within management, such that equivocal results are often observed across investigations due to both Type I and Type II errors (Cashen & Geiger, 2004; Scherbaum & Ferreter, 2009). Further, each competing theory is expected to make predictions about both significant relationships and weak to minimal relationships among the host of included variables and conditions. Thus, no theory has the unfair advantage of predicting only null effects, which can be confounded by problems with the measures or samples.

Epistemologists have long argued that falsification tests play a critical role in advancing scientific knowledge (Kuhn, 1962; Popper, 1959). Although management has lagged behind some other sciences in doing so, strong inference comparisons between theories have long been an acknowledged goal of organizational science (Davis, 2006). Tests which allow for the immediate support of one theory and rejection of the core arguments of another are likely to remain uncommon for myriad reasons (Leavitt et al., 2010), but the creative destruction approach may accelerate the ultimate abandonment of comparatively weaker theories. Science can generally not prove a theory correct or incorrect, but it can falsify propositions or statements which emerge from the theory (Lakatos, 1970; Popper, 1959). Lakatos (1970) argued that, as emergent propositions are falsified, the core of a theory becomes surrounded by a “protective belt” of boundary conditions, exceptions, and qualifying conditions. Although the core itself may not appear directly in jeopardy, the predictive belt of a questionable theory becomes dense and heavy enough over time to reduce its practical usefulness, leading scholars to abandon it in favor of less burdened theories. We suggest that a creative destruction approach can accelerate the accumulation of protective belts, and accordingly orient scholars toward theories without such constraints. Although neither direct nor conceptual replications can easily disprove a theory, when multiple theories are tested against one together, the accumulating evidence can suggest one theory has greater explanatory power to another and should be preferred. To illustrate this, we describe below the results of three recent creative destruction replication initiatives.

6. Example 1: Culture and work morality

Management scholars have long noted that work centrality and work values vary across countries, as a function of both differences in organizational forms (Parboteeah & Cullen, 2003), and deeply embedded cultural assumptions (Bond & Smith, 1996; Hofstede, 2001; Schwartz, 1999). Tierney et al. (2019) recently applied the creative destruction approach to past experimental research on *Implicit Puritanism* in American work morality (Poehlman, 2007; Uhlmann, Poehlman, & Bargh, 2009; Uhlmann, Poehlman, Tannenbaum, & Bargh, 2011). Unlike other religious faiths, traditional Puritan-Protestantism valorizes work as an end unto itself and path to divine salvation (Weber, 1904/1958). The theory of Implicit Puritanism argues for a founder effect in U.S. culture, such that the traditional values of the Puritan-Protestant settlers continue to shape contemporary Americans' moral intuitions and behaviors related to work. The theory draws both on cross-disciplinary scholarship on U.S. culture (Baker, 2005; Tocqueville, 1840/1990; Landes, 1998; Lipset, 1996) and contemporary research on implicit social cognitive processes (Greenwald & Banaji, 1995). Just as cultural racial stereotypes implicitly influence individuals exposed to the social context creating those stereotypes in the first place (Payne, Vuletich, & Brown-Iannuzii, 2019), traditional Puritan-Protestant values are hypothesized to implicitly influence not only devout American Protestants, but also non-Protestant and less religious Americans.

Relevant experimental research (Poehlman, 2007; Uhlmann et al., 2009) finds that moral character inferences about a lottery winner who continues to work in the absence of any material need are highly favorable. Further, among Americans but not Mexicans, this "needless work" effect is sensitive to target age, such that a 23 year old lottery winner who continues to work is praised more than a 46 year old who does the same. Presumably it is more legitimate, from the standpoint of the Protestant work ethic, to retire after already contributing decades of hard work. Another theoretically expected moderator of moral judgments based on needless work is the social perceiver's mindset. Specifically, thoughtless, automatic processing should promote the expression of implicit cultural work values. Consistent with this idea, American participants are especially likely to morally praise a person who continues to work after a windfall lottery win when making judgments intuitively rather than deliberatively.

Further supporting the subtle and even nonconscious nature of Implicit Puritanism are the tacit inferences drawn by Americans (Poehlman, 2007; Uhlmann et al., 2009). Specifically, American but not Chinese participants falsely remember a target person who violates traditional work morality (e.g., by contributing less work than others at their job) as sexually promiscuous, and vice versa. This implicit link between American work and sex values is theoretically forged, via cognitive balance (Greenwald et al., 2002; Heider, 1958), by their mutual links with American identity. In other words, since implicit U.S. work values and implicit U.S. sex values are both automatically linked with U.S. identity, they tend to be automatically linked to one another as well.

The theory of Implicit Puritanism predicts and finds in a series of empirical tests (Poehlman, 2007; Uhlmann et al., 2009, 2011) that U.S. work morality is distinct not only from Latin and East Asian comparison cultures, but also other Western nations such as Canada and the United Kingdom. The theory thus makes strong, readily testable predictions regarding work morality effects expected to be solely present in the United States.

As shown in Table 1, there are also a number of alternative theories of work morality across cultures. The *Explicit American Moral Exceptionalism* perspective concurs that Americans exhibit a unique moral orientation towards work, but postulates that this is fully conscious (Baker, 2005; Lipset, 1996) as reflected for example in explicit endorsement of the Protestant work ethic (Katz & Hass, 1988).

Since the original experimental demonstrations of Implicit Puritanism relied on relatively small samples, it is possible the reported effects (e.g., tacit inferences drawn from work behaviors, moral judgments based on needless work) are all false positives. Alternatively, the experimental effects could be reliable, but the originally observed cultural differences (i.e., between the U.S. and other Western and non-Western nations) may not be. Of particular interest, work could be intuitively moralized across cultures, with nothing special about U.S. work morality in this respect. This *General Moralization of Work* hypothesis is indirectly supported by research on thirty-party punishment of noncontributors to group efforts (Dreber, Rand, Fudenberg, & Nowak, 2008; Jordan, Hoffman, Bloom, & Rand, 2016), and predicts that the experimental effects originally predicted by the theory of Implicit Puritanism will replicate in any society.

A distinct pattern of national differences is anticipated by studies of the effects of economic prosperity on national work values. Research relying on the World Values Survey (WVS) identifies a developmental sequence such that people in economically poorly off countries tend to endorse survival values, among these working strictly for material gain (Inglehart, 1997; Inglehart & Welzel, 2005). As a society becomes wealthier, there is a shift from materialism to post-materialistic values such as treating work as a source of meaning, self-expression, and fulfillment. This *Self-Expression Values* account suggests individuals from relatively prosperous nations, not only the U.S. but also for example Australia or the United Kingdom, should moralize work as an end unto itself. In contrast, individuals from less economically well-off nations characterized by survival values (e.g., India) should not.

Yet another competing theoretical perspective argues that subregions within nations are often just as, if not more, important than national borders when it comes to delineating cultural boundaries (Harrington & Gelfand, 2014; Kitayama, Ishii, Imada, Takemura, & Ramaswamy, 2006; Nisbett & Cohen, 1996; Talhelm et al., 2014; Vandello & Cohen, 1999). Of particular relevance here, the *Regional Folkways* perspective (Fisher, 1989) argues there are multiple U.S. cultures—Puritan influenced New England, the plantation culture of the South (shaped by English gentry), the industrial culture of the Midwest (shaped by Quaker influence), and the ranch culture of the American West (shaped by Scotch-Irish migration). If so, then Puritan-Protestant morality effects originally predicted by the theory of Implicit Puritanism should be strongest in the New England region of the United States.

It is also possible that individual differences in ideologies are more important in driving moral judgments of work than broader culture mores. For example, personally held religious beliefs, rather than a nation or region's religious history, may best predict upholding traditional work morality. This *Religious Differences* perspective predicts that religious Protestants should be more likely than non-Protestants, and religious persons more likely than atheists, to moralize needless work—regardless of what country or countries the individuals in question are from.

With regard to cultural divides within national borders, research highlights the importance of social class differences (Snibbe & Markus, 2005; Stephens, Fryberg, & Markus, 2011). Both within the United States and other nations (e.g., Italy, Poland, Ukraine, Russia, and Japan), low socio-economic status (SES) individuals are more relationally oriented and deferent to authority than individuals with a higher income and more formal education (Grossmann & Varnum, 2011). Particularly relevant here, low-SES people also tend to regard work instrumentally, in other words as a means of earning income rather than a source of meaning and fulfillment (Argyle, 1994; Williams, 2012). This *Social Class* perspective thus suggests the tendency to valorize needless work may characterize high-SES individuals across societies. The original investigations of Implicit Puritanism (Poehlman,

Table 1
Empirical predictions of competing perspectives on culture and work values.

THEORY	NEEDLESS WORK EFFECT	TACIT INFERENCES EFFECT	INTUITIVE WORK MORALITY EFFECT
Description of key effect: *The experimental finding the theories make competing predictions about*	A postal worker who continues to work after winning the lottery is perceived as a morally good person, especially if she is young (23) rather than older (46). In other words, target age moderates the effects of working for no reason on judgments of moral character.	Women and men who fail to uphold traditional work morality are misremembered as violating traditional sex morality, and vice versa.	The needless work effect is exhibited in an intuitive mindset, but not a deliberative mindset.
Implicit Puritanism perspective: *Americans unconsciously moralize work*	Americans, but not non-Americans, are sensitive to the age of a target who works needlessly. No moderation by individual differences in religion (Protestant or not), religiosity, social class, sub-region within the United States (New England states vs. other states), or explicit endorsement of the Protestant Work ethic (PWE).	Americans, but not non-Americans, exhibit the tacit inferences effect. No moderation by individual differences in religion, religiosity, social class, sub-region of the U.S., or explicit PWE endorsement.	Americans, but not non-Americans, exhibit the intuitive work morality effect. No moderation by individual differences in religion, religiosity, social class, sub-region of the U.S., or explicit PWE endorsement.
Religious differences perspective: *Religious Protestants moralize work*	Protestant and religious participants should be more likely to exhibit the needless work effect than non-Protestants and less religious individuals.	Protestant and religious participants should be more likely to exhibit the tacit inferences effect than non-Protestants and less religious individuals.	Protestant and religious participants should be more likely to exhibit the intuitive work morality effect than non-Protestants and less religious individuals.
Regional folkways perspective: *New Englanders moralize work*	Participants from the New England U.S. states should be more likely than others to exhibit the needless work effect.	Participants from the New England U.S. states should be more likely than others to exhibit the tacit inferences effect.	Participants from the New England U.S. states should be more likely than others to exhibit the intuitive work morality effect.
Explicit American exceptionalism perspective: *Americans consciously moralize work*	Americans, but not non-Americans, are sensitive to the age of a target who works needlessly. The effect is observed more strongly among individuals who explicitly endorse the Protestant Work Ethic.	Americans, but not non-Americans, exhibit the tacit inferences effect. The effect is observed more strongly among individuals who explicitly endorse the Protestant Work Ethic.	Americans, but not non-Americans, exhibit the intuitive work morality effect. The effect is observed more strongly among individuals who explicitly endorse the Protestant Work Ethic.
General moralization of work perspective: *People across cultures moralize work*	Both Americans and non-Americans exhibit the needless work effect and are sensitive to target age.	Both Americans and non-Americans exhibit the tacit inferences effect.	Both Americans and non-Americans exhibit the intuitive work morality effect.
False positives perspective: *The original findings are spurious*	No needless work effect or sensitivity to target age, and no moderation by individual differences in religion, religiosity, or sub-region.	No tacit inferences effect and no moderation by individual differences in religion, religiosity, or sub-region.	No intuitive work morality effect and no moderation by individual differences in religion, religiosity, or sub-region.
Self-expression values perspective: *Individuals from wealthy nations moralize work*	Participants from the USA, UK, and Australia should exhibit the needless work effect, whereas Indian participants should not.	This theory does not anticipate the tacit inferences effect.	Participants from the USA, UK, and Australia should exhibit the intuitive work morality effect, whereas Indian participants should not.
Social class perspective: *High-SES persons moralize work*	High socioeconomic status participants should exhibit the needless work effect more than low socioeconomic status participants.	This theory does not anticipate the tacit inferences effect.	High socioeconomic status participants should exhibit the intuitive work morality effect more than low socioeconomic status participants.

Note. The table entries represent the extreme case in which a given theory's empirical predictions hold to the exclusion of all other theories.

2007; Uhlmann et al., 2009, 2011) did not observe any reliable individual differences based on religion, religiosity, or socioeconomic status, but relying on small samples were potentially underpowered to detect them. The creative destruction replications conducted by Tierney et al. (2019) allowed for high-powered tests of all these plausible accounts of work morality across cultures (see Table 1 for an overview).

Tierney et al.'s (2019) replication initiative re-examined the aforementioned set of work-morality findings predicted by the theory of Implicit Puritanism (Poehlman, 2007; Uhlmann et al., 2009, 2011). These included the previously observed patterns that (1) Americans are more likely to laud a young (rather than an older) person who continues to work after winning the lottery, (2) that this needless work effect observed among Americans is especially strong in an intuitive mindset, and finally (3) tacit inferences reflecting an intuitive link between work and sex morality in American moral cognition. These new data collections encompassed novel populations, including large samples from not only the United States and United Kingdom (as in Uhlmann et al., 2011), but also Australia and India. Unlike the original investigations, participants were systematically recruited from all nine of the U.S. census districts, with the New England states strategically oversampled to facilitate high powered tests of the regional folkways account (Fisher, 1989). Further included were novel measures, such as the Protestant Work Ethic scale (Katz & Hass, 1988) to allow for tests of the explicit American exceptionalism thesis (Baker, 2005; Lipset, 1996) and the validated Duke University Religion Index (DUREL) assessment of religious beliefs (Koenig & Büssing, 2010). The design thus encompassed not only direct replications of the original findings in the original U.S. samples, but also conceptual replications with new populations and measures, allowing us to test eight theoretical accounts of culture and work.

The results of the cross-national data collection, encompassing over 5000 research participants sampled from the constituent regions of four nations, were highly informative in terms of adjudicating between the competing theories. As summarized in Table 2, as a direct consequence

Table 2
Implications of the replication results for competing theories of culture and work values.

THEORY	NEEDLESS WORK EFFECT	TACIT INFERENCES EFFECT	INTUITIVE WORK MORALITY EFFECT	OVERALL ASSESSMENT
Implicit Puritanism perspective	The theory of Implicit Puritanism's original prediction that a younger person is praised more than an older person for continuing to work after winning the lottery is not supported. In other words, target age does not reliably moderate the "needless work" effect that continuing to work rather than retiring elicits favorable character judgments.	As predicted by the theory of Implicit Puritanism, women and men who fail to uphold traditional work morality are misremembered as violating traditional sex morality (and vice versa). However, sharply contradicting the original theory, the tacit inferences effect is observed not only in the United States, but also Australia, the United Kingdom, and India (although Indians exhibit the effect less strongly than Americans).	As predicted by the original theory, the needless work effect is stronger in an intuitive mindset than in a deliberative mindset. Sharply contradicting the original theory, the intuitive work morality effect is observed not only in the U.S., but also Australia and the United Kingdom. Consistent with the original theory, Indians do not appear to exhibit the effect.	The theory of Implicit Puritanism suffers a theoretical core breach due to the empirical results of the replication initiative. One of the three key effects predicted by the theory (target age and needless work) failed to replicate entirely. Two other effects (tacit inferences and intuitive work morality) did replicate, but were also found across several other nations, contrary to the theory's core claim of a unique American work morality.
Religious differences perspective	Contrary to this theory's predictions, religion (Protestant or not) and religiosity did not make participants more sensitive than others to target age in their judgments of needless work.	No moderating effect of religion (Protestant or not). Religiosity either predicts in the wrong direction (i.e., more religious participants exhibit the tacit inferences effect *less*), or not at all, depending on the sample.	No moderating effect of religion (Protestant or not). Religiosity either predicts in the wrong direction (i.e., more religious participants exhibit the intuitive work morality effect *less*), or not at all, depending on the sample.	No support for the prediction that religious Protestants exhibit the work morality effects targeted for replication more strongly.
Regional folkways perspective	Contrary to this theory's predictions, New Englanders are not more sensitive than others to target age in their judgments of needless work.	Contrary to this theory's predictions, New Englanders are not any more likely to exhibit the tacit inferences effect.	Contrary to this theory's predictions, New Englanders are not any more likely to exhibit the intuitive work morality effect.	No empirical support for the idea that individuals from the New England states exhibit the work morality effects targeted for replication more strongly.
Explicit American exceptionalism perspective	Contrary to the theory's predictions, Americans do not differ from others in terms of sensitivity to target age in judgments of needless work.	Contradicting this theory, Americans are not consistently more likely to exhibit the tacit inferences effect than members of other Western cultures. Further, individuals who explicitly endorse the Protestant work ethic exhibit the tacit inferences effect *less*, directly contrary to predictions.	Contradicting this theory, Americans are not consistently more likely to exhibit the intuitive work morality effect than members of other Western cultures. Further, explicit endorsement of the Protestant work ethic either predicted this effect in the wrong direction or not at all, depending on the sample.	No empirical support for a unique American response to the specific work morality effects studied. The expected pattern of national differences did not emerge, and explicit endorsement of the Protestant work ethic either predicted effects in the wrong direction or not at all.
General moralization of work perspective	No support for the original prediction that target age moderates moral judgments based on needless work; no such pattern is observed across four nations. Strong support for the prediction that across cultures, working in the absence of material need is morally praised.	Strong support for the prediction that across cultures, women and men who fail to uphold traditional work morality are misremembered as violating traditional sex morality (and vice versa). The effect is observed in all four nations studied, although Indian participants make weaker tacit inferences than Americans.	Fairly strong support for the prediction that across cultures, the needless work effect is stronger in an intuitive mindset than in a deliberative mindset. The effect is observed in three out of four nations studied (the US, UK, and Australia, but not India).	Strong empirical support for the prediction work is moralized across cultures, with the caveat that the intuitive work morality effect did not emerge reliably in India (see below under "self-expression values perspective").
False positives perspective	The original finding that target age moderates the needless work effect appears to be a false positive. There is however a robust main effect of needless work on judgments of moral character that replicates across cultures.	The tacit inferences effect is robust across four out of four societies and not a false positive.	The intuitive work morality effect is robust across three out of four societies and not a false positive.	The false positives perspective is supported for one of the original effects targeted for replication. Specifically, the moderating effect of target age on character judgments based on needless work is not reliable. In contrast, the main effect of needless work on moral judgments, tacit inferences effect, and intuitive work morality effect are true positives that emerge in most samples.
Self-expression values perspective	No support for cultural differences in the effects of target age on moral judgments of needless work.	Did not anticipate the tacit inferences effect would emerge, when in fact it replicated across four out of four societies.	Consistent with this theory, while the intuitive work morality effect is robust in U.S., U.K., and Australian samples, Indians do not exhibit this pattern of judgments.	Partial empirical support for the prediction that nations high in self-expression values (USA, UK, Australia) intuitively moralize work more than a nation characterized by survival values (India). This theory's predictions are supported for one of three effects targeted for replication (intuitive work morality effect). Further research comparing

(*continued on next page*)

Table 2 (*continued*)

THEORY	NEEDLESS WORK EFFECT	TACIT INFERENCES EFFECT	INTUITIVE WORK MORALITY EFFECT	OVERALL ASSESSMENT
				more cultures high and low in self-expression orientation, and measuring individual differences in such values, is needed before drawing strong conclusions.
Social class perspective	No support for the idea that social class moderates the effects of target age on moral judgments of needless work.	Did not anticipate the tacit inferences effect would emerge, when in fact it replicated across four societies.	Formally educated participants were not consistently more likely to exhibit the intuitive work morality effect, failing to support the predictions of this theory.	No support for the social class perspective. Socioeconomic status did not consistently moderate the effect in the expected direction for the intuitive work morality effect, or the target age and needless work effect. A third effect, not anticipated by this theory, emerged as replicable across cultures.

of the replication initiative, Implicit Puritanism suffers a theoretical core breach. One of the key original findings predicted by the theory (target age moderating judgments of needless work) fails to replicate entirely and is identified as a likely false positive. Two further effects (intuitive mindset moderating judgment of needless work, and tacit inferences based on work behaviors) replicate not only in the United States, but also in other nations, sharply contradicting the theory's core claim of a unique American work morality. Due in no small part to the inclusion of additional measures and populations, we were able to identify alternative theories of culture and work values that better capture the observed pattern of empirical results. Specifically, strong evidence was obtained that work is moralized intuitively across cultures. At the same time, partial support emerged for the prediction that needless work is moralized to a greater extent in self-expression cultures (U.S., Australia, U.K.) than in a culture characterized by survival values (India).

Further studies of implicit and explicit work morality across a larger number of countries are needed to adjudicate between the general moralization of work and self-expression values perspectives. A theoretical integration, such that work is moralized across cultures but significantly more so in self-expression cultures than in survival values cultures, seems viable. Regardless, scholars of culture and work can set aside the Implicit Puritanism thesis with confidence, and theorize anew. We believe this outcome underscores the utility and generative nature of the creative destruction approach to replication. Below, we describe another such initiative, testing different theories of how people reason about scientific evidence.

7. Example 2: Working parents' reasoning about child care choices

Are we dispassionate information processors, drawing rational inferences from the available data using a bottom-up approach? Or are we theory driven, accepting or rejecting new information in a top-down manner based on pre-existing schemas and expectations? Finally, is human reasoning distorted by directional motives to reach desired conclusions?

An experimental approach is uniquely suited to addressing age-old philosophical questions regarding the extent to which reasoning is data driven, theory driven, and motive driven. By holding constant extraneous factors, measuring key individual differences, and manipulating critical features of the situation between subjects, investigators can empirically distinguish whether participants are objectively weighting the relevant evidence, confirming pre-existing theories, or striving for hoped-for conclusions. Using a now classic paradigm, Lord, Ross, and Lepper (1979) provide evidence that people with strong opinions on a controversial issue (e.g., the death penalty) evaluate scientific evidence in light of their prior beliefs. Specifically, when participants were randomly assigned to read about studies with different methodologies and conclusions, their assessments of study quality were driven by the studies' results (e.g., pro-deterrence vs. anti-deterrence) not the objective methodology (e.g., pretest–posttest vs. correlational design). A host of related findings speak to the influence of prior convictions on information processing (Koriat, Lichtenstein, & Fischhoff, 1980; Mahoney, 1977; Pitz, 1969; Ross, Lepper, & Hubbard, 1975), which is arguably rationally defensible in Bayesian terms (Baron & Jost, 2019; Krueger & Funder, 2004).

The cognitive vs. motivational underpinnings of such information processing are extremely difficult to parse—in fact, Tetlock and Levi (1982) pronounced the motivation-cognition debate potentially intractable. Are participants, again potentially quite rationally (Baron & Jost, 2019; Krueger & Funder, 2004), less likely to cognitively accept new information that contradict their priors? Or, are they truly contorting the evidence and standards in order to believe what they want to believe? For example, decisions about parenting and family arrangements impact the attitudes and behaviors of employees at work

Table 3
Empirical predictions of different theoretical perspectives on working parents' reasoning about child care.

EFFECT	MOTIVATED REASONING PERSPECTIVE	COGNITIVE SCHEMA-BASED PROCESSING PERSPECTIVE	ACCURACY-DRIVEN REASONING PERSPECTIVE
Prior beliefs and the processing of evidence	Beliefs only appear to influence reasoning because they are aligned with desires; when misaligned, desires trump beliefs in driving reasoning.	Desires only appear to influence reasoning because they are aligned with beliefs; when misaligned, beliefs trump desires in driving reasoning.	Prior beliefs do not influence reasoning about scientific evidence.
Prior desires and the processing of evidence	Desired conclusions influence reasoning about scientific evidence.	Desired conclusions do not influence reasoning about scientific evidence.	Desired conclusions do not influence reasoning about scientific evidence.
Effects of pre-commitment to criteria	Commitment to criteria should constrain motivated reasoning, and reduce the effects of desired outcomes on the processing of scientific evidence.	Commitment to criteria should reduce ambiguity and constrain the application of cognitive schemas, and therefore reduce the extent to which prior beliefs drive the processing of scientific evidence.	People already apply criteria in an objective manner, hence pre-commitment to criteria should not affect their judgments.
Effects of being an actual parent vs. intended parent	Actual parents should exhibit stronger assimilation effects than would-be-parents, since the psychological need to rationalize actual (rather than intended) child care decisions is greater.	No predicted difference between intended parents and actual parents in assimilation to prior beliefs, so long as they hold the same cognitive beliefs about child care.	If both are sufficiently accuracy motivated, neither actual nor intended parents will exhibit assimilation effects. If anything, actual parents should exhibit more objective reasoning about child care than intended parents. The stakes are higher for the former group, activating accuracy goals.

Notes. The table entries represent the extreme case in which a given theory's empirical predictions hold to the exclusion of all other theories.

(Desai, Chugh, & Brief, 2014), and work experiences similarly spill over into parenting behaviors (Stewart & Barling, 1996). Satisfaction with child care arrangements are a critical predictor of work-family conflict and consequent absenteeism (Goff, Mount, & Jamison, 1990). Thus, child care represents a critical domain in which employees should be motivated to invest substantial cognitive resources and seek to optimize their outcomes, but how such decisions are made would be differentially predicted by various theories of reasoning.

One admittedly imperfect approach to disentangling these processes, introduced by Bastardi, Uhlmann, and Ross (2011), is to identify individuals whose factual beliefs and emotional desires are misaligned with one another, then examine how they engage with ambiguous evidence. Such situations in which what a person wants to be true and what they believe is factually true are diametrically opposed are highly theoretically informative, but also rare. One such case is parents-to-be who believe home care is better for children, yet intend to place their own future children in day care (e.g., in order to pursue a professional career outside the home). For such individuals, the cognitive expectancy that rigorous scientific research will support the developmental advantages of home care conflict with their earnest hope that the science will find day care to be just as good for children as home care. Adapting the Lord et al. (1979) paradigm, Bastardi et al. (2011) find that such "conflicted" participants, when presented with the methods and results of purported scientific studies on the topic, favor whichever methodology (random assignment versus statistical matching) suggests day care is not disadvantageous for children. When motivational factors (hoped-for and feared outcomes) were placed in conflict with cognitive priors, the hopes and fears won. The wishful thinking paradigm has limitations, such as the difficulty of accurately measuring prior beliefs and desires, as well as changes in beliefs in response to new evidence. However, we believe it is informative regarding the motivation-cognition debate.

At the same time, other work supports the importance of accuracy-driven reasoning (Devine, Hirt, & Gehrke, 1990; Funder, 1987; Jussim, 1991; Trope & Bassok, 1982). From the standpoint of evolutionary adaptiveness, it follows that humans come equipped with reasoning abilities to help us construct a fairly veridical internal representation of the external world. If so then accuracy goals, either chronic or situationally activated in important situations, should explain the bulk of the variance in how human beings process evidence.

Ebersole (2019, Study 6) recently conducted a large sample replication-and-extension using the Bastardi et al. (2011) materials as a starting point, and further including an experimental manipulation of *a priori* commitment to criteria. Specifically, some participants were asked to indicate which scientific method (random assignment vs. statistical matching) they considered most valid before learning the results of scientific studies of the effects of home care vs. day care that employed those methodologies. Pre-commitment to criteria should constrain reasoning (whether based on cognitive beliefs or motivated desires), promoting accuracy-based, bottom-up consideration of the evidence.

In another extension of the original Bastardi et al., 2011 design, Ebersole (2019) expanded the populations sampled to include not only would-be-parents (as in Bastardi et al., 2011), but also actual parents who have made the choice to use home care or day care for their children. This allows for novel tests of the effects of hypothetical vs. real situations on assimilation effects. From an accuracy-based perspective, the higher stakes in actual situations should attenuate any irrational departures from the logical maximization of accuracy and realized value (Armor & Sackett, 2006; Carpenter, Verhoogen, & Burks, 2005; Levitt & List, 2007; List, 2006). This suggests parents may process new information about the efficacy of their child care practices more rigorously and dispassionately than non-parents.

In contrast, theories of motivated reasoning make the directly opposing prediction, postulating that rationalizations for child care choices should be more evident among actual parents than would-be parents. Festinger's (1957) theory of cognitive dissonance suggests that having already committed to a course of action in a consequential domain should increase the desire to justify one's decisions. This suggests that parents who have already entrusted their children to day care should be more, not less, prone to motivated reasoning in this domain.

Table 3 displays the theoretical predictions of the *Motivated Reasoning*, *Cognitive Schema*, and *Accuracy-Driven* perspectives on reasoning in the wishful thinking paradigm (Bastardi et al., 2011; Ebersole, 2019). While conducting direct/literal replications of the original method, we thus at the same time attempt to achieve what Köhler and Cortina (in press) call generalizability tests, in this case specifically testing moderators about which competing theories make opposing predictions (e.g., parental status). The pre-registered analysis plans and study materials are available on the OSF (https://osf.io/9fy8m) and in Supplement 1, and the data and code are likewise posted online (data: https://osf.io/fhq45/, analysis code: https://osf.io/rphwv/). Notably, the creative destruction analyses were formulated and pre-registered after the Ebersole (2019) data collections were carried out, thus this constitutes a secondary analysis of the dataset (Van den Akker et al., 2019).

The results of this re-analysis (1) reproduced the pre-registered predictions of Ebersole (2019) regarding the effects of pre-commitment on assimilation to prior beliefs, and (2) pitted theories of motivated reasoning, cognitive schema-based processing, and accuracy-based reasoning against each other in a highly informative manner. Conceptually replicating the assimilation-to-beliefs effect (Lord et al., 1979), participants who had not committed to methodological standards rejected the methodology and findings of a scientific study whose results challenged their cognitive beliefs about the efficacy of home vs. day care. As hypothesized, the commitment condition eliminated cognitive assimilation (Ebersole, 2019).

The wishful thinking paradigm’s approach to teasing apart cognitive and motivational explanations for assimilation effects focuses on “conflicted” participants who either have children in day care or expect to one day, yet believe home care is better for children’s development. Such individuals’ cognitive beliefs in the superiority of home care are in conflict with their motivated desire to find out that day care is just as good. Our re-analyses of Ebersole (2019, Study 6) failed to replicate the original wishful thinking effect that desired outcomes trump factual beliefs in the assimilation paradigm. Directly contrary to the striking pattern reported by Bastardi, Uhlmann, & Ross, 2011, prior beliefs rather than desired outcomes predicted evaluations of the methodology of the scientific studies. Further, actual parents and intended parents were similarly likely to display assimilation effects regarding child care practices, failing to support theories predicting that high-stakes situations would be associated with stronger (or weaker) assimilation effects. Table 4 summarizes the implications of the creative destruction analyses for different theories of reasoning. Overall, the results most strongly support the cognitive schema perspective, in which new evidence is evaluated in light of prior beliefs, not desires. Such cognitive confirmation effects are arguably compatible with Bayesian thinking and human rationality (Baron & Jost, 2019; Krueger & Funder, 2004).

What drives human reasoning—do we follow the evidence where it leads us, tend to confirm pre-existing theories and expectations, or believe what we want to believe? A definitive answer to this very old question is beyond the scope of any original study or replication. The field could use further empirical approaches, for example experimentally creating new beliefs and desires, varying the strength of arguments and looking at belief updating, or using longitudinal designs examining the dynamic interplay between beliefs and the processing of evidence. We believe the creative destruction approach, encompassing new conditions and measures and direct as well as conceptual replications, can add value for future research on the nature of the reasoning process across topics. On that point, we report the results of a novel empirical study re-examining prior work on motivated gender stereotyping in hiring contexts.

8. Example 3: Motivated gender discrimination

Gender-based selection decisions have long been a topic of interest to organizational scholars (Harvie, Marshall-Mcaskey, & Johnston, 1998; Olian, Schwab, & Haberfeld, 1988; Perry, Davis-Blake, & Kulick, 1994). In an empirical study conducted for this paper, we apply the creative destruction approach to earlier findings regarding the roles of psychological rationalizations and illusions of personal objectivity in discrimination against women. The original series of experiments finds that evaluators shift the hiring criteria for the position in favor of male applicants for stereotypically male jobs, but do not exhibit the same favoritism toward female applicants (Uhlmann & Cohen, 2005, 2007). If evaluators were applying cognitive schemas based on gender stereotypes to the descriptions of the applicants, then this should have affected the impressions formed of their traits and characteristics (e.g., perceived toughness or communication skills). However, candidate gender instead affected endorsement of hiring criteria (e.g., are

Table 4
Implications of the replication project’s results for different theories of reasoning.

THEORY	BELIEFS VS. DESIRES AND REASONING	PRE-COMMITMENT TO CRITERIA	EFFECTS OF PARENTAL STATUS	OVERALL ASSESSMENT
Cognitive schema perspective	Participants rejected the methods of scientific studies that disconfirmed their *a priori* beliefs and embraced the methods of studies that supported their beliefs. This conceptually replicates Lord et al. (1979). Further, when desires and beliefs were placed in opposition, beliefs drove assimilation, strongly supporting the cognitive schema account.	Pre-commitment to criteria attenuated assimilation to cognitive beliefs, supporting this theoretical account. Reducing ambiguity in the target of judgment diminished a reliance on pre-existing schemas such as cognitive beliefs.	Supported, in that actual parents were no different in their reasoning from intended parents. This is consistent with assimilation to a simple cognitive schema, as opposed to the boosts and reductions in assimilation predicted by the motivated reasoning and accuracy perspectives.	Very strongly supported by the replication initiative. The original assimilation to cognitive beliefs effect (Lord et al., 1979) was conceptually replicated in a large-sample, pre-registered study. Further, higher stakes did not moderate the processing of ambiguous scientific evidence, attesting to the robustness of the assimilation effect.
Motivated reasoning perspective	“Conflicted” parents, who believed home care is better for kids but expected to rely on day care themselves, exhibited assimilation towards their prior beliefs, not their hoped-for outcome. This fails to replicate the Bastardi et al. (2011) wishful thinking effect, and contradicts the motivated reasoning account.	Pre-commitment to criteria did not moderate assimilation towards desired outcomes. This is unsurprising, given that assimilation was towards beliefs, not desires, hence there was nothing for the commitment intervention to moderate.	Not supported, in that actual parents were just as likely to confirm desired outcomes as intended parents. The motivated reasoning account had predicted that actual parents would be more driven to rationalize desired outcomes.	Strongly contradicted by the creative destruction analyses. The Bastardi et al. (2011) effect that desires trump beliefs in responses to new evidence failed to replicate in a sample much larger than the original demonstration. Indeed, the original pattern was completely reversed, such that beliefs trumped desires in predicting the processing of scientific evidence.
Accuracy perspective	Not supported, in that prior beliefs influenced interpretations of ambiguous scientific evidence.	Not supported. This perspective holds that decision makers endorse scientific standards in an objective manner. Thus commitment to standards before or after knowing the results of the studies should not have made any difference.	Not supported, in that actual parents were no different in their reasoning from intended parents. The accuracy perspective had predicted that actual parents would be less influenced by prior beliefs, given their goal to make the most informed choice possible for their children.	The creative destruction project failed to support the prediction that participants would process evidence in a bottom-up, evidence-based manner. Rather, participants exhibited assimilation towards prior cognitive beliefs that was not corrected in high-stakes situations.

toughness or communication skills more important for the job of police chief?), with no effects on perceived applicant characteristics.

Further consistent with a motivated reasoning account, decisions makers who flexibly change their hiring criteria to rationalize selecting male candidates believe themselves to be more objective (Uhlmann & Cohen, 2005). Providing evidence of a causal relationship, Uhlmann & Cohen (2007) show that experimentally inducing a sense of objectivity leads decision makers to rely more on their sexist beliefs, as well as use temporarily accessible gender stereotypes in their judgments. Seeing oneself as rational and objective may engender an "I think it, therefore it's true" mindset that licenses individuals to act on their beliefs. At the same time, rationalizing judgments may reinforce an illusion of personal objectivity.

Utilizing the creative destruction approach to replication, we conducted a high-powered data collection combining key materials from both (Uhlmann & Cohen, 2005, Study 1) and (Uhlmann & Cohen, 2007, Study 3). Building on the original designs, we added conditions and measures testing competing theories of the effects of candidate gender on hiring judgments for male-typed jobs. To further test the original theory that hiring criteria and a sense of personal objectivity are constructed and maintained in a motivated manner, we included a manipulation of self-affirmation vs. self-threat (Steele, 1988; Uhlmann & Nosek, 2012). If the effects observed in Uhlmann and Cohen (2005, 2007) are "hot" processes, they should be amplified under psychological threat and ameliorated when an unrelated but important identity has been affirmed (Sherman & Cohen, 2006, 2010; cf. Dee, 2015; Protzko & Aronson, 2016; Hanselman et al., in press).

Although the original Uhlmann and Cohen (2005, 2007) findings are consistent with a motivated account of gender discrimination, the experiments were based on small samples, and moreover conducted over 15 years ago. Studies of gender discrimination are a special case of replication as there are theoretical and empirical reasons to expect (and moral reasons to deeply hope for) change over time. While the rate of change in gender gaps in pay and leadership representation has slowed (Bar-Haim, Chauvel, Gornick, & Hartung, 2018), gender stereotypes about competence have changed over time (Eagly, Nater, Miller, Kaufmann, & Sczesny, 2020), and the #MeToo movement (Garber, 2017; Johnson & Hawbaker, 2018) may have heightened awareness of mistreatment against women and the desire to take corrective steps.

In contemporary times, ideological movements and social sensitivities may potentially lead to hiring preferences in favor of female candidates for traditionally male jobs. Thus, we examined whether participants with high levels of exposure to the #MeToo movement on social media, and who strongly reject sexism and believe that gender limits women's workplace opportunities, tend to render pro-female decisions (McCormick-Huhn & Shields, 2019). To the extent that such reverse discrimination effects are based on motivated ideologies (Ditto et al., 2019; Greenberg & Jonas, 2003), they may be associated with constructing job criteria in *favor* of women, especially when threatened rather than affirmed.

Finally, a related but distinct hypothesis posits that the lay public are increasingly study-savvy and wary of "falling for" experimental manipulations. If so, individuals who have participated in more research studies, have taken a course in psychology, or are for any reason suspicious of the topic of study may exhibit overcompensation effects. In other words, they may prefer women over men for stereotypically male jobs, and provide female candidates with more favorable evaluations in general, in order to avoid appearing sexist.

Table 5 summarizes the predictions of the *Motivated Discrimination*, *Cognitive Assimilation*, *Motivated Liberalism*, and *Study Savviness* perspectives on gender and hiring decisions in experimental contexts. Supplements 2–4 contain a detailed report of a creative destruction replication study putting these ideas to an empirical test. As summarized in Table 6, the creative destruction effort yielded empirical patterns in many ways directly opposite to those in the original studies targeted for replication. The original studies observed discrimination in selection decisions against female candidates that was most evident among male evaluators whose sense of their own objectivity was activated (Uhlmann & Cohen, 2005, 2007). In contrast, the replication found overall favoritism towards female candidates among male evaluators, especially if those participants were made to feel objective. In the replication study, only female evaluators exhibited the pattern of stereotype-based discrimination against women familiar from the 2005 and 2007 papers, and this effect was not robust to alternative analytic approaches (see Supplement 4 and Table S4-1).

In terms of explaining the observed pattern of reverse discrimination among male evaluators, the study savviness explanation and motivated ideologies explanations both received some empirical support. Participants who had previously completed similar studies, or strongly rejected sexist beliefs, tended to favor female over male applicants. Although the two can be difficult to parse (Tetlock & Manstead, 1985), it is more consistent with an impression management than ideological explanation that it was male rather than female evaluators who exhibited reverse discrimination. Men are more likely than women to express a fear of appearing sexist (Soklaridis et al., 2018), yet less supportive of the #MeToo movement and feminism (Kirkman & Oswald, 2019; Kunst, Bailey, Prendergast, & Gundersen, 2019). Gender differences in self-presentation concerns in this domain track the pattern of hiring judgments, whereas gender differences in ideological commitments do not.

The original findings reflecting the motivated rationalization of discrimination against women did not directly replicate (Uhlmann & Cohen, 2005, 2007). Indeed, participants who perceived themselves as highly objective tended to construct hiring criteria favorable to *female* candidates, the mirror-opposite pattern of results to the original findings. However, a novel conceptual test did partly support the motivated discrimination against women account. Specifically, male evaluators who experienced a self-threat (relative to a self-affirmation) became less likely to favor female over male candidates for the stereotypically male-typed job of police chief. This effect of the threat-affirmation manipulation suggests the tantalizing possibility of a theoretical integration. Specifically, contemporary male participants in hiring simulations who are more experienced and knowledgeable regarding academic research may overcorrect their judgments, exhibiting reverse gender discrimination out of a fear of appearing sexist. Yet, after receiving a blow to their identity, ego-protection motives are activated and counteract this effect, so that their evaluations of female candidates become no better than those for male candidates. This mixed-motives account is highly speculative, and awaits systematic testing and empirical confirmation or disconfirmation.

A complementary forecasting survey examined whether independent scientists were able to anticipate these replication results (see https://osf.io/nz48k, and Supplements 7–9 for the forecasting survey materials, pre-registered analysis plan, and detailed report). Prior work finds that scientists are able to accurately predict simple condition differences by merely reading the study abstract or examining the study materials (Camerer et al., 2016; DellaVigna & Pope, 2018; Dreber et al., 2015; Forsell et al., 2019). We tested, for the first time, whether scientists can likewise anticipate complex interactions between variables. In this politically charged context (Tetlock, 2005), we further examined whether scientists' beliefs and values regarding gender moderate the accuracy of their predictions. Consistent with past research, in our primary pre-registered hypothesis test, we found a positive association between the observed effect sizes and the individual predictions (beliefs) of the forecasters ($\beta = 0.027$, $p < 0.001$). In a pre-registered robustness test, aggregated predictions, computed as mean predicted effect size of each of the 24 effects replicated, were directionally positively associated with the observed effect sizes, although this zero-order correlation was no longer statistically significant, $r = 0.193$, $p = 0.366$. A notable discrepancy between forecasts about selection decisions by male evaluators and the actual study outcomes was also apparent. Forecasters expected that both male and female

Table 5
Empirical predictions of different perspectives on gender discrimination in hiring decisions.

RESEARCH QUESTION	MOTIVATED DISCRIMINATION PERSPECTIVE	COGNITIVE ASSIMILATION PERSPECTIVE	MOTIVATED LIBERALISM PERSPECTIVE	STUDY SAVVINESS PERSPECTIVE
Do hiring decisions favor men or women?	*Hiring decisions favor men for stereotypically male jobs.	*Hiring decisions favor men for stereotypically male jobs.	*Hiring decisions favor female candidates.	*Hiring decisions favor female candidates.
Are perceived characteristics influenced by candidate gender?	*No target gender effect in impression formation when descriptions of candidates' characteristics are clear and unambiguous.	*Impressions of male candidates' traits and characteristics should be more favorable than for identically described female candidates, due to assimilation to stereotypes.	Either no difference, or more favorable impressions of female candidates' characteristics.	*Yes, female candidates' characteristics are rated favorably relative to male candidates.
Are hiring criteria constructed in favor of one gender over another?	*Yes, hiring criteria are shifted in favor of male candidates.	No, since stereotypes shape impressions of social targets, not judgmental standards.	*Yes, hiring criteria are shifted in favor of female candidates.	*Yes, hiring criteria are shifted in favor of female candidates.
What are the effects of affirmation-threat on hiring judgments?	*Relative to a self-threat, a self-affirmation reduces the tendencies to construct hiring criteria that favor men, choose male candidates, and act on sexist beliefs and accessible stereotypes.	*No effect of self-affirmation or threat, since hiring discrimination is cognitive not motivational in nature.	Relative to a self-threat, a self-affirmation reduces ideologically based tendencies to construct hiring criteria that favor women, choose female candidates, and act based on feminist beliefs.	No effect, since pro-female judgments are based on public impression management not intrapsychic processes.
What are the effects of experimentally inducing a sense of objectivity?	*Making a sense of personal objectivity salient increases discrimination against female candidates and reliance on sexist beliefs and accessible stereotypes.	No causal effect of such self-views on judgments, since hiring discrimination is due to the operation of cognitive expectations about targets.	Making a sense of personal objectivity salient increases reliance on ideologies that promote positive judgments of female candidates.	No effect, since hiring decisions are for public consumption not about personal identity.
What are the correlates of individual differences in self-perceived objectivity?	*Seeing oneself as objective is correlated with constructing hiring criteria that favor male over female candidates.	No relationship between such self-views and hiring judgments. Discrimination in hiring is due to the operation of cognitive expectations about targets.	A sense of personal objectivity correlates with increased reliance on ideologies that promote positive judgments of female candidates.	No effect, since hiring decisions are for public consumption and not about personal identity.
What are the effects of individual differences in feminist media exposure and beliefs about gender in the workplace?	Either no effect, or such beliefs partly compensate for motivated discrimination against female candidates.	Either no effect, or such beliefs partly compensate for cognitive discrimination against female candidates.	*Greater exposure to feminist social media and the belief that workplaces are gendered predicts pro-female judgments in selection contexts.	Either no effect, or exposure to feminist media increases the desire to avoid appearing sexist and therefore favor female candidates.
What are the effects of prior experience participating in studies and suspicions about the hypothesis?	Selecting out suspicious and non-naïve participants should increase discrimination against female candidates.	Selecting out suspicious and non-naïve participants should increase discrimination against female candidates.	No strong directional prediction.	*Individuals with greater degrees of experience participating in research studies or who are otherwise suspicious about the topic will favor female candidates.

Notes. The table entries represent the extreme case in which a given theory's empirical predictions hold to the exclusion of all other theories. An asterisk (*) indicates a key theoretical prediction. In all instances, predictions are regarding hiring decisions between male and female candidates for traditionally male jobs.

Table 6
Implications of the replication project's results for different theories of gender discrimination.

EMPIRICAL RESULT	MOTIVATED DISCRIMINATION PERSPECTIVE	COGNITIVE ASSIMILATION PERSPECTIVE	MOTIVATED LIBERALISM PERSPECTIVE	STUDY SAVVINESS PERSPECTIVE
Gender and hiring evaluations. Overall preference for female candidates among male evaluators. No consistent preference among female evaluators.	Highly contrary results to past research supporting the motivated reasoning account. These studies found discrimination against female candidates, especially among male evaluators (Uhlmann & Cohen, 2005, 2007).	Highly contrary results to the predictions of the cognitive assimilation to stereotypes perspective, which predicted discrimination against female candidates among both male and female evaluators.	The hypothesized reverse gender discrimination pattern is supported for male evaluators but not female evaluators.	The hypothesized reverse gender discrimination pattern is supported for male evaluators but not female evaluators.
Process measures. No overall effects of candidate gender on perceived characteristics or hiring criteria. However, participants high in self-perceived objectivity constructed hiring criteria that favored female candidates.	No support for the key prediction of motivated construction of hiring criteria that favor men over women (Uhlmann & Cohen, 2005). At least among evaluators high in self-perceived objectivity, the pattern of results is directly contrary, with hiring criteria favoring female candidates.	No support for key prediction of stereotype-based perceptions of candidates.	Support for the prediction that evaluators high in self-perceived objectivity construct hiring criteria that favor female over male candidates.	Not supported, since the theory expected no effect of self-perceived objectivity on judgments.
Affirmation-threat intervention. When threatened, male evaluators rate female candidates less positively.	Supported, in that male evaluators became less positive towards female candidates under threat. However, male evaluators did not outright favor male over female candidates under threat.	Not supported since the theory predicts no effect of affirmation-threat. Identity threat and affirmation effects are "hot" motivated processes, not "cold" cognitive ones.	Directly contradicted, since the theory predicted the opposite result. Specifically, it was expected that an affirmation would reduce pro-female judgments by deactivating ideological motives.	Not supported, since the theory predicted no affirmation-threat effect.
Objectivity mindset intervention. Making a sense of personal objectivity salient led to more favorable ratings of female candidates by male evaluators.	Opposite result to the prediction of this theory, i.e., that an objectivity mindset would exacerbate discrimination against women (Uhlmann & Cohen, 2005, 2007).	Not supported, since this theory predicted no effects of objectivity mindset.	Supported for male evaluators, who became more favorable towards female candidates when led to feel objective.	Not supported, since this theory predicted no effects of objectivity mindset.
Individual differences. Rejection of sexist beliefs, having participated in a similar study before, and self-perceived objectivity predict favoritism towards female candidates.	Not supported. The theory predicted that self-perceived objectivity would correlate with constructing hiring criteria in favor of male candidates (Uhlmann & Cohen, 2005); objectivity beliefs instead predicted hiring criteria that favored female candidates. This theory also failed to anticipate the other individual differences moderators that emerged.	Made no strong prediction regarding individual differences.	Supported, in that rejection of sexist beliefs predicted preferring female over male candidates in hiring decisions, and seeing oneself as objective predicted hiring criteria that favored women.	Supported, in that participating in a similar study previously predicted favoritism towards female candidates.
Summary assessment of each theory. What does the overall pattern of empirical results mean for this theory?	The original effects predicted by the motivated reasoning perspective (Uhlmann & Cohen, 2005, 2007) all failed to replicate, with the observed patterns in the opposite direction in several cases. The effects of the affirmation-threat intervention on male evaluators are broadly consistent with the motivated discrimination account. Overall, minimal support for this account of hiring evaluations of female and male candidates.	None of the predicted effects from this perspective were obtained. The cognitive schema account of gender discrimination receives no empirical support from the creative destruction initiative.	Several key predictions of the motivated liberalism account were supported. Some evaluators (men) exhibited favoritism towards female candidates, and rejection of sexist beliefs predicted such favoritism.	A number of key predictions of the savviness perspective were supported. Some evaluators (men) favored female over male candidates, and previous experience with research studies predicted pro-female hiring decisions. Although not a predicted pattern, male evaluators exhibiting more favoritism towards female candidates than female evaluators is more consistent with impression management concerns than with ideological motives.
Overall assessment. What broad conclusions can be drawn from the replication initiative?	The overall project results are most strongly supportive of the study savviness account, followed by motivated liberalism. Some novel evidence for motivated discrimination against women was observed in the effects of the threat manipulation on men's evaluations of female candidates.			

evaluators would prefer male job candidates (forecasted $d = 0.357$ for male evaluators; forecasted $d = 0.110$ for female evaluators, mean of the differences = 0.248, $p < 0.0001$). However, only the aggregate forecasts about selection decisions by female evaluators were in the same direction as the realized results (realized $d = -0.128$ for male evaluators; realized $d = 0.018$ for female evaluators). As a consequence, forecasters were less accurate at anticipating gender discrimination by male evaluators relative to female evaluators ($p < 0.0001$). A non-preregistered follow-up analysis revealed that 184 of 194 forecasters predicted that male evaluators would discriminate against female job candidates, directionally contrary to the replication results reported earlier (mean of the differences = 0.485, $p < 0.001$). Thus, although the expected positive association between forecasts and outcomes emerged for the moderator effects, for some simple effects the association is in the wrong direction (negative) and significant. Among forecasters, individual differences in beliefs about gender did not moderate accuracy (see Supplement 9). Further research should continue to examine whether scientists can predict the results of complex experiments addressing socially sensitive topics, and what factors might facilitate (or impede) their accuracy.

9. When the creative destruction approach will be most (and least) useful

The creative destruction approach to replication seeks to not just support or cast doubt on the original finding (Dreber et al., 2015; Open Science Collaboration, 2015), but also to potentially supersede the previous theoretical account with positive evidence for a revised and improved theory (Tierney et al., 2019). Consistent with the results of other replication initiatives (e.g., Klein et al., 2014, 2018; Open Science Collaboration, 2015) our recent efforts to repeat the methodology of previous experimental studies in new samples failed to support the original theoretical predictions regarding Implicit Puritanism in American work values, motivated processing of scientific evidence in order to reach desired conclusions, and motivated discrimination against women. Increasing the information gain from these new investigations, the novel conditions, measures, and populations allowed not only for supporting or not supporting the original theorizing, but also generating positive evidence for alternative theoretical accounts. Specifically, this process of creative destruction supports the general moralization of work (especially in self-expression-oriented cultures), assimilation to cognitive priors regarding child care practices, and study savviness and motivated liberalism accounts of male evaluators' decisions in hiring simulations. Testing multiple theories against one another with pre-registered analyses and both conceptual and direct replications facilitates strong inferences (Mayo, 2018; Platt, 1964).

Although the present empirical applications are in organizational research and psychology, we see the creative destruction method as generally applicable across academic fields. We hope the three empirical examples discussed here illustrate the novelty of our approach (see Fig. 1). Past replication efforts have typically compared the original theory to the null (e.g., Klein et al., 2014; Open Science Collaboration, 2015), rather than adding new measures, conditions, and populations to test multiple theories against each other. Further, past theory pruning efforts in the management literature have generally not relied on direct replication, pre-registration of analyses, and complete data transparency.

As with all research methodologies, the creative destruction paradigm has important limitations, and is no "silver bullet" for generating scientific knowledge. Further, theory pruning is not necessary or desirable in all circumstances. Accordingly, certain limits may inform when creative destruction may be most (versus least) appropriate and useful as a tool for theoretical competition. First, while creative destruction involves collecting data on "neutral ground" for all relevant theories, underlying differences in populations will always limit generalizability from any research sample (Hanel & Vione, 2016). Scholars must be aware of the very real influence of context in organizational research (Bamberger, 2008), and no single replication will be sufficient to cover all domains where a theory may be relevant. That said, research within the creative destruction paradigm may develop a set of theoretical predictions and methods that can be applied across different topics and populations.

The creative destruction approach requires that theories be well positioned for theoretical competition within a given phenomenological space. Theories to be tested against one another should be carefully examined to verify that they specify equivalent terms and conditions (i.e., sufficiently similar IVs and DVs), describe a shared context and population, and describe similar sets of unfolding events (Leavitt et al., 2010; Mitchell & James, 2001). Moreover, competing theories should be considered for their methodological compatibility.

The creative destruction approach is most useful as follow-up research to an initial set of published findings—in other words, in the context of replicating or re-examining established research. This approach is meant to create a series of severe tests (Mayo, 2018) for competing theories. Severe tests often require a great deal of resources, both in terms of study design and participant recruitment. As such, the creative destruction approach will be most effective when there is a set of competing theories with each having an empirical basis of support. Such a basis will allow researchers to effectively design tests of each theory and will hopefully limit wasting resources on theories that were, *a priori*, unlikely to find support.

At the same time, the creative destruction approach is most useful when each competing theory predicts significant and, on some level, conflicting effects. Theories can vary in their number of predictions in a given testing content, but each theory should make at least one positive prediction (that is, predict the existence of a significant effect). Theories can certainly make predictions of some null effects. However, a theory that only makes null predictions may in some circumstances be unfairly advantaged in a replication context, such that underpowered or otherwise deficient studies (e.g., use of methods that do not generalize to the new sample population) will be more likely to support that theory. Overall, the creative destruction approach will provide the most diagnostic information when competing theories make clear, non-overlapping, and ideally directionally opposed predictions.

The creative destruction approach, then, is most effective within the context of well-developed theories. Whereas many theories within organizational sciences merely predict directional associations between pairs of variables (Vancouver, Wang, & Li, 2018), more precise theories are defined by their boundaries and limitations, including reducing the number of outcomes that would be considered consistent with that theory (Byrd, 2019; Edwards, 2010). Creative destruction, then, will be most useful when theories are already sufficiently bounded, such that the scope of their predictions can be reasonably captured within a short series of studies. Notably, mature areas of research inquiry, which are often those with the most well-developed theories, are also the most likely to suffer from theoretical proliferation. This makes them especially good candidates for strong inference comparisons (e.g., Thau & Mitchell, 2010). For highly advanced theories associated with large numbers of published empirical investigations, the creative destruction approach can be employed not only in novel data collections, but also in the context of meta-analytic tests for publication bias and evidentiary value in competing sets of findings (see Supplement 5). The ideal context, however, is likely to be Registered Reports, in which the methods, predictions, and analytic plan for a study are peer reviewed prior to data collection (He & Côté, 2019).

10. Conclusion

We propose that issues germane to the problem of theoretical proliferation are intimately coupled with practices which contribute to low replicability. That is, the combination of incentives for theoretical novelty, sub-optimal research practices and a lack of replication efforts have led to myriad (often contradictory) theories populating a given space. The need for solutions which simultaneously give us confidence in scientific findings while also circumscribing their theoretical limits is increasingly clear. As we have argued and demonstrated, the creative destruction approach allows for the application of strong-inference tests (theory pruning) leveraging best practices for open science. Creative destruction offers the strengths of both direct and conceptual replications, testing theories with multiple methods and measures, high statistical power, pre-registration of analysis plans, and novel samples for testing the key terms and propositions from multiple theories simultaneously. As Kuhn (1962) noted, faster-moving sciences are characterized by their tendency to create critical tests of their own proposed findings. By boldly testing our own theories using the best open science practices and subjecting them to creative destruction, management scholars may have the opportunity to not only increase confidence in our theories, but rapidly accelerate their development in the process.

Author contributions statement

The first four and last authors collaboratively ideated and drafted the paper and contributed equally. WT, JH, CE, & EU designed the gender and hiring decisions study and WT, JH, & CE carried out the statistical analyses. CE, WT, & EU designed the re-analysis of the Ebersole (2019) dataset on working parents and child care choices, and CE & WT carried out the analyses. DV, EC, MG, AD, MJ, & TP designed, ran, analyzed, and wrote the report of the forecasting study. Members of the "Hiring Decisions Forecasting Collaboration" lent their expertise as forecasters, and are listed with full names and affiliation in Appendix A. All authors collaboratively edited the paper.

Funding acknowledgments

Eric Luis Uhlmann is grateful for an R&D grant from INSEAD in support of this research. Anna Dreber is grateful for generous financial support from the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser), the Knut and Alice Wallenberg Foundation and the Marcus and Marianne Wallenberg Foundation (Anna Dreber is a Wallenberg Scholar), and the Swedish Foundation for Humanities and Social Sciences.

CRediT authorship contribution statement

Warren Tierney: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Jay Hardy III:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Charles R. Ebersole:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Keith Leavitt:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Domenico Viganola:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Elena Giulia Clemente:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Michael Gordon:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Anna Dreber:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Magnus Johannesson:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Thomas Pfeiffer:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Eric Luis Uhlmann:** Conceptualization, Writing - original draft, Writing - review & editing. **Hiring Decisions Forecasting Collaboration:** Forecasting results of formal analysis, Writing - review & editing.

Appendix A. Names and affiliations for the Hiring Decisions Forecasting Collaboration

The following collaborators lent their time and expertise as forecasters:

Ajay T. Abraham, Seattle University
Matus Adamkovic, Institute of Social Sciences CSPS, Slovak Academy of Sciences & Institute of Psychology, Faculty of Arts, University of Presov
Jais Adam-Troian, American University of Sharjah
Rahul Anand, Aarhus BSS
Kelly J. Arbeau, Trinity Western University
Eli C. Awtrey, University of Cincinnati
Ofer H. Azar, Ben-Gurion University of the Negev
Štěpán Bahník, Prague College of Psychosocial Studies
Gabriel Baník, University of Presov
Ana Barbosa Mendes, ITEC, Faculty of Psychology and Educational Sciences, KU Leuven
Michael M. Barger, University of Georgia
Ernest Baskin, Saint Joseph's University
Jozef Bavolar, Pavol Jozef Safarik University in Kosice
Ruud M.W.J. Berkers, Max Planck Research Group: Adaptive Memory, Max Planck Institute for Human Cognitive and Brain Sciences
Randy Besco, University of Toronto
Michał Białek, Institute of Psychology, University of Wrocław
Michael M. Bishop
Helena Bonache, Universidad de La Laguna
Sabah Boufkhed, King's College London
Mark J. Brandt, Department of Social Psychology, Tilburg University
Max E. Butterfield, Point Loma Nazarene University
Nick Byrd, Stevens Institute of Technology
Neil R. Caton, The University of Queensland
Michelle L. Ceynar, Pacific Lutheran University
Mike Corcoran, University of Missouri
Thomas H. Costello, Emory University
Leslie D. Cramblet Alvarez, University of Denver
Jamie Cummins, Ghent University
Oliver S. Curry, University of Oxford
David P. Daniels, National University of Singapore
Lea L. Daskalo, Ben-Gurion University of the Negev
Liora Daum-Avital, Ben-Gurion University of the Negev
Martin V. Day, Memorial University of Newfoundland
Matthew D. Deeg, University of Kansas
Tara C. Dennehy, University of British Columbia
Erik Dietl, Loughborough University
Eugen Dimant, University of Pennsylvania
Artur Domurat, Centre for Economic Psychology and Decision Sciences, Kozminski University
Christilene du Plessis, Singapore Management University
Dmitrii Dubrov, National Research University Higher School of Economics
Mahmoud M. Elsherif, University of Birmingham
Yuval Engel, University of Amsterdam
Martin R. Fellenz, Trinity College Dublin
Sarahanne M. Field, University of Groningen
Mustafa Firat, University of Alberta
Raquel M. K. Freitag, Federal University of Sergipe
Enav Friedmann, Ben-Gurion University of the Negev
Omid Ghasemi, Department of Cognitive Science, Macquarie University
Matthew H. Goldberg, Yale University
Amélie Gourdon-Kanhukamwe, Kingston University London
Lorenz Graf-Vlachy, ESCP Business School
Jennifer A. Griffith, University of New Hampshire
Dmitry Grigoryev, National Research University Higher School of Economics
Sebastian Hafenbrädl, IESE Business School
David Hagmann, Harvard Kennedy School
Andrew H. Hales, University of Virginia

Hyemin Han, University of Alabama

Jason L. Harman, Louisiana State University

Andree Hartanto, Singapore Management University

Benjamin C. Holding, Department of Clinical Neuroscience, Karolinska Institutet

Astrid Hopfensitz, Toulouse School of Economics

Joachim Hüffmeier, Institute of Psychology, TU Dortmund University

Jeffrey R. Huntsinger, Loyola University Chicago

Katarzyna Idzikowska, Kozminski University

Åse H. Innes-Ker, Lund University

Bastian Jaeger, Tilburg University

Kristin Jankowsky, University of Kassel

Shoshana N. Jarvis, Haas School of Business, University of California, Berkeley

Nilotpal Jha, Singapore Management University

David Jimenez-Gomez, Fundamentos de Análisis Económico (FAE), University of Alicante

Daniel Jolles, University of Essex

Bibiana Jozefiakova, Olomouc University Social Health Institute, Palacky University

Olomouc, Olomouc, Czech Republic

Pavol Kačmár, Department of Psychology, Faculty of Arts, Pavol Jozef Šafárik University in Košice

Mariska Kappmeier, University of Otago

Matthias Kasper, Tulane University & University of Vienna

Lucas Keller, Department of Psychology, University of Konstanz

Viktorija Knapic, University of Rijeka

Mikael Knutsson, Linköping University

Olga Kombeiz, Loughborough University

Marta Kowal, Institute of Psychology, University of Wroclaw

Goedele Krekels, IESEG

Tei Laine

Daniel Lakens, Eindhoven University of Technology

Bingjie Li, Warwick Business School

Ronda F. Lo, York University

Jonas Ludwig, University of Würzburg

James C. Marcus, Evidera

Melvin S. Marsh, Georgia Southern University

Mario Martinoli, DiECO, Università degli Studi dell'Insubria

Marcel Martončik, University of Presov, Faculty of Arts, Institute of Psychology

Allison Master, University of Washington & University of Houston

Theodore C. Masters-Waage, Singapore Management University

Lewend Mayiwar, Department of Leadership and Organizational Behavior, BI Norwegian Business School

Jens Mazei, TU Dortmund University

Randy J. McCarthy, Northern Illinois University

Gemma S. McCarthy, University of Limerick

Stephanie Mertens, Swiss Center for Affective Sciences, University of Geneva

Leticia Micheli, Maastricht University

Marta Miklikowska, Umeå University

Talya Miron-Shatz, Ono Academic College

Andres Montealegre, Cornell University

David Moreau, The University of Auckland

Carmen Moret-Tatay, Universidad Católica de Valencia San Vicente Mártir

Marcello Negrini, Maastricht University

Philip W. S. Newall, CQUniversity

Gustav Nilsonne, Karolinska Institutet, Department of Clinical Neuroscience & Stockholm University, Department of Psychology

Paweł Niszczota, Poznań University of Economics and Business

Nurit Nobel, Stockholm School of Economics

Aoife O'Mahony, School of Psychology, Cardiff University

Mehmet A. Orhan, PSB Paris School of Business

Deirdre O'Shea, University of Limerick

Flora E. Oswald, The Pennsylvania State University

Miriam Panning

Peter C. Pantelis

Mariola Paruzel-Czachura, Institute of Psychology, University of Silesia in Katowice

Mogens Jin Pedersen, University of Copenhagen

Gordon Pennycook, University of Regina

Ori Plonsky, Technion - Israel Institute of Technology

Vince Polito, Macquarie University

Paul C. Price, California State University, Fresno

Maximilian A. Primbs, Radboud University

John Protzko, Department of Psychological & Brain Sciences, University of California, Santa Barbara

Michael Quayle, University of Limerick

Rima-Maria Rahal, Tilburg University

Md. Shahinoor Rahman, University of Chittagong

Liz Redford, Healthy Minds Innovations, Inc.

Niv Reggev, Department of Psychology & Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev

Caleb J. Reynolds, Department of Psychology, Florida State University

Marta Roczniewska, SWPS University of Social Sciences and humanities & LIME Department, Karolinska Institutet

Ivan Ropovik, Faculty of Education, Institute for Research and Development of Education, Charles University & Faculty of Education, University of Presov

Robert M. Ross, Department of Philosophy, Macquarie University

Thomas J. Roulet, University of Cambridge

Andrea May Rowe

Silvia Saccardo, Carnegie Mellon University

Margaret Samahita, University College Dublin

Michael Schaerer, Singapore Management University

Joyce Elena Schleu, TU Dortmund University

Brendan A. Schuetze, The University of Texas at Austin

Ulrike Senftleben, Technische Universität Dresden

Raffaello Seri, DiECO, Università degli Studi dell'Insubria & CORG, University of Southern Denmark

Zeev Shtudiner, Ariel University, Israel

Jack Shuai, University of Toledo

Ray Sin, Early Warning Services

Varsha Singh, Humanities and Social Sciences, Indian Institute of Technology Delhi

Aneeha Singh, International Research & Exchanges Board (IREX)

Tatiana Sokolova, Tilburg University

Victoria Song, Fordham University

Tom Stafford, University of Sheffield

Natalia Stanulewicz, De Montfort University

Samantha M. Stevens, The Pennsylvania State University

Eirik Strømland, University of Bergen

Samantha Stronge, University of Auckland

Kevin P. Sweeney, Western Kentucky University

David Tannenbaum, University of Utah

Stephanie J. Tepper, Cornell University

Kian Siong Tey, INSEAD

Hsuchi Ting, Goldman Sachs

Ian W. Tingen, Tingen Industries

Ana Todorovic, Department of Experimental Psychology, University of Oxford

Hannah M.Y. Tse, University of Hong Kong

Joshua M. Tybur, Vrije Universiteit Amsterdam

Gerald H. Vineyard, Independent Researcher

Alisa Voslinsky, Department of Industrial Engineering and Management, Sami Shamoon Academic College of Engineering

Marek A. Vranka, Charles University

Jonathan Wai, University of Arkansas

Alexander C. Walker, University of Waterloo
Laura E. Wallace, Ohio State University
Tianlin Wang, University at Albany-SUNY
Johanna M. Werz, RWTH Aachen University
Jan K. Woike, University of Plymouth, UK, Max Planck Institute for Human Development
Conny E. Wollbrant, University of Stirling
Joshua D. Wright, Simon Fraser University
Sherry J. Wu, University of California-Los Angeles
Qinyu Xiao, University of Hong Kong
Paolo Barretto Yaranon, University of Limerick
Siu Kit Yeung, The University of Hong Kong
Sangsuk Yoon, University of Dayton
Karen Yu, Sewanee: The University of the South
Meltem Yucel, University of Virginia
Ignazio Ziano, Grenoble Ecole de Management, F-38000 France
Ro'i Zultan, Ben-Gurion University of the Negev
Camilla S. Øverup, Department of Public Health, University of Copenhagen, Denmark

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2020.07.002.

References

Aguinis, H., & Solarino, A. M. (in press). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*.

Albertini, D. F. (2017). On strong inferences and irreproducibility in reproductive medicine. *Journal of Assisted Reproduction and Genetics*, *34*, 695–696.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of organizational research methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods, 12*(1), 9–34.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., et al. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*(5), 556–578.

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS One, 12*(3), e0172792.

Argyle, M. (1994). *The psychology of social class.* New York: Psychology Press.

Armor, D. A., & Sackett, A. M. (2006). Accuracy, error, and bias in predictions for real versus hypothetical events. *Journal of Personality and Social Psychology, 91*(4), 583–600.

Baker, W. (2005). *America's crisis of values.* Princeton, NJ: Princeton University Press.

Bamberger, P. A. (2019). On the replicability of abductive research in management and organizations: Internal replication and its alternatives. *Academy of Management Discoveries, 5*(2), 103–108.

Bamberger, P. (2008). Beyond contextualization: Using context theories to narrow the micro-macro gap in management research. *Academy of Management Journal, 51*(5), 839–846.

Bar-Haim, E., Chauvel, L., Gornick, J., & Hartung, A. (2018). *The persistence of the gender earnings gap: Cohort trends and the role of education in twelve countries* (LISWorkingPaper737). Esch-Belval: Cross National Data Center in Luxembourg.

Barger, P. B., & Grandey, A. A. (2006). Service with a smile and encounter satisfaction: Emotional contagion and appraisal mechanisms. *Academy of Management Journal, 49*(6), 1229–1238.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., et al. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences, 115*(11), 2607–2612.

Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? *Perspectives on Psychological Science, 14*(2), 292–303.

Barrick, M. R., & Zimmerman, R. D. (2005). Reducing voluntary, avoidable turnover through selection. *Journal of Applied Psychology, 90*(1), 159–166.

Bastardi, A., Uhlmann, E. L., & Ross, L. (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological Science, 22*(6), 731–732.

Bedeian, A. G., Taylor, S. G., & Miller, A. N. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education, 9*(4), 715–725.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization, 15*(3), 423–436.

Bond, M. H., & Smith, P. B. (1996). Cross-cultural social and organizational psychology. *Annual Review of Psychology, 47*(1), 205–235.

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*(3), 709–750.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., … van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Brainerd, C. J., & Reyna, V. F. (2018). Replication, registration, and scientific creativity. *Perspectives on Psychological Science, 13*(4), 428–432.

Byington, E. K., & Felps, W. (2017). Solutions to the credibility crisis in management science. *Academy of Management Learning & Education, 16*(1), 142–162.

Byrd, N. (2019). What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese,* 1–29.

Cairo, A.H., Green, J.D., Forsyth, D.R., Behler, A.M.C., & Raldiris, T.L. (in press). Gray (Literature) Matters: Evidence of Selective Hypothesis Reporting in Social Psychological Research. *Personality and Social Psychology Bulletin*.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*, 637–644.

Carpenter, J., Verhoogen, E., & Burks, S. (2005). The effects of stakes in distribution experiments. *Economics Letters, 86*(3), 393–398.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods, 7*(2), 151–167.

Chang, A. C., & Li, P. (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review, 107*(5), 60–64.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93–99.

Davis, G. F. (2006). Mechanisms and the theory of organizations. *Journal of Management Inquiry, 15*(2), 114–118.

Dee, T. S. (2015). Social identity and achievement gaps: Evidence from an affirmation intervention. *Intervention, Evaluation, and Policy Studies, 8*(2), 149–168.

Devine, P. G., Hirt, E. R., & Gehrke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology, 58*(6), 952–963.

DellaVigna, S., & Pope, D. G. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy, 126*(6), 2410–2456.

Desai, S. D., Chugh, D., & Brief, A. P. (2014). The implications of marriage structure for men's workplace attitudes, beliefs, and behaviors toward women. *Administrative Science Quarterly, 59*(2), 330–365.

de Tocqueville, A. (1840/1990). *Democracy in America*. New York: Vintage Books.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science, 14*(2), 273–291.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y. Nosek, et al. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*(50), 15343–15347.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*, 348–351.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist, 75*(3), 301–315.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*(621), 1–11.

Ebersole, C. (2019). *Pre-commitment and updating beliefs*. Unpublished doctoral dissertation.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82.

Edwards, J. R. (2010). Reconsidering theoretical progress in organizational and management research. *Organizational Research Methods, 13*(4), 615–619.

Evangelou, E., Siontis, K. C., Pfeiffer, T., & Ioannidis, J. P. (2012). Perceived information gain from randomized trials correlates with publication in high–impact factor journals. *Journal of Clinical Epidemiology, 65*(12), 1274–1281.

Fanelli, D. (2010). Positive results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), e10068.

Festinger, L. (1957). *A theory of cognitive dissonance.* Evanston, IL: Row, Peterson.

Fisher, D. H. (1989). *Albion's seed: Four British folkways in America.* New York, NY: Oxford University Press.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., et al. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology, 75*, 102–117.

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS ONE, 13*(7), e0200303.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*(1), 75–90.

Garber, M. (2017). All the angry ladies. *The Atlantic*. Available at: https://www.theatlantic.com/entertainment/archive/2017/11/all-the-angryladies/545042/.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management, 41*(2), 632–643.

Greenberg, J., & Jonas, E. (2003). Psychological motives and political orientation—The left, the right, and the rigid: Comment on Jost et al. (2003). *Psychological Bulletin*, *129*(3), 376–382.

Goff, S. J., Mount, M. K., & Jamison, R. L. (1990). Employer supported child care, work/

family conflict, and absenteeism: A field study. *Personnel Psychology, 43*(4), 793–809.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellot, D. S. (2002). A unified theory of implicit attitudes, beliefs, self-esteem and self-concept. *Psychological Review, 109*, 3–25.

Grossmann, I., & Varnum, M. E. W. (2011). Social class, culture, and cognition. *Social Psychological and Personality Science, 2*(1), 81–89.

Hanselman, P., Rozek, C.S., Grigg, J., & Borman, G.D. (in press). New evidence on self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *Journal of Educational Psychology*.

Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal, 50*(6), 1346–1352.

Hanel, P. H., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PloS One, 11*(12), e0168354.

Harrington, J. R., & Gelfand, M. J. (2014). Tightness–looseness across the 50 United States. *Proceedings of the National Academy of Sciences, 111*(22), 7990–7995.

Harvie, K., Marshall-Mcaskey, J., & Johnston, L. (1998). Gender-based biases in occupational hiring decisions. *Journal of Applied Social Psychology, 28*(18), 1698–1711.

He, J. C., & Côté, S. (2019). Self-insight into emotional and cognitive abilities is not related to higher adjustment. *Nature Human Behavior, 3*, 867–884.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.* London: Sage Publications.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies.* Princeton, NJ: Princeton University Press.

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence.* Cambridge, MA: Cambridge University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*, 473–476.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review, 98*(1), 54–73.

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology, 55*(6), 893–905.

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15*(4), 624–662.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Kirkman, M. S., & Oswald, D. L. (2019). Is it just me, or was that sexist? The role of sexism type and perpetrator race in identifying sexism. *The Journal of Social Psychology, 160*(2), 1–12.

Kitayama, S., Ishii, K., Imada, T., Takemura, K., & Ramaswamy, J. (2006). Voluntary settlement and the spirit of independence: Evidence from Japan's "northern frontier". *Journal of Personality and Social Psychology, 91*(3), 369–384.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr, Alper, S., et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490.

Kluger, A. N., & Tikochinsky, J. (2001). The error of accepting the "'theoretical'" null hypothesis: The rise, fall, and resurrection of commonsense hypotheses in psychology. *Psychological Bulletin, 127*(3), 408–423.

Johnson, C.A., & Hawbaker, KT. (2018). #MeToo: A timeline of events. *Chicago Tribune*. Retrieved at: https://www.chicagotribune.com/lifestyles/ct-me-too-timeline-20171208-htmlstory.html.

Köhler, T., & Cortina, J. M. (in press). Play it again, Sam! An analysis of constructive replication in the organizational sciences. *Journal of Management*.

Koenig, H. G., & Büssing, A. (2010). The Duke University Religion Index (DUREL): A five-item measure for use in epidemological studies. *Religions, 1*(1), 78–85.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107–118.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27*(3), 313–327.

Kuhn, T. S. (1962). *The structure of scientific revolutions* (1st ed.). Chicago, IL: University of Chicago Press.

Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic Bulletin and Review, 23*(5), 1631–1638.

Kunst, J. R., Bailey, A., Prendergast, C., & Gundersen, A. (2019). Sexism, rape myths and feminist identification explain gender differences in attitudes toward the #metoo social media campaign in two countries. *Media Psychology, 22*(5), 818–843.

Kvarven, A., Strømland, E., & Johannesson, M. (in press). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos, & A. Musgrave (Eds.). *Criticism and the growth of knowledge* (pp. 91–195). London: Cambridge University Press.

Landes, D. S. (1998). *The wealth and poverty of nations: Why some are so rich and some so poor.* New York, NY: W.W. Norton & Co.

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., et al. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin, 146*(5), 451–479.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists – Application to the Erez-Latham dispute regarding participation in goal setting. *Journal of Applied Psychology, 73*(4), 753–772.

Leavitt, K., Mitchell, T., & Peterson, J. (2010). Theory pruning: Strategies for reducing our dense theoretical landscape. *Organizational Research Methods, 13*(4), 644–667.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives, 21*(2), 153–174.

Lipset, S. M. (1996). *American exceptionalism: A double edged sword.* New York, NY: W.W. Norton & Co.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy, 114*(1), 1–37.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109.

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology, 69*, 178–183.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3), 151–159.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 7*, 161–175.

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. A. (2019). *Questionable and open research practices in education research*. Unpublished manuscript. Available at: https://edarxiv.org/f7srb/.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge, MA: Cambridge University Press.

McCormick-Huhn, K., & Shields, S. A. (2019). *Can angry Black and White women get ahead in the era of #MeToo? Social dynamics in emotion appropriateness*. Unpublished manuscript.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Lessons from the JMCB archive. *Journal of Money, Credit and Banking, 38*(4), 1093–1107.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834.

Mischel, W. (2008). The toothbrush problem. *APS Observer*. Retrieved September 7, 2019 at https://www.psychologicalscience.org/observer/the-toothbrush-problem.

Mitchell, T. R., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review, 26*(4), 530–547.

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology, 34*(1), 1–17.

Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the South.* Boulder, CO: Westview Press.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631.

O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*(2), 376–399.

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.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Parboteeah, K. P., & Cullen, J. B. (2003). Social institutions and work centrality: Explorations beyond national culture. *Organization Science, 14*(2), 137–148.

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences, 116*(24), 11693–11698.

Perry, E. L., Davis-Blake, A., & Kulik, C. T. (1994). Explaining gender-based selection decisions: A synthesis of contextual and cognitive approaches. *Academy of Management Review, 19*(4), 786–820.

Petty, R. E., & Cacioppo, J. T. (2016). Methodological choices have predictable consequences in replicating studies on motivation to think: Commentary on Ebersole et al. (2016). *Journal of Experimental Social Psychology*, *67*, 86–87.

Peteraf, M., Di Stefano, G., & Verona, G. (2013). The elephant in the room of dynamic capabilities: Bringing two diverging conversations together. *Strategic Management Journal, 34*(12), 1389–1410.

Pitz, G. F. (1969). An inertia effect (resistance to change) in the revision of opinion. *Canadian Journal of Psychology, 23*(1), 24–33.

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

Poehlman, T. A. (2007). *Ideological inheritance: Implicit Puritanism in American moral cognition.* Doctoral dissertation, Yale University.

Popper, K. (1959/2002). *The logic of scientific discovery*. London and New York: Routledge.

Popper, K. R. (1963). *Conjectures and Refutations: The growth of scientific knowledge* (5th ed.). London and New York: Routledge.

Porter, L. W. (1996). Forty years of organization studies: Reflections from a micro perspective. *Administrative Science Quarterly, 41*(2), 262–269.

Pratt, M. G., Kaplan, S., & Whittington, R. (2019). Editorial essay: The tumult over transparency: Decoupling transparency from replication in establishing trustworthy qualitative research. *Administrative Science Quarterly, 41*, 262–269.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews. Drug Discovery, 10*, 712.

Protzko, J., & Aronson, J. (2016). Context moderates affirmation effects on the ethnic achievement gap. *Social Psychological & Personality Science, 7*(6), 500–507.

Ramagopalan, S., Skingsley, A. P., Handunnetthi, L., Klingel, M., Magnus, D., Pakpoor, J., et al. (2014). Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: A cross-sectional study. *F1000Research, 3*, 77.

Reynolds, S. J., Dang, C. T., Yam, K. C., & Leavitt, K. (2014). The role of moral knowledge in everyday immorality: What does it matter if I know what is right? *Organizational Behavior and Human Decision Processes, 123*(2), 124–137.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perserverance in self perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*(5), 880–892.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*(2), 347–367.

Schlaegel, C., & Koenig, M. (2014). Determinants of entrepreneurial intent: A meta-analytic test and integration of competing models. *Entrepreneurship Theory and Practice, 38*(2), 291–332.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100.

Schnall, S. (2014). *Social media and the crowd-sourcing of social psychology*. Retrieved at: https://www.psychol.cam.ac.uk/cece/blog.

Schumpeter, J. A. (1942/1994). *Capitalism, socialism and democracy*. London: Routledge. pp. 82–83.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Uhlmann, E. L., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55–67.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In M. P. Zanna (Vol. Ed.), *Advances in Experimental Social Psychology: Vol. 38*, (pp. 183–242). San Diego, CA: Academic Press.

Schwartz, S. H. (1999). A theory of cultural values and some implications for work. *Applied Psychology, 48*(1), 23–47.

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Social Psychology, 45*(4), 305–306.

Sherman, D. K., & Cohen, G. L. (2010). Self-affirmation theory. In R. L. Jackson (Ed.). *Encyclopedia of identity* (pp. 669–672). Thousand Oakes, CA: Sage Publications.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80.

Snibbe, A. C., & Markus, H. R. (2005). You can't always get what you want: Social class, agency, and choice. *Journal of Personality and Social Psychology, 88*(4), 703–720.

Soklaridis, S., Zahn, C., Kuper, A., Gillis, D., Taylor, V. H., & Whitehead, C. (2018). Men's fear of mentoring in the #MeToo era – What's at stake for academic medicine? *New England Journal of Medicine, 379*(23), 2270–2274.

Stewart, W., & Barling, J. (1996). Fathers' work experiences effect children's behaviors via job-related affect and parenting behaviors. *Journal of Organizational Behavior, 17*(3), 221–232.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Vol. Ed.), *Advances in Experimental Social Psychology: Vol. 21*, (pp. 261–302). New York: Academic Press.

Stephens, N. M., Fryberg, S. A., & Markus, H. R. (2011). When choice does not equal freedom: A sociocultural analysis of agency in working-class American contexts. *Social Psychological and Personality Science, 2*(1), 33–41.

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science, 11*(6), 929–930.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59–71.

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., et al. (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science, 344*, 603–608.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Levi, A. (1982). Attribution bias: On the inconclusiveness of the cognition-motivation debate. *Journal of Experimental Social Psychology, 18*, 68–88.

Tetlock, P. E., & Manstead, A. S. R. (1985). Impression management versus intrapsychic explanations in social psychology: A useful dichotomy? *Psychological Review, 92*(1), 59–77.

Tierney, W., Hardy, J. H., III., Ebersole, C., Viganola, D., Clemente, E., Gordon, M., et al. (2020). *A creative destruction approach to replication: Implicit work and sex morality across cultures*. Registered Report proposal accepted in principle at the Journal of Experimental Social Psychology.

Thau, S., & Mitchell, M. S. (2010). Self-gain or self-regulation impairment: Competitive tests of the relationship between abuse and deviance through distributive justice perceptions. *Journal of Applied Psychology, 95*(6), 1009–1031.

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology, 43*(1), 22–34.

Tsang, E. W., & Kwan, K. M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review, 24*(4), 759–780.

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

Uhlmann, E. L., & Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes, 104*(2), 207–223.

Uhlmann, E. L., & Nosek, B. A. (2012). My culture made me do it: Lay theories of responsibility for automatic prejudice. *Social Psychology, 43*(2), 108–113.

Uhlmann, E. L., Poehlman, T. A., & Bargh, J. A. (2009). American moral exceptionalism. In J. T. Jost, A. C. Kay, & H. Thorisdottir (Eds.). *Social and psychological bases of ideology and system justification* (pp. 27–52). New York, NY: Oxford University Press.

Uhlmann, E. L., Poehlman, T. A., Tannenbaum, D., & Bargh, J. A. (2011). Implicit Puritanism in American moral cognition. *Journal of Experimental Social Psychology, 47*, 312–320.

Van de Ven, A. H., & Johnson, P. E. (2006). Knowledge for theory and practice. *Academy of Management Review, 31*(4), 802–821.

Vancouver, J. B., Wang, M., & Li, X. (2018). Translating informal theories into formal theories: The case of the dynamic computational model of the integrated model of work motivation. *Organizational Research Methods, 23*(2), 238–274.

Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., et al., (2019). *Preregistration of secondary data analysis: A template and tutorial*. Unpublished manuscript.

Vandello, J. A., & Cohen, D. (1999). Patterns of individualism and collectivism across the United States. *Journal of Personality and Social Psychology, 77*(2), 279–292.

Vandenberg, R. J., & Grelle, D. M. (2008). Alternative model specifications in structural equation modeling: Fact, fictions and truth. In C. E. Lance, & R. J. Vandenberg (Eds.). *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 165–192). New York: Taylor & Francis Group.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638.

Weber, M. (1904/1958). *The protestant ethic and the spirit of capitalism*. New York, NY: Charles Scribner's Sons.

Williams, J. C. (2012). The class culture gap. In S. T. Fiske, & H. R. Markus (Eds.). *Facing social class: How societal rank influences interaction* (pp. 39–57). New York: Russell Sage Foundation.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral & Brain Sciences. 41*, e120.

Supplements for “Creative Destruction in Science”

Table of Contents

Supplement 1: Pre-Registered Plan for "Wishful Predictions" Re-Analysis of Ebersole (2019)

We will apply the creative destruction approach to replication (Tierney et al., 2019) to a re-analysis of the data from Study 6 of Ebersole (2019). This large-sample experiment with a lay adult sample (N = 1,514; Mage= 51.27, SD = 11.66; 65.3% female) found that pre-commitment to criteria reduced biased assimilation to prior beliefs, relying in part on materials from Bastardi, Uhlmann, and Ross (2011). We will repeat some of those analyses here for completeness.

Our novel analyses will attempt to directly replicate the original Bastardi et al. (2011) "wishful thinking" effect that desired outcomes trump factual beliefs in driving the biased assimilation of scientific evidence. To do this, we will select intended parents who believe home care is better than day care for children, yet intend to use day care for their own kids. These "conflicted" individuals cognitively believe day care is inferior, but hope to find out day care is just as effective as home care. Since the theoretical goal of his work was to examine pre-commitment and biased assimilation to beliefs, not pit beliefs against desires, Ebersole (2019) did not carry out these replication analyses.

Expanding on Ebersole's (2019) analysis of belief confirmation, we will further examine whether commitment to criteria reduces the effects of desired outcomes on the processing of evidence. In other words, are "conflicted" participants less likely to dismiss studies finding day care is harmful when they have previously evaluated the studies' methods while blind to the results? If so, this would suggest an important boundary condition to the "wishful thinking" effect (Bastardi et al., 2011).

Finally, we will directly compare the reasoning processes of actual parents who have made real childcare decisions to intended parents who have not yet carried out such decisions. Theories of motivated reasoning predict that a personal stake in the issue will exacerbate biased rationalizations. In contrast, accuracy-based theories expect that personally important issues activate the goal to be correct and therefore reduce bias (see Table S1-1).

The study materials are provided at https://osf.io/n83ks/, and the pre-registered analysis plan and exclusion criteria from the first phase of analyses reported in Ebersole (2019) are available at: https://osf.io/bv6uy/. As in Ebersole (2019), only participants who pass both attention checks (att.Check and att.Check2) and indicate that they paid attention throughout the study and that we should therefore use their data (PersonCheck) will be used in these new analyses.

Table S1-1 below summarizes the predictions of the competing theoretical perspectives on working parents' reasoning about child care choices. Table S1-2 outlines the planned statistical analyses.

Table S1-1. Empirical predictions of different theoretical perspectives on working parents' reasoning about child care.

EFFECT	**MOTIVATED REASONING PERSPECTIVE**	**COGNITIVE SCHEMA-BASED PROCESSING PERSPECTIVE**	**ACCURACY-DRIVEN REASONING PERSPECTIVE**
Prior beliefs and the biased processing of evidence	Beliefs only appear to bias reasoning because they are aligned with desires; when misaligned, desires trump beliefs in driving reasoning	Desires only appear to bias reasoning because they are aligned with beliefs; when misaligned, beliefs trump desires in driving reasoning	Beliefs do not bias reasoning about scientific evidence
Prior desires and the biased processing of evidence	Desired conclusions bias reasoning about scientific evidence	Desired conclusions do not bias reasoning about scientific evidence	Desired conclusions do not bias reasoning about scientific evidence
Effects of pre-commitment to criteria	Commitment to criteria should constrain motivated reasoning, and reduce the effects of desired outcomes on the processing of scientific evidence.	Commitment to criteria should reduce ambiguity and constrain the application of cognitive schemas, and therefore reduce the extent to which prior beliefs drive the processing of scientific evidence	People do not generally use criteria in a biased manner, hence pre-commitment to criteria should not affect their judgments of scientific evidence.
Effects of being an actual parent vs. intended parent	Actual parents should exhibit more biased assimilation than would-be-parents, since the psychological need to rationalize actual (rather than intended) child care decisions is greater.	No predicted difference between intended parents and actual parents in biased assimilation, so long as they hold the same cognitive beliefs about child care.	If both are sufficiently accuracy motivated, neither actual nor intended parents will exhibit biased assimilation. If anything, actual parents should exhibit less biased reasoning about child care than intended parents. The stakes are higher for the former group, activating accuracy goals.

Notes. The table entries represent the extreme case in which a given theory's empirical predictions hold to the exclusion of all other theories.

Table S1-2. Planned "creative destruction" analyses testing competing theories of reasoning about evidence.

Notes. An asterisk "*" in the code indicates that the models will produce a main effect and interaction. Statistically significant (*p* < .05) interactions will be broken down by their constituent components (e.g., if variable A interacts with variable B, the main effect of variable B will be tested separately within each of the two conditions of variable A). Analyses (1) and (3) were previously pre-registered and reported by Ebersole (2019, Study 6), and are repeated here for completeness. As in Ebersole (2019), only participants who pass both attention checks (att.Check and att.Check2) and indicate that their data should be used (PersonCheck) will be included in the analyses.

RESEARCH QUESTION	**PARTICIPANTS SELECTED FOR ANALYSES**	**DESCRIPTION OF ANALYSES**	**DEPENDENT MEASURE**	**CODE**
1) Do participants exhibit biased assimilation to pre-existing beliefs?	Participants in the no-commitment condition only.	Conceptually, what is of interest is the relation between individual differences in pre-existing beliefs about home care vs. day care and evaluations of the studies and the post-measure of belief.	Study evaluation composite (ratings of the convincingness of the study and the quality of its method)	DV ~ Pre-Beliefs, data = NoCommitment
		To simplify the analyses, the measures will be scored such that higher scores are positive for home care. For study evaluations, this means that higher scores indicate more positive evaluations of the study that supported home care, regardless of which study that was.	Post-measure of beliefs about the relative efficacy of home care vs. day care	
2) How does parental status affect biased assimilation to pre-existing beliefs?	Participants in the no-commitment condition only.	Interaction between parental status (actual parent vs. intended parent vs. no intention to be a parent), and individual differences in pre-existing beliefs about home care vs. day care.	Study evaluation composite	DV ~ Pre-Beliefs * Parental_Status, data = NoCommitment
			Post-measure of beliefs about home care vs. day care	
3) Does pre-commitment to criteria reduce biased assimilation to pre-existing beliefs?	Participants in both the commitment and no-commitment conditions.	Interaction between 2 (commitment to criteria vs. no commitment) x individual differences in pre-existing beliefs about home care vs. day care. The relationship between pre-existing beliefs about the efficacy of home vs. day care and post-beliefs is then tested separately for the committed condition and non-committed condition.	Post-measure of beliefs about home care vs. day care	DV ~ Pre-Beliefs * Commitment_Condition, data = All

RESEARCH QUESTION	**PARTICIPANTS SELECTED FOR ANALYSES**	**DESCRIPTION OF ANALYSES**	**DEPENDENT MEASURE**	**CODE**
4) Does the pattern of results in Bastardi, Uhlmann, & Ross (2011) directly replicate, following the original approach as closely as possible?	Only participants in the non-committed condition are selected for these analyses. Further, only non-parents who intend to be parents and believe home care to be better for children than day care are selected. This sub-sample of participants are further sorted into two groups based on the alignment of their pre-existing beliefs and desires. “Conflicted” would-be-parents intend to use day care for their own children in the future. “Unconflicted” would-be parents intend to use home care.	Note – for this analysis and analysis 5, we will analyze the DVs in two ways. Main strategy: Rescored such that that higher scores mean more positive views of home care, to maintain comparability with the analyses outlined above. Of interest is the relationship between belief/desires group (conflicted would-be parent vs. unconflicted would-be parent) and the outcomes. Alternative strategy: Not rescored, to increase comparability with the original study (Bastardi et al., 2011). In these models, we will include study results as a predictor (Cummings study supports day care vs. Cummings study supports home care). Of interest here is the interaction between 2 (belief/desires group: conflicted would-be parent vs. unconflicted would-be parent) x 2 (study results: Cummings study supports day care vs. Cummings study supports home care). For both approaches, the effect of study results on study evaluations is then tested separately for conflicted would-be parents and unconflicted would-be-parents.	Study evaluation composite	DV ~ Conflicted_Status, data = NoCommitment, IntendedParents Alternative analysis DV ~ Conflicted_Status * Study_Results, data = NoCommitment, IntendedParents
			Post-measure of beliefs about home care vs. day care	

RESEARCH QUESTION	**PARTICIPANTS SELECTED FOR ANALYSES**	**DESCRIPTION OF ANALYSES**	**DEPENDENT MEASURE**	**CODE**
5) When they are in conflict with one another, do pre-existing beliefs or desires drive reasoning?	This analysis expands on #4 above by including both actual and intended parents. Only participants in the non-committed condition are selected for these analyses. Only participants who believe home care to be better for children than day care are selected. This sub-sample of participants are further sorted into two groups based on the alignment of their pre-existing beliefs and desires. "Conflicted" actual and intended parents have used or will use day care for their own children in the future. "Unconflicted" actual and intended parents have selected home care.	As in #4 above, we analyze the data using both the recoded and non-recoded variables. The two analytic strategies are the same as above in #4, only now the sample is expanded to include both actual and intended parents.	Study evaluation composite Post-measure of beliefs about home care vs. day care	DV ~ Conflicted_Status, data = NoCommitment, IntendedandActualParents Alternative analysis DV ~ Conflicted_Status * Study_Results, data = NoCommitment, IntendedandActualParents
6) Does parental status influence biased assimilation to desired outcomes?	Same as #5 above, but actual parents who have used day care vs. home care for their kids are added to the analysis.	Interaction between 2 (parental status: parent vs. intended parent) x 2 (beliefs/desires group: conflicted vs. unconflicted) The effect of study results on study evaluations is then tested separately for conflicted and unconflicted participants who are intended parents vs. actual parents.	Study evaluation composite Post-measure of beliefs about home care vs. day care	DV ~ Parent_Status * Conflicted_Status, data = NoCommitment, IntendedandActualParents

RESEARCH QUESTION	**PARTICIPANTS SELECTED FOR ANALYSES**	**DESCRIPTION OF ANALYSES**	**DEPENDENT MEASURE**	**CODE**
7) Does commitment to criteria reduce biased assimilation to desired outcomes?	Same as #6 above, but both participants in the committed and non-committed conditions are selected for these analyses.	Interaction between 2 (commitment to criteria vs. no commitment) x 2 (conflicted vs. unconflicted participant).	Post-measure of beliefs about home care vs. day care	DV ~ Commitment_Condition * Conflicted_Status, data = IntendedandActualParents

References for Supplement 1

Bastardi, A., Uhlmann, E.L., & Ross, L. (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological Science*, *22*, 731–732.

Ebersole, C. R. (2019, April 27). Pre-commitment and Updating Beliefs. https://doi.org/10.31234/osf.io/5vsq3

Tierney, W., Ebersole, C., Hardy, J., Chapman, H., Gantman, A., Vanaman, M., DeMarree, K., Wylie, J., Storbeck J., Andreychik, M.R., McPhetres, J., Vaughn, L.A., & Uhlmann, E. L. (2019). *A creative destruction approach to replication: Implicit work and sex morality across cultures.* Registered Report proposal accepted in principle at the Journal of Experimental Social Psychology, with data collection in progress.

Supplement 2: Pre-Registered Analysis Plan and Materials for Motivated Discrimination Study

Overview

We will apply the creative destruction approach to replication (Tierney et al., 2019) to earlier findings from our research group regarding the roles of psychological rationalizations and illusions of personal objectivity in discrimination against women. Specifically, we will add new conditions, measures, and subject populations to facilitate pitting competing theories of group-based discrimination against one another (Brainerd & Reyna, 2018; Leavitt, Mitchell, & Peterson, 2010).

The previously published studies in question find that decisions makers who flexibly change their hiring criteria to rationalize selecting male candidates believe themselves to be less biased, when in fact they are more biased (Uhlmann & Cohen 2005). Providing evidence of a causal relationship, Uhlmann and Cohen (2007) show that experimentally inducing a sense of objectivity leads decision makers to use temporarily accessible (i.e., primed) gender stereotypes in their judgments, and to rely more on sexist beliefs. Our theoretical explanation in the original research was that seeing oneself as rational and objective licenses individuals to act on biased cognitions and beliefs. At the same time, rationalizing judgments likely assists in maintaining an illusion of personal objectivity.

In this first phase of the initiative, we will report the results of a large-sample replication combining key materials from both Uhlmann and Cohen (2007, Study 3) and Uhlmann and Cohen (2005, Study 1), as well as further manipulations and measures. To maximize statistical power, we will collect thousands of participants online via a professional survey firm. In a later and phase, an accompanying crowd initiative with a separate pre-registration plan, we will conduct further data collections among college students and lay adults using partner laboratories.

Consistent with the creative destruction approach, we will include additional conditions and measures testing competing theories of the effects of candidate gender on hiring judgments. For example, as a further test of the idea that hiring criteria and a sense of personal objectivity are constructed and maintained in a motivated manner, we will include a manipulation of self-affirmation vs. self-threat (Steele, 1988). If the effects observed in Uhlmann and Cohen (2005, 2007) are “hot” motivated processes, they should be amplified under psychological threat and ameliorated when an unrelated but important identity has been affirmed.

On the other hand, discrimination against female candidates may be attributable to a cognitive assimilation effect based on cultural knowledge of gender stereotypes. If so, a candidate’s gender should affect social perceivers’ impressions of her or his characteristics (rather than leading to shifts in the hiring criteria used), affirmation-threat should be irrelevant, and illusions of personal objectivity should not moderate discriminatory judgments.

We will additionally test the competing theory that in contemporary times, ideological movements and social sensitivities may lead to hiring biases in favor of female candidates for traditionally male jobs. Thus, we will examine whether participants with high levels of exposure to feminist media messaging, or who strongly endorse the belief that gender limits women's workplace opportunities, tend to render pro-female decisions. To the extent that such reverse discrimination effects are based on motivated ideologies (Ditto et al., 2018; Greenberg, & Jonas, 2003), they may be associated with hiring criteria biased in favor of women and exacerbated by the threat manipulation.

Finally, a related but distinct hypothesis posits that the lay public are increasingly study-savvy. If so, individuals who have participated in more research studies, or are otherwise suspicious of the hypothesis, may overcompensate and favor women over men for stereotypically male jobs in order to avoid appearing sexist.

Note that the use of an online context of this first data collection, with some relatively naïve participants and others who have participated in many research surveys and studies, favors the study-savviness hypothesis. If online participants favor female over male candidates due to awareness of the hypothesis and/or prior experience taking part in experiments, further research with less savvy participants (e.g., college students and lay adults with little experience with research studies) is called for.

Prior research has reported priming and affirmation effects in online samples (e.g., Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009; Uhlmann, Poehlman, Tannenbaum, & Bargh, 2011; Uhlmann & Nosek, 2012) in addition to laboratory experiments. If these manipulations fail to produce the hypothesized effects in the online sample, it will be useful to follow-up with crowdsourced laboratory data collections, as already planned for the second phrase of this project.

Sample, Design, and Measures

Sample:

Through the online survey firm PureProfile, we will collect data with 3,000 U.S. based participants whom are 18 years of age or older. The final sample size for some statistical tests will likely be smaller than this, due to a subset of respondents skipping items (e.g., demographics such as self-reported gender).

The cover page will include the captcha item, "I am not a robot," to avoid contamination of the experiments by bots. Following best practices with online studies, we will also screen out participants with duplicate GPS coordinates.

Design:

The online study will employ a 2 (prime condition: gender stereotypes or neutral concepts) x 4 (mindset manipulation: affirmation essay, threat essay, objectivity questions, neutral questions) x 2 (applicant characteristics: streetwise vs. educated applicant) x 2 (candidate gender: female or male) x 2 (participant gender: female or male) between-subjects design.

Materials:

Manipulations will include:

- Applicant gender (via applicant name: Karen or Brian; Uhlmann & Cohen, 2005, 2007)
- Applicant characteristics (streetwise or educated; scenarios from Uhlmann & Cohen, 2005, Study 1)
- Affirmation vs. threat essay (online version used in Uhlmann & Nosek, 2012; adapted from earlier work on self-affirmations, see Steele, 1988)
- Objectivity questions vs. Neutral questions (from Uhlmann & Cohen, 2007, Study 3)
- Stereotype priming (gender stereotype vs. neutral concepts scrambled-sentences task; from Uhlmann & Cohen, 2007, Study 3; adapted from Srull & Wyer, 1979)

Dependent measures will include:

- Hiring evaluation composite (Uhlmann & Cohen, 2005, Studies 1-3)
- Perceived streetwise characteristics (Uhlmann & Cohen, 2005, Study 3)
- Perceived educated characteristics (Uhlmann & Cohen, 2005, Study 3)
- Rated importance of streetwise characteristics (Uhlmann & Cohen, 2005, Study 3)
- Rated importance of educated characteristics (Uhlmann & Cohen, 2005, Study 3)

Moderator measures will include:

- Sexist beliefs (Uhlmann & Cohen, 2007, Study 2)
- Exposure to feminist media messages
- Beliefs about gender in the workplace
- Number of studies previously completed (indicator of study-savviness)
- Having completed a similar study before (indicator of study-savviness)
- Having previously taken a course in Psychology (indicator of study-savviness)
- Suspicion the study is about gender (indicator of study-savviness). Participant is coded as “aware” the study was about gender if she/he 1) reports the belief the study was about gender in an open-ended probe, and 2) further indicates she became suspicious before or while evaluating the candidate.

The complete study materials are provided at the end of this pre-registered analysis plan.

Theoretical Predictions and Planned Analyses

Table S2-1 below summarizes the predictions of the competing theoretical perspectives on the role of gender in hiring decisions. Table S2-2 outlines the planned analyses for the online data collection. Table S2-3 outlines the data exclusions for our second wave of analyses of the online data. Finally, we describe our test-holdout sample approach for exploring the data from the online study while minimizing false positives.

Table S2-1. Theoretical predictions of different perspectives on gender and hiring decisions.

Notes. The table entries represent the extreme case in which a given theory's empirical predictions hold to the exclusion of all other theories. An asterisk (*) indicates a key theoretical prediction. In all instances, predictions are regarding hiring decisions between male and female candidates for traditionally male jobs.

RESEARCH QUESTION	**MOTIVATED DISCRIMINATION PERSPECTIVE**	**COGNITIVE ASSIMILATION PERSPECTIVE**	**MOTIVATED LIBERALISM PERSPECTIVE**	**STUDY-SAVVINESS PERSPECTIVE**
Do hiring decisions favor men or women?	*Hiring decisions favor men for stereotypically male jobs	*Hiring decisions favor men for stereotypically male jobs	*Hiring decisions favor female candidates	*Hiring decisions favor female candidates
Are perceived characteristics biased by candidate gender?	*No bias in impression formation when descriptions of candidates' characteristics are clear and unambiguous	*Impressions of male candidates' traits and characteristics should be more favorable than for identically described female candidates, due to assimilation to stereotypes	Either no difference, or more favorable impressions of female candidates' characteristics	*Yes, female candidates' characteristics are rated favorably relative to male candidates
Are hiring criteria constructed in a biased manner?	*Yes, hiring criteria are shifted in favor of male candidates	No, since stereotypes bias impressions of social targets, not judgmental standards	*Yes, hiring criteria are shifted in favor of female candidates	*Yes, hiring criteria are shifted in favor of female candidates
What are the effects of affirmation-threat on hiring judgments?	*Relative to a self-threat, a self-affirmation reduces the tendencies to construct hiring criteria that favor men, choose male candidates, and act on sexist beliefs and accessible stereotypes	*No effect of self-affirmation or threat, since hiring biases are cognitive not motivational in nature	Relative to a self-threat, a self-affirmation reduces ideologically based tendencies to construct hiring criteria that favor women, choose female candidates, and act based on feminist beliefs	No effect, since pro-female judgments are based on public impression management not intrapsychic processes
What are the effects of experimentally inducing a sense of objectivity?	*Making a sense of personal objectivity salient increases bias against female candidates and reliance on sexist beliefs and accessible stereotypes.	No causal effect of such self-views on judgments, since hiring biases are due to the operation of cognitive expectations about targets.	Making a sense of personal objectivity salient increases reliance on ideologies that promote positive judgments of female candidates.	No effect, since hiring decisions are for public consumption not about personal identity.

RESEARCH QUESTION	**MOTIVATED DISCRIMINATION PERSPECTIVE**	**COGNITIVE ASSIMILATION PERSPECTIVE**	**MOTIVATED LIBERALISM PERSPECTIVE**	**STUDY-SAVVINESS PERSPECTIVE**
What are the correlates of individual differences in self-perceived objectivity?	*Seeing oneself as objective is correlated with constructing hiring criteria biased against women	No relationship between such self-views and hiring judgments. Biases in hiring are due to the operation of cognitive expectations about targets.	A sense of personal objectivity correlates with increased reliance on ideologies that promote positive judgments of female candidates.	No effect, since hiring decisions are for public consumption and not about personal identity.
What are the effects of individual differences in feminist media exposure and beliefs about gender in the workplace?	Either no effect, or such beliefs partly compensate for motivated biases against female candidates.	Either no effect, or such beliefs partly compensate for cognitive biases against female candidates.	*Greater exposure to feminist social media and the belief that workplaces are gendered predicts pro-female judgments in selection contexts.	Either no effect, or exposure to feminist media increases the desire to avoid appearing sexist and therefore favor female candidates
What are the effects of prior experience participating in studies and suspicions about the hypothesis?	Selecting out suspicious and non-naïve participants should increase empirical support for the predicted biases against women (e.g., hiring criteria and hiring decisions).	Selecting out suspicious and non-naïve participants should increase empirical support for the predicted biases against women (e.g., trait impressions and hiring decisions).	No strong directional prediction	*Individuals with greater degrees of experience participating in research studies or who are otherwise suspicious about the topic will favor female candidates.

Table S2-2. Planned analyses for the motivated discrimination online data collection.

Notes. Statistically significant (*p* < .05) interactions will be broken down by their constituent components (e.g., if objectivity condition interacts with stereotyping priming, the main effect of the stereotype prime will be tested separately within each of the two objectivity conditions). The potential moderating role of gender of the evaluator will be assessed by further including the main effect and interactions involving participant gender in each analysis. An asterisk "*" in the code indicates that the models will produce a main effect and interaction (e.g., DV ~ Candidate_Gender*Participant_Gender, will result in a main effect of Candidate_Gender on the DV, a main effect of Participant_Gender on the DV, and the interaction between Candidate_Gender and Participant_Gender on the DV).

RESEARCH QUESTION	DESCRIPTION OF ANALYSIS	DEPENDENT MEASURE	CODE
Do hiring decisions favor men or women?	Main effect of candidate gender (female or male)	Hiring evaluations composite	DV ~ Candidate_Gender
Are perceived characteristics biased by candidate gender?	Main effect of candidate gender (female or male)	Perceived streetwise characteristics	DV ~ Candidate_Gender
		Perceived educated characteristics	DV ~ Candidate_Gender
Are hiring criteria constructed in a biased manner?	Interaction between candidate gender (female or male) and candidate characteristics (educated or streetwise)	Rated importance of streetwise characteristics	DV ~ Candidate_Gender*Characteristics
		Rated importance of educated characteristics	DV ~ Candidate_Gender*Characteristics

RESEARCH QUESTION	**DESCRIPTION OF ANALYSIS**	**DEPENDENT MEASURE**	**CODE**
Does priming stereotypes affect gender discrimination?	Interaction between stereotype prime condition (gender stereotypes or neutral concepts) and candidate gender (female or male)	Hiring evaluations composite	DV ~ Candidate_Gender*Stereotype_Prime
What are the effects of affirmation-threat on hiring judgments?	Interaction between affirmation vs. threat condition and candidate gender (female or male)	Hiring evaluations composite	DV ~ Candidate_Gender*Affirmation
	Interaction between affirmation vs. threat condition, candidate gender (female or male), and stereotype prime condition (stereotypes or neutral concepts)	Hiring evaluations composite	DV ~ Candidate_Gender*Affirmation*Stereotype_Prime
	Interaction between affirmation vs. threat condition, candidate gender (female or male), and individual differences in endorsement of sexist beliefs	Hiring evaluations composite	DV ~ Candidate_Gender*Affirmation*Endorsement_of_sexist_beliefs
	Interaction between affirmation vs. threat condition, candidate gender (female or male), and individual differences in beliefs about gender in the workplace	Hiring evaluations composite	DV ~ Candidate_Gender*Affirmation*Beliefs_about_gender_in_the_workplace
	Interaction between affirmation vs. threat condition, candidate gender (female or male), and candidate characteristics (educated or streetwise)	Rated importance of streetwise characteristics	DV ~ Candidate_Gender*Affirmation*Characteristics
		Rated importance of educated characteristics	DV ~ Candidate_Gender*Affirmation*Characteristics

RESEARCH QUESTION	**DESCRIPTION OF ANALYSIS**	**DEPENDENT MEASURE**	**CODE**
What are the effects of experimentally inducing a sense of objectivity?	Interaction between objectivity questions vs. neutral questions manipulation, and candidate gender (female or male)	Hiring evaluations composite	DV ~ Candidate_Gender*Objectivity_Condition
	Interaction between objectivity questions vs. neutral questions, candidate gender (female or male), and stereotype prime condition (stereotypes or neutral concepts)	Hiring evaluations composite	DV ~ Candidate_Gender*Objectivity_Condition*Stereotype_Prime
	Interaction between objectivity questions vs. neutral questions, candidate gender (female or male), and individual differences in endorsement of sexist beliefs	Hiring evaluations composite	DV ~ Candidate_Gender*Objectivity_Condition*Endorsement_of_sexist_beliefs
	Interaction between objectivity questions vs. neutral questions, candidate gender (female or male), and individual differences in beliefs about gender in the workplace	Hiring evaluations composite	DV ~ Candidate_Gender*Objectivity_Condition*Beliefs_about_gender_in_the_workplace

RESEARCH QUESTION	**DESCRIPTION OF ANALYSIS**	**DEPENDENT MEASURE**	**CODE**
What are the correlates of individual differences in self-perceived objectivity?	Interaction between candidate gender (female or male) and individual differences in self-perceived objectivity	Hiring evaluations composite	DV ~ Candidate_Gender*Objectivity_Beliefs
	Interaction between candidate gender (female or male) and individual differences in self-perceived objectivity	Within-subjects correlation between perceived characteristics and rated importance of characteristics, calculated at the level of individual participant (see Uhlmann & Cohen, 2005)	DV ~ Candidate_Gender*Objectivity_Beliefs
What are the effects of individual differences in feminist ideology?	Interaction between candidate gender (female or male) and individual differences in beliefs about gender in the workplace	Hiring evaluations composite	DV ~ Candidate_Gender*Beliefs_about_gender_in_the_workplace
	Interaction between candidate gender (female or male) and individual differences in exposure to feminist media	Hiring evaluations composite	DV ~ Candidate_Gender*Feminist_Media

RESEARCH QUESTION	**DESCRIPTION OF ANALYSIS**	**DEPENDENT MEASURE**	**CODE**
What are the effects of study-savviness?	Interaction between candidate gender (female or male) and number of studies previously completed	Hiring evaluations composite	DV ~ Candidate_Gender*Number_of_studies
	Interaction between candidate gender (female or male) and having done a similar study before	Hiring evaluations composite	DV ~ Candidate_Gender*Similar_study
	Interaction between candidate gender (female or male) and having taken a course in psychology before	Hiring evaluations composite	DV ~ Candidate_Gender*Psy_course
	Interaction between candidate gender (female or male) and suspicion the study is about gender issues on the free response item, as coded by independent raters blind to condition. Only participants who report becoming aware before or while evaluating the candidate will be coded as "aware" for the purposes of this analysis.	Hiring evaluations composite	DV ~ Candidate_Gender*Aware

Data Exclusions

To maximize power, we will first carry out the analyses above on the full sample. Then, to maximize data quality, we will re-analyze the data with the following exclusions.

Table S2-3. Data exclusions in the second round of analyses.

Relevant analyses	Excluded participants or data
All analyses	Participants who answered incorrectly (i.e., other than "strongly disagree") on the attention check item.
All analyses	Participants with less than five years of experience with the language of study administration (English).
All analyses	While blind to condition, we will code written responses to the free response awareness probe ("What do you think this survey was about?") for nonsensical and incoherent written comments and remove the relevant participants. We will likewise screen out participants with duplicate written comments (e.g., two supposedly different participants write word-for-word identical free responses to the same open-ended query).
All analyses	Participants who "straightline" in the survey, in other words give the same numeric response to all items in a scale (e.g., always putting "3" on a scale from 1-9).
All analyses	Participants who finish the survey too quickly, at a speed that would require reading an unrealistic 675 words per minute (wpm). This suggests insufficient effort responding (Huang, 2014).
Stereotype Prime vs. Neutral Prime manipulation	Participants who score 5 or above on the awareness of influence item (1-9 scale) for the scrambled-sentences manipulation
Stereotype Prime vs. Neutral Prime manipulation	Participants who failed to respond to all the scrambled-sentences items.
Objectivity questions vs. Neutral questions manipulation	Participants who failed to respond to all the objectivity questions or neutral questions.
Affirmation-Threat manipulation	Participants who write less than two sentences for the affirmation or threat essay.
Manipulation of candidate gender (female or male)	Participants who do not correctly remember the candidate's gender on the manipulation check item.

Manipulation of candidate gender	Participants who score 5 or above on awareness of being influenced by the candidate's gender (1-9 scale).
Manipulation of candidate characteristics (streetwise or educated)	Participants who do not correctly remember the candidate's characteristics (streetwise or educated) on the manipulation check item.
Items that reduce scale reliability	If a multi-item scale exhibits an alpha reliability below .40, we will drop the items with the lowest inter-item correlations one-by-one until reliability exceeds .40. If at the end of this process the most highly correlated items do not exhibit an alpha reliability above .40, we will rely on the single highest loading item.

Data-Dependent vs. Data-Independent Decisions

The resulting dataset will provide a rich opportunity for further analyses beyond the pre-specified ones. For example, demographic variables such as political conservatism or nation of citizenship, or certain process measures (e.g., above vs. below the median response times for the stereotype priming effect; see Huang, 2014), may help explain certain results.

In order to provide verification for any interesting patterns, we will divide the dataset into two parts: a data-dependent-decision sample (i.e., initial test sample) and a data-independent-decision sample (i.e., holdout sample). We will randomly divide the dataset within experimental condition in order to ensure representation of important variables in each subset. The initial test sample will be used for data-dependent analyses. Any promising analyses will then be preregistered and applied to the holdout sample (i.e., data-independent-decision sample). Ultimately any promising analyses from the test sample will be pre-registered and applied to the holdout sample.

Further, any analyses from this online data collection that return theoretically promising results will be pre-registered and applied to the crowdsourced data collections in partner laboratories in the second phase of the project.

References for Supplement 2

Brainerd, C. J., & Reyna, V. F. (2018). Replication, registration, and scientific creativity. *Perspectives on Psychological Science, 13*, 428–432.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., Celniker, J. B., & Zinger, J. F. (2018). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science, 14(2),* 273-291.

Greenberg, J., & Jonas, E. (2003). Psychological Motives and Political Orientation—The Left, the Right, and the Rigid: Comment on Jost et al. (2003). *Psychological Bulletin, 129*(3), 376–382.

Huang, J. L. (2014). Does cleanliness influence moral judgments? Response effort moderates the effect of cleanliness priming on moral judgment. *Frontiers in Psychology, 5*(1276), 1-8.

Leavitt, K., Mitchell, T., & Peterson, J. (2010). Theory pruning: Strategies for reducing our dense theoretical landscape. *Organizational Research Methods, 13*, 644-667.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). New York: Academic Press.

Tierney, W., Ebersole, C., Hardy, J., Chapman, H., Gantman, A., Vanaman, M., DeMarree, K., Wylie, J., Storbeck J., Andreychik, M.R., McPhetres, J., Vaughn, L.A., & Uhlmann, E. L. (2019). *A creative destruction approach to replication: Implicit work and sex morality across cultures.* Registered Report proposal accepted in principle at the Journal of Experimental Social Psychology, with data collection in progress.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37*, 1660–1672.

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

Uhlmann, E.L., & Cohen, G.L. (2007). “I think it, therefore it’s true”: Effects of self perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes, 104*, 207-223.

Uhlmann, E.L., & Nosek, B.A. (2012). My culture made me do it: Lay theories of responsibility for automatic prejudice. *Social Psychology, 43*, 108-113.

Uhlmann, E.L., Pizarro, D.A., Tannenbaum, D., & Ditto, P.H. (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*, 476–491.

Uhlmann, E.L., Poehlman, T.A., Tannenbaum, D., & Bargh, J.A. (2011). Implicit Puritanism in American moral cognition. *Journal of Experimental Social Psychology, 47*, 312-320.

Materials for Motivated Discrimination Online Data Collection

Material in red is notes to the study programmer, and is not seen by the research participant.

OVERALL DESIGN

The study will use a 2 (prime: stereotype primes vs. neutral primes) x 4 (mindset: objectivity questions vs. neutral questions vs. affirmation essay vs. threat essay) x 2 (applicant gender: male or female) x 2 (applicant characteristics: streetwise or educated) x 2 (participant gender: female or male) between-subjects design.

Order in which the study contents are administered:

1. Cover page. Seen by all participants.
2. Stereotype prime vs. control prime manipulation (2 conditions). The prime manipulation always comes first, with each participant completing 1 of 2 conditions.
3. Mindset manipulation (4 conditions). Then, the mindset manipulation of objectivity questions vs. neutral questions vs. affirmation essay vs. threat essay (each participant completes one of 4 conditions).
4. Hiring scenario. Seen by all participants.
5. Candidates (assignment to 1 of 4 candidates). Finally, participants are exposed to the male streetwise, male educated, female streetwise, or female educated candidates (each participant evaluates 1 of 4 candidates).
6. Dependent measures. Seen by all participants.
7. First round of moderator measures. Seen by all participants, in fixed order.
8. Second round of moderator measures. Seen by all participants, with the three measures —sexist beliefs, news exposure, and beliefs about gender— appearing in counterbalanced order, with order of administration recorded
9. Demographics. Seen by all participants.
10. Debriefing. Seen by all participants.

1. COVER PAGE (SEEN BY ALL PARTICIPANTS)

THANKS FOR HELPING US OUT!

THIS SET OF UNRELATED TASKS AND QUESTIONNAIRES TAKES ABOUT 10 MINUTES TO COMPLETE

YOU WILL COMPLETE A PUZZLE, FILL OUT SOME QUESTIONS ABOUT YOUR BELIEFS, AS WELL AS READ SCENARIOS AND MAKE DECISIONS

You must be at least 18 years old to participate in this study.

CONSENT STATEMENT:

I understand that my responses to this survey are completely anonymous, and that my participation is strictly voluntary. I may withdraw from the study at any time. Also, I am free to skip any questions I prefer not to answer.

[Page break]

2.STEREOTYPE PRIMING MANIPULATION

Stereotype prime condition

In each of the following scrambled sentences one word does not belong. Please remove that word and form a sentence with the remaining words.

world the welcomes is complex the world is complex ~~welcomes~~

homework pillows are pink nice

walk please olives dog the

timeless together group the gossiped

store appreciation the is nearby

barbie restaurant doll is a

drink topography water gallons of

is convenient sky very make-up

are very dogs university furry

the quickly tree came nurse

people some emotional are list

ate house the new is

Control prime condition

In each of the following scrambled sentences one word does not belong. Please remove that word and form a sentence with the remaining words.

world the welcomes is complex the world is complex ~~welcomes~~

walk please olives dog the

store appreciation the is nearby

drink topography water gallons of

the was composition dark forest

are very dogs university fuzzy

the brown television chair is

train nobody that does anymore

challenging always chair is homework

the unlocked rapid building was

ate house the new is

the blue look is curtain

3. MINDSET MANIPULATION

OBJECTIVITY QUESTIONS CONDITION

[Page break]

QUESTIONS ABOUT YOUR BELIEFS:

In most situations, I try to do what seems reasonable and logical.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

When forming an opinion, I try to objectively consider all of the facts.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

My judgments are based on a logical analysis of the facts.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

My decisions are rational and objective.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

NEUTRAL QUESTIONS CONDITION

[Page break]

QUESTIONS ABOUT YOUR BELIEFS:

I consider myself a morning person.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

I prefer light colors to dark colors.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

I enjoy listening to the radio.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

I usually get a full night's sleep.

0--------1--------2--------3--------4--------5--------6--------7--------8--------9--------10

Strongly Disagree Neutral Strongly Agree

AFFIRMATION ESSAY CONDITION

[Page break]

Which of these values is the most personally important to you? *(select one)*:

Artistic skills/appreciation
Relations with friends/family
Social skills
Musical ability/appreciation
Creativity
Romantic values
Sense of humor
Living life in the moment
Athletics
Physical attractiveness
Business/managerial skills

Please write about a time when you <u>succeeded</u> in living up to your #1 value or characteristic. Focus on expressing your memory of the event and the feelings that you had at the time.

____

____

____

____

____

THREAT ESSAY CONDITION

[Page break]

Which of these values is the most personally important to you? *(select one)*:

Artistic skills/appreciation
Relations with friends/family
Social skills
Musical ability/appreciation
Creativity
Romantic values
Sense of humor
Living life in the moment
Athletics
Physical attractiveness
Business/managerial skills

Please write about a time when you <u>failed</u> to live up to your #1 value or characteristic. Focus on expressing your memory of the event and the feelings that you had at the time.

____

____

____

____

____

____

4. HIRING SCENARIO (SEEN BY ALL PARTICIPANTS)

[Page break]

STUDY OF DECISION-MAKING IN HIRING

Thank you for agreeing to participate in this study. Complete this study as privately as possible. All of your responses are completely anonymous.

INSTRUCTIONS: We are interested in decision making processes in a hiring context. You will read about the traits and credentials of a job applicant. These traits may or may not be relevant to the decision of whether or not to hire the applicant.

After viewing the applicant's record, you will then decide if the person should be hired or not.

You may not always feel you have enough information to make a decision, but please do the best you can with the information provided.

"HIRING A NEW POLICE CHIEF"

Imagine that you have just been elected mayor of an urban town in the United States. Historically, the town's police department has had severe problems with scandals, inefficiency, corruption, lack of discipline, and skyrocketing crime rates. In fact, you were elected mayor primarily because you promised to appoint a new police chief that would clean up the department and enforce the law.

The time has come to hire this new police chief. The new chief must be able to ensure the quality and training of all officers, respond to and act upon citizen complaints, and above all keep property and violent crimes under control.

Remember that this is a critical decision: whether or not the person you decide to hire succeeds or fails as police chief will have a large impact on whether or not you are viewed as competent and ultimately re-elected to office.

[Page break]

5. CANDIDATE DESCRIPTION (ONE OF FOUR BELOW)

MALE, STREETWISE

DESCRIPTION OF APPLICANT FOR POLICE CHIEF:

BRIAN ROSNO

Brian has a great deal of street experience as a police officer. He has worked for 15 years as a police officer in town, and was involved in tough assignments. For example, he served on the homicide squad for 5 years. As a result, he has an excellent understanding of the local criminal elements, the police department, and the townspeople. He has personally arrested a large number of perpetrators of violent and property crimes. An outgoing person with a good sense of humor, Brian gets along very well with his fellow officers. Every year, he throws a holiday party that almost everybody in the department attends. He is a single male who lives alone in an apartment. Within the department, he is considered a straight-talker, tough and streetwise. He also has a reputation as an energetic leader and risk-taker. For example, he successfully pushed to increase prosecutions for car break-ins, which the department had tended to ignore. Finally, Brian is free and open in expressing his enthusiasm, both for his work and for his colleagues.

However, Brian is not very well educated, having only a 2-year degree from a community college. As a result, he does not have an in-depth understanding of criminal law, police administration or scientific theories of crime. Nor does he have much experience as an administrator. He is a weak public speaker and writer, finds it difficult to communicate well with the media, and is poorly connected to local and state politicians. Unskillful as a diplomat, he sometimes says the wrong things and offends important people. Finally, he is a bit disorganized and not very detail-oriented.

MALE, EDUCATED

DESCRIPTION OF APPLICANT FOR POLICE CHIEF:

BRIAN ROSNO

Brian is well-educated, with an undergraduate degree from Dartmouth and a law degree from the University of Washington. As a result, he has an excellent understanding of the intricacies of criminal law, police administration and scientific theories of crime. He also has 20 years of experience as an administrator in police departments in other towns. His family (a wife and two teenagers) lives in a nearby town. A good public speaker and writer, he is able to communicate effectively with the media. Recently, when his department had a potential scandal on their hands due to police officers taking bribes, he was able to communicate to the public that it was only a few "bad apples," not a problem with the whole department. Brian also has excellent political connections and is a skilled diplomat, able to avoid saying the wrong things and offending important people. His networking skills were critical to a successful lobbying campaign in the state senate to avoid cuts in police salaries. Finally, Brian is very well organized and pays careful attention to details.

However, Brian has only 3 years of street experience as a police officer. He has never worked a tough assignment like a homicide squad and does not currently have a strong understanding of the local criminal elements, of the personalities and politics within the department, or of the local townspeople. During his brief career as a street cop, he made few arrests for violent and property crimes. Within his department, Brian is a somewhat introverted person, and he has not consistently formed quality relationships with his fellow officers. He also has a reputation for being reserved and cautious, and somewhat humorless. Finally, Brian tends to refrain from expressing his enthusiasm for his work and for his co-workers.

FEMALE, STREETWISE

DESCRIPTION OF APPLICANT FOR POLICE CHIEF:

KAREN ROSNO

Karen has a great deal of street experience as a police officer. She has worked for 15 years as a police officer in town, and was involved in tough assignments. For example, she served on the homicide squad for 5 years. As a result, she has an excellent understanding of the local criminal elements, the police department, and the townspeople. She has personally arrested a large number of perpetrators of violent and property crimes. An outgoing person with a good sense of humor, Karen gets along very well with her fellow officers. Every year, she throws a holiday party that almost everybody in the department attends. She is a single female who lives alone in an apartment. Within the department, she is considered a straight-talker, tough and streetwise. She also has a reputation as an energetic leader and risk-taker. For example, she successfully pushed to increase prosecutions for car break-ins, which the department had tended to ignore. Finally, Karen is free and open in expressing her enthusiasm, both for her work and for her colleagues.

However, Karen is not very well educated, having only a 2-year degree from a community college. As a result, she does not have an in-depth understanding of criminal law, police administration or scientific theories of crime. Nor does she have much experience as an administrator. She is a weak public speaker and writer, finds it difficult to communicate well with the media, and is poorly connected to local and state politicians. Unskillful as a diplomat, she sometimes says the wrong things and offends important people. Finally, she is a bit disorganized and not very detail-oriented.

FEMALE, EDUCATED

DESCRIPTION OF APPLICANT FOR POLICE CHIEF:

KAREN ROSNO

Karen is well-educated, with an undergraduate degree from Dartmouth and a law degree from the University of Washington. As a result, she has an excellent understanding of the intricacies of criminal law, police administration and scientific theories of crime. She also has 20 years of experience as an administrator in police departments in other towns. Her family (a husband and two teenagers) lives in a nearby town. A good public speaker and writer, she is able to communicate effectively with the media. Recently, when her department had a potential scandal on their hands due to police officers taking bribes, she was able to communicate to the public that it was only a few “bad apples,” not a problem with the whole department. Karen also has excellent political connections and is a skilled diplomat, able to avoid saying the wrong things and offending important people. Her networking skills were critical to a successful lobbying campaign in the state senate to avoid cuts in police salaries. Finally, Karen is very well organized and pays careful attention to details.

However, Karen has only 3 years of street experience as a police officer. She has never worked a tough assignment like a homicide squad and does not currently have a strong understanding of the local criminal elements, of the personalities and politics within the department, or of the local townspeople. During her brief career as a street cop, she made few arrests for violent and property crimes. Within her department, Karen is a somewhat introverted person, and she has not consistently formed quality relationships with her fellow officers. She also has a reputation for being reserved and cautious, and somewhat humorless. Finally, Karen tends to refrain from expressing her enthusiasm for her work and for her co-workers.

6. DEPENDENT MEASURES (ALL PARTICIPANTS)

APPLICANT RATINGS

[Page break]

WHAT IS THE APPLICANT LIKE?

	Extremely **WEAK** in this area									Extremely **STRONG** in this area	
Streetwise	1	2	3	4	5	6	7	8	9	10	11
Educated	1	2	3	4	5	6	7	8	9	10	11
Tough	1	2	3	4	5	6	7	8	9	10	11
Experienced as an administrator	1	2	3	4	5	6	7	8	9	10	11
Organizational skills	1	2	3	4	5	6	7	8	9	10	11
Has made a large number of arrests	1	2	3	4	5	6	7	8	9	10	11
Computer skills	1	2	3	4	5	6	7	8	9	10	11
Detail-oriented	1	2	3	4	5	6	7	8	9	10	11
Administrative skills	1	2	3	4	5	6	7	8	9	10	11
Can communicate with the media well	1	2	3	4	5	6	7	8	9	10	11
Has kids	1	2	3	4	5	6	7	8	9	10	11

IMPORTANCE RATINGS (HIRING CRITERIA)

[Page break]

NOW WE WANT YOU TO DO SOMETHING DIFFERENT.

HOW IMPORTANT ARE THESE CHARACTERISTICS TO BEING A POLICE CHIEF?

	Makes success as a police chief LESS likely				**Makes No Difference**				**Essential to success as a police chief**		
Being streetwise	1	2	3	4	5	6	7	8	9	10	11
Being well educated	1	2	3	4	5	6	7	8	9	10	11
Toughness	1	2	3	4	5	6	7	8	9	10	11
Experience as an administrator	1	2	3	4	5	6	7	8	9	10	11
Organizational skills	1	2	3	4	5	6	7	8	9	10	11
Having made a large number of arrests	1	2	3	4	5	6	7	8	9	10	11
Computer skills	1	2	3	4	5	6	7	8	9	10	11
Being detail-oriented	1	2	3	4	5	6	7	8	9	10	11
Administrative skills	1	2	3	4	5	6	7	8	9	10	11
Ability to communicate with the media well	1	2	3	4	5	6	7	8	9	10	11
Having kids	1	2	3	4	5	6	7	8	9	10	11

HIRING DECISIONS

[Page break]

Please answer the following questions honestly and accurately.
Remember all your answers are in no way linked to your identity.

How successful would this applicant be as Police Chief?

Not successful at all								Extremely successful
1	2	3	4	5	6	7	8	9

How much of a good fit is the applicant for this position?

An extremely BAD fit								An extremely GOOD fit
1	2	3	4	5	6	7	8	9

Should this applicant be hired?

Should definitely NOT be hired								Should definitely be hired
1	2	3	4	5	6	7	8	9

7. FIRST-ROUND OF MODERATOR MEASURES (ALL PARTICIPANTS)

[Note: the first-round moderator measures appear in the following fixed order]

STUDY-SAVVINESS ITEMS (ALWAYS FIRST AMONG FOLLOW-UP MEASURES)

[Page break here]

What do you think this study was about?: ________________________________

____

When did you decide what the study was about? (for example, while you were rating the candidate, or after you made your ratings)? *(Please select one)*

Before I rated the candidate
While I was rating the candidate
After I rated the candidate

How many research studies have you previously completed? Number: _______________

Have you done a study similar to this one in the past? Yes No

If so, please describe it: ____

____

Have you ever taken a course in Psychology? Yes No

POST-MEASURE OF SELF-PERCEIVED OBJECTIVITY (ALWAYS 2ND)

[Page break]

	strongly DISAGREE		**strongly AGREE**
My judgments in this study were based on a logical analysis of the facts.	1……2…...3…...4…...5…...6…...7		
My decision-making in this study was rational and objective.	1……2…...3…...4…...5…...6…...7		

AWARENESS OF INFLUENCE (ALWAYS THIRD)

[Page break]

Did the sentence unscrambling task you completed influence your applicant ratings in any way?

NO				Not Sure				YES
1	2	3	4	5	6	7	8	9

If yes, please explain how and why it influenced you in your own words?

____

____

____

[Page break]

Did the gender of the candidate influence your ratings in any way?

NO				Not Sure				YES
1	2	3	4	5	6	7	8	9

If yes, please explain how and why it influenced you in your own words?

____

____

____

8. SECOND-ROUND OF MODERATOR MEASURES

[Note: the second-round moderator measures—sexist beliefs, news exposure, and beliefs about gender— appear in ***counterbalanced order****, with order of administration recorded]*

SEXIST BELIEFS

[Page break]

	strongly **DISAGREE**	**strongly** **AGREE**
It's a fact that men are better suited for some jobs than are women.	1…...2…...3…...4…...5…...6…...7	
Sometimes it's the objective thing to do to hire a man rather than a woman.	1…...2…...3…...4…...5…...6…...7	
It's a fact that men are better suited for the job of police chief than are women.	1…...2…...3…...4…...5…...6…...7	

FEMINIST MEDIA EXPOSURE MEASURE

[Page break]

How frequently do you read news articles? (Likert-type scale from 1 = not at all frequently to 7 = extremely frequently)

To what extent are you familiar with the #MeToo movement? (Likert-type scale from 1 = not at all familiar to 7 = extremely familiar)

How often have you come across news articles about gender discrimination in the workplace? (Likert-type scale from 1 = not at all frequently to 7 = extremely frequently)

How much exposure have you had to online commentary (e.g., Twitter, Facebook, etc) alleging biases against women in professional settings?
(1 = no exposure at all, 7 = a great deal of exposure)

How much exposure have you had to mainstream news coverage (e.g., newspapers, television news programs) alleging biases against women in professional settings?
(1 = no exposure at all, 7 = a great deal of exposure)

To what extent have you been actively following the #MeToo movement?
(1= not at all, 7 = following very closely)

BELIEFS ABOUT GENDER IN THE WORKPLACE MEASURE

[Page break]

Women are more likely to be passed over for assignments in the workplace than men are
(Likert-type scale from 1 = Strongly disagree to 7 = Strongly agree).

Women experience more instances of bias in the workplace than men do
(Likert-type scale from 1 = Strongly disagree to 7 = Strongly agree).

Men tend to get more opportunities than women do in the workplace
(Likert-type scale from 1 = Strongly disagree to 7 = Strongly agree).

Do you believe there is more bias against women or against men in professional settings, limiting their chances for advancement?
(1 = much more bias against men, 4 = men and women treated about the same, 7 = much more bias against women)

Female managers face systematic gender discrimination in today's workplaces.
(1= strongly disagree, 7 = strongly agree)

9. DEMOGRAPHIC MEASURES (ALL PARTICIPANTS)

[Page break here]

My gender is (*select one*): Male Female Other (please indicate):___________

My ethnicity is: White Asian Hispanic Black
Other (please indicate): __________________

My age is: __________ years

Politically, I am *(please circle one)*
Very Liberal
Liberal
Somewhat Liberal
Moderate
Somewhat Conservative
Conservative
Very Conservative

My occupation is: ______________________________

What country/region do you live in? __________________________________

Of what nation are you a citizen?______________________________

How many years have you lived in the United States? __________________

How many years of experience do you have with the English language? _______________

My educational level is:
Some high school/secondary school
High school degree/completed secondary school
Some university
University degree
Some graduate/postgraduate education
Graduate/postgraduate degree (e.g., doctoral degree)

A you currently a student at a university?
Yes
No

My yearly household income level is:

1= Less than $10,000 United States dollars (USD) a year
2= USD $10,000-$20,000
3= USD $20,000-$40,000
4= USD $40,000-$60,000
5= USD $60,000-$80,000
6= USD $80,000-$100,000
7= USD $100,000 a year or more

What is the education level of your most educated parent?

Some high school/secondary school
High school degree/completed secondary school
Some university
University degree
Some graduate/postgraduate education
Graduate/postgraduate degree (e.g., doctoral degree)

ATTENTION CHECK

Please select "strongly disagree" on the scale below:

strongly disagree
moderately disagree
neither disagree nor agree
moderately agree
strongly agree

MANIPULATION CHECKS

Without looking back, was the candidate you evaluated male or female?

Male
Female
Do not remember

Without looking back, was the candidate you evaluated stronger in terms of formal education or street experience?

Strongest in formal education
Strongest in street experience
Do not remember

[Page break]

10. DEBRIEFING (ALL PARTICIPANTS)

DEBRIEFING

Thanks for participating in this study. Your participation will help us to study the ways in which people make hiring decisions.

Previous research has shown that people prefer to hire women for some jobs (for example, a secretary but not a janitor) and prefer to see men in others (e.g. a janitor but not a secretary). Such gender-based hiring decisions tend to result from unconscious, culturally ingrained stereotypes of which the person doing the hiring is often unaware.

We are hypothesizing that one reason such hiring decisions occur is that people tend to unconsciously shift their hiring criteria. For example, if a man applies for a counter-stereotypical job such as a secretary), the person doing the hiring may find that they see the areas in which the man is strong (such as typing) as relatively less important for the job, and those in which he is weak (such an interpersonal skills) as more important. This is why participants are asked, in addition to their judgments of the applicants qualifications, how important they believe those qualifications are for the job.

We are additionally investigating the role of beliefs about objectivity in people's decisions. We are hypothesizing that the more people believe they are objective, the more likely they are to act on their attitudes, or stereotypes that have been subtly activated. People may also be less likely to act on stereotypes, or shift their hiring criteria, when their values have been recently affirmed, or when they are motivated to be accurate.

All of your responses in this experiment are completely anonymous— it is impossible to link your name to your questionnaire responses.

Thank you again for your participation in this study. If you have further questions or would like to hear about the results of the study, please talk to your experimenter and/or contact Eric Uhlmann (eric.uhlmann@insead.edu).

PLEASE DON'T DISCUSS THE RESULTS OF THE STUDY WITH OTHERS, EITHER ONLINE OR IN PERSON—IT'S IMPORTANT FOR OUR RESEARCH THAT PARTICIPANTS COME IN TO THE STUDY NOT KNOWING THE HYPOTHESIS. THANKS!

Supplement 3: Deviations from Pre-Registered Analysis Plan for the "Motivated Discrimination" Replication

Below we outline instances in which the analyses reported in the paper departed in meaningful ways from those specified in the preregistered analysis plan.

Sexism as a predictor of hiring decisions. As seen in Table S2-2, we preregistered analyses examining whether the threat-affirmation and objectivity mindset manipulations moderated the relationship between individual differences in sexism and hiring evaluations for female and male candidates. However, by accidental omission, we did not pre-register the simple and straightforward analysis looking at whether endorsement of sexist beliefs predicts hiring evaluations of women vs. men, as we did for beliefs about gender in the workplace and exposure to feminist ideologies. Parallel analyses were still conducted interacting candidate gender, participant gender, and each of these individual differences in predicting hiring evaluations (Supplement 4). Notably, sexist beliefs were used as predictors in the original research we were attempting to directly replicate (Uhlmann & Cohen, 2007), and the omission of the simple interaction between sexist beliefs and candidate gender from the table of planned analyses was a complete oversight.

Supplement 4: Methods and Results for the Motivated Discrimination Study

Below, we provide the methods and results for the creative destruction replication of Uhlmann and Cohen (2005, 2007). The replication study is described narratively in the main text. The methods and results are followed by Table S4-1 with more detailed analyses for the pre-registered variables of interest.

Participants

A sample of 3251 U.S. based participants (71% female, 28% male, 0.40% other, 0.74% no response) was recruited via the professional survey firm Pure Profile. Participants ranged from 18 to 87 years of age (*M* = 45.23, *SD* = 16.29). In terms of self-identified ethnicity, 72.50% were White, 4.46% Asian, 7.14% Hispanic, 12.33% Black, and 2.65%, selected "Other." Politically, 32.27% identified as liberals, 34.08% as moderates, and 22.85% as conservatives. With regard to education level, 4.46% of participants had completed some high school, 27.01% had completed a high school degree, 26.91% had some university education, 23.99% had graduated from university, 5.97% had some graduate education, and 10.3% had a postgraduate degree. The typical respondent's income was in the USD $20,000 to $40,000 bracket.

Design

The experiment employed a 2 (prime condition: gender stereotypes or neutral concepts) x 4 (mindset manipulation: affirmation essay, threat essay, objectivity questions, neutral questions) x 2 (applicant characteristics: streetwise vs. educated applicant) x 2 (candidate gender: female or male) x 2 (participant gender: female or male) between-subjects design.

Materials

Participants were informed they would be completing a set of unrelated tasks and questionnaires. These would include a puzzle, questions about their beliefs, and decision scenarios. The complete study materials are provided at the end of Supplement 2.

Stereotype priming manipulation. Participants completed one of two versions of a sentence-unscrambling task (Srull & Wyer, 1979). Embedded in the task were either words representing gender stereotypes (e.g., *pink, Barbie, make-up*) or neutral concepts (e.g., *gallons, chair, building*).

Mindset manipulation. Next, participants were assigned to one of four conditions designed to shift their general mindset going into the hiring simulation. In the objectivity mindset condition, they completed survey items designed to increase the salience of their sense of personal objectivity (e.g., "My judgments are based on a logical analysis of the facts"), and in the neutral mindset condition they completed nondescript items (e.g., "I consider myself a morning person"). In the affirmation condition, they selected their most important value from a list (e.g., *relationships with family, creativity, managerial skills*) and wrote a brief essay about a time they lived up to that value. In the threat condition, they wrote about a time they had failed to live up to their most important value.

Hiring scenario. All participants were told they would read about the traits and credentials of a job applicant and then decide if that person should be hired. In the simulation scenario, they were the mayor of a town dealing with skyrocketing crime and a police department in disarray due to inefficiency and corruption. The time had come to make a critical decision: hiring a new police chief that would clean up the department and enforce the law.

Applicant descriptions. Each participant read about one candidate for police chief, who was either female (Karen Rosno) or male (Brian Rosno) and either streetwise or formally educated. The streetwise candidate had made numerous arrests and got along very well socially with her/his fellow officers, among other characteristics. The educated candidate had a law degree and strong political and public speaking skills, among other characteristics.

Applicant ratings. On a scale ranging from 1 (*extremely weak in this area*) to 11 (*extremely strong in this area*), participants rated each applicant along a series of streetwise characteristics (e.g., *tough, has made a large number of arrests*) (α = 0.89) and educated characteristics (e.g., *ability to communicate with the media, administrative skills*) (α = 0.95).

Importance ratings. Next, participants separately rated the importance of each characteristic to the job of police chief (*1 = makes success much less likely, 11 = essential to success*). This produced composites for the importance of streetwise (α = 0.79) and educated (α = 0.91) characteristics.

Hiring evaluations. The applicant was also rated on how successful she/he would be as a police chief, whether she/he was a good fit for the position, and whether she/he should be hired (α = 0.93).

Self-perceived objectivity. A two-item post-measure of perceived personal objectivity (Uhlmann & Cohen, 2005) asked “My judgments in this study were based on a logical analysis of the facts” and “My decision-making in this study was rational and objective” (*1 = strongly disagree, 7 = strongly agree*) (α = 0.73).

Study-savviness measures. In a free response item, participants were asked what they thought the study was about, and in a follow-up item when they realized this (*before, while,* or *after* they made their candidate evaluations). They were further asked how many total studies they had previously completed, whether they had completed a similar study in the past, and whether they had taken a course in psychology.

Awareness of influence. Two separate probe items asked “Did the sentence unscrambling task you completed influence your applicant ratings in any way?” and “Did the gender of the candidate influence your ratings in any way?” (*1 = no, 4 = not sure, 9 = yes*).

Gendered ideologies. A set of three measures assessed sexist beliefs (e.g., “It’s a fact that men are better suited for some jobs than are women”; Uhlmann & Cohen, 2005) (α = 0.82), exposure to feminist social media (e.g., “How often have you come across news articles about gender discrimination in the workplace?”; McCormick-Huhn & Shields, 2019) (α = 0.87), and beliefs about gender in the workplace (e.g., “Women are more likely to be passed over for assignments in the workplace than men are”; McCormick-Huhn & Shields, 2019) (α = 0.91). The three gender ideology measures appeared in randomized order.

Demographics. Finally, participants completed a battery of demographics including their age, gender, ethnicity, nationality, income, education level, and political orientation, among other questions (see Supplement 2 for the complete materials).

Results

The pre-registered analysis plan is available at [https://osf.io/snbyg/] and in Supplement 2, and deviations from the plan are outlined in Supplement 3. The data and code are publicly posted online at [https://osf.io/xvs37/]. Notably, we pre-registered that we would analyze the data in two ways: 1) with the full sample, to maximize statistical power, and 2) with a set of pre-specified exclusion criteria to maximize data quality. These exclusion criteria were in some cases specific to certain variables, and included attention checks, completion checks, and awareness checks (see Table S2-3 of Supplement 2 for a detailed summary).

The reporting of the results below is organized around our pre-registered research questions (see Table S2-2 of Supplement 2). Below, F-tests underscored "full" refer to analyses on the entire sample (*N* = 3251 to 1593, depending on the analysis), and F-tests underscored "restricted" refer to analyses with the exclusion criteria in Table S2-3 (*N* = 2153 to 737, depending on the analysis).

Do hiring decisions favor men or women? A 2 (candidate gender) x 2 (participant gender) ANOVA with hiring evaluations as the dependent measure revealed a significant or marginally significant interaction depending on whether the full or restricted sample was used, F_{full}(3, 3218)=3.51, *p*=0.061, $F_{\text{restricted}}$ (3, 2147)=5.141, *p*=0.023. Directly contrary to the pattern in the original studies (Uhlmann & Cohen, 2005, 2007), male evaluators directionally favored female over male candidates, F_{full}(1, 919)=3.774, *p*=0.052, $F_{\text{restricted}}$(1, 506)=2.785, *p*=0.096. In contrast, female evaluators were either impartial to candidate gender or preferred male over female candidates, depending on the analysis, F_{full}(1, 2286)=0.192, *p*=0.661, $F_{\text{restricted}}$(1, 1634)=3.951, *p*=0.047.

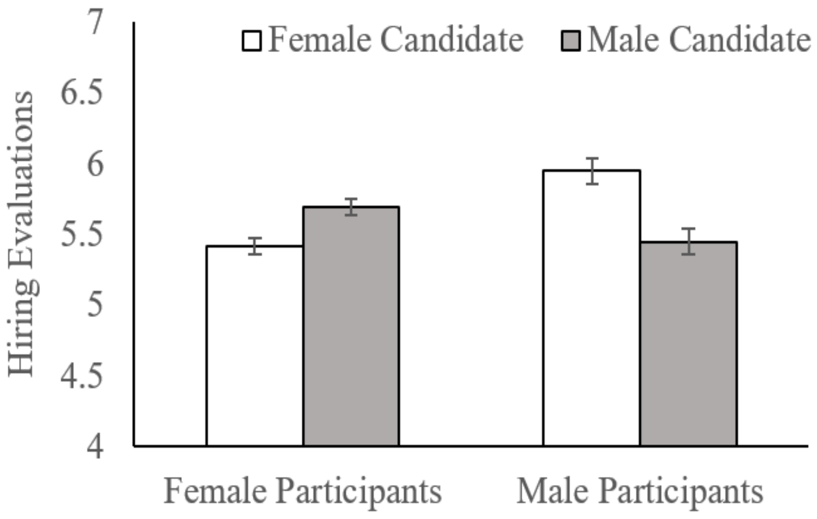


Figure S4-1: Hiring decisions for female and male candidates, among female and male evaluators

Are perceived characteristics influenced by candidate gender? Contrary to the cognitive schema account, no main effect differences emerged between female and male candidates for perceived streetwise characteristics, $F_{\text{full}}(1, 3217)=1.096$, $p=0.295$, $F_{\text{restricted}}(1, 2147)=0.57$, $p=0.45$; or perceived educated characteristics, $F_{\text{full}}(1, 3189)=0.303$, $p=0.582$, $F_{\text{restricted}}(1, 2139)=0.002$, $p=0.961$. In other words, inconsistent with cognitive assimilation to stereotypes, female and male candidates were not seen differently along these dimensions.

Are hiring criteria constructed to favor male or female candidates? 2 (candidate gender) x 2 (candidate characteristics: educated or streetwise) ANOVAs with streetwise and educated ratings as the dependent measures revealed no evidence of constructed criteria, for either female or male participants. Neither streetwise, $F_{\text{full}}(3, 3219)=0.093$, $p=0.76$, $F_{\text{restricted}}(3, 1966)=0.349$, $p=0.555$, nor educated characteristics, $F_{\text{full}}(3, 3201)=2.81$, $p=0.094$, $F_{\text{restricted}}(3, 1961)=1.915$, $p=0.167$, were shifted in favor of or against female or male candidates. Below, however, we report some evidence of constructed criteria among participants high in self-perceived objectivity based on within-subject correlations between their perceptions of the candidates and ratings of the importance of those same traits.

Does a self-affirmation vs. threat affect gender discrimination? A 2 (affirmation vs. threat) x 2 (candidate gender) x 2 (participant gender) ANOVA revealed a significant three-way interaction in the full sample only, $F_{\text{full}}(7, 1566)=6.105$, $p=0.014$, $F_{\text{restricted}}(7, 790)=0.278$, $p=0.598$. In the full-sample analyses, male participants were affected by the affirmation-threat manipulation, $F_{\text{full}}(3, 429)=6.519$, $p=0.011$, $F_{\text{restricted}}(3, 167)=1.134$, $p=0.288$, whereas female participants were not regardless of the subject-selection criteria, $F_{\text{full}}(3, 1130)=0.044$, $p=0.834$, $F_{\text{restricted}}(3, 620)=0.66$, $p=0.417$. In the full sample, among male participants who were affirmed, evaluations of female candidates were more positive than for male candidates, $F_{\text{full}}(1, 219)=4.848$, $p=0.029$, $F_{\text{restricted}}(1, 49)=1.391$, $p=0.244$. In contrast, among male participants who were threatened, evaluations of female and male candidates were similar, $F_{\text{full}}(1, 210)=2.01$, $p=0.158$, $F_{\text{restricted}}(1, 118)=0.019$, $p=0.89$.

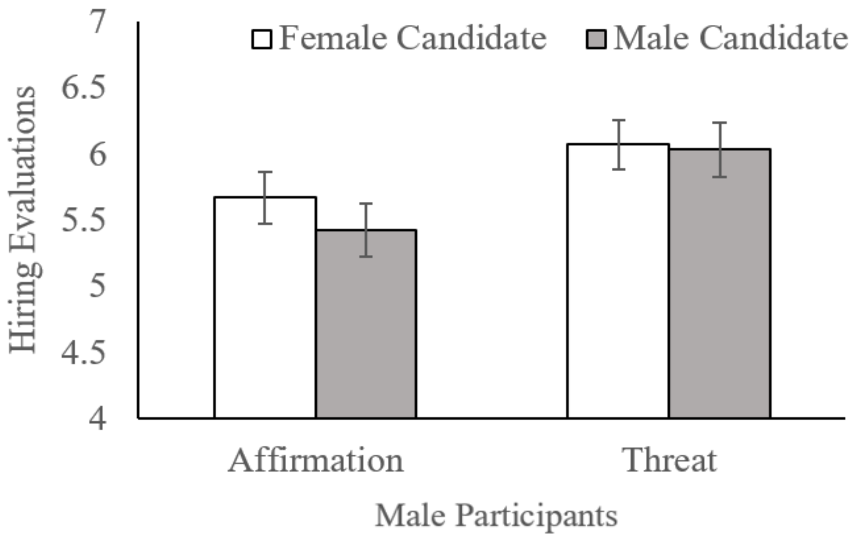


Figure S4-2: Hiring decisions by male evaluators in the threat vs. affirmation condition for female and male candidates.

Does activating a sense of objectivity affect gender discrimination? No two-way interaction emerged between objectivity vs. neutral mindset and candidate gender, $F_{\text{full}}(3, 1647)=0.466$, $p=0.495$, $F_{\text{restricted}}(3, 1088)=0.458$, $p=0.499$. There was also no three-way interaction between objectivity mindset, candidate gender, and participant gender, $F_{\text{full}}(7, 1640)=2.305$, $p=0.129$, $F_{\text{restricted}}(7, 1082)=0.014$, $p=0.905$. However, in the full sample of male evaluators, a marginally significant 2 (objectivity mindset vs. neutral mindset) x 2 (candidate gender) interaction emerged. Directly opposite to the originally observed pattern (Uhlmann & Cohen, 2007), an objectivity mindset if anything made male participants' hiring evaluations of female candidates more favorable relative to male candidates, $F_{\text{full}}(3, 484)=3.412$, $p=0.065$, $F_{\text{restricted}}(3, 275)=0.272$, $p=0.602$. In the full sample, male evaluators led to feel objective favored female over male candidates in their hiring judgments, $F_{\text{full}}(1, 246)=8.178$, $p=0.005$, $F_{\text{restricted}}(1, 151)=3.061$, $p=0.082$, whereas male evaluators in a neutral mindset did not, $F_{\text{full}}(1, 238)=0.037$, $p=0.848$, $F_{\text{restricted}}(1, 124)=0.782$, $p=0.378$. Failing to replicate Uhlmann and Cohen (2007), objectivity mindset condition did not interact with the stereotype priming condition or sexist attitudes to predict hiring decisions, Fs < 1.695, ps > .19 (see Table S4-1).

Do individual differences in self-perceived objectivity moderate the effect of gender on judgments? In the opposite pattern to that observed by Uhlmann and Cohen (2005), evaluators who perceived themselves as highly objective were if anything more likely to select female over male candidates. This interaction was marginally significant only in the restricted sample, $F_{\text{full}}(3, 3218)=2.643$, $p=0.104$, $F_{\text{restricted}}(3, 2149)=3.798$, $p=0.051$. We also looked at whether seeing oneself as objective correlated with constructing hiring criteria, captured by the within-subjects correlation between candidate trait ratings and the perceived importance of those traits for the job (see Uhlmann & Cohen, 2005, Study 1). A significant effect of objectivity beliefs on the construction of hiring criteria influenced by candidate gender emerged in both samples, $F_{\text{full}}(3, 2965)=3.977$, $p=0.046$, $F_{\text{restricted}}(3, 2079)=8.414$, $p=0.004$. In a reversal of the pattern observed by Uhlmann and Cohen (2005), seeing oneself as low in objectivity predicted constructing hiring criteria favorable to male candidates relative to female candidates. In contrast, high self-perceived objectivity participants did not set standards based on candidate gender.

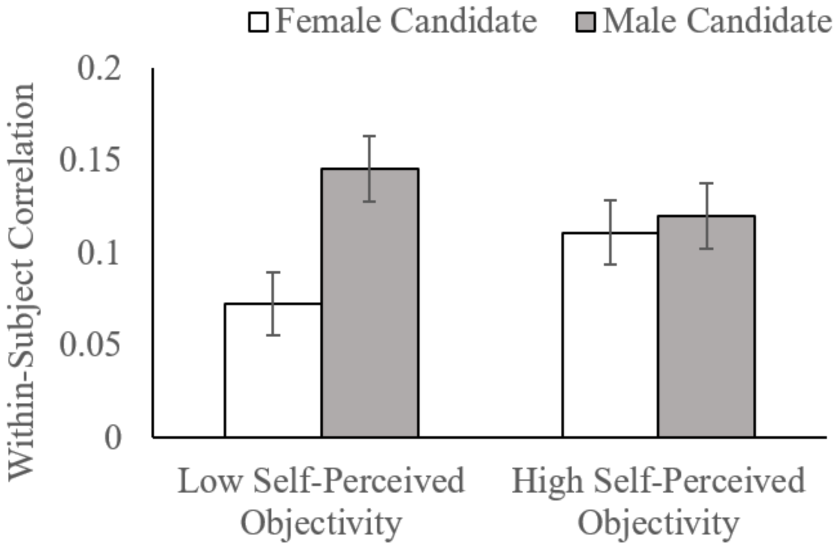


Figure S4-3: Self-perceived objectivity and favoritism in hiring criteria towards female vs. male candidates. Higher numbers reflect a stronger within-subjects correlation between

perceived candidate characteristics and the rated importance of such characteristics for the job, i.e., criteria constructed in a manner favorable to the candidate.

Do individual differences in gender ideologies moderate hiring decisions? Beliefs about gender and workplace opportunities did not moderate evaluations of female relative to male job candidates $F_{\text{full}}(3, 3221)=0.03$, $p=0.862$, $F_{\text{restricted}}(3, 2150)=0.238$, $p=0.626$. However, the sexist beliefs measure did interact with candidate gender to predict hiring evaluations in both samples, $F_{\text{full}}(3, 3220)=6.669$, $p=0.01$, $F_{\text{restricted}}(3, 2149)=12.572$, $p<.001$. As seen in Figure S4-4, strong rejection of sexist beliefs was associated with favoring female over male candidates, whereas relatively higher scores on sexist beliefs were associated with evaluating female and male candidates similarly.

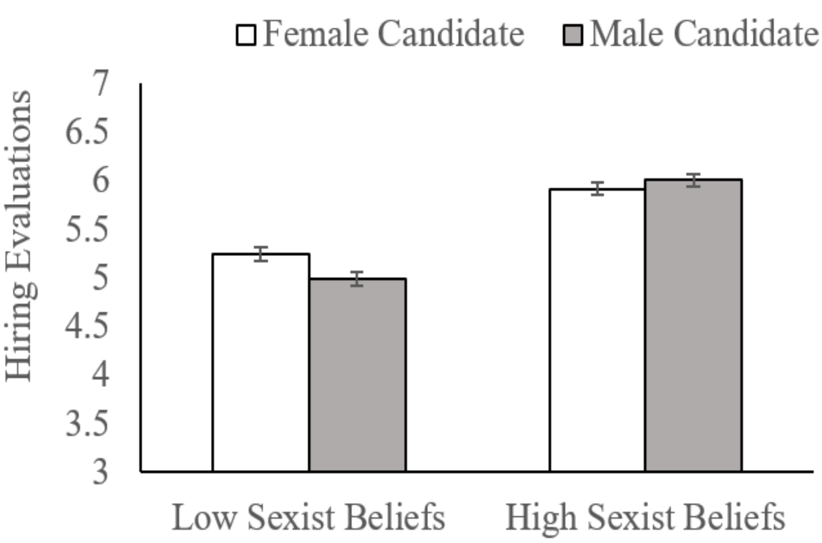


Figure S4-4: Sexist beliefs and hiring evaluations of female and male candidates.

In addition, exposure to feminist social media significantly interacted with candidate gender and participant gender in an unexpected pattern, $F_{\text{full}}(7, 3212)=3.954$, $p=0.047$, $F_{\text{restricted}}(7, 2143)=4.529$, $p=0.033$. For male evaluators, low levels of exposure to feminist social media was directionally associated with more favorable evaluations of female candidates relative to male candidates, $F_{\text{full}}(3, 917)=2.641$, $p=0.104$, $F_{\text{restricte1}}(3, 504)=2.386$ $p=0.123$. In contrast, for female evaluators, greater exposure to feminist social media was directionally associated with a preference for female over male candidates, $F_{\text{full}}(3, 2282)=2.794$, $p=0.095$, $F_{\text{restricted}}(3, 1632)=2.65$, $p=0.104$. This pattern is somewhat difficult to interpret. If it proves robust in subsequent research, it suggests women may be more receptive to #MeToo messaging than men are. Specifically, higher levels of #MeToo exposure predicted more favorable evaluations of female candidates among female evaluators, but relatively less favorable evaluations of female candidates by male evaluators.

Does study-savviness matter? Neither having completed a psychology course nor having done a larger number of studies before moderated the effects of candidate gender on hiring decisions, *F*s < 2.489. Very few participants (N = 47 in total) expressed suspicion the study was about gender on the free response item and further indicated they became suspicious before or while evaluating the candidate, rendering this measure not particularly useful for statistical tests of moderation. However, in the full sample, having done a similar study before did moderate the effect of candidate gender on hiring evaluations, F_{full}(3, 3203)=4.798, *p*=0.029, $F_{\text{restricted}}$(3, 2145)=0.391, *p*=0.532. Participants who had completed a similar study before tended to favor female over male applicants, F_{full}(1, 269)=4.293, *p*=0.039, $F_{\text{restricted}}$(1, 76)=0.181, *p*=0.672, whereas more naive participants tended to evaluate applicants of either gender similarly, F_{full}(1, 2934)=0.049, *p*=0.825, $F_{\text{restricted}}$(1, 2069)=1.076, *p*=0.30.

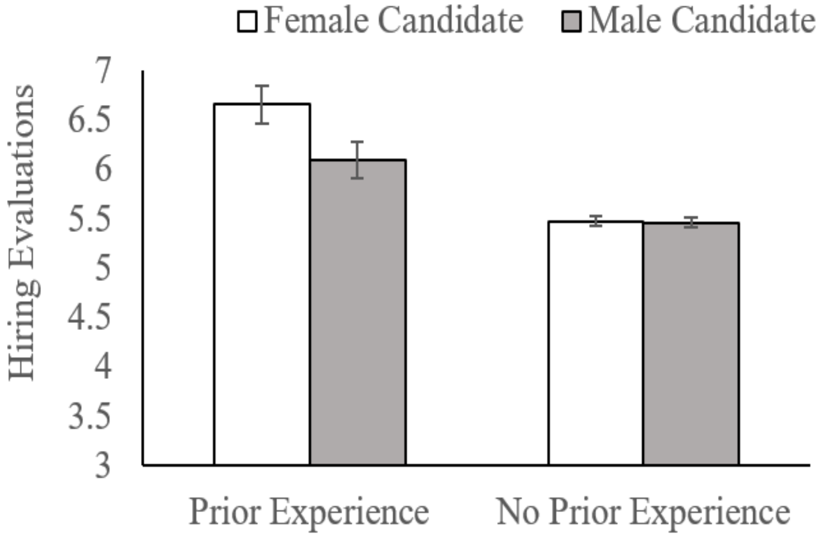


Figure S4-5: Prior experience with similar studies and hiring evaluations of female and male candidates.

Highlighting the contingency of research results on data analytic approaches (Silberzahn et al., 2018; Silberzahn & Uhlmann, 2015), several of these results were not robust to our two distinct pre-registered analytic strategies (full sample vs. restricted samples of participants), underscoring the need for further investigation of these topics. Further circumscribing the observed patterns, the replication sample was recruited online by a professional survey firm, likely oversampling more experienced and knowledgeable research participants. As noted in the pre-registration plan (see Supplement 2), the online context favors the study-savviness account, in that such respondents may be especially likely to accurately guess the hypothesis during the experiment. We are currently organizing a crowdsourced data collection that will repeat past experiments on gender discrimination in both college student and lay adult samples in the laboratory and field settings. This next phase of the replication initiative will again compete the motivated discrimination, cognitive assimilation to stereotypes, motivated liberalism, and study savviness accounts of participants' choices in hiring simulations involving female and male job candidates.

The implications of the replication project's results for the competing theories of gender discrimination are discussed narratively in the main article, and summarized in Table 2 of the main text.

References for Supplement 4

McCormick-Huhn, K., & Shields, S.A. (2019). *Can angry Black and White women get ahead in the era of #MeToo? Social dynamics in emotion appropriateness.* Unpublished manuscript.

Silberzahn, R., & Uhlmann, E.L. (2015). Many hands make tight work: Crowdsourcing research can balance discussions, validate findings and better inform policy. *Nature, 526,* 189-191.

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E., et al., & Nosek, B.A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science, 1,* 337–356.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37*, 1660-1672.

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

Uhlmann, E.L., & Cohen, G.L. (2007). "I think it, therefore it's true": Effects of self perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes, 104*, 207-223.

Table S4-1. Detailed results of the gender and hiring study

The table below reports the statistics for the models created to analyse the data from the motivated discrimination replication, as per the pre-registered analysis plan (Supplement 2). For each research question and dependent measure, we report:

A) The primary effect of interest
B) Whether this effect is moderated by evaluator gender
C) Whether the effect is present among male evaluators only
D) Whether the effect is present among female evaluators only
E-F) Additional analyses.

Unless stated otherwise, the dependent variable (DV) is the composite variable of hiring evaluations.

The descriptive statistics for each of the models are available on OSF website in flat file format. Simply use the model number in the first cell of the table row to find the associated descriptive statistics. For instance, the descriptive statistics for the primary effect model for “Do hiring decisions favor male or female candidates?” is in the file labelled “1a.csv”.

#	Full sample	Restricted sample
Do hiring decisions favor male or female candidates?		
1a	*F*(1, 3229)=0.61, *p*=0.435)	*F*(1, 2152)=0.81, *p*=0.368)
1b	*F*(3, 3218)=3.51, *p*=0.061)	*F*(3, 2147)=5.141, *p*=0.023)
1c	*F*(1, 919)=3.774, *p*=0.052)	*F*(1, 506)=2.785, *p*=0.096)
1d	*F*(1, 2286)=0.192, *p*=0.661)	*F*(1, 1634)=3.951, *p*=0.047)
Are perceived characteristics influenced by candidate gender? (DV= streetwise trait ratings)		
2a	*F*(1, 3217)=1.096, *p*=0.295)	*F*(1, 2147)=0.57, *p*=0.45)
2b	*F*(3, 3207)=0.864, *p*=0.353)	*F*(3, 2142)=1.286, *p*=0.257)
2c	*F*(1, 916)=0.029, *p*=0.865)	*F*(1, 505)=0.388, *p*=0.534)
2d	*F*(1, 2278)=1.412, *p*=0.235)	*F*(1, 1630)=1.195, *p*=0.274)
Are perceived characteristics influenced by candidate gender? (DV= educated trait ratings)		
3a	*F*(1, 3189)=0.303, *p*=0.582)	*F*(1, 2139)=0.002, *p*=0.961)
3b	*F*(3, 3179)=0.033, *p*=0.857)	*F*(3, 2134)=0.38, *p*=0.538)
3c	*F*(1, 907)=0.503, *p*=0.478)	*F*(1, 501)=0.447, *p*=0.504)
3d	*F*(1, 2260)=0.025, *p*=0.874)	*F*(1, 1626)=0.126, *p*=0.722)

Are hiring criteria constructed based on candidate gender? (DV= streetwise importance ratings)

4a	*F*(3, 3219)=0.093, *p*=0.76)	*F*(3, 1966)=0.349, *p*=0.555)
4b	*F*(7, 3207)=1.378, *p*=0.24)	*F*(7, 1959)=1.561, *p*=0.212)
4c	*F*(3, 916)=0.456, *p*=0.5)	*F*(3, 459)=0.766, *p*=0.382)
4d	*F*(3, 2278)=1.46, *p*=0.227)	*F*(3, 1493)=2.1, *p*=0.147)

Are hiring criteria constructed based on candidate gender? (DV = educated importance ratings)

5a	*F*(3, 3201)=2.81, *p*=0.094)	*F*(3, 1961)=1.915, *p*=0.167)
5b	*F*(7, 3189)=1.559, *p*=0.212)	*F*(7, 1954)=0.679, *p*=0.41)
5c	*F*(3, 912)=0.048, *p*=0.826)	*F*(3, 457)=0, *p*=0.989)
5d	*F*(3, 2265)=5.65, *p*=0.018)	*F*(3, 1490)=2.564, *p*=0.11)

Does priming stereotypes affect gender discrimination?

6a	*F*(3, 3227)=0.01, *p*=0.921)	*F*(3, 1730)=0.01, *p*=0.92)
6b	*F*(7, 3214)=0.103, *p*=0.748)	*F*(7, 1723)=0.023, *p*=0.879)
6c	*F*(3, 917)=0.046, *p*=0.83)	*F*(3, 399)=0.049, *p*=0.825)
6d	*F*(3, 2284)=0.02, *p*=0.887)	*F*(3, 1319)=0, *p*=0.996)

Interaction between affirmation vs. threat condition and candidate gender

7a	*F*(3, 1576)=1.092, *p*=0.296)	*F*(3, 795)=1.286, *p*=0.257)
7b	*F*(7, 1566)=6.105, *p*=0.014)	*F*(7, 790)=0.278, *p*=0.598)
7c	*F*(3, 429)=6.519, *p*=0.011)	*F*(3, 167)=1.134, *p*=0.288)
7d	*F*(3, 1130)=0.044, *p*=0.834)	*F*(3, 620)=0.66, *p*=0.417)
7e	*F*(1, 219)=4.848, *p*=0.029)	*F*(1, 49)=1.391, *p*=0.244)
7f	*F*(1, 210)=2.01, *p*=0.158)	*F*(1, 118)=0.019, *p*=0.89)

Interaction between affirmation vs. threat condition, candidate gender, and stereotype prime condition

8a	*F*(7, 1572)=0.811, *p*=0.368)	*F*(7, 791)=1.571, *p*=0.21)
8b	*F*(15, 1558)=1.527, *p*=0.217)	*F*(15, 782)=0.244, *p*=0.621)
8c	*F*(7, 425)=0.244, *p*=0.622)	*F*(7, 163)=1.41, *p*=0.237)
8d	*F*(7, 1126)=2.226, *p*=0.136)	*F*(7, 616)=0.543, *p*=0.461)

Interaction between affirmation vs. threat condition, candidate gender, and individual differences in endorsement of sexist beliefs

9a	*F*(7, 1567)=0.014, *p*=0.907)	*F*(7, 790)=0.016, *p*=0.899)
9b	*F*(15, 1556)=3.729, *p*=0.054)	*F*(15, 781)=4.351, *p*=0.037)
9c	*F*(7, 425)=3.877, *p*=0.05)	*F*(7, 163)=4.108, *p*=0.044)
9d	*F*(7, 1124)=0.475, *p*=0.491)	*F*(7, 615)=0.914, *p*=0.339)

Interaction between affirmation vs. threat condition, candidate gender, and individual differences in beliefs about gender in the workplace

10a	*F*(7, 1567)=0.814, *p*=0.367)	*F*(7, 791)=0.022, *p*=0.883)
10b	*F*(15, 1556)=1.049, *p*=0.306)	*F*(15, 782)=0.395, *p*=0.53)
10c	*F*(7, 425)=0.014, *p*=0.905)	*F*(7, 163)=0.397, *p*=0.53)
10d	*F*(7, 1124)=1.93, *p*=0.165)	*F*(7, 616)=0.137, *p*=0.711)

Interaction between affirmation vs. threat condition, candidate gender, and candidate characteristics (DV= streetwise importance ratings)

11a	*F*(7, 1566)=2.657, *p*=0.103)	*F*(7, 729)=0.955, *p*=0.329)
11b	*F*(15, 1552)=0.121, *p*=0.728)	*F*(15, 720)=0.291, *p*=0.59)
11c	*F*(7, 424)=0.792, *p*=0.374)	*F*(7, 147)=0.817, *p*=0.367)
11d	*F*(7, 1121)=1.598, *p*=0.206)	*F*(7, 570)=0.595, *p*=0.441)

Interaction between affirmation vs. threat condition, candidate gender, and candidate characteristics (DV = educated importance ratings)

12a	*F*(7, 1560)=0.017, *p*=0.895)	*F*(7, 728)=2.455, *p*=0.118)
12b	*F*(15, 1546)=0.827, *p*=0.363)	*F*(15, 719)=0.005, *p*=0.941)
12c	*F*(7, 422)=0.725, *p*=0.395)	*F*(7, 146)=0.438, *p*=0.509)
12d	*F*(7, 1117)=0.237, *p*=0.627)	*F*(7, 570)=1.689, *p*=0.194)

Interaction between objectivity questions vs. neutral questions manipulation, and candidate gender

13a	*F*(3, 1647)=0.466, *p*=0.495)	*F*(3, 1088)=0.458, *p*=0.499)
13b	*F*(7, 1640)=2.305, *p*=0.129)	*F*(7, 1082)=0.014, *p*=0.905)
13c	*F*(3, 484)=3.412, *p*=0.065)	*F*(3, 275)=0.272, *p*=0.602)
13d	*F*(3, 1150)=0.303, *p*=0.582)	*F*(3, 804)=0.062, *p*=0.803)
13e	*F*(1, 246)=8.178, *p*=0.005)	*F*(1, 151)=3.061, *p*=0.082)
13f	*F*(1, 238)=0.037, *p*=0.848)	*F*(1, 124)=0.782, *p*=0.378)

Interaction between objectivity questions vs. neutral questions, candidate gender, and stereotype prime condition

14a	*F*(7, 1643)=0.183, *p*=0.669)	*F*(7, 1084)=0.615, *p*=0.433)
14b	*F*(15, 1632)=0.119, *p*=0.731)	*F*(15, 1074)=0.131, *p*=0.718)
14c	*F*(7, 480)=0.015, *p*=0.903)	*F*(7, 271)=0.479, *p*=0.49)
14d	*F*(7, 1146)=0.621, *p*=0.431)	*F*(7, 800)=0.192, *p*=0.661)

Interaction between objectivity questions vs. neutral questions, candidate gender, and individual differences in endorsement of sexist beliefs

13a	*F*(7, 1641)=1.695, *p*=0.193)	*F*(7, 1084)=0.968, *p*=0.326)
13b	*F*(15, 1631)=0.502, *p*=0.479)	*F*(15, 1074)=0.646, *p*=0.422)
13c	*F*(7, 480)=1.364, *p*=0.243)	*F*(7, 271)=1.097, *p*=0.296)
13d	*F*(7, 1145)=0.159, *p*=0.69)	*F*(7, 800)=0.193, *p*=0.661)

Interaction between objectivity questions vs. neutral questions, candidate gender, and individual differences in beliefs about gender in the workplace

14a	F(7, 1642)=0.733, p=0.392)	F(7, 1084)=0.291, p=0.59)
14b	F(15, 1632)=0.166, p=0.684)	F(15, 1074)=0.657, p=0.418)
14c	F(7, 480)=0.042, p=0.839)	F(7, 271)=0.05, p=0.823)
14d	F(7, 1146)=0.205, p=0.651)	F(7, 800)=1.054, p=0.305)

Interaction between objectivity questions vs. neutral questions, candidate gender, and candidate characteristics (DV = streetwise importance ratings)

15a	F(7, 1641)=5.259, p=0.022)	F(7, 996)=2.794, p=0.095)
15b	F(15, 1631)=0.229, p=0.632)	F(15, 986)=0, p=0.998)
15c	F(7, 480)=3.288, p=0.07)	F(7, 252)=0.666, p=0.415)
15d	F(7, 1145)=2.181, p=0.14)	F(7, 731)=2.138, p=0.144)

Interaction between objectivity questions vs. neutral questions, candidate gender, and candidate characteristics (DV = educated importance ratings)

16a	F(7, 1629)=0.151, p=0.698)	F(7, 992)=0.368, p=0.544)
16b	F(15, 1619)=0.081, p=0.776)	F(15, 982)=0.103, p=0.748)
16c	F(7, 478)=0.285, p=0.594)	F(7, 251)=0.408, p=0.523)
16d	F(7, 1136)=0.015, p=0.902)	F(7, 728)=0.381, p=0.537)

Interaction between candidate gender and individual differences in beliefs about gender in the workplace

17a	F(3, 3221)=0.03, p=0.862)	F(3, 2150)=0.238, p=0.626)
17b	F(7, 3212)=0.716, p=0.398)	F(7, 2143)=0.008, p=0.928)
17c	F(3, 917)=0.335, p=0.563)	F(3, 504)=0.214, p=0.644)
17d	F(3, 2282)=0.641, p=0.423)	F(3, 1632)=0.493, p=0.483)

Interaction between candidate gender and individual differences in exposure to feminist media

18a	F(3, 3221)=0.434, p=0.51)	F(3, 2150)=0.643, p=0.423)
18b	F(7, 3212)=3.954, p=0.047)	F(7, 2143)=4.529, p=0.033)
18c	F(3, 917)=2.641, p=0.104)	F(3, 504)=2.386, p=0.123)
18d	F(3, 2282)=2.794, p=0.095)	F(3, 1632)=2.65, p=0.104)

Interaction between candidate gender and individual differences in endorsement of sexist beliefs

19a	F(3, 3220)=6.669, p=0.01)	F(3, 2149)=12.572, p<0.00)
19b	F(7, 3211)=3.424, p=0.064)	F(7, 2142)=2.635, p=0.105)
19c	F(3, 917)=14.522, p<0.00)	F(3, 504)=13.399, p<0.00)
19d	F(3, 2281)=1.964, p=0.161)	F(3, 1631)=5.178, p=0.023)

Interaction between candidate gender and number of studies previously completed

20a	*F*(3, 3145)=0.601, *p*=0.438)	*F*(3, 2124)=0.204, *p*=0.652)
20b	*F*(7, 3137)=0.575, *p*=0.448)	*F*(7, 2118)=5.022, *p*=0.025)
20c	*F*(3, 889)=1.194, *p*=0.275)	*F*(3, 495)=5.507, *p*=0.019)
20d	*F*(3, 2236)=0.009, *p*=0.925)	*F*(3, 1616)=0.03, *p*=0.862)

Interaction between candidate gender and having done a similar study before

21a	*F*(3, 3203)=4.798, *p*=0.029)	*F*(3, 2145)=0.391, *p*=0.532)
21b	*F*(7, 3194)=1.612, *p*=0.204)	*F*(7, 2138)=2.474, *p*=0.116)
21c	*F*(3, 910)=0, *p*=0.993)	*F*(3, 501)=1.215, *p*=0.271)
21d	*F*(3, 2271)=5.58, *p*=0.018)	*F*(3, 1630)=1.892, *p*=0.169)
21e	*F*(1, 269)=4.293, *p*=0.039)	*F*(1, 76)=0.181, *p*=0.672)
21f	*F*(1, 2934)=0.049, *p*=0.825)	*F*(1, 2069)=1.076, *p*=0.3)

Interaction between candidate gender and having taken a course in psychology before

22a	*F*(3, 3211)=0.549, *p*=0.459)	*F*(3, 2148)=0.571, *p*=0.45)
22b	*F*(7, 3202)=2.489, *p*=0.115)	*F*(7, 2141)=1.465, *p*=0.226)
22c	*F*(3, 914)=3.124, *p*=0.077)	*F*(3, 503)=1.82, *p*=0.178)
22d	*F*(3, 2275)=0.043, *p*=0.835)	*F*(3, 1631)=0.003, *p*=0.954)

Interaction between candidate gender and individual differences in self-perceived objectivity

24a	*F*(3, 3218)=2.643, *p*=0.104)	*F*(3, 2149)=3.798, *p*=0.051)
24b	*F*(7, 3209)=0.14, *p*=0.708)	*F*(7, 2142)=0.638, *p*=0.425)
24c	*F*(3, 915)=1.895, *p*=0.169)	*F*(3, 504)=0.078, *p*=0.78)
24d	*F*(3, 2281)=1.349, *p*=0.246)	*F*(3, 1631)=4.077, *p*=0.044)

Interaction between candidate gender and individual differences in self-perceived objectivity (DV = within-subject correlation between trait and importance ratings)

25a	*F*(3, 2965)=3.977, *p*=0.046)	*F*(3, 2079)=8.414, *p*=0.004)
25b	*F*(7, 2956)=0.166, *p*=0.684)	*F*(7, 2072)=0.61, *p*=0.435)
25c	*F*(3, 808)=0.484, *p*=0.487)	*F*(3, 477)=0.51, *p*=0.476)
25d	*F*(3, 2137)=3.431, *p*=0.064)	*F*(3, 1588)=6.722, *p*=0.01)

Supplement 5: Creative Destruction and Tests for Publication Bias

The creative destruction ethos applies not only to new experiments and re-analyses of existing datasets, but also to meta-analytic tests for publication bias. Consider the test for excess significance (Ioannidis, 2005) which calculates whether a set of studies report too many statistically significant ($p < .05$) findings given the statistical power of the studies. Given the ever-present publication filter, this test will almost inevitably conclude bias in a large enough set of articles on a topic. New tools such as *p*-uniform and *p*-curve can also be used to test for publication bias and evidentiary value in a sub-literature (Simonsohn, Nelson, & Simmons, 2014; van Aert, Wicherts, & van Assen, 2016). Such tests may conclude a body of empirical evidence, for example in favor of ego depletion effects (Carter & McCullough, 2014) or money priming (Lodder, Ong, Grasman, and Wicherts, in press) is high in publication bias and low in evidentiary value. However, such results do not point to which alternative theory of human motivation or materialism might be more robust, reliable, and useful.

The informational value of publication bias tests is much higher, we suggest, when multiple sub-literatures, or competing effects within the same literature, are simultaneously tested for publication bias. For example Simonsohn et al. (2014) *p*-curve both studies reporting significant choice overload effects (i.e., giving people more choices reduces post-choice satisfaction), as well as studies finding a broader array of choices is associated with increased satisfaction. The resulting pattern, such that the choice overload effects are heavily contaminated by publication bias whereas the more-choice-is-good effects are not, suggests providing decision makers with more options will generally make them happier with their final selection.

Ongoing research by Tey et al. (2019) adopts a similar approach, comparing publication bias in experiments finding hiring discrimination against women and underrepresented minorities (stereotype-based discrimination effects) and experiments finding that selection and promotion decisions favor women and minorities (reverse discrimination effects). Of further interest is which category of studies is more cited by other scholars, and receives the most media coverage. Comparatively greater publication and attentional biases in favor of evidence consistent with the liberal vs. conservative narrative on group inequalities may reflect pre-existing ideological commitments (Baron & Jost, 2019; Ditto et al., 2018; Duarte et al., 2015; Jelveh et al., 2015).

Another politically charged debate concerns the extent to which Implicit Association Test (IAT) measures predict relevant judgments and behaviors, with different meta-analytic investigations reporting aggregated correlations of .24, .14, and .10 in the domain of racial attitudes and beliefs (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Kurdi et al., 2019; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2015). Notably, none of these investigations fully leveraged advanced tools such as *p*-uniform and *p*-curve. Crawford, Inbar, Van Bavel, and Uhlmann (2019) are systematically comparing the relative levels of publication bias in studies finding IAT measures and explicit self-report measures predict behavior across politically sensitive domains (stereotype and prejudice) and non-sensitive domains (e.g., consumer choices). If the liberal worldview of most scientists affects our research (Duarte et al., 2015) then publication bias should be greatest in studies fitting the “pervasive prejudice” narrative that implicit biases are held by practically everyone and contribute to widespread systematic discrimination. Conversely, if evidence for the predictive validity of implicit and explicit measures exhibits similar statistical properties across topic domains, then perhaps the role of politics is more interpretive— for instance in the terminology used (e.g., different definitions of "prejudice"; Arkes & Tetlock, 2004; Banaji, Nosek, & Greenwald, 2004) or conclusions drawn from the evidence (Jussim, Crawford, Anglin, Stevens, & Duarte, 2016), rather than in the production of the science itself.

Testing contrasting sets of evidence for *relative* publication bias moves us away from the unsurprising conclusion that publication bias is present to assessing the relative robustness of the evidence for competing theories of what drives intergroup judgments and behaviors. It can also help address important meta-scientific questions regarding the roles played by researchers' ideological (Eitan et al., 2018; Jelveh et al., 2015) and intellectual commitments (Munder et al., 2013) in the reported empirical results.

References for Supplement 5

Arkes, H., & Tetlock, P.E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry*, *15*(4), 257-278.

Banaji, M. R., Nosek, B. A., & Greenwald. A. G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry, 15,* 279–289.

Baron, J., & Jost, J.T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? *Perspectives on Psychological Science, 14*(2), 292–303.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823.

Crawford, J., Inbar, Y., Van Bavel, J., & Uhlmann, E..L. (2019). *Relative publication bias in studies of the predictive validity of implicit and explicit measures.* Research in progress.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., Celniker, J. B., & Zinger, J. F. (2018). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science, 14*(2), 273-291.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. (2015). Political diversity will improve social and personality psychological science. *Behavioral and Brain Sciences, 38*, 1-13.

Eitan, O., Viganola, D., Inbar, Y., Dreber, A., Johanneson, M., Pfeiffer, T., Thau, S., & Uhlmann, E. L. (2018). Is scientific research politically biased? Systematic empirical tests and a forecasting tournament to address the controversy. *Journal of Experimental Social Psychology, 79*, 188-199.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17–41.

Ioannidis, J.P. (2005). Why most published research findings are false. PLoS Medicine. http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

Jelveh, Z., Kogut, B., & Naidu, S. (2015). *Political language in economics.* Unpublished manuscript. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2535453

Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. T., & Duarte, J. L. (2016). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology, 66*, 116-133.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*(5), 569–586.

Lodder, P., Ong, H. H., Grasman, R. P. P. P., & Wicherts, J. (in press). A comprehensive meta-analysis of money priming. *Journal of Experimental Psychology: General.*

Munder, T., Brütsch, O., Leonhart., R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. *Clinical Psychology Review, 33*, 501–511.

Oswald, F., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P.E. (2015). Revisiting the predictive validity of the Implicit Association Test. *Journal of Personality and Social Psychology, 105*(2), 171-192.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666–681.

Tey, K.S., Schaerer, M., van Aert, R., van Assen, M., Thau, S., & Uhlmann E.L. (2019). *Politics and p-values: Does ideology contribute to publication bias in research studies?* Meta-analysis in progress.

van Aert, R. C. M., Wicherts, J. M., and van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*, 713-729.

Supplement 6: Examples of Different Theory Pruning Approaches

As noted in the main text, there are five general categories of theory pruning strategies (Leavitt et al., 2010). Each of the successive approaches described below allows for stronger inferences (Platt, 1964).

Adding predictive variance

First, investigators can demonstrate that new constructs from one theory explain additional variance beyond that captured by another theory. While many scholars may use control variables to manage concerns of alternative explanations or endogeneity, scholars should more often consider collecting controls to demonstrate theoretical necessity of new constructs or measures (Leavitt et al., 2010). For example, Barrick and Zimmerman (2005) found that "clear purpose" scales fail to predict turnover variance, when disguised purpose scales and biodata are considered.

Nesting models

Second, researchers have compared two models which "nest" with regard to total propositions required for an explanation, showing that one theory is more parsimonious than the other. For example, Barger and Grandey (2006) argued that a signaling perspective, rather than a more complex emotional contagion perspective, is likely sufficient to explain the effects of smiling on customer service satisfaction. Specially, they reasoned that an emotional contagion argument linking smiling to customer satisfaction requires and subsumes all of the necessary positions of a signaling explanation (i.e., the customer must recognize the behavior and decode its intentions), but also requires the addition of an affective mediator. But demonstrating that the affective mediator was not necessary (or significant), they supported the more parsimonious explanation that was "nested" within the more complex one, and accordingly added an important boundary condition to emotional contagion theory.

Comparing magnitudes of effect sizes

Third, investigators can compare the magnitude of effect sizes associated with the predictions of two competing theories. The set of five studies conducted by Thau and Mitchell (2010) with regard to competing explanations for responses to abusive supervision are an example of this approach, demonstrating that a self-regulation impairment explanation consistently out-predicted a self-gain (i.e., distributive justice as mediator) perspective across multiple samples, measures, and designs. Although single-manuscript attempts at comparing effect sizes are laudable, the use of meta-analytic comparisons likely provides stronger tests of the relative explanatory power of two theories.

Comparison of predictive robustness

Fourth, scholars may apply a comparison of the predictive robustness of two theories, favoring the theory which best describes a stable relationship across a greater range of the predictors and criteria. For example, a key criticism of the moral disengagement theory of unethical behavior (Bandura, 1999; Bandura, Barbaranelli, Caprara, & Pastorelli, 1996; Bandura, Underwood, & Fromson, 1975) is that while it likely explains how individuals maintain their self-concept while committing significant transgressions, it does not appear to explain why individuals engage in relatively minor, everyday moral transgressions compared to theories of moral awareness (Reynolds, Dang, Yam, & Leavitt, 2014). Specifically, while Bandura (1999) theorizes that war crimes and other abhorrent acts occur through a cognitive process in which actors excuse their own behavior from commonly accepted moral standards (e.g., by derogating a victim, arguing for a greater moral good, or relinquishing responsibility to powerful others), Reynolds and colleagues (2014) argued that such extensive cognitive processing was not necessary to explain small transgressions (such as "checking" an exam answer off of a classmate). To this end, scholars should consider comparing theories across a broad range of conditions, favoring theories that best predict across a wide range of circumstances and placing clear boundaries around those which predict only for more extreme instances.

Contrasting incompatible theories

Finally, the most definitive approach to theory pruning involves carefully constructing tests where two truly incompatible theories are introduced in the same space. Notably, while this type of comparison represents the idealized prototype for strong inference described by Plat (1964), such contests are only appropriate when two theories are fully *comparable* and truly incompatible (see Leavitt et al., 2010, for considerations of comparability and compatibility).

Whereas the goal of contrasting incompatible theories is to vanquish one in favor of the other, such critical tests in the behavioral sciences may result in the discovery of omitted boundary conditions rather than identifying a clear winner. Latham and colleagues (1988) and Peteraf and colleagues (2013) provide illustrative examples. For example, Latham and colleagues (1988) created a series of critical studies attempting to compare the effectiveness of self-set versus other-set goals on performance. Through the careful construction of these studies, a critical boundary condition was discovered, such that both types of goals could be equally effective if they are internalized by the focal individual. This set of studies led to a more unified theory relating goals to performance, incorporating goal internalization as a key process variable. In the area of strategic management, research by Peteraf and colleagues (2013) attempted a similar undertaking an effort to explain contradictory findings in the dynamic capabilities literature. Ultimately, these authors utilized cocitation analysis to uncover two very different sets of assumptions from competing models within the literature, resulting in a (narrative) attempt to reconcile underlying boundary conditions between the two source models.

Supplement 6 References (Not Cited in Main Text)

Bandura, A., Barbaranelli, C., Caprara, G.V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*(2), 364–374.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193–209.

Bandura, A., Underwood, B., & Fromson, M.E. (1975). Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. *Journal of Research in Personality, 9*(4), 253–269.

Barrick, M. R., & Zimmerman, R. D. (2005). Reducing voluntary, avoidable turnover through selection. *Journal of Applied Psychology*, *90*(1), 159.

Peteraf, M., Di Stefano, G., & Verona, G. (2013). The elephant in the room of dynamic capabilities: Bringing two diverging conversations together. *Strategic Management Journal*, *34*(12), 1389-1410.

Supplement 7: Pre-Registered Analysis Plan for the Forecasting Survey

GENDER AND HIRING DECISIONS:
PRE-ANALYSIS PLAN FOR THE FORECASTING SURVEY

Contributors to analysis plan: Domenico Viganola, Elena Giulia Clemente, Anna Dreber, Michael Gordon, Magnus Johannesson, Thomas Pfeiffer, Warren Tierney, Eric Luis Uhlmann.

Summary: In this survey, we will examine whether researchers can predict the results of a set of direct and conceptual replications of experimental research on gender and hiring decisions. We are targeting researchers with training in judgment and decision making/social psychology research to participate in the forecasting survey, with no exclusion based on seniority or any other demographic characteristic.

Each participant (also referred to as forecaster in the rest of this pre-analysis plan) makes a total of $p = 24$ predictions. These will focus on the experimental effect sizes of the replications of hypotheses from Uhlmann & Cohen, 2005, 2007, as well as several novel effects derived from theories of gender discrimination. The predictions are subdivided into three groups:

- 2 predictions focusing on the simple effects (separately by evaluator gender)
- 6 predictions focusing on interaction effects (separately by evaluator gender)
- 16 predictions focusing on moderator effects

In addition to making these predictions, the participants are asked to answer a set of questions aimed at eliciting their personal beliefs on gender-related topics as well as assessing their demographics.

Prior to data collection, the forecasting survey was piloted with a few colleagues to provide feedback on the clarity of the questions and design. The data for these pilot participants (N = 8) was not included in the final report as it occurred prior to the final preregistration of the methods and analyses.

In this forecasting study we use both the more conservative significance threshold of $p <$ 0.005 (Benjamin et al., 2018; Secchi & Seri, 2017) and the traditional threshold for statistical significance of $p < 0.05$. All the tests in this pre-analysis plan are two-sided tests.

Primary hypotheses

Hypothesis 1

Hypothesis 1: There is a positive association between the predictions (beliefs) of the forecasters and the observed effect size

Individual-level regression to test whether forecasters' beliefs are significantly related to the realized effect sizes after controlling for individual fixed effects:

$$(1) \qquad RES_p = \beta_0 + \beta_1 PES_{ip} + FE_i + \varepsilon_{ip}$$

where:

- RES_p is a continuous variable indicating the realized effect size of the hypothesis p object of the prediction;
- PES_{ip} is a continuous variable indicating the predicted effect size of the effect of hypothesis p of forecaster i;
- FE_i is a set of individual fixed effects.

In equation (1) we plan to cluster standard errors at the individual level (number of clusters determined by the number of forecasters with $N = 24$ observations per cluster), since doing so allows us to take into account the fact that the predictions elicited from the same forecaster might be correlated.

Tests: t-test on coefficient β_1 in regression equation (1); t-test on coefficient β_0 in (1).

Robustness test of Hypothesis 1: we will estimate regression (1) separately for the three sets of predictions - predictions on simple effects, on interaction effects, and on moderator effects. Moreover, we will also carry out a robustness test where we estimate the Pearson correlation between the two vectors ($N = 24$ each) with the mean predicted effect size (PES_p) of each of the 24 effects replicated and the realized effect sizes RES_p.

Hypothesis 2

Can participants predict complex experimental results, such as interaction effects between conditions and individual differences moderators? To answer this question, first we compute the *accuracy* achieved in forecast p by each survey-taker i in terms of squared prediction error (Brier score), according to the formula:

$$BS_{ip} = (PES_{ip} - RES_p)^2$$

where RES_p and PES_{ip} should be interpreted as specified above. Then, we regress the variable BS_{ip} on 2 dummy variables identifying the forecasts regarding interactions $(INTES_{ip})$and the forecasts regarding the effects of the moderators $\left(IDMES_{ip}\right)$ and on the individual fixed effects FE_i, clustering the standard errors at the individual level in line with model (1):

$$(2) \qquad BS_{ip} = \beta_0 + \beta_1 INTES_{ip} + \beta_2 IDMES_{ip} + FE_i + \varepsilon_{ip}$$

Tests: *t*-test on coefficient β_1 in regression equation (2); *t*-test on coefficient β_2 in (2); Wald test on coefficient β_1 being different from β_2. Under the assumption that the forecasts on the interactions and on the moderators effects are more demanding, we expect both β_1 and β_2 to be positive.

Exploratory hypotheses

Introducing the ideological piece: how do scientists' political beliefs and convictions about gender relate to the accuracy of their forecasts? We exploit the individual accuracy measure introduced in hypothesis (2) and relate it to the forecasters' beliefs (sexist beliefs measure; beliefs about gender in the workplace; feminist media exposure measure; internal motivation to respond without sexism; external motivation to respond without sexism; political liberalism-conservatism on social issues) and to the forecasters' demographic characteristics (gender, academic seniority). The following tests are exploratory.

Individual-level regression to test whether forecasters' demographics and their convictions about gender relate to their accuracy in predicting the effect sizes. We plan to regress BS_{ip} on the following variables:

- Sexist beliefs measure (SBM_i)
- Feminist media exposure measure $(FMEM_i)$
- Beliefs about gender in workplace measure $(BGWM_i)$
- Internal motivation to respond without sexism $(IMSM_i)$
- External motivation to respond without sexism $(EMSM_i)$
- Political orientation on social issues measure (POL_i)
- Gender (G_i)
- Years from obtaining doctoral degree (SEN_i)

Please refer to the pre-registration document for the overall project (https://osf.io/snbyg/) and Supplements 2 and 4 for more details on these measures, most of which were also administered to the participants in the experiments whose results are being predicted.

Note that for these forecasts, we will again cluster the standard errors at the individual level to take into account potential correlations across forecasts made by the same forecaster:

(3) $BS_{ip} = \beta_0 + \sum_{k=1}^{8} \beta_k \; IC_{ik} + \varepsilon_{ip}$ for $k = 1, \ldots, 8$

where $IC = \{SBM_i;\ FMEM_i;\ BGWM_i;\ IMSM_i;\ EMSM_i;\ POL_i;\ G_i;\ SEN_i\}$

Test: *t*-tests on coefficients $\beta_1$to $\beta_8$in regression equation (3).

As a robustness check for hypothesis 3, we will analyze the accuracy of predictions on simple effects, on interaction effects, and on moderators effects separately. Therefore, we will estimate the models in equation (3) on mutually exclusive subsets of all the predictions, namely:

- Predictions on gender discrimination patterns in hiring with $2 \times n$ observations, n being the total number of forecasters
- Predictions on interaction effects of experimental manipulations with $6 \times n$ observations
- Predictions on the moderators effect sizes with $16 \times n$ observations

Do predictions regarding gender discrimination in hiring by male evaluators differ from those regarding gender discrimination in hiring by female evaluators? Are the predictions regarding the hiring evaluations made by women or men more accurate? We plan to answer this question by exploiting the fact that in the forecasting survey we ask exactly the same type of question for the two evaluator genders separately (e.g., 'What do you predict will be the effect size for the influence of candidate gender on hiring evaluations among male participants?' and 'What do you predict will be the effect size for the influence of candidate gender on hiring evaluations among female participants?'). In order to test whether the predictions regarding discrimination by female and male evaluators differ significantly, we focus on the predictions of the simple effects as main test (1 test), and on the predictions of the interaction effects as secondary tests (3 tests). In the spirit of avoiding over-testing, we restrict the domain of these exploratory tests to the simple and the interaction effects, and to the differences in terms of predictions' levels and predictions' accuracy only.

Do the predictions about female and male evaluators differ significantly?
Test: paired *t*-test comparing the predictions regarding the simple effects about male evaluators and about female evaluators.
Test: paired *t*-test comparing the predictions regarding the interactions effects for male evaluators and for female evaluators, for a total of 3 different tests.

Do the predictions about female and male evaluators differ in terms of accuracy?
Test: paired *t*-test comparing the Brier score (BS_{ip} as defined for hypothesis 2) for predictions regarding the simple effects for male evaluators and for female evaluators.

Test: paired *t*-test comparing the Brier score for the predictions regarding the interactions effects for male evaluators and for female evaluators, for a total of 3 different tests.

Incentive scheme

The incentive scheme to participate in this study is composed of two parts: the first one is co-authorship on the study report and it is granted to all the forecasters; the second one is a monetary incentive granted to two forecasters who are randomly selected.

Co-authorship. Upon completion of the prediction survey in all its parts, the participants qualify to be listed as co-authors on the final manuscript reporting the results of this study, which will be submitted for publication in a scientific journal. The forecasters may join via a consortium credit (e.g., "Hiring Decisions Forecasting Collaboration").

Monetary incentives. We will randomly select two of the participants and reward them with a bonus payout determined as a function of the accuracy of their forecasts. The bonus payoffs will be computed according to the following scoring rule:

$$\$200 - (\underline{Sq.Error} \times 200)$$

where $\underline{Sq.Error}$ is the average of the squared errors for all the 24 forecasts of the 'Gender and Hiring Decisions Forecasting Study' made by the forecasters.

Reference for Supplement 7

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ...Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10.

Secchi, D., & Seri, R. (2017), Controlling for false negatives in agent-based models. A review of power analysis in organizational research, *Computational and Mathematical Organization Theory*, *23*(1), 94-121.

Supplement 8: Forecasting Survey Materials

GENDER AND HIRING DECISIONS: FORECASTING SURVEY

We are scientists at the Stockholm School of Economics, University of Limerick, and INSEAD conducting an investigation of forecasting accuracy. We are interested in whether researchers can predict the results of experimental research on candidate gender and hiring decisions. We are recruiting researchers with training in judgment and decision making/social psychology research to participate in this study. All levels of expertise are welcome, from graduate students to senior professors. In addition to providing your forecasts, you will also complete a brief demographic questionnaire.

Consortium authorship. By completing the entire survey, you qualify to be listed as a co-author on the manuscript reporting the results. This will take the form of a consortium credit "Hiring Decisions Forecasting Collaboration" in the first page/author string, with all forecasters listed by name and affiliation in an appendix. Notably, the investigators who carried out the project will be listed by name in the author string, whereas forecasters will be grouped together in a consortium credit, as per the preferences of previous journal editors.

Monetary payments. In addition, as described in greater detail later, you may receive monetary rewards for completing the survey. This reward, if you are randomly chosen, is based on the accuracy of your predictions.

All data collected in this study are for research purposes only. We may share the data we collect in this study with other researchers doing future studies – if we share your data, we will not link your responses with your name or any identifying information.

Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study. Partial data will not be analyzed. For additional questions about this research, you may contact Anna Dreber Almenberg at: anna.dreber@hhs.se.

Please indicate, in the box below, that you are at least 18 years old, have read and understand this consent form, and you agree to participate in this online research study.

○ I am at least 18 years old, have read and understand this consent form, and agree to participate in this online research study.

[Page break here]

Your Contact Information

Please provide your complete email so we can deliver any payment [Free response text box]

Then click "next" to complete the survey.

[Page break here]

Forecasting Survey: Candidate Gender and Hiring Decisions

About the initiative

This initiative tested four competing theories of candidate gender and hiring decisions against one another, by directly and conceptually replicating previously observed gender discrimination effects with large sample sizes and measuring a number of theoretically important individual differences moderators. Of particular interest is the previously observed tendency for evaluators to engage in motivated rationalizations for discriminating in favor of male job candidates over female job candidates (Uhlmann & Cohen, 2005, 2007). This motivated discrimination account was pitted against three alternative accounts of gender and selection decisions in hiring simulations.

The four competing theories of candidate gender and hiring decisions are the following:

Motivated discrimination perspective: Evaluators change their hiring criteria to rationalize choosing male over female job applicants, preserving a sense of personal objectivity despite being biased in their selection decisions.

Cognitive assimilation perspective: Biased perceptions based on cognitive schemas lead evaluators to select men over women for traditionally male jobs.

Motivated liberalism perspective: Due to an increasing awareness of workplace gender inequalities and exposure to feminist ideologies such as the #MeToo movement, evaluators favor female over male job candidates.

Study-savviness perspective: Participants who have greater prior experience with research studies, and thus are more likely to be suspicious the study is about gender, overcompensate to avoid appearing sexist and therefore favor female over male job candidates.

Format of predictions

We will ask you to make predictions about the effect sizes associated with a set of research predictions, separately for female and male evaluators (i.e., participants in the hiring experiment). We will also ask for your forecasts regarding potential individual-differences moderators of gender discrimination in hiring decisions. We will ask you about the expected effect sizes in terms of Cohen's d (Cohen, 1988; Sawilowsky, 2009). For more on Cohen's d please see this link: https://en.wikipedia.org/wiki/Effect_size#Cohen.27s_d

Quoting Wikipedia on effect sizes: "*an effect size is a quantitative measure of the strength of a phenomenon. Examples of effect sizes are the correlation between two variables, the regression coefficient in a regression, the mean difference, or even the risk with which something happens, such as how many people survive after a heart attack for every one person that does not survive. For each type of effect-size, a larger absolute value always indicates a stronger effect.*"

In the social sciences, a Cohen's d of 0.20 is considered to be a small effect, 0.50 is considered to be a medium effect, and 0.80 is considered to be a large effect.

Please note

- Your answers are saved in real time, so you can complete the survey in more than one session. To do this simply click on the survey link: the survey will automatically continue where you stopped at the end of your previous session.
- The "back button" on the bottom right allows you to go back and update the answers that you submitted previously.
- Please complete this survey on a sufficiently large screen.
- Please do not clear cookies or browsing history of your browser, especially if you are planning to complete the survey in multiple sittings.
- Please do not complete the survey in private/incognito mode on your browser, as your progress will not be saved then.

Incentives for accuracy

As a reward for your time, you will be listed as a co-author on the final manuscript as described earlier. In addition, we will randomly select 2 participants and reward them with a bonus payout determined as a function of the accuracy of their forecasts: more accurate forecasts in terms of lower average squared prediction error (i.e., the absolute difference between the prediction and the realized outcome) lead to higher bonuses. The bonus payment is determined according to the following scoring rule:

$$\$200\ -(\underline{Sq.Error} \times 200)$$

where $\underline{Sq.Error}$ is the average of the squared prediction errors for all the forecasts you are asked to submit. The bonus payment ranges between $200 (if you get all the predictions equal to the realized output) and $0 (if the $\underline{Sq.Error}$ computed on your forecasts exceeds 1, or if you are not selected for the bonus payout).

You will make predictions about effects of experimental manipulations and individual differences moderators of gender discrimination, for a total of 24 predictions. You will also complete measures of your personal beliefs and demographic items (total of 36 questions). In all, you will complete 60 questions in this survey.

Please click the “forward” button to read about the original studies targeted for replication, the design and methods of the replication study, and provide your forecasts about the replication results.

[Page break here]

OVERVIEW OF ORIGINAL STUDIES TARGETED FOR REPLICATION

The direct and conceptual replication initiative re-examined earlier findings on the roles of psychological rationalizations and illusions of personal of objectivity in discrimination against women (Uhlmann & Cohen, 2005, 2007). The references and abstracts for the two papers are below.

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

Abstract: This article presents an account of job discrimination according to which people redefine merit in a manner congenial to the idiosyncratic credentials of individual applicants from desired groups. In three studies, participants assigned male and female applicants to gender-stereotypical jobs. However, they did not view male and female applicants as having different strengths and weaknesses. Instead, they redefined the criteria for success at the job as requiring the specific credentials that a candidate of the desired gender happened to have. Commitment to hiring criteria prior to disclosure of the applicant's gender eliminated discrimination, suggesting that bias in the construction of hiring criteria plays a causal role in discrimination.

Full text UC2005:
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf

Uhlmann, E.L., & Cohen, G.L. (2007). "I think it, therefore it's true": Effects of self perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, *104*, 207-223.

Abstract: A sense of personal objectivity may prompt an ''I think it, therefore it's true'' mindset, in which people assume that their own beliefs and introspections are, by definition, valid and therefore worthy of being acted on. In the present studies, priming a sense of personal objectivity increased gender discrimination, particularly among decision-makers who endorsed stereotypic beliefs or who had stereotypic thoughts made cognitively accessible through implicit priming. Implications for discrimination in organizational contexts, and for theories of attitude–behavior consistency, are discussed.

Full text UC2007:
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf

[Page break here]

OVERVIEW OF REPLICATION STUDY

The replication study design included key aspects of Uhlmann and Cohen (2005) and (2007), as well as further manipulations and measures to allow for testing the competing theories against one another (motivated discrimination, cognitive assimilation, motivated liberalism, study-savviness). Below we provide a summary of the methods for the replication.

Methods

Sample characteristics

A total of 3251 U.S. based participants (71% female, 28% male, 0.40% other, 0.74% no response) were recruited online via the professional survey firm Pure Profile. Participants ranged from 18 to 87 years of age (*M* = 45.23, *SD* = 16.29). In terms of self-identified ethnicity, 72.50% were White, 4.46% Asian, 7.14% Hispanic, 12.33% Black, and 2.65%, selected "Other." Politically, 32.27% identified as liberals, 34.08% as moderates, and 22.85% as conservatives. With regard to education level, 4.46% of participants had completed some high school, 27.01% had completed a high school degree, 26.91% had some university education, 23.99% had graduated from university, 5.97% had some graduate education, and 10.3% had a postgraduate degree. The typical respondent's income was in the USD $20,000 to $40,000 bracket.

Design

The replication combined key aspects of the Uhlmann and Cohen (2005) and (2007) studies as well as additional conditions and measures. Thus, the replication study featured a 2 (prime condition: gender stereotypes or neutral concepts) x 4 (mindset manipulation: affirmation essay, threat essay, objectivity questions, neutral questions) x 2 (applicant characteristics: streetwise vs. educated applicant) x 2 (candidate gender: female or male) x 2 (participant gender: female or male) between-subjects design.

Materials

Participants were informed they would be completing a set of unrelated tasks and questionnaires. These would include a puzzle, questions about their beliefs, and decision scenarios. The complete study materials are available here https://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0

Stereotype priming manipulation. Participants completed one of two versions of a sentence-unscrambling task (Srull & Wyer, 1979). Embedded in the task were either words representing gender stereotypes (e.g., *pink, Barbie, make-up*) or neutral concepts (e.g., *gallons, chair, building*).

Mindset manipulation. Next, participants were assigned to one of four conditions designed to shift their general mindset going into the hiring simulation. In the objectivity mindset condition, they completed survey items designed to increase the salience of their sense of personal objectivity (e.g., "My judgments are based on a logical analysis of the facts"), and in the neutral mindset condition they completed nondescript items (e.g., "I consider myself a morning person"). In the affirmation condition, they selected their most important value from a list (e.g., *relationships with family, creativity, managerial skills*) and wrote a brief essay about a time they lived up to that value. In the threat condition, they wrote about a time they had failed to live up to their most important value. The idea behind including this new manipulation was that a self-threat, relative to a self-affirmation, should activate motivated biases.

Hiring scenario. All participants were told they would read about the traits and credentials of a job applicant and then decide if that person should be hired. In the simulation scenario, they were the mayor of a town dealing with skyrocketing crime and a police department in disarray due to inefficiency and corruption. The time had come to make a critical decision: hiring a new police chief that would clean up the department and enforce the law.

Applicant descriptions. Each participant read about one candidate for police chief, who was either female (Karen Rosno) or male (Brian Rosno) and either streetwise or formally educated. The streetwise candidate had made numerous arrests and got along very well socially with her/his fellow officers, among other characteristics. The educated candidate had a law degree and strong political and public speaking skills, among other characteristics.

Applicant ratings. On a scale ranging from 1 (*extremely weak in this area*) to 11 (*extremely strong in this area*), participants rated each applicant along a series of streetwise characteristics (e.g., *tough, has made a large number of arrests*) (α = 0.89) and educated characteristics (e.g., *ability to communicate with the media, administrative skills*) (α = 0.95).

Importance ratings. Next, participants separately rated the importance of each characteristic to the job of police chief (*1 = makes success much less likely, 11 = essential to success*). This produced composites for the importance of streetwise (α = 0.79) and educated (α = 0.91) characteristics.

Hiring evaluations. The applicant was also rated on how successful she/he would be as a police chief, whether she/he was a good fit for the position, and whether she/he should be hired (α = 0.93).

Self-perceived objectivity. A two-item post-measure of perceived personal objectivity (Uhlmann & Cohen, 2005) asked “My judgments in this study were based on a logical analysis of the facts” and “My decision-making in this study was rational and objective” (*1 = strongly disagree, 7 = strongly agree*) (α = 0.73).

Study-savviness measures. Participants were asked how many total studies they had previously completed, whether they had completed a similar study in the past, and whether they had taken a course in psychology.

Gendered ideologies. A set of three measures assessed sexist beliefs (e.g., “It’s a fact that men are better suited for some jobs than are women”; Uhlmann & Cohen, 2005) (α = 0.82), exposure to feminist social media (e.g., “How often have you come across news articles about gender discrimination in the workplace?”; McCormick-Huhn & Shields, 2019) (α = 0.87), and beliefs about gender in the workplace (e.g., “Women are more likely to be passed over for assignments in the workplace than men are”; McCormick-Huhn & Shields, 2019) (α = 0.91). The three gender ideology measures appeared in randomized order.

[Page break here; participants should be able to go backwards to review the methods]

YOUR FORECASTS OF THE REPLICATION RESULTS

The pre-registered analysis plan for the replication is available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0. Below, we ask you to predict a set of key results from the replication.

A brief few points before you start:

- Due to the complexity of the design, some analyses focused on subsets of conditions. For example, analyses of the effects of experimentally activating a sense of personal objectivity focused on participants in the objectivity questions and neutral questions conditions. For this reason, sample sizes vary considerably across different analyses.
- For your predictions, please assume the analyses involved all participants who completed the relevant measures and conditions (i.e., without selecting out participants based on manipulation and attention checks).
- The outcome measure is hiring evaluations unless otherwise stated.
- When the question focuses on evaluations of female candidates, this is meant relative to male candidates unless otherwise stated. Likewise, when the question focuses on evaluations of male candidates, this is meant relative to female candidates unless otherwise stated.
- When question focuses on a specific experimental condition (e.g., self-threat), this is meant relative to the comparison condition (e.g., self-affirmation).
- We will ask you to make forecasts separately for female and male evaluators, to accommodate your predictions about interactions with participant gender.
- For each result, first we will ask you your predictions of the effect size in terms of Cohen's d, then we will ask you the direction of the effect.

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

YOUR PREDICTED PATTERN OF GENDER DISCRIMINATION IN HIRING:

What do you predict will be the effect size for the influence of <u>candidate gender on hiring evaluations</u> among <u>male participants</u>? Here we ask about the effect size in terms of Cohen’s d, across the other conditions (e.g., stereotype priming, mindset manipulation). The replication sampled **920** male participants who evaluated either a female or male candidate. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the <u>direction</u> of the effect:

- Male evaluators will favor <u>**male** over **female** candidates.</u>

- Male evaluators will favor **female** over **male** candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the influence of candidate gender on hiring evaluations among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming, mindset manipulation). The replication sampled **2,287** female participants who evaluated either a female or male candidate.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Female evaluators will favor **male** over **female** candidates.
- Female evaluators will favor **female** over **male** candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

EFFECTS OF AFFIRMATION-THREAT MANIPULATION ON GENDER DISCRIMINATION:

What do you predict will be the effect size for the interaction between affirmation-threat and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **432** male participants who evaluated either a female or male candidate and were either affirmed or threatened beforehand.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Self-threat will make male evaluators give relatively less favorable hiring evaluations to **female** candidates.
- Self-threat will make male evaluators give relatively less favorable hiring evaluations to **male** candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between affirmation-threat and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **1,133** female participants who evaluated either a female or male candidate and were either affirmed or threatened beforehand.

[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Self-threat will make female evaluators give relatively less favorable hiring evaluations to **female** candidates.
- Self-threat will make female evaluators give relatively less favorable hiring evaluations to **male** candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

EFFECTS OF OBJECTIVITY MINDSET EXPERIMENTAL MANIPULATION ON GENDER DISCRIMINATION

What do you predict will be the effect size for the interaction between objectivity vs. neutral mindset and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **487** male participants who evaluated either a female or male candidate and either completed questions about their personal objectivity or neutral questions beforehand.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- An objectivity mindset will make male evaluators give relatively less favorable hiring evaluations to **female** candidates.
- An objectivity mindset will make male evaluators give relatively less favorable hiring evaluations to **male** candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between objectivity vs. neutral mindset and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **1,153** female participants who evaluated either a female or male candidate and either completed questions about their personal objectivity or neutral questions beforehand.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- An objectivity mindset will make female evaluators give relatively less favorable hiring evaluations to **female** candidates.

- An objectivity mindset will make female evaluators give relatively less favorable hiring evaluations to **male** candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

EFFECTS OF STEREOTYPE PRIMING ON GENDER DISCRIMINATION

What do you predict will be the effect size for the interaction between priming stereotypes vs. neutral concepts and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., affirmation-threat, objectivity mindset). The replication sampled **920** male participants who were primed with either stereotypes or neutral concepts and then evaluated either a female or male candidate.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Priming stereotypes will make male evaluators give relatively less favorable hiring evaluations to **female** candidates.
- Priming stereotypes will make male evaluators give relatively less favorable hiring valuations to **male** candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between priming stereotypes vs. neutral concepts and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., affirmation-threat, objectivity mindset). The replication sampled **2,287** female participants who were primed with either stereotypes or neutral concepts and then evaluated either a female or male candidate.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Priming stereotypes will make female evaluators give relatively less favorable hiring evaluations to **female** candidates.
- Priming stereotypes will make female evaluators give relatively less favorable hiring evaluations to **male** candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here
http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here
https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here
https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

BELIEFS ABOUT A SENSE OF PERSONAL OBJECTIVITY AS AN INDIVIDUAL DIFFERENCES MODERATOR

Hiring decisions as the DV:

What do you predict will be the effect size for the interaction between individual differences in a sense of personal objectivity and candidate gender predicting the hiring evaluations of male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **918** male participants who evaluated either a female or male candidate and completed a scale of their conviction in their own objectivity.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- A sense of personal objectivity will be associated with male evaluators giving relatively more **positive** hiring evaluations to female candidates.
- A sense of personal objectivity will be associated with male evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between individual differences in a sense of personal objectivity and candidate gender predicting the hiring evaluations of female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,284** female participants who evaluated either a female or male candidate and completed a scale of their conviction in their own objectivity.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- A sense of personal objectivity will be associated with female evaluators giving relatively more **positive** hiring evaluations to female candidates.
- A sense of personal objectivity will be associated with female evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.

Original study: Uhlmann and Cohen 2007 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here
http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here
https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here
https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

Constructing biased criteria as the DV:

Note: Our measure of biased hiring criteria is the within-subjects correlation between applicant ratings as streetwise vs. educated and the rated importance of streetwise and educated characteristics to the job of police chief. High within-subjects correlations reflect setting hiring criteria that favor the specific applicant being evaluated. Please see Uhlmann and Cohen (2005) for more details on this measure of favoritism in criteria (Full text UC2005: http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf)

What do you predict will be the effect size for the interaction between individual differences in a sense of personal objectivity and candidate gender predicting hiring criteria favorable to the applicant among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **811** male participants who evaluated either a female or male candidate and completed applicant and importance ratings used to calculate the within-subjects index of criteria favorable to the applicant.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- A sense of personal objectivity will be associated with male evaluators setting criteria biased **in favor of male** relative to female candidates.
- A sense of personal objectivity will be associated with male evaluators setting criteria biased **in favor of female** relative to male candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between individual differences in a sense of personal objectivity and candidate gender predicting hiring criteria favorable to the applicant among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,140** female participants who evaluated either a female or male candidate and completed applicant and importance ratings used to calculate the within-subjects index of criteria favorable to the applicant.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- A sense of personal objectivity will be associated with female evaluators setting criteria biased **in favor of male** relative to female candidates.
- A sense of personal objectivity will be associated with female evaluators setting criteria biased **in favor of female** relative to male candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here
http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here
https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here
https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

GENDER IDEOLOGIES AS MODERATORS OF GENDER DISCRIMINATION

Sexist beliefs:

What do you predict will be the effect size for the interaction between sexist beliefs and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **920** male participants who evaluated either a female or male candidate and completed a sexism scale.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Sexist beliefs will be associated with male evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Sexist beliefs will be associated with male evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between sexist beliefs and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,284** female participants who evaluated either a female or male candidate and completed a sexism scale.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Sexist beliefs will be associated with female evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Sexist beliefs will be associated with female evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.

Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

Beliefs about gender in the workplace:

What do you predict will be the effect size for the interaction between the belief that workplaces are biased against women and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **920**male participants who evaluated either a female or male candidate and completed a scale assessing their beliefs about gender in the workplace.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- The belief that workplaces are biased against women will be associated with male evaluators giving relatively more **positive** hiring evaluations to female candidates.
- The belief that workplaces are biased against women will be associated with male evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between the belief that workplaces are biased against women and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,285** female participants who evaluated either a female or male candidate and completed a scale assessing their beliefs about gender in the workplace.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- The belief that workplaces are biased against women will be associated with female evaluators giving relatively more **positive** hiring evaluations to female candidates.
- The belief that workplaces are biased against women will be associated with female evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.

Replication: Pre-analysis plan available here
https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here
https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

Feminist messaging exposure:

What do you predict will be the effect size for the <u>interaction between exposure to feminist messaging and candidate gender</u> among <u>male participants?</u> Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **920** male participants who evaluated either a female or male candidate and completed questions about their exposure to feminist messaging.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the <u>direction</u> of the effect:

- Exposure to feminist messaging will be associated with male evaluators giving relatively <u>more **positive** hiring evaluations to female candidates.</u>
- Exposure to feminist messaging will be associated with male evaluators giving relatively <u>more **negative** hiring evaluations to female candidates.</u>

[Multiple choice with two options]

What do you predict will be the effect size for the <u>interaction between exposure to feminist messaging and candidate gender</u> among <u>female participants?</u> Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,285**female participants who evaluated either a female or male candidate and completed questions about their exposure to feminist messaging.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the <u>direction</u> of the effect:

- Exposure to feminist messaging will be associated with female evaluators giving relatively <u>more **positive** hiring evaluations to female candidates.</u>
- Exposure to feminist messaging will be associated with female evaluators giving relatively <u>more **negative** hiring evaluations to female candidates.</u>

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here
http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here
http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here
https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here
https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

EFFECTS OF STUDY-SAVVINESS ON HIRING DECISIONS INVOLVING FEMALE AND MALE CANDIDATES

Having done a similar study before as the moderator:

What do you predict will be the effect size for the interaction between having done a similar study before and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **913** male participants who evaluated either a female or male candidate and completed a question about whether they had done a similar study before. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Having done a similar study before will be associated with male evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Having done a similar study before will be associated with male evaluators giving relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between having done a similar study before and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,274** female participants who evaluated either a female or male candidate and completed a question about whether they had done a similar study before. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Having done a similar study before will be associated with female evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Having done a similar study before will be associated with female evaluators giving relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

Having taken a course in psychology as the moderator:

What do you predict will be the effect size for the interaction between having taken a course in psychology before and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **917** male participants who evaluated either a female or male candidate and completed a question about whether they had taken a course in psychology before.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Having taken a course in psychology before will be associated with male evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Having taken a course in psychology before will be associated with male evaluators give relatively **more negative** hiring evaluations to female candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between having taken a course in psychology before and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,278** female participants who evaluated either a female or male candidate and completed a question about whether they had taken a course in psychology before.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Having taken a course in psychology before will be associated with female evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Having taken a course in psychology before will be associated with female evaluators give relatively **more negative** hiring evaluations to female candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

Number of studies done previously as the moderator:

What do you predict will be the effect size for the interaction between number of studies previously completed and candidate gender among male participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming).

The replication sampled **892** male participants who evaluated either a female or male candidate and completed a question about the number of studies they had previously completed.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Having participated in more studies before will be associated with male evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Having participated in more studies before will be associated with male evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

What do you predict will be the effect size for the interaction between number of studies previously completed and candidate gender among female participants? Here we ask about the effect size in terms of Cohen's d, across the other conditions (e.g., stereotype priming). The replication sampled **2,239** female participants who evaluated either a female or male candidate and completed a question about the number of studies they had previously completed.
[Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].
Please specify the direction of the effect:

- Having participated in more studies before will be associated with female evaluators giving relatively more **positive** hiring evaluations to female candidates.
- Having participated in more studies before will be associated with female evaluators give relatively more **negative** hiring evaluations to female candidates.

[Multiple choice with two options]

Original study: Uhlmann and Cohen 2005 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202005.pdf.
Original study: Uhlmann and Cohen 2007 full text available here http://socialjudgments.com/docs/Uhlmann%20and%20Cohen%202007.pdf.
Replication: Complete study materials available here http://www.dropbox.com/s/wrf7cgrkx47ips4/1.Study%20Materials%20Gender%20and%20Hiring.pdf?dl=0.
Replication: Pre-analysis plan available here https://www.dropbox.com/s/vj2pq6st2kw2zw0/2.Planned%20Analyses%20Gender%20and%20Hiring.pdf?dl=0.
Instructions to the survey available here https://www.dropbox.com/s/bu02we7uf5gzv2j/gender%20and%20hiring%20decisions%20instructions.pdf?dl=0.

Please note you will no longer be able to go back and change your predictions after proceeding to the next page.

Measures of Your Beliefs

[Note: The following measures are shown to forecasters in randomized order. These measures parallel those completed by research participants in the replication, with the exception of the internal and external motivation scales.]

[NOT SHOWN TO RESPONDENTS: “SEXIST BELIEFS MEASURE”]

	strongly DISAGREE strongly AGREE
It’s a fact that men are better suited for some jobs than are women.	1…..2…...3…...4…...5…...6…...7
Sometimes it’s the objective thing to do to hire a man rather than a woman.	1…..2…...3…...4…...5…...6…...7
It’s a fact that men are better suited for the job of police chief than are women.	1…..2…...3…...4…...5…...6…...7

[NOT SHOWN TO RESPONDENTS: “FEMINIST MEDIA EXPOSURE MEASURE”]

How frequently do you read news articles? (Likert-type scale from 1 = not at all frequently to 7 = extremely frequently)

To what extent are you familiar with the #MeToo movement? (Likert-type scale from 1 = not at all familiar to 7 = extremely familiar)

How often have you come across news articles about gender discrimination in the workplace? (Likert-type scale from 1 = not at all frequently to 7 = extremely frequently)

How much exposure have you had to online commentary (e.g., Twitter, Facebook, etc) alleging biases against women in professional settings?
(1 = no exposure at all, 7 = a great deal of exposure)

How much exposure have you had to mainstream news coverage (e.g., newspapers, television news programs) alleging biases against women in professional settings?
(1 = no exposure at all, 7 = a great deal of exposure)

To what extent have you been actively following the #MeToo movement?
(1= not at all, 7 = following very closely)

[NOT SHOWN TO RESPONDENTS: “BELIEFS ABOUT GENDER IN THE WORKPLACE MEASURE”]

Women are more likely to be passed over for assignments in the workplace than men are
(Likert-type scale from 1 = Strongly disagree to 7 = Strongly agree).

Women experience more instances of bias in the workplace than men do
(Likert-type scale from 1 = Strongly disagree to 7 = Strongly agree).

Men tend to get more opportunities than women do in the workplace
(Likert-type scale from 1 = Strongly disagree to 7 = Strongly agree).

Do you believe there is more bias against women or against men in professional settings, limiting their chances for advancement?
(1 = much more bias against men, 4 = men and women treated about the same, 7 = much more bias against women)

Female managers face systematic gender discrimination in today’s workplaces.
(1= strongly disagree, 7 = strongly agree)

[NOT SHOWN TO RESPONDENTS: “INTERNAL MOTIVATION TO RESPOND WITHOUT SEXISM”]

I am personally motivated by my beliefs to be nonsexist toward women.
1 (strongly disagree) to 7 (strongly agree)

Being nonsexist toward women is important to my self-concept.
1 (strongly disagree) to 7 (strongly agree)

Because of my personal values, I believe that using stereotypes about women is wrong.
1 (strongly disagree) to 7 (strongly agree)

[NOT SHOWN TO RESPONDENTS: “EXTERNAL MOTIVATION TO RESPOND WITHOUT SEXISM”]

Because of today’s PC (politically correct) standards I try to appear nonsexist toward women.
1 (strongly disagree) to 7 (strongly agree)

I try to hide any negative thoughts about women in order to avoid negative reactions from others.
1 (strongly disagree) to 7 (strongly agree)

I attempt to appear nonsexist toward women in order to avoid disapproval from others.
1 (strongly disagree) to 7 (strongly agree)

[NOT SHOWN TO RESPONDENTS: “POLITICAL ORIENTATION MEASURE”]

In general, how would you rate your political views regarding social issues?

1 Very Left-Wing
2
3
4 Moderate
5
6
7 Very Right-Wing

Demographic Questions

What is your age? [Free response]

What is your gender?

1= Male
2= Female
3= Other: [Free response text box]
4= Prefer not to tell

In which country/region were you born in? [Pulldown menu with numerous options, including Taiwan]

In which country/region do you currently reside? [Pulldown menu with numerous options, including Taiwan]

How many years of experience with English do you have? [Pulldown menu with numeric responses]

What department are you in at your institution (e.g., psychology, organizational behavior, statistics)? [Free response]

If relevant, what year did you receive, or do you expect to receive, your doctoral degree? [Pulldown menu with numeric responses]

What is your job rank? (please select one)

○ Research assistant (1)

○ Graduate student (2)

○ Postdoctoral researcher (3)

○ Assistant Professor (4)

○ Associate Professor (5)

○ Full Professor (6)

○ Other (please indicate) (7)

Other job rank, please indicate: [Free response]

Please specify whether you want to withdraw from the study. Recall that you will be anonymous to the researchers, and that when the data in this study will become "open data", we will NOT include your name or demographic questions in the public data uploaded.

o Yes, you may use my anonymized data in this research
o No, please do NOT use my data in this research

How should we deliver your payment in the event you are selected for the monetary bonus? (please select one)

○ Amazon US voucher (2)

○ Amazon UK voucher (3)

○ Amazon DE voucher (4)

○ Paypal account (1)

[Page Break]

Consortium Co-authorship

Completing the entire survey qualifies you to be listed as a consortium co-author on the manuscript reporting the results. Would you like to be listed as a co-author on the final project report?

○ Yes, I would like to be listed as a co-author.

○ No, I would not like to be listed as a co-author.

First name as you would like it to appear on the final project report: [Free response text box]

Last name as you would like it to appear on the final project report: [Free response text box]

Middle initial as you would like it to appear on the final project report: [Free response text box]

Institutional affiliation as you would like it to appear on the final project report: [Free response text box]

[Page break]

Feedback

If you have any feedback on this forecasting survey, please provide it using the space below. [Free response text box]

Supplement 9: Detailed Report of the Forecasting Results

Methodological details

Materials. We asked the respondents to the forecasting survey to each make a total of 24 predictions about effect sizes in terms of Cohen's d as well as the direction of the effect: two predictions focusing on simple effects of target gender (separately by evaluator gender), six predictions focusing on interaction effects (separately by evaluator gender), and 16 predictions focusing on moderator effects. Effect sizes were bounded between -3 and 3. The forecasters were also asked to answer a set of questions capturing their personal beliefs on gender-related topics as well as assessing their demographics.

All the relevant study materials were fully disclosed to the forecasters, including detailed information about the sample sizes, sample characteristics, study design and materials (including links to complete study materials and pre-analysis plans), and links to the original articles targeted for replication.

Recruiting forecasters. We targeted researchers with training in judgment and decision making/social psychology research to participate in the forecasting survey, with no exclusion based on seniority or any other demographic characteristic. We posted the link to a signup page for the forecasting survey on various academic websites, and online platforms and Facebook pages aimed at researchers in psychology, judgment and decision making and research methodology (e.g., Psych Map, Psych Methods Discussion Group, Judgment and Decision Making list). We also asked colleagues on Twitter with many followers to post the link to the signup page. Once signing up, respondents received an individualized link to the forecasting survey. This link allowed them to start and continue with the survey at multiple occasions. Respondents also received at least two reminders to finish the survey.

Respondents were incentivized to participate in two ways: they were offered coauthorship on the study report via a consortium credit, and two randomly selected forecasters were rewarded with a bonus payment determined as a function of the accuracy of their forecasts using the following scoring rule:

$$\$200 -(\text{Sq.Error } 200)$$

where Sq.Error is the average of the squared errors for all the 24 forecasts of the 'Gender and Hiring Decisions Forecasting Study' made by the forecasters.

An initial group of 354 individuals signed up for the forecasting survey, out of which 194 completed the survey, while 111 started but did not complete the survey. 59.8% of the forecasters reported that they were men, 37.1% that they were women, and 1.5% chose ´Other´ and 1.5% chose ´Prefer not to tell.´ The average number of years after the PhD was 4.9 years (SD = 6.4). Note that the sample size and composition in an online survey of this kind is not under the control of the investigators. One has to accept whatever sample size and statistical power is achieved. Our final sample size was comparable to past academic forecasting surveys (e.g., Landy et al., 2020).

Results

Hypothesis tests. The planned analyses are outlined in our pre-analysis plan on https://osf.io/nz48k/ and in Supplement 7. In the report below, we follow the pre-analysis plan unless otherwise specified.

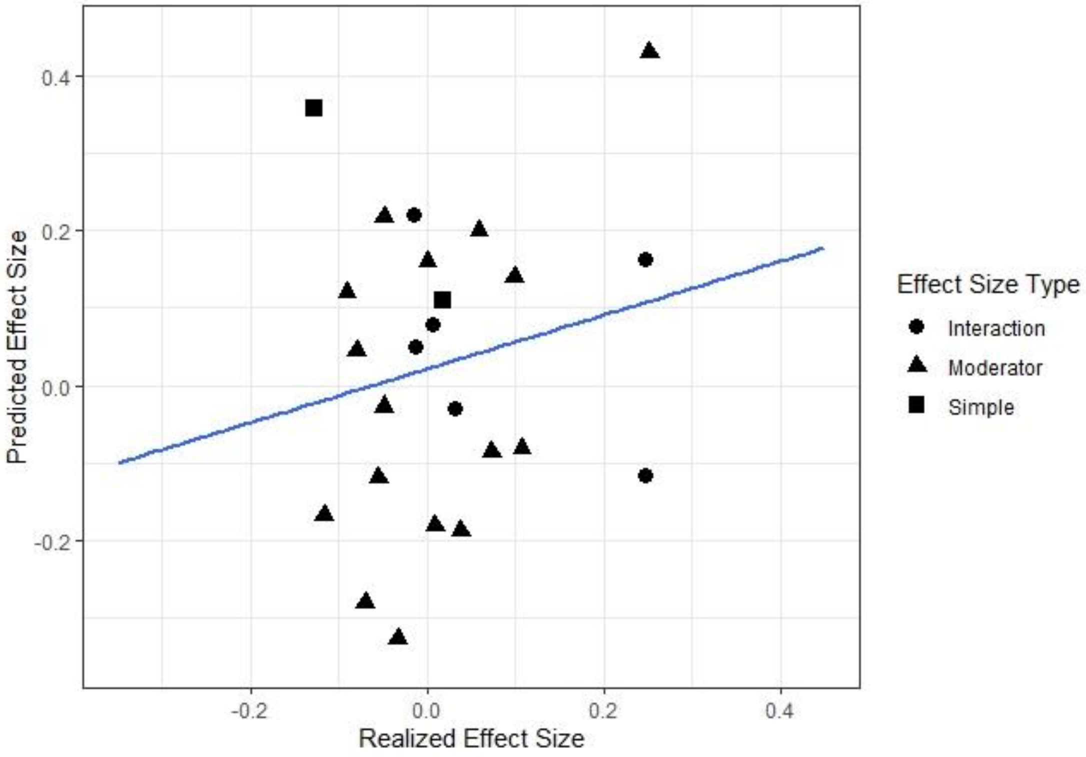
Our primary hypothesis 1 for the forecasting survey was that there would be a positive association between the predictions (beliefs) of the forecasters and the observed effect sizes. The individual-level regression and the t-test confirm that there is a positive and statistically significant association between the predictions of the forecasters and the observed effect sizes, with $\beta_1 = 0.027$ and $p < 0.0001$. See Table S9-1 for the individual-level regression estimates and Figure S9-1 for the correlation ($r = 0.193$, $p = 0.366$) between the average predicted effect sizes and the realized effect size.

Table S9-1. Correlation between forecasted and observed effect sizes.

	Dependent variable: Realized effect size
Forecasted effect size	0.027** (0.004)
Observations	4656
R^2	0.009

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at individual level.

Figure S9-1: Correlation between realized effect sizes and mean predicted effect sizes.



Our primary hypothesis 2 was that forecasters could predict complex experimental results, such as interaction effects between conditions and individual differences moderators. For this we compute the *accuracy* achieved in each forecast by each forecaster in terms of squared prediction error (Brier score). In the regression of the Brier score we find that both coefficients on the forecasts regarding interactions and the effects of the moderators are statistically significant but, contrary to expectations, negative, relative to predictions for simple effects. The coefficient on the variable identifying the forecasts regarding interaction effects is β = -0.079 with *p* = 0.0002 and that of the variable identifying the forecasts regarding the effects of the moderators is β = -0.094 with *p* = 0.0036. See Table S9-2. Surprisingly, the results suggest that forecasters are able to predict experimental results and their accuracy is higher (lower Brier Score) for complex results such as interaction and moderator effects compared to simple effects. The Wald test cannot reject the null hypothesis that the two coefficients are equal (*p* =0.395).

Table S9-2: Forecasts of interaction effects and moderators in terms of squared prediction error (Brier score).

	Dependent variable:
	Brier Score
Forecasts regarding interactions	-0.079** (0.017)
Forecasts regarding the effects of the moderators	-0.094** (0.016)
Observations	4656
R^2	0.008

Note: ∗*p* < 0.05; ∗∗*p* < 0.005. Standard errors clustered at individual level.

Additional analyses. We preregistered several ancillary exploratory hypotheses, all reported below in addition to one test that was not preregistered. As reported in the main text, we explore whether the forecasters' political beliefs and convictions about gender (sexist beliefs measure; beliefs about gender in the workplace; feminist media exposure measure; internal motivation to respond without sexism; external motivation to respond without sexism; political liberalism-conservatism on social issues; see supplements 2, 4, and 8 for more details on the measures) and the forecasters' demographic characteristics (gender where female is coded as 1 and the other three categories as 0, academic seniority measured by years since PhD) relate to the accuracy of their forecasts using the individual accuracy measure from hypothesis 2 (the Brier Score). Because there are so many of these individual-differences measures, we consider these analyses exploratory even though they were preregistered.

See Table S9-3 for the summary statistics of the individual differences variables in the sample of forecasters.

Table S9-3: Summary statistics of measures in the exploratory hypotheses.

Variable	Mean	SD
Sexist beliefs measure	2.90	1.33
Feminist media exposure measure	5.05	1.13
Beliefs about gender in the workplace measure	5.52	1.06
Internal motivation to respond without sexism	5.785	1.11
External motivation to respond without sexism	3.10	1.67
Political orientation measure	2.57	1.20
Years since PhD	4.88	6.36

Further analyses indicate that none of the variables above are statistically significantly related to the accuracy of the forecast: sexist beliefs measure β = - 0.035, *p* = 0.275, feminist media exposure β = -0.015, *p* = 0.415, beliefs about gender in the workplace measure β = -0.014, *p* = 0.612, internal motivation to respond without sexism measure β = -0.002, *p* = 0.813, external motivation to respond without sexism measure β = -0.011, *p* = 0.182, political orientation measure β = 0.022, *p* = 0.095, gender in the workplace measure β = 0.028, *p* = 0.636, and years since PhD measure β = -0.006, *p* = 0.183. See Table S9-4.

Table S9-4: Forecaster beliefs and demographics on squared prediction error (Brier Score).

	Dependent variable:
	Brier Score
Sexist beliefs measure	-0.035 (0.032)
Feminist media exposure measure	-0.015 (0.018)
Beliefs about gender in the workplace measure	-0.014 (0.028)
Internal motivation to respond without sexism	-0.002 (0.010)
External motivation to respond without sexism	-0.011 (0.008)
Political orientation measure	0.022 (0.013)
Female forecaster	0.028 (0.060)
Years since PhD	-0.006 (0.004)
Constant	0.412 (0.396)
Observations	4656
R^2	0.013

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at individual level.

We also test whether predictions regarding gender discrimination in hiring by male evaluators differ from those regarding gender discrimination in hiring by female evaluators, in terms of levels and accuracy. This allows us to test whether the predictions about the hiring evaluations made by men or women are more accurate. In this analysis we only look at the predictions of the simple effect of candidate gender as the main test (one test), and on the predictions of the interaction effects as secondary tests (three tests). The results suggest that the predictions of simple effects and interactions effects are different for male and female evaluators (simple effect of candidate gender mean of the differences = 0.248 and $p < 0.0001$, affirmation-threat mean of the differences = 0.112, $p = 0.002$, objectivity vs. neutral mindset mean of the differences = -0.085, $p = 0.007$, priming stereotypes vs. neutral concepts mean of the differences = 0.140, $p = 0.0003$). In terms of accuracy, respondents have less accurate predictions regarding the simple effect of candidate gender for male evaluators vs. female evaluators ($p < 0.0001$), and forecasters are again less accurate for male evaluators relative to female evaluators for two of the three interaction effects (affirmation-threat $p = 0.191$, objectivity vs. neutral mindset $p < 0.0001$, priming stereotypes vs. neutral concepts $p = 0.0005$).

Robustness tests. We estimate hypothesis 1 separately for the three sets of predictions: predictions on simple effects, on interaction effects, and on moderator effects. For the predictions of simple effects there is a statistically significant negative correlation (β = -0.150 and *p* = 0.0007) with realized effect sizes, as well as for the interaction effects (β = -0.034, *p* = 0.010), while for the moderator effects the correlation remains positive and statistically significant (β = 0.064, *p* < 0.0001 respectively). See Table S9-5.

Table S9-5: Robustness test for hypothesis 1 for predictions on simple effects (1), interaction effects (2), and moderator effects (3) separately.

	Dependent variable: Realized effect size		
	(1)	(2)	(3)
Forecasted effect size	-0.150** (0.019)	-0.034** (0.011)	0.064** (0.004)
Observations	388	1164	3104
R^2	0.253	0.010	0.005

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at individual level.

For hypothesis 1 we also pre-registered a robustness test where we estimate the Pearson correlation between the mean predicted effect size of each of the 24 effects replicated and the realized effect sizes. As noted in the main text, this correlation is positive (0.193) but not significant (*p* = 0.366).

For the exploratory hypothesis on whether forecasters' demographics and their convictions about gender relate to their accuracy in predicting the effect sizes we also estimate it separately for the three sets of predictions (predictions on simple effects, on interaction effects, and on moderator effects). We again find that none of the forecasters' characteristics is statistically significantly associated with their accuracy. See Table S9-6.

Table S9-6: Forecaster beliefs and demographics on squared prediction error (Brier Score) for predictions on simple effects, interaction effects and moderator effects separately.

	Dependent variable: Brier Score		
	(1)	(2)	(3)
Sexist beliefs measure	-0.026 (0.024)	-0.041 (0.030)	-0.033 (0.034)
Feminist media exposure measure	-0.028 (0.032)	-0.014 (0.022)	-0.014 (0.018)
Beliefs about gender in the workplace measure	0.017 (0.036)	-0.002 (0.028)	-0.022 (0.029)
Internal motivation to respond without sexism	-0.006 (0.015)	-0.006 (0.014)	0.000 (0.009)
External motivation to respond without sexism	-0.017 (0.016)	-0.009 (0.009)	-0.011 (0.008)
Political orientation measure	0.042 (0.041)	0.037 (0.020)	0.013 (0.010)
Female	0.160* (0.077)	0.081 (0.064)	-0.007 (0.063)
Years since PhD	-0.002 (0.004)	-0.004 (0.004)	-0.007 (0.005)
Constant	0.281 (0.268)	0.316 (0.366)	0.464 (0.429)
Observations	388	1164	3104
R^2	0.032	0.018	0.013

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at individual level.

We also carried out a regression that was not specified in the pre-analysis plan, where the focus is on whether forecasters' demographics and their convictions about gender relate to their accuracy in predicting the effect sizes on the simple effect of candidate gender among male evaluators only. Again we find no statistically associations with accuracy. In particular, forecasters' accuracy regarding gender discrimination by male evaluators was not associated with any of the following: forecasters' own sexist beliefs ($p = 0.380$), the feminist media exposure measure ($p = 0.939$), beliefs about gender in the workplace measure ($p = 0.897$), internal/external motivation to respond without sexism ($p = 0.478$ / $p = 0.735$), and political orientation ($p = 0.566$). See Table S9-7.

Table S9-7: Forecaster beliefs and demographics on squared prediction error (Brier Score) for main effect of candidate gender on male evaluators only.

	Dependent variable:
	Brier Score
Sexist beliefs measure	-0.023 (0.026)
Feminist media exposure measure	-0.002 (0.023)
Beliefs about gender in the workplace measure	0.004 (0.030)
Internal motivation to respond without sexism	-0.015 (0.021)
External motivation to respond without sexism	-0.005 (0.016)
Political orientation measure	0.016 (0.027)
Female forecaster	0.200** (0.061)
Years since PhD	-0.005 (0.004)
Constant	0.387 (0.311)
Observations	194
R^2	0.060

Note: ∗p < 0.05; ∗∗p < 0.005. Standard errors clustered at individual level.