Creative destruction in science

Warren Tierneya,⁎, Jay H. Hardy IIIb, Charles R. Ebersolec, Keith Leavittd, Domenico Viganolae, Elena Giulia Clementef, Michael Gordonf, Anna Dreberh, Magnus Johannessonf, Thomas Pfeifferg, Hiring Decisions Forecasting Collaboration, Eric Luis Uhlmanna,⁎

a INSEAD, Singapore
b Oregon State University, United States
c University of Virginia, United States
d Oregon State University, United States
f The World Bank
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
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 (Berg, 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 (Berg et al., 2017; Bosco, Aguinis, Field, Pierce, & Dalton, 2016). Accordant, 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 spaces. Within this approach, a finding in support of propositions 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, & Donnelian, 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, Schlang, & Asadullah, 2011), economics (Camerer et al., 2016; Chang & Li, 2017; McCullough, McGeeary, & 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 (Agnoi, 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 resource 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
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 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 and weak relationships and qualifying conditions. Although the core itself may not appear directly in jeopardy, the predictive belt of a questionable 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 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). Although management has lagged behind 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 these stereotypes in the first place (Payne, Vuletic, & Brown-Iannuzzi, 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 socialreceiver’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 sub-regions within nations are often just as, if not more, important than national borders when it comes to delineating cultural boundaries (Harrington & Geland, 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 (Fischer, 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,
Only the United States and United Kingdom (as in 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 (Baker, 2005; Lipset, 1996) and the validated Duke University Religion Index (DUREL) assessment of religious beliefs (Koenig & Büssing, 2010).

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 of such efforts, a new set of theoretical predictions emerged.
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. | Contrary to this theory’s predictions, 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. | Contrary to 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. |

(continued on next page)
Table 2 (continued)

| THEORY | NEEDLESS WORK EFFECT | INTUITIVE WORK MORALITY EFFECT | TACIT INFERENCE EFFECT | OVERALL ASSESSMENT |
|--------|-----------------------|-------------------------------|-----------------------|-------------------|
| Self-expression values perspective | No support | Consistent with the theory, while the intuitive work morality effect would normally be associated with self-expression values (USA, UK, Australia), this theory's predictions are supported for one of the three effects targeted for replication (intuitive work morality effect). | Did not anticipate the tacit inferences effect would replicate across four societies. | 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 the three effects targeted for replication (intuitive work morality effect). |
| Social class perspective | No support | No support for the social class perspective. | Formally educated participants were not consistently more likely to exhibit the intuitive mindset moderating judgment of needless work. | No support for the idea that social class moderates the effects of target age on moral judgments of needless work. |

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 contradicts 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 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).
response to new evidence. However, we believe it is informative re
motive factors (hoped-for and feared outcomes) were placed in
matching) suggests day care is not disadvantageous for children. When
whichever methodology (random assignment versus statistical
methods and results of purported scientific studies on the topic, favor
ment advantages of home care conflict with their earnest hope that
pectancy that rigorous scientific research will support the develop
who believe home care is better for children, yet intend to place their
what they believe is factually true are diametrically opposed are highly
such situations in which what a person wants to be true and
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 expec-
tancy that rigorous scientific research will support the develop-
mental 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
whichsoever 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 re-
garding the motivation-cognition debate.

At the same time, other works 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 situ-
tionally 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 re-
plication-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. sta-
tistical matching) they considered most valid before learning the results
of scientific studies of the effects of home care vs. day care that em-
ployed those methodologies. Pre-commitment to criteria should con-
strain reasoning (whether based on cognitive beliefs or motivated de-
sires), 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 rig-
orous and dispassionately than non-parents.

In contrast, theories of motivated reasoning make the directly op-
posing 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 do-
main 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 Rea-
soning, Cognitive Schema, and Accuracy-Driven perspectives on reasoni-
ging 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 mod-
erators about which competing theories make opposing predictions
(e.g., parental status). The pre-registered analysis plans and study ma-
terials are available on the OSF (https://osf.io/9fjy8m) and in
Supplement 1, and the data and code are likewise posted online (data:
https://osf.io/hq45v/, analysis code: https://osf.io/rphv7/). Notably,
the creative destruction analyses were formulated and pre-registered
after the Ebersole (2019) data collections were carried out, thus this

| 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 trumps beliefs in driving reasoning. | Desires only appear to influence reasoning because they are aligned with beliefs; when misaligned, beliefs trumps 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.
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 do 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.

### 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...
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; Hanselmann 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; DellàVigna & 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 (β = 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 SAVINESS 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 hiring criteria in favor of female candidates, and previous experience with research studies predicted pro-female hiring decisions. | 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 replying 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 context, 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 Uhmann 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 Forskningsstiftelse), 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. Domenico Viganola: Conceptualization, Methodology, Formal analysis, Writing - review & editing. Anna Dreber: Conceptualization, Methodology, Formal analysis, Writing - review & editing. Michael Gordon: Conceptualization, Methodology, Formal analysis, Writing - review & editing. Elena Giulia Clemente: 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 Uhmann: 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 Báhník, 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 Bavoler, Pavol Jozef Safárik 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
- Michal Bialek, Institute of Psychology, University of Wroclaw
- Michael M. Bishop
- Helena Bonache, Universidad de La Laguna
- Sabah Boulkhaled, 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. Crumblit 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-Avalit, Ben-Gurion University of the Negev
- Martin V. Day, Memorial University of Newfoundland
- Matthew D. Deeg, University of Kansas
- Tara C. Dennehly, University of British Columbia
- Erik Dietl, Loughborough University
- Eugen Dimant, University of Pennsylvania
- Artur Domurat, Centre for Economic Psychology and Decision Sciences, Kozminski University
- Christielen de Plessis, Singapore Management University
- Dmitrii Dubrov, National Research University Higher School of Economics
- Mahmood 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 Friedman, Ben-Gurion University of the Negev
- Omid Ghasemi, University of Birmingham
- Ajay T. Abraham, Seattle 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 Hakenbrädl, IESE Business School
- David Hagmann, Harvard Kennedy School
- Andrew H. Hales, University of Virginia

W. Tierney, et al. Organizational Behavior and Human Decision Processes 161 (2020) 291–309

305
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 reproducing 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.
Agnina, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., et al. (2014). Registered replication report: Schoolder & Engstoller-Schloffer (1990). Perspectives on Psychological Science, 9(5), 556–578.
Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Aliberti, 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.
Armstrong, 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), 639–660.
Baker, W. (2005). America’s crisis of values. Princeton, NJ: Princeton University Press.
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., Goeckner, 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.
Barth, B. W., & Grant, A. A. (2006). Service with a smile and encounter satisfaction: Emotional contagion and appraisal mechanisms. Academy of Management Journal, 49(6), 1229–1238.
Barbier, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., et al. (2018). Meta-studies for robust tests of theory. Proceedings of the National Academy of Sciences, 115(21), 5561–5564.
Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? Science, 364(6442), 93–99.
Brandt, M. J., Uzerman, H., Dijksterhuis, A., Farach, P., 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.
Braicred, C. A., & Reyna, V. F. (2018). Replication, registration, and scientific creativity. Perspectives on Psychological Science, 13(4), 428–432.
Byrington, E. K., & Felps, W. (2017). Solutions to the credibility crisis in management science. Academy of Management Learning & Education, 16(1), 142–166.
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., Forthor, D. R., Beiler, A. M. C., & Baldaris, T. L. (in press). Gray (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, 1453–1456.
Camerer, C. F., Dreber, A., Holmztein, 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., Verhoeven, E., & Burks, S. (2005). The effects of stakes in distribution experiments. Economics Letters, 86(3), 393–398.
Cash, I. 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 conspiratorial 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., Hirn, E. R., & Gerhke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. Journal of Personality and Social Psychology, 58(6), 929–963.
DellaVigna, S., & Pope, D. G. (2018). Predicting experimental results: Who knows what? Journal of Political Economy, 126(6), 2410–2456.
Desi, 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. L., Grady, R. H., & Ziegler, 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., Issakson, S., Wilson, B., Chen, Y., Nosek, B. A., 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.
Eagle, J., Nazer, C., Miller, D. I., Kaufmann, M., & Sazensy, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. American Psychological Association, 75(1), 301–315.
Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. Perspectives on Psychological Science, 14(2), 330–351.
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., & Elster, J. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology, 65(1), 1–31.
Eysenck, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. Perspectives on Psychological Science, 14(2), 273–291.
Fisher, D. H. (1989). Albinon’s seed: Four British folkways in America. New York, NY: Oxford University Press.
Forsell, E., Vigano, J., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., et al. (2019). Predicting replication outcomes in the Many Labs 2 study. Journal of Experimental Social Psychology, 75, 102–117.
Fraser, M. R., Parker, T., Nakagawa, S., Barnett, A., & Filid, F. (2018). Questionable research practices in ecology and evolution. PloS ONE, 13(7), e0200033.
Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. Psychological Bulletin, 101(1), 90–117.
Garber, M. (2017). All the angry ladies. The Atlantic. Available at: https://www.theatlantic.com/entertainment/archive/2017/11/all-the-angryladies/545042/.
Garner, A. (2015). The connection between varying treatment effects and the crisis of unreliable research: A Bayesian perspective. Journal of Management, 41(2), 222–243.
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
Protzel, J., & Aronsen, J. (2016). Context moderates affirmation effects on the ethnic achievement gap. Social Psychological & Personality Science, 7(6), 500-507.

Ragagopalan, 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. PloS One, 3(7), 1-7.

Reed, S. J., Dang, C. T., Yen, K. C., & Leavitt, K. (2014). Sample size considerations for structural equation modeling: Fact, fictions and truth. In C. E. Lance, & R. J. Vandenberg (Eds.). Advances in Experimental Social Psychology: Vol. 38, pp. 169–242. New York, NY: Academic Press.

Reichert, J. P., & Reilly, B. (2016). The role of moral knowledge in everyday immorality: What does it matter if I know what is right? Organizational Behavior and Human Decision Processes, 132(2), 124-137.

Ronsin, J., 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., & Ferretti, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. Organizational Research Methods, 12(2), 347-367.

Schlagel, C., & Koenig, M. (2014). Determinants of entrepreneurial intent: A meta-analytic test and integration of competing models. Entrepreneurship Theory and Practice, 38(2), 291-322.

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.

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

Schumpeter, J. A. (1942/1994). Capitalism, socialism and democracy. London: Routledge.

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.

Sokladis, 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(22), 2270-2274.

Stewart, W., & Barling, J. (1996). Fathers’ work experiences effect children’s behaviors via work motivation. Organizational Research Methods, 23(2), 238-274.

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

Van de Walle, J. A., & 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.

Vandeweghe, D., Vanderschelden, A., & Van Offelt, 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.

Vandenbosch, R., & Grell, 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., Wanner, H., & Vosgerau, J. (2017). Making replication a routine part of psychological research. Psychological Science, 28(6), 825–828.