Creative Destruction in Science

Warren Tierney
INSEAD

Jay H. Hardy III
Oregon State University

Charles R. Ebersole
University of Virginia

Keith Leavitt
Oregon State University

Domenico Viganola
The World Bank

Elena Giulia Clemente
Stockholm School of Economics

Michael Gordon
Massey University

Anna Dreber
Stockholm School of Economics, University of Innsbruck

Magnus Johannesson
Stockholm School of Economics

Thomas Pfeiffer
Massey University

Hiring Decisions Forecasting Collaboration

Eric Luis Uhlmann

INSEAD

Corresponding Authors:
Warren Tierney & Eric Luis Uhlmann, INSEAD, Organisational Behaviour Area, 1 Ayer Rajah Avenue, 138676 Singapore, Phone: 65 8468 5671, E-mails: warrentierney@hotmail.com, eric.luis.uhlmann@gmail.com.

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 1. All authors collaboratively edited the paper.

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

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.

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

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.

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, p. 649).

Concerns about theory proliferation are compounded by the limited number of studies focusing on replication (Bergh, Sharp, Aguinis, & Li, 2017; Brandt, Ijzerman, Dijksterhuis, Farach, Geller, Giner-Sorolla, Grange, Perugini, Spies, & van't Veer, 2014; Earp & Trafimow, 2015; Lykken, 1968; Tsang & Kwan, 1999), and new findings regarding a general lack of replicability within organizational scholarship (Bergh et al., 2017; Bosco, Aguinis, Field, Pierce, & Dalton, 2016). Accordingly, commentators have recently described the risk of a crisis of confidence in organizational research (Gelman, 2015; Köhler & Cortina, in press). Thus, while scholars continue to generate new theory at an accelerated pace, their propositions typically enjoy preliminary rather than definitive support, and are rarely

subjected to attempts at direct replication (Schmidt, 2009; Simons, 2014) or placed in competition against adjacent (and sometimes contradictory) theories.

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

In each case, high powered and in some cases cross-national samples, combined with pre-registered (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) empirical predictions from each theoretical perspective, allow for strong inferences (Platt, 1964) in the

absence of publication bias (Kvarven et al., in press). In addition to repeating the original design, we systematically include further measures, conditions, and populations, allowing for novel tests of competing theoretical accounts operating in the same domains. We suggest that the creative destruction paradigm can serve the long sought goal of encouraging the development of new theories and insights for the study of management and organizations, while also rigorously pruning and bounding theories as they emerge (Porter, 1996).

The need for theory pruning in management scholarship

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

Pitting competing empirical predictions against one another in the same experimental paradigm provides the opportunity to bound, qualify, and reduce theory (Aguinis, Pierce, Bosco, & Muslin, 2009; Hambrick, 2007; Kluger & Tikochinsky, 2001; Van de Ven & Johnson, 2006; Vandenberg & Grelle, 2008). By directly considering and testing theories in tandem, scholars are able to determine the necessity of additional constructs introduced by a novel theory, or identify which of two theories provides predictive validity across a broader range of criteria (Leavitt et al., 2010). Such an approach may generate support for one theoretical explanation over another (Schlaegel & Koenig, 2014), reconcile apparent

contradictions that are later explained by differences in assumptions underlying divergent theoretical orientations (Peteraf et al., 2013), or facilitate new discovery by identifying previously hidden moderators that emerge when one theory directly antagonizes another (Latham, Erez, & Locke, 1988).

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

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

The crisis of confidence in science

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

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

and other fields (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Bedian, Taylor, & Miller, 2010; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; John, Loewenstein, & Prelec, 2012; Makel, Hodges, Cook, & Plucker, 2019; Ramagopalan et al., 2014).

These concerns surrounding replication and research practices appear similarly relevant within myriad organizational literatures and across management research methodologies (Aguinis & Solarino, 2019; Bamberger, 2019; Bergh et al., 2017; Pratt, Kaplan, & Whittington, 2019). While our search was unable to identify a systematic assessment of the direct replicability of organizational behavior or human resources research, a survey by Bedian, Taylor, and Miller (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.

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 (Petty & Cacioppo, 2016; Schnall, 2014; Schwarz & Strack, 2014; Strack, 2016; Stroebe & Strack, 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.

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

Fourth, creative destruction draws conclusions from especially large sample sizes, as per the lessons of recent replication initiatives (Alogna et al., 2014; Klein et al., 2018). The problem of under-powered studies is well known within management, such that equivocal results are often observed across investigations due to both Type I and Type II errors (Cashen & Geiger, 2004; Scherbaum & Ferreter, 2009). Further, each competing theory is expected to

make predictions about both significant relationships and weak to minimal relationships among the host of included variables and conditions. Thus, no theory has the unfair advantage of predicting only null effects, which can be confounded by problems with the measures or samples.

Epistemologists have long argued that falsification tests play a critical role in advancing scientific knowledge (Kuhn, 1962; Popper, 1959). Although management has lagged behind some other sciences in doing so, strong inference comparisons between theories have long been an acknowledged goal of organizational science (Davis, 2006). Tests which allow for the immediate support of one theory and rejection of the core arguments of another are likely to remain uncommon for myriad reasons (Leavitt et al., 2010), but the creative destruction approach may accelerate the ultimate abandonment of comparatively weaker theories. Science can generally not prove a theory correct or incorrect, but it can falsify propositions or statements which emerge from the theory (Lakatos, 1978; Popper, 1959). Lakatos (1978) 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.

Example 1: Culture and work morality

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

Relevant experimental research (Poehlman, 2007; Uhlmann et al., 2009) finds that moral character inferences about a lottery winner who continues to work in the absence of any material need are highly favorable. Further, among Americans but not Mexicans, this "needless work" effect is sensitive to target age, such that a 23 year old lottery winner who continues to work is praised more than a 46 year old who does the same. Presumably it is more legitimate, from the standpoint of the Protestant work ethic, to retire after already contributing decades of hard work. Another theoretically expected moderator of moral

judgments based on needless work is the social perceiver's mindset. Specifically, thoughtless, automatic processing should promote the expression of implicit cultural work values. Consistent with this idea, American participants are especially likely to morally praise a person who continues to work after a windfall lottery win when making judgments intuitively rather than deliberatively.

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

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

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

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

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

Yet another competing theoretical perspective argues that subregions within nations are often just as, if not more, important than national borders when it comes to delineating cultural boundaries (Harrington & Gelfand, 2014; Kitayama, Ishii, Imada, Takemura, & Ramaswamy, 2006; Nisbett & Cohen, 1996; Talhelm et al., 2014; Vandello & Cohen, 1999).

Of particular relevance here, the *Regional Folkways* perspective (Fisher, 1989) argues there are multiple U.S. cultures—Puritan influenced New England, the plantation culture of the South (shaped by English gentry), the industrial culture of the Midwest (shaped by Quaker influence), and the ranch culture of the American West (shaped by Scotch-Irish migration). If so, then Puritan-Protestant morality effects originally predicted by the theory of Implicit Puritanism should be strongest in the New England region of the United States.

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

With regard to cultural divides within national borders, research highlights the importance of social class differences (Snibbe & Markus, 2005; Stephens, Fryberg, & Markus, 2011). Both within the United States and other nations (e.g., Italy, Poland, Ukraine, Russia, and Japan), low socio-economic status (SES) individuals are more relationally oriented and deferent to authority than individuals with a higher income and more formal education (Grossmann & Varnum, 2011). Particularly relevant here, low-SES people also tend to regard work instrumentally, in other words as a means of earning income rather than a source of meaning and fulfillment (Argyle, 1994; Williams, 2012). This *Social Class* perspective thus suggests the tendency to valorize needless work may characterize high-SES individuals across societies. The original investigations of Implicit Puritanism (Poehlman, 2007; Uhlmann et al., 2009, 2011) did not observe any reliable individual differences based on religion, religiosity, or socioeconomic status, but relying on small samples were

potentially underpowered to detect them. The creative destruction replications conducted by Tierney et al. (2019) allowed for high powered tests of all these plausible accounts of work morality across cultures (see Table 1 for an overview).

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

The results of the cross-national data collection, encompassing over 5,000 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 the replication initiative, Implicit Puritanism suffers a theoretical core breach.

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

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

Example 2: Working parents' reasoning about child care choices

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

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

The cognitive vs. motivational underpinnings of such information processing are extremely difficult to parse—in fact, Tetlock and Levi (1982) pronounced the motivation-cognition debate potentially intractable. Are participants, again potentially quite rationally (Baron & Jost, 2019; Krueger & Funder, 2004), less likely to cognitively accept new information that contradict their priors? Or, are they truly contorting the evidence and standards in order to believe what they want to believe? For example, decisions about parenting and family arrangements impact the attitudes and behaviors of employees at work (Desai, Chugh, & Brief, 2014), and work experiences similarly spill over into parenting behaviors (Stewart & Barling, 1996). Satisfaction with child care arrangements are a critical predictor of work-family conflict and consequent absenteeism (Goff, Mount, & Jamison,

1990). Thus, child care represents a critical domain in which employees should be motivated to invest substantial cognitive resources and seek to optimize their outcomes, but how such decisions are made would be differentially predicted by various theories of reasoning.

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

At the same time, other work supports the importance of accuracy driven reasoning (Devine, Hirt, & Gehrke, 1990; Funder, 1987; Jussim, 1991; Trope & Bassok, 1982). From the standpoint of evolutionary adaptiveness, it follows that humans come equipped with reasoning abilities to help us construct a fairly veridical internal representation of the external

world. If so then accuracy goals, either chronic or situationally activated in important situations, should explain the bulk of the variance in how human beings process evidence.

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

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

In contrast, theories of motivated reasoning make the directly opposing prediction, postulating that rationalizations for child care choices should be more evident among actual parents than would-be parents. Festinger's (1962) theory of cognitive dissonance suggests that having already committed to a course of action in a consequential domain should increase the desire to justify one's decisions. This suggests that parents who have already

entrusted their children to day care should be more, not less, prone to motivated reasoning in this domain.

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

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

The wishful thinking paradigm's approach to teasing apart cognitive and motivational explanations for assimilation effects focuses on "conflicted" participants who either have children in day care or expect to one day, yet believe home care is better for children's development. Such individuals' cognitive beliefs in the superiority of home care are in

conflict with their motivated desire to find out that day care is just as good. Our re-analyses of Ebersole (2019, Study 6) failed to replicate the original wishful thinking effect that desired outcomes trump factual beliefs in the assimilation paradigm. Directly contrary to the striking pattern reported by Bastardi et al. (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 and 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 and Cohen (2005, Study 1) and Uhlmann and 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; Hanselman, Rozek, Grigg, & Borman, in press; Protzko & Aronson, 2016).

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 et al., 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., 2018; 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. Supplement 2, 3, and 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, 8, and 9 for the forecasting survey materials, pre-registered analysis plan, and detailed report). Prior work finds that scientists are able to accurately predict simple condition differences by merely reading the study abstract or examining the study materials (Camerer et al., 2016;

DellaVigna & Pope, 2018; Dreber et al., 2015; Forsell et al., 2019). We tested, for the first time, whether scientists can likewise anticipate complex interactions between variables. In this politically charged context (Tetlock, 2015), we further examined whether scientists' beliefs and values regarding gender moderate the accuracy of their predictions. Consistent with past research, in our primary pre-registered hypothesis test, we found a positive association between the observed effect sizes and the individual predictions (beliefs) of the forecasters ($\beta = 0.027$, $p < 0.001$). In a pre-registered robustness test, aggregated predictions, computed as mean predicted effect size of each of the 24 effects replicated, were directionally positively associated with the observed effect sizes, although this zero-order correlation was no longer statistically significant, $r = 0.193$, $p = 0.366$. A notable discrepancy between forecasts about selection decisions by male evaluators and the actual study outcomes was also apparent. Forecasters expected that both male and female 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.

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 Figure 1). Past replication efforts have typically compared the original theory to the null (e.g., Klein et al., 2014; Open Science Collaboration, 2015), rather than adding

new measures, conditions, and populations to test multiple theories against each other. Further, past theory pruning efforts in the management literature have generally not relied on direct replication, pre-registration of analyses, and complete data transparency.

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

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

The creative destruction approach is most useful as follow up research to an initial set of published findings—in other words, in the context of replicating or re-examining established research. This approach is meant to create a series of severe tests (Mayo, 2018) for competing theories. Severe tests often require a great deal of resources, both in terms of study design and participant recruitment. As such, the creative destruction approach will be

most effective when there is a set of competing theories with each having an empirical basis of support. Such a basis will allow researchers to effectively design tests of each theory and will hopefully limit wasting resources on theories that were, *a priori*, unlikely to find support.

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

The creative destruction approach, then, is most effective within the context of well developed theories. Whereas many theories within organizational sciences merely predict directional associations between pairs of variables (Vancouver, Wang, & Li, 2018), more precise theories are defined by their boundaries and limitations, including reducing the number of outcomes that would be considered consistent with that theory (Byrd, 2019; Edwards, 2010). Creative destruction, then, will be most useful when theories are already sufficiently bounded, such that the scope of their predictions can be reasonably captured within a short series of studies. Notably, mature areas of research inquiry, which are often those with the most well developed theories, are also the most likely to suffer from theoretical proliferation. This makes them especially good candidates for strong inference comparisons (e.g., Thau & Mitchell, 2010). For highly advanced theories associated with large numbers of published empirical investigations, the creative destruction approach can be employed not

only in novel data collections, but also in the context of meta-analytic tests for publication bias and evidentiary value in competing sets of findings (see Supplement 5). The ideal context, however, is likely to be Registered Reports, in which the methods, predictions, and analytic plan for a study are peer reviewed prior to data collection (He & Côté, 2019).

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, suboptimal 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.

References

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*, 9-34.

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

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

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

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

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

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

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

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

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

Barger, P. B., & Grandey, A. A. (2006). Service with a smile and encounter satisfaction:

Emotional contagion and appraisal mechanisms. *Academy of Management Journal, 49,* 1229-1238.

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

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

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

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

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

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

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

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

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

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's

threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*(3), 709–750.

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

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J., & van ‘t Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224.

Butler, N., Delaney, H., & Spoelstra, S. (2017). The grey zone: Questionable research practices in the business school. *Academy of Management Learning & Education, 16*(1), 94–109.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Garber, M. (2017, November 6). All the angry ladies. *The Atlantic*. Available at: https://www.theatlantic.com/entertainment/archive/2017/11/all-the-angry ladies/545042/

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

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

Goff, S. J., Mount, M. K., & Jamison, R. L. (1990). Employer supported child care, work/family conflict, and absenteeism: A field study. *Personnel Psychology, 43*(4), 793-809.

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

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

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

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

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

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

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

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

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

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

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8): e124. https://doi.org/10.1371/journal.pmed.0020124

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

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

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

James, W. (1890/1950). *The principles of psychology, vols. I and II.* New York: Dover Publications.

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science, 23*(5), 524-532.

Johnson, C.A., & Hawbaker, KT. (2018, May 25). #MeToo: A timeline of events. *Chicago Tribune*. Retrieved at: http://www.chicagotribune.com/lifestyles/ct-me-too-timeline20171208-htmlstory.html

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Leavitt, K. (2013). Publication bias might make us untrustworthy, but the solutions may be worse. *Industrial and Organizational Psychology, 6*(3), 290-295.

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

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

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

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

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

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

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

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

Makel, M.C., Hodges, J., Cook, B.G., & Plucker, J.A. (2019). *Questionable and Open Research Practices in Education Research.* Unpublished manuscript. Available at: https://edarxiv.org/f7srb/

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False– positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366.

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

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

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

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

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

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

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

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

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

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

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

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

Tierney, W., Hardy, J. H., III., Ebersole, C., Viganola, D., Clemente, E., Gordon, M., Hoogeveen, S., Haaf, J., Dreber, A.A., Johannesson, M., Pfeiffer, T., Chapman, H., Gantman, A., Vanaman, M., DeMarree, K., Igou, E., Wylie, J., Storbeck J., Andreychik, M.R., McPhetres, J., Vaughn, L.A., Work Morality Forecasting Collaboration, & Uhlmann, E. L. (2020). A creative destruction approach to replication: Implicit work and sex morality across cultures. Registered Report proposal accepted in principle at the *Journal of Experimental Social Psychology*.

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

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

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

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

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

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

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

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

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

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

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

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

Van't Veer, A., & Giner-Sorolla, R. (2016). Pre-registration in social psychology: A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12.

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

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

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

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

Figure

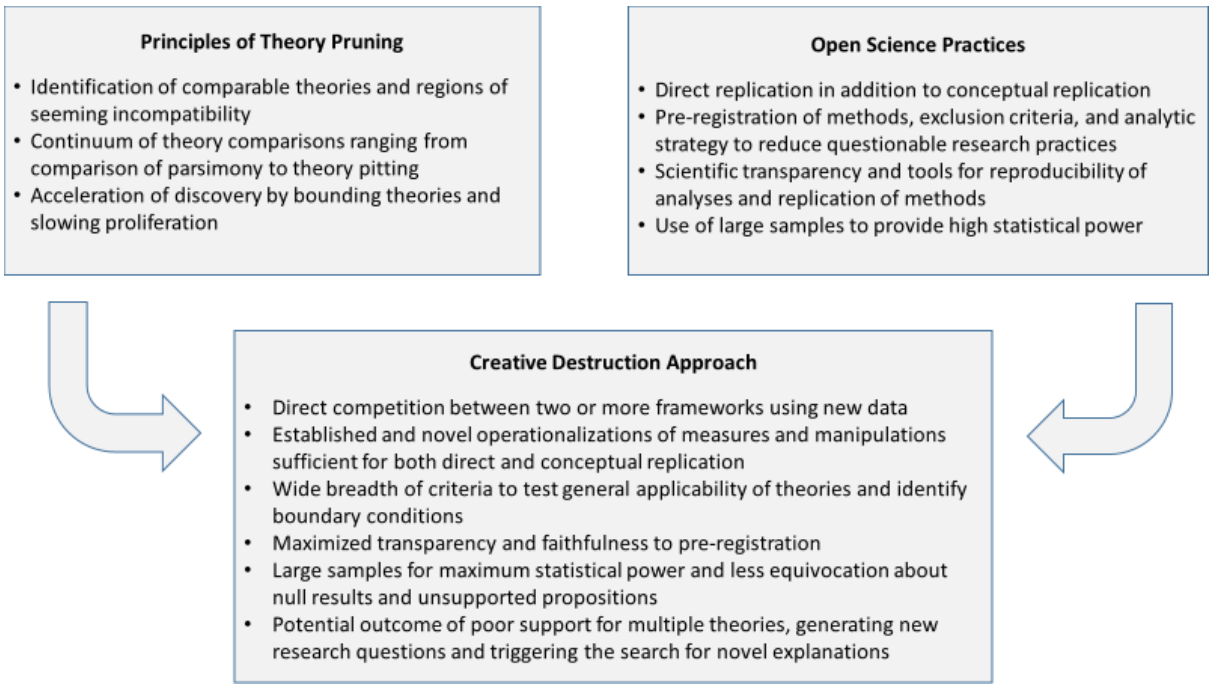

Figure 1. The creative destruction approach to replication, and its roots in theory pruning methods and open science practices.

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

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

THEORY	NEEDLESS WORK EFFECT	TACIT INFERENCES EFFECT	INTUITIVE WORK MORALITY EFFECT
General moralization of work perspective: *People across cultures moralize work*	Both Americans and non-Americans exhibit the needless work effect and are sensitive to target age.	Both Americans and non-Americans exhibit the tacit inferences effect.	Both Americans and non-Americans exhibit the intuitive work morality effect.
False positives perspective: *The original findings are spurious*	No needless work effect or sensitivity to target age, and no moderation by individual differences in religion, religiosity, or sub-region.	No tacit inferences effect and no moderation by individual differences in religion, religiosity, or sub-region.	No intuitive work morality effect and no moderation by individual differences in religion, religiosity, or sub-region.
Self-expression values perspective: *Individuals from wealthy nations moralize work*	Participants from the USA, UK, and Australia should exhibit the needless work effect, whereas Indian participants should not.	This theory does not anticipate the tacit inferences effect.	Participants from the USA, UK, and Australia should exhibit the intuitive work morality effect, whereas Indian participants should not.
Social class perspective: *High-SES persons moralize work*	High socioeconomic status participants should exhibit the needless work effect more than low socioeconomic status participants.	This theory does not anticipate the tacit inferences effect.	High socioeconomic status participants should exhibit the intuitive work morality effect more than low socioeconomic status participants.

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

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

THEORY	NEEDLESS WORK EFFECT	TACIT INFERENCES EFFECT	INTUITIVE WORK MORALITY EFFECT	OVERALL ASSESSMENT
Implicit Puritanism perspective	The theory of Implicit Puritanism's original prediction that a younger person is praised more than an older person for continuing to work after winning the lottery is not supported. In other words, target age does not reliably moderate the "needless work" effect that continuing to work rather than retiring elicits favorable character judgments.	As predicted by the theory of Implicit Puritanism, women and men who fail to uphold traditional work morality are misremembered as violating traditional sex morality (and vice versa). However, sharply contradicting the original theory, the tacit inferences effect is observed not only in the United States, but also Australia, the United Kingdom, and India (although Indians exhibit the effect less strongly than Americans).	As predicted by the original theory, the needless work effect is stronger in an intuitive mindset than in a deliberative mindset. Sharply contradicting the original theory, the intuitive work morality effect is observed not only in the U.S., but also Australia and the United Kingdom. Consistent with the original theory, Indians do not appear to exhibit the effect.	The theory of Implicit Puritanism suffers a theoretical core breach due to the empirical results of the replication initiative. One of the three key effects predicted by the theory (target age and needless work) failed to replicate entirely. Two other effects (tacit inferences and intuitive work morality) did replicate, but were also found across several other nations, contrary to the theory's core claim of a unique American work morality.
Religious differences perspective	Contrary to this theory's predictions, religion (Protestant or not) and religiosity did not make participants more sensitive than others to target age in their judgments of needless work.	No moderating effect of religion (Protestant or not). Religiosity either predicts in the wrong direction (i.e., more religious participants exhibit the tacit inferences effect *less*), or not at all, depending on the sample.	No moderating effect of religion (Protestant or not). Religiosity either predicts in the wrong direction (i.e., more religious participants exhibit the intuitive work morality effect *less*), or not at all, depending on the sample.	No support for the prediction that religious Protestants exhibit the work morality effects targeted for replication more strongly.
Regional folkways perspective	Contrary to this theory's predictions, New Englanders are not more sensitive than others to target age in their judgments of needless work.	Contrary to this theory's predictions, New Englanders are not any more likely to exhibit the tacit inferences effect.	Contrary to this theory's predictions, New Englanders are not any more likely to exhibit the intuitive work morality effect.	No empirical support for the idea that individuals from the New England states exhibit the work morality effects targeted for replication more strongly.
Explicit American exceptionalism perspective	Contrary to the theory's predictions, Americans do not differ from others in terms of sensitivity to target age in judgments of needless work.	Contradicting this theory, Americans are not consistently more likely to exhibit the tacit inferences effect than members of other Western cultures. Further, individuals who explicitly endorse the Protestant work ethic exhibit the tacit inferences effect *less*, directly contrary to predictions.	Contradicting this theory, Americans are not consistently more likely to exhibit the intuitive work morality effect than members of other Western cultures. Further, explicit endorsement of the Protestant work ethic either predicted this effect in the wrong direction or not at all, depending on the sample.	No empirical support for a unique American response to the specific work morality effects studied. The expected pattern of national differences did not emerge, and explicit endorsement of the Protestant work ethic either predicted effects in the wrong direction or not at all.

THEORY	NEEDLESS WORK EFFECT	TACIT INFERENCES EFFECT	INTUITIVE WORK MORALITY EFFECT	OVERALL ASSESSMENT
General moralization of work perspective	No support for the original prediction that target age moderates moral judgments based on needless work; no such pattern is observed across four nations. Strong support for the prediction that across cultures, working in the absence of material need is morally praised.	Strong support for the prediction that across cultures, women and men who fail to uphold traditional work morality are misremembered as violating traditional sex morality (and vice versa). The effect is observed in all four nations studied, although Indian participants make weaker tacit inferences than Americans.	Fairly strong support for the prediction that across cultures, the needless work effect is stronger in an intuitive mindset than in a deliberative mindset. The effect is observed in three out of four nations studied (the US, UK, and Australia, but not India).	Strong empirical support for the prediction work is moralized across cultures, with the caveat that the intuitive work morality effect did not emerge reliably in India (see below under "self-expression values perspective").
False positives perspective	The original finding that target age moderates the needless work effect appears to be a false positive. There is however a robust main effect of needless work on judgments of moral character that replicates across cultures.	The tacit inferences effect is robust across four out of four societies and not a false positive.	The intuitive work morality effect is robust across three out of four societies and not a false positive.	The false positives perspective is supported for one of the original effects targeted for replication. Specifically, the moderating effect of target age on character judgments based on needless work is not reliable. In contrast, the main effect of needless work on moral judgments, tacit inferences effect, and intuitive work morality effect are true positives that emerge in most samples.
Self-expression values perspective	No support for cultural differences in the effects of target age on moral judgments of needless work.	Did not anticipate the tacit inferences effect would emerge, when in fact it replicated across four out of four societies.	Consistent with this theory, while the intuitive work morality effect is robust in U.S., U.K., and Australian samples, Indians do not exhibit this pattern of judgments.	Partial empirical support for the prediction that nations high in self-expression values (USA, UK, Australia) intuitively moralize work more than a nation characterized by survival values (India). This theory's predictions are supported for one of three effects targeted for replication (intuitive work morality effect). Further research comparing more cultures high and low in self-expression orientation, and measuring individual differences in such values, is needed before drawing strong conclusions.

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

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

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

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

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

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

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

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

RESEARCH QUESTION	**MOTIVATED DISCRIMINATION PERSPECTIVE**	**COGNITIVE ASSIMILATION PERSPECTIVE**	**MOTIVATED LIBERALISM PERSPECTIVE**	**STUDY SAVVINESS PERSPECTIVE**
What are the correlates of individual differences in self-perceived objectivity?	*Seeing oneself as objective is correlated with constructing hiring criteria 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.

EMPIRICAL RESULT	**MOTIVATED DISCRIMINATION PERSPECTIVE**	**COGNITIVE ASSIMILATION PERSPECTIVE**	**MOTIVATED LIBERALISM PERSPECTIVE**	**STUDY SAVVINESS PERSPECTIVE**
Individual differences. Rejection of sexist beliefs, having participated in a similar study before, and self-perceived objectivity predict favoritism towards female candidates.	Not supported. The theory predicted that self-perceived objectivity would correlate with constructing hiring criteria in favor of male candidates (Uhlmann & Cohen, 2005); objectivity beliefs instead predicted hiring criteria that favored female candidates. This theory also failed to anticipate the other individual differences moderators that emerged.	Made no strong prediction regarding individual differences.	Supported, in that rejection of sexist beliefs predicted preferring female over male candidates in hiring decisions, and seeing oneself as objective predicted hiring criteria that favored women.	Supported, in that participating in a similar study previously predicted favoritism towards female candidates.
Summary assessment of each theory. What does the overall pattern of empirical results mean for this theory?	The original effects predicted by the motivated reasoning perspective (Uhlmann & Cohen, 2005, 2007) all failed to replicate, with the observed patterns in the opposite direction in several cases. The effects of the affirmation-threat intervention on male evaluators are broadly consistent with the motivated discrimination account. Overall, minimal support for this account of hiring evaluations of female and male candidates.	None of the predicted effects from this perspective were obtained. The cognitive schema account of gender discrimination receives no empirical support from the creative destruction initiative.	Several key predictions of the motivated liberalism account were supported. Some evaluators (men) exhibited favoritism towards female candidates, and rejection of sexist beliefs predicted such favoritism.	A number of key predictions of the savviness perspective were supported. Some evaluators (men) favored female over male candidates, and previous experience with research studies predicted pro-female hiring decisions. Although not a predicted pattern, male evaluators exhibiting more favoritism towards female candidates than female evaluators is more consistent with impression management concerns than with ideological motives.
Overall assessment. What broad conclusions can be drawn from the replication initiative?	The overall project results are most strongly supportive of the study savviness account, followed by motivated liberalism. Some novel evidence for motivated discrimination against women was observed in the effects of the threat manipulation on men's evaluations of female candidates.			

Appendix 1 – Names and Affiliations for the Hiring Decisions Forecasting Collaboration

The following collaborators lent their time and expertise as forecasters:

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

Sarahanne M. Field, University of Groningen
Mustafa Firat, University of Alberta
Raquel M. K. Freitag, Federal University of Sergipe
Enav Friedmann, Ben-Gurion University of the Negev
Omid Ghasemi, Department of Cognitive Science, Macquarie University
Matthew H. Goldberg, Yale University
Amélie Gourdon-Kanhukamwe, Kingston University London
Lorenz Graf-Vlachy, ESCP Business School
Jennifer A. Griffith, University of New Hampshire
Dmitry Grigoryev, National Research University Higher School of Economics
Sebastian Hafenbrädl, IESE Business School
David Hagmann, Harvard Kennedy School
Andrew H. Hales, University of Virginia
Hyemin Han, University of Alabama
Jason L. Harman, Louisiana State University
Andree Hartanto, Singapore Management University
Benjamin C. Holding, Department of Clinical Neuroscience, Karolinska Institutet
Astrid Hopfensitz, Toulouse School of Economics
Joachim Hüffmeier, Institute of Psychology, TU Dortmund University
Jeffrey R. Huntsinger, Loyola University Chicago
Katarzyna Idzikowska, Kozminski University
Åse H. Innes-Ker, Lund University
Bastian Jaeger, Tilburg University
Kristin Jankowsky, University of Kassel
Shoshana N. Jarvis, Haas School of Business, University of California, Berkeley
Nilotpal Jha, Singapore Management University
David Jimenez-Gomez, Fundamentos de Análisis Económico (FAE), University of Alicante
Daniel Jolles, University of Essex
Bibiana Jozefiakova, Olomouc University Social Health Institute, Palacky University Olomouc, Olomouc, Czech Republic
Pavol Kačmár, Department of Psychology, Faculty of Arts, Pavol Jozef Šafárik University in Košice
Mariska Kappmeier, University of Otago
Matthias Kasper, Tulane University & University of Vienna
Lucas Keller, Department of Psychology, University of Konstanz
Viktorija Knapic, University of Rijeka
Mikael Knutsson, Linköping University
Olga Kombeiz, Loughborough University
Marta Kowal, Institute of Psychology, University of Wroclaw
Goedele Krekels, IESEG
Tei Laine
Daniel Lakens, Eindhoven University of Technology
Bingjie Li, Warwick Business School
Ronda F. Lo, York University
Jonas Ludwig, University of Würzburg
James C. Marcus, Evidera
Melvin S. Marsh, Georgia Southern University
Mario Martinoli, DiECO, Università degli Studi dell'Insubria
Marcel Martončik, University of Presov, Faculty of Arts, Institute of Psychology
Allison Master, University of Washington & University of Houston

Theodore C. Masters-Waage, Singapore Management University
Lewend Mayiwar, Department of Leadership and Organizational Behavior, BI Norwegian Business School
Jens Mazei, TU Dortmund University
Randy J. McCarthy, Northern Illinois University
Gemma S. McCarthy, University of Limerick
Stephanie Mertens, Swiss Center for Affective Sciences, University of Geneva
Leticia Micheli, Maastricht University
Marta Miklikowska, Umeå University
Talya Miron-Shatz, Ono Academic College
Andres Montealegre, Cornell University
David Moreau, The University of Auckland
Carmen Moret-Tatay, Universidad Católica de Valencia San Vicente Mártir
Marcello Negrini, Maastricht University
Philip W. S. Newall, CQUniversity
Gustav Nilsonne, Karolinska Institutet, Department of Clinical Neuroscience & Stockholm University, Department of Psychology
Paweł Niszczota, Poznań University of Economics and Business
Nurit Nobel, Stockholm School of Economics
Aoife O'Mahony, School of Psychology, Cardiff University
Mehmet A. Orhan, PSB Paris School of Business
Deirdre O'Shea, University of Limerick
Flora E. Oswald, The Pennsylvania State University
Miriam Panning
Peter C. Pantelis
Mariola Paruzel-Czachura, Institute of Psychology, University of Silesia in Katowice
Mogens Jin Pedersen, University of Copenhagen
Gordon Pennycook, University of Regina
Ori Plonsky, Technion - Israel Institute of Technology
Vince Polito, Macquarie University
Paul C. Price, California State University, Fresno
Maximilian A. Primbs, Radboud University
John Protzko, Department of Psychological & Brain Sciences, University of California, Santa Barbara
Michael Quayle, University of Limerick
Rima-Maria Rahal, Tilburg University
Md. Shahinoor Rahman, University of Chittagong
Liz Redford, Healthy Minds Innovations, Inc.
Niv Reggev, Department of Psychology & Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev
Caleb J. Reynolds, Department of Psychology, Florida State University
Marta Roczniewska, SWPS University of Social Sciences and humanities & LIME Department, Karolinska Institutet
Ivan Ropovik, Faculty of Education, Institute for Research and Development of Education, Charles University & Faculty of Education, University of Presov
Robert M. Ross, Department of Philosophy, Macquarie University
Thomas J. Roulet, University of Cambridge
Andrea May Rowe
Silvia Saccardo, Carnegie Mellon University
Margaret Samahita, University College Dublin

Michael Schaerer, Singapore Management University
Joyce Elena Schleu, TU Dortmund University
Brendan A. Schuetze, The University of Texas at Austin
Ulrike Senftleben, Technische Universität Dresden
Raffaello Seri, DiECO, Università degli Studi dell'Insubria & CORG, University of Southern Denmark
Zeev Shtudiner, Ariel University, Israel
Jack Shuai, University of Toledo
Ray Sin, Early Warning Services
Varsha Singh, Humanities and Social Sciences, Indian Institute of Technology Delhi
Aneeha Singh, International Research & Exchanges Board (IREX)
Tatiana Sokolova, Tilburg University
Victoria Song, Fordham University
Tom Stafford, University of Sheffield
Natalia Stanulewicz, De Montfort University
Samantha M. Stevens, The Pennsylvania State University
Eirik Strømland, University of Bergen
Samantha Stronge, University of Auckland
Kevin P. Sweeney, Western Kentucky University
David Tannenbaum, University of Utah
Stephanie J. Tepper, Cornell University
Kian Siong Tey, INSEAD
Hsuchi Ting, Goldman Sachs
Ian W. Tingen, Tingen Industries
Ana Todorovic, Department of Experimental Psychology, University of Oxford
Hannah M.Y. Tse, University of Hong Kong
Joshua M. Tybur, Vrije Universiteit Amsterdam
Gerald H. Vineyard, Independent Researcher
Alisa Voslinsky, Department of Industrial Engineering and Management, Sami Shamoon Academic College of Engineering
Marek A. Vranka, Charles University
Jonathan Wai, University of Arkansas
Alexander C. Walker, University of Waterloo
Laura E. Wallace, Ohio State University
Tianlin Wang, University at Albany-SUNY
Johanna M. Werz, RWTH Aachen University
Jan K. Woike, University of Plymouth, UK, Max Planck Institute for Human Development
Conny E. Wollbrant, University of Stirling
Joshua D. Wright, Simon Fraser University
Sherry J. Wu, University of California-Los Angeles
Qinyu Xiao, University of Hong Kong
Paolo Barretto Yaranon, University of Limerick
Siu Kit Yeung, The University of Hong Kong
Sangsuk Yoon, University of Dayton
Karen Yu, Sewanee: The University of the South
Meltem Yucel, University of Virginia
Ignazio Ziano, Grenoble Ecole de Management, F-38000 France
Ro'i Zultan, Ben-Gurion University of the Negev
Camilla S. Øverup, Department of Public Health, University of Copenhagen, Denmark