A creative destruction approach to replication: Implicit work and sex morality across cultures


a INSEAD, Singapore
b Oregon State University, United States of America
c University of Virginia, United States of America
d The World Bank
e Stockholm School of Economics, Sweden
f Massey University, New Zealand
g University of Amsterdam, Netherlands
h Stockholm School of Economics, Sweden, University of Innsbruck, Austria
i Michigan State University, United States of America
j Ithaca College, United States of America
k University at Buffalo, The State University of New York, United States of America
l University of Limerick, Ireland
m Brooklyn College CUNY, United States of America
n Queens College CUNY, United States of America
o Fairfield University, United States of America
p Durham University, United Kingdom
q Many Institutions

ARTICLE INFO

Keywords:
Replication
Theory testing
Falsification
Implicit social cognition
Priming
Work values
Culture

ABSTRACT

How can we maximize what is learned from a replication study? In the creative destruction approach to replication, the original hypothesis is compared not only to the null hypothesis, but also to predictions derived from multiple alternative theoretical accounts of the phenomenon. To this end, new populations and measures are included in the design in addition to the original ones, to help determine which theory best accounts for the results across multiple key outcomes and contexts. The present pre-registered empirical project compared the Implicit Puritanism account of intuitive work and sex morality to theories positing regional, religious, and social class differences; explicit rather than implicit cultural differences in values; self-expression vs. survival values as a key cultural fault line; the general moralization of work; and false positive effects. Contradicting Implicit Puritanism’s core theoretical claim of a distinct American work morality, a number of targeted findings replicated across multiple comparison cultures, whereas several failed to replicate in all samples and were identified as likely false positives. No support emerged for theories predicting regional variability and specific individual-differences moderators (religious affiliation, religiosity, and education level). Overall, the results provide evidence that work is intuitively moralized across cultures.

The present initiative aimed to assess the robustness, generality, and cultural boundedness of prior findings on Implicit Puritanism, an

Received 24 September 2018; Received in revised form 12 September 2020; Accepted 13 September 2020

https://doi.org/10.1016/j.jesp.2020.104060

This paper has been recommended for acceptance by Joris Lammers.

Corresponding authors at: INSEAD, Organisational Behaviour Area, 1 Ayer Rajah Avenue, 138676, Singapore.
E-mail addresses: warrentierney@hotmail.com (W. Tierney), eric.luis.uhlmann@gmail.com (E.L. Uhlmann).
account of the role of the United States’ cultural and religious history on the moral intuitions of contemporary Americans (Poehlman, 2007; Uhlmann, Poehlman, & Bargh, 2008, 2009; Uhlmann, Poehlman, Tannenbaum, & Bargh, 2011). The theory of Implicit Puritanism draws on research on automatic and unconscious social cognition (Banaji, 2001; Greenwald & Banaji, 1995; Haidt, 2001; Nisbett & Wilson, 1977) and cross-disciplinary scholarship on America’s religious roots (Baker, 2005; de Tocqueville, 1840/1996; Landes, 1998; Lipset, 1996) to form testable empirical predictions about national differences in intuitive work and sex morality. According to the theory, a history of Puritan-Protestant influence has led traditional work and sex values to implicitly permeate U.S. culture, shaping the moral intuitions and unconscious reactions of even non-Protestant and less religious Americans. In contrast to cultural frameworks focused on East-West differences (e.g., Nisbett, Peng, Choi, & Norenzayan, 2001; Oyserman, Coon, & Kemmelmeier, 2002) or comparisons between Western, Educated, Industrialized, Rich, and Democratic (WEIRD) and non-WEIRD populations (Henrich, Heine, & Norenzayan, 2010), Implicit Puritanism focuses on cultural variability within Western societies. The implicit values of Americans—as elicited via moral scenarios, mindset manipulations, and priming paradigms—are contrasted with those of individuals from ostensibly similar Western societies with different religious histories (e.g., Canada, Australia, or the United Kingdom).

Employing what we term a “creative destruction” approach to replication, we leveraged the complex set of experimental results and cultural differences hypothesized by Implicit Puritanism to further prespecify alternative results predicted by competing accounts of work and sex morality. A number of these alternative frameworks posit that religious, regional, and social class differences are more important than national differences. Another perspective argues that cultural differences in the relevant values are explicit and conscious rather than implicit and unconscious. Yet another competing theory proposes that implicit orientations towards work and sexuality are consistent across cultures, perhaps due to common evolutionary roots. In addition to directly replicating the original study designs (Simons, 2014), this initiative strategically included new measures and samples—permitting not only a comparison of the original theoretical predictions (Poehlman, 2007; Uhlmann et al., 2008, 2009, 2011) with the null hypothesis of no condition or group differences, but also tests of further ideas. We were then able to examine which theory best accounts for the results across multiple key outcomes and contexts. The goal, in the specific case of work morality across cultures but also more generally, was to identify ways to maximize the generativity and information gain from a replication initiative.

1. Creative destruction in science

The scientific community’s shaken faith in original effects that do not emerge in a single direct replication (same method, new observations; Simons, 2014) has been documented in the context of a prediction market (Dreber et al., 2015). More generally, debate and discussion regarding replications centers largely on the existence or nonexistence of a given finding, as opposed to testing competing predictions of positive effects against one another. Consider, however, that a replication could broaden its scope beyond the original design and theorizing, including further measures and conditions testing additional ideas (Brainerd & Reyna, 2018). Large scale replications can and should be leveraged to simultaneously test multiple competing and complementary ideas that operate in the same theoretical space (Tierney et al., in press).

The inspiration is Schumpeter’s (1942/1994) concept of the “gale of creative destruction” in a capitalistic economy, the “process of industrial mutation that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one.” Schumpeter characterizes capitalism as a cyclical process through which outmoded products, approaches, and organizations are destroyed and supplanted by stronger ones. The destruction is both healthy and necessary for improved institutions to emerge. The notion of creative destruction or a “Schumpeter’s gale” has a clear parallel in natural selection in evolutionary biology. In the Origin of Species, Darwin (1872) noted that “extinction of old forms is the almost inevitable consequence of the production of new forms.”

For too long, psychological theories have been sheltered and protected from disconfirmation, rather than subjected to the type of survival pressures Darwin outlined. Historically, approximately 1% of articles published in the fields of psychology and marketing are direct replications of prior work (Bozarth & Roberts, 1972; Hubbard & Armstrong, 1994; Makel, Plucker, & Hegarty, 2012). Most of the research questions examined in the many thousands of papers published yearly are only ever pursued by the original laboratory, who are biased to confirm their own theories (Berman & Reich, 2010; Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; Kuhn, 1962; Manzoli et al., 2014; Mynatta, Dohertya, & Tweney, 1977). The recent movement to reexamine published findings suggests replication rates of 36% in psychology (Open Science Collaboration, 2015), 11–25% in biomedicine (Begley & Ellis, 2012; Prinz, Schlang, & Asadullah, 2011), 61% in experimental economics (Camerer et al., 2016), 70% in experimental philosophy (Cova et al., 2018), and 62% for behavioral experiments published in elite journals (i.e., Science and Nature; Camerer et al., 2018). Yet it is also worth considering what is left in the wake of a gale of failed replications. The original theory has been cast into doubt, but has a new, stronger theory emerged in its place?

In the creative destruction approach to replication, the original hypothesis is compared not only to the null hypothesis, but also to pre-registered (Van’t Veer & Giner-Sorolla, 2016; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) predictions derived from multiple additional theories (Tierney et al., in press). This may involve administering new measures, adding further conditions, and testing new populations in addition to the original ones (what Brainerd & Reyna, 2018, refer to as a Registered Report plus or RR+ approach). Which theoretical framework best accounts for the variance in outcomes is then rigorously assessed. This may lead to the conclusion that multiple complementary theories are needed to fully explain the phenomenon under study (Jussim, Coleman, & Lerch, 1987).

The aim is to provide critical tests (Kahneman & Klein, 2009; Lakatos, 1970; Mayo, 2018; Mellers, Hertwig, & Kahneman, 2001; Platt, 1964; Popper, 1959/2002) that maximize the yield of scientific knowledge from the investigation. The present effort complements broader calls to engage in “theory pruning” by testing competing theories against one another (Aguinis, Pierce, Bosco, & Muslin, 2009; Kluger & Tikchinsky, 2001) in order to reduce the dense theoretical landscape of the sciences (Hambrock, 2007; Leavitt, Mitchell, & Peterson, 2010). As previous commentators have noted, “one has a much greater likelihood of making important knowledge advances to theory and practice if the study is designed so that it juxtaposes and compares competing plausible explanations of the phenomenon being investigated” (Van de Ven & Johnson, 2006, p. 814), and “The greatest scientific value emerges when at least two models are specified representing competing conceptualizations and one emerges the strongest” (Vandenbergh & Grele, 2008).

2. Implicit puritanism

Scholars across fields have traced aspects of contemporary U.S. culture to the nation’s history of religious migration (Baker, 2005; de Tocqueville, 1840/1990; Lipset, 1996; Schafer, 1991; Voss, 1993). Among the New England region’s earliest European settlers were devout Puritan-Protestants fleeing religious persecution in England. Although eventually dwarfed numerically by settlers seeking economic opportunities, these early colonists had a disproportionate influence on the cultural values of the emerging nation. This is analogous to founder effects in organizations (Schein, 1996; Weeks, 2004) and biology
(Mayr, 1942, 1954; Thompson, 1978): the earliest members of a group may strongly impact the characteristics and behaviors of later generations of members. Consider for instance that the Southern culture of honor in the United States can be traced back to settlement from herding communities in the United Kingdom, where a reputation for violent retribution served as a deterrent against theft of one’s flock (Nisbett & Cohen, 1996).

Historical patterns of religious migration may be one reason why the United States today remains deeply religious and traditional despite sharing in the economic growth that has contributed to the secularization of other Western countries (Inglehart, 1997; Inglehart & Welzel, 2005). The values of contemporary Americans with regards to sexuality, suicide, divorce, and abortion resemble prior generations much more so than in ostensibly similar nations such as the United Kingdom, Canada, and Australia. A related legacy of America’s Puritan-Protestant heritage may be a distinctive orientation towards work (Poehlman, 2007; Uhlmann et al., 2008, 2009, 2011). Although most of the world’s faiths moralize sexuality, Calvinist Protestantism is distinctive in the religious significance accorded to everyday labor. Theologian John Calvin believed that material wealth accumulated meritously through hard work indicated that a person was among God’s chosen (Weber, 1904/1958). Other national cultures encourage long work hours out of secular concerns such as duty to family or country; the Protestant work ethic is truly special in linking work to divine salvation.

These unique historical and religious roots hold continuing relevance in part due to the unconscious internalization and operation of pervasive cultural mores. Dual process models propose that in addition to explicit, deliberately endorsed attitudes and beliefs, people also have implicit, automatic associations that they may not consciously recognize (Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995). Whereas explicit beliefs are at least somewhat responsive to logical argumentation, automatic associations are ingrained by the broader culture or other environmental conditioning (Banaji, 2001; Gregg, Selbst, & Banaji, 2006). As a result, implicit associations and explicit beliefs can diverge sharply (Nosek, 2005). For instance, even individuals who deliberately reject pernicious stereotypes about Black criminality nonetheless associate Black targets with crime more so than White targets (Correll, Park, Judd, & Wittenbrink, 2002; Greenwald, Oakes, & Hoffman, 2003). Without drawing any moral comparison between racism and religion, a similar divergence may come into play with regard to Americans’ work and sex morality. Even non-Protestant and non-religious Americans may, by virtue of their exposure to U.S. culture, unconsciously absorb associations based in traditional Puritan-Protestant values. At times, these associations lead contemporary Americans to show some of the same tendencies as the Puritan colonists. This includes intuitively condemning sexual promiscuity, lauding individuals who work in the absence of any material need to do so, and working harder on an assigned task when thoughts about religion are accessible.

The theory of Implicit Puritanism further expects Americans to link work and sex values together in an overarching ethos. American moral cognition, these novel experimental findings represent critical building blocks of the theory (Poehlman, 2007; Uhlmann et al., 2009, 2011), capturing the unique predictions that distinguish Implicit Puritanism from alternative accounts of American values (e.g., Fisher, 1989; Hofstede, 2001; Inglehart & Welzel, 2005; Lipset, 1996).

### 3. Key empirical evidence

The primary empirical support for Implicit Puritanism stems from a series of studies comparing the responses of Americans and non-Americans to experimental manipulations. Although far from an exhaustive list of all the evidence consistent with Implicit Puritanism in American moral cognition, these novel experimental findings represent critical building blocks of the theory (Poehlman, 2007; Uhlmann et al., 2009, 2011), demonstrating that intuitive, gut feelings lacked justification (we will refer to this as the “Intuitive Mindset Effect”).

#### 3.1. Moralization of needless work

Two of these key studies examined the moralization of work in the absence of any material need, what Snir and Harpaz (2009) refer to as “work devotion” (Poehlman, 2007; Uhlmann et al., 2009). In the first of these experiments, participants read about a postal worker who won the lottery and either retired early or stayed-on-the-job, and was either relatively young (23 years of age) or comparatively older (46 years) at the time. Americans, but not Mexicans, particularly praised a young person who continued to work at a low-ranked job despite becoming a multi-millionaire (henceforth referred to as the “Target Age and Needless Work Effect”). A follow-up experiment demonstrated that intuitive processes underlie this pattern of judgments. American participants read about two potato peelers who shared a winning lottery ticket. One retired young, and the other continued working in the restaurant kitchen. Following on prior research on rational-experiential framing (Epstein, 1998), participants were asked for both their “intuitive, gut feeling” and “most rational, objective” response as to which of the two was the better person. Americans significantly preferred the target who persisted in needless work, but only in an intuitive mindset. When it came to their logically reasoned beliefs, Americans seemed to realize their gut feelings lacked justification (we will refer to this as the “Intuitive Mindset Effect”).

#### 3.2. Linking work with salvation

Another key experiment used a priming paradigm ( Bargh, 2014; Bargh, Chen, & Burrows, 1996; Srull & Wyer, 1979) to examine whether traditional Puritan-Protestant values operate outside of conscious awareness. Prior empirical studies suggest that direct activation of concepts can influence downstream judgments and behaviors absent any mediation by conscious intentions (see Weingarten, Hepler, Chen, McAdams, Yi, & Albarracin, 2016, for a meta-analysis). A priming manipulation was therefore employed to test the hypothesized implicit link between work and divine salvation in American minds (Uhlmann et al., 2007; Uhlmann et al., 2011).
et al., 2011). Participants from the United States and Canada first completed a sentence unscrambling puzzle in which either words representing salvation (e.g., redeem, divine, heaven) or similarly valanced concepts unrelated to religion (e.g., flowers, rainbow, happiness) were subtly embedded. After completing one of the two versions of the scrambled-sentences task, all participants were presented with an anagram task framed as a work assignment. American, but not Canadian participants responded to activation of religious concepts with improved work performance (i.e., greater number of anagrams solved; we will refer to this as the “Salvation Prime Effect”).

3.3. Linking work and sex values

The final study key to the theory of Implicit Puritanism provides evidence of the hypothesized link between work and sex morality in American moral cognition. This experiment adapted a false memory paradigm from cognitive psychology (Barrett & Keil, 1996) to examine the tacit inferences drawn about social targets. American participants read a series of vignettes about women and men who either upheld or violated traditional sex or work values (Poehlman, 2007; Uhlmann et al., 2009). In one scenario, a high school (secondary school) student named Ann was described as either sexually promiscuous or abstinent. In both conditions, Anne scored poorly on her history quiz. After a brief distractor task, participants were tested on their memory of the vignettes. Embedded among the memory items were target statements that were in fact false (i.e., did not reflect the information provided). Yet at the same time, they represented inferences flowing from the assumption that a good person is both sexually restrained and hard-working, whereas a bad person is neither. As hypothesized, Americans falsely remembered sexually promiscuous individuals as lazy, and vice versa. For example, when Anne was promiscuous, participants were significantly more likely to misremember her having failed to study hard for the quiz. (This overall pattern of results, obtained across four such scenarios, is henceforth referred to as the “Tacit Inferences Effect”).

Across each of these investigations, individual differences in religiosity and religion (of particular interest, whether the research participant was a Protestant or not) did not significantly moderate the effects. Not only devout American Protestants, but also members of other religious faiths and even atheists appear to moralize work and sexuality in a manner consistent with the faith of the early Puritan-Protestant colonists. This is consistent with the idea that such beliefs are implicitly absorbed from the broader culture context of the United States (Boyd et al., 2011; Sinclair et al., 2005), rather than deliberatively chosen through a process of careful reflection. This streak of Implicit Puritanism, the original research suggests, coexists with the multifold other influences on American culture over the centuries.

4. Alternative accounts of work and sex morality

Consistent with the creative destruction approach to replication (Tierney et al., in press), rather than re-examine the predictions of Implicit Puritanism theory in isolation, we will leverage the same data collections to simultaneously test other theories. Some of these alternative accounts of work and sex morality are competing, or in other words formulate predictions in direct opposition to those tested in the original research (Poehlman, 2007; Uhlmann et al., 2009, 2011). Others are potentially reconcilable with the original theorizing, positing individual-differences or demographic moderators that might coexist with the basic patterns of effects core to Implicit Puritanism.

4.1. False positives

The false positives perspective adopts a skeptical stance towards the original studies, which were conducted prior to the crisis of confidence and subsequent methodological reforms in the field of psychology (Nelson, Simmons, & Simonsohn, 2018). Like most research investigations conducted before 2011, they were underpowered to detect the reported effects (Fanelli, 2010; Ioannidis, 2005) and the analyses were not pre-registered (Vanh Voorst & Giner-Sorolla, 2016; Wagenmakers et al., 2012). In addition, one key experiment—the salvation prime study—relied on nonconscious priming methods (Bargh et al., 1996; Srull & Wyer, 1979), which have been subject to a wave of replication failures (e.g., Caruso, Shapiro, & Landy, 2017; Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013; Klein et al., 2014; McCarthy et al., 2018; D’Onofrio et al., 2018; Olsson-Collentine, Wicherts, & van Assen, in press; Pashler, Coburn, & Harris, 2012; Pashler, Rohrer, & Harris, 2013; Rohrer, Pashler, & Harris, 2015). Thus, the original Implicit Puritanism findings may simply reflect false positive effects (Simmons, Nelson, & Simonsohn, 2011). It may not be the case that needless work elicits intuitive admiration, religion primes hard work, and work and sex morality are implicitly linked—either in the United States or in other societies. If the original effects are false positives, effect sizes should be negligible across cultures, and variability across locations (e.g., different laboratories, regions, and nations) should not exceed what would be expected based on chance (Klein, Vianello, Hasselman, et al., 2018, 2014; McCarthy et al., 2018; Olsson-Collentine, et al., in press).

4.2. Religious differences

Another possibility is that the original effects hold only for some Americans, but not others. It seems straightforward that traditional Puritan-Protestant moral attitudes towards work and sexuality would be most evident among individuals who are themselves devout, practicing Protestants. That an implicit association is pervasive in a culture does not preclude individual differences, such that people who deliberately endorse the association show its effects most strongly (Gawronski & Bodenhausen, 2006; Nosek, 2005). Notably, U.S. Protestants and Catholics exhibit important differences in the tendency to behave impersonally at work, including on indirect and implicit measures (Sanchez-Burks, 2002, 2005; Sanchez-Burks & Lee, 2007).

Although the original research on Implicit Puritanism obtained no support for religion and religiosity as moderators of the reported effects, methodological limitations warrant caution. First, the original studies relied on relatively small samples, and may have failed to detect the signal of important moderators amid the noise caused by imprecise estimates. Second, only a single-item assessment of religiosity was used, making it impossible to calculate the reliability of the measure. The present replications therefore used a validated multi-item measure of religiosity (Roenig & Büssing, 2010) and collected thousands rather than hundreds of participants to allow for more confident conclusions.

4.3. Regional differences

A wealth of evidence indicates that variability within different regions of a society can be just as meaningful as cross-national comparisons (Cohen & Varnum, 2016; Muthukrishna et al., 2020). Historical patterns of rice cultivation, which requires high levels of cooperation, predict contemporary endorsement of collectivism within China (Talhelm et al., 2014), and U.S. states vary in their individualism and tight adherence to norms (Harrington & Gelfand, 2014; Vandello & Cohen, 1999). Regions of Japan settled under frontier conditions are characterized by levels of individualism comparable to those in the United States (Kitayama, Ishii, Imada, Takemura, & Ramaswamy, 2006). And as noted earlier, Northern and Southern U.S. states differ dramatically in their norms regarding insult-based violence (Nisbett & Cohen, 1996).

Influential historical scholarship proposes that four major regions of the United States were shaped in distinct ways by migration from different populations within Great Britain, or “Albion” (Fisher, 1989). The religious values of the Pilgrims and Puritans most strongly influenced the New England region, English gentry played an important role in the
planted the cultural tradition of the South, Quakers shaped the cultural tradition of the Midwest, and Scotch-Irish migration contributed to the rural culture of the American West. In contrast to the theory of Implicit Puritanism, the regional folkways perspective predicts that Puritan-Protestant moral intuitions should manifest themselves primarily in the New England states, the U.S. region most influenced by Puritan migration.

In the original research (Poehlman, 2007; Uhllmann et al., 2009, 2011) regional comparisons within the United States based on state of origin yielded only null results, yet were based on small samples of participants and potentially underpowered to detect real differences. Another limitation of the original investigations is that the U.S. samples were recruited primarily, although not exclusively, from the New England region. Several experiments were conducted with undergraduates at Yale university, most of whom were studying outside their home state, in contrast to a state school which would be attended mostly by locally based individuals. Nonetheless, these Yale students had at a minimum a few months of exposure to New England culture, if not several years or more. Such samples make it more difficult to tease apart the effects of regional cultural mores and those of the broader U.S. culture. Although perhaps doubtful, one cannot rule out the possibility that Yale students from other areas of the U.S. only exhibited Implicit Puritanism due to their recent exposure to New England culture.

The replications therefore recruited large samples of respondents from both the New England states and other U.S. states to allow for a fairer test of regional variability. The “Albion’s seed” hypothesis suggests the effects outlined by Implicit Puritanism theory should be confined largely to the New England region, rather than characteristic of the nation as a whole. This is again in contrast to the theory of Implicit Puritanism, which proposes that traditional Puritan-Protestant work and sex morality characterizes U.S. culture in general—i.e., not only New England but all the U.S. states and regions. Implicit Puritanism is postulated to have seeped into the broader American culture, not just New England culture (Poehlman, 2007; Uhllmann et al., 2009, 2011). Further, rather than being conditioned in a matter of months, the underlying associations with work and sexuality are thought to be socialized from a relatively early age (Poehlman, 2007; Uhllmann et al., 2009, 2011), again similar to cultural stereotypes of groups (Banaji, Baron, Dunham, & Olson, 2008; Baron & Banaji, 2006; Dunham, Baron, & Banaji, 2006, 2008, 2016). Our large-sample replications provided much greater power to detect regional differences than in the original studies, providing direct tests of the opposing predictions of the Implicit Puritanism and regional folkways accounts of American values.

4.4. Social class differences

Experimental, survey, and archival research converges in identifying profound differences in values and cognitive tendencies based on social class (Cohen & Varnum, 2016). Relative to high socioeconomic status (SES) persons from the same society, low-SES individuals are more likely to take into account situational constraints when forming judgments of others; valorize steadfastness in the face of adversity and obedience to authorities over personal agency; and are more relational, obedient to authorities over personal agency; and are more relational and family-oriented (Snibbe & Markus, 2005; Stephens, Fryberg, & Markus, 2011; Stephens, Fryberg, Markus, Johnson, & Covarrubias, 2012; Varnum, Na, Murata, & Kitayama, 2012). Such demographic differences have been observed only within the United States, but also other cultures, among these India, Poland, the Ukraine, Russia, and Japan (Grossmann & Varnum, 2011; Kohn, 1969; Kohn et al., 2002; Kohn, Naoi, Schoenbach, Schooler, & Slomczynski, 1990).

In surveys, working class people generally report viewing work as a job and means to an end—to them, the purpose of work is to earn wages to support themselves and their family. In contrast, middle and upper-class respondents are more likely to see work as an end unto itself and in the context of a long-term career (Argyle, 1994; Corney & Richards, 2005; King & Bu, 2005; Williams, 2012; cf. Adigun, 1997). This suggests that within any given culture, indices of social class (i.e., educational attainment and income) should be associated with intuitively moralizing needless work, as in the Target Age and Needless Work effect, and Intuitive Mindset effect. The social class perspective makes no strong predictions for the Tacit Inferences or Salvation Prime effects. However, the strong version of the theory, in which social class differences exclusively drive moral cognition, anticipates null findings. The literature on class differentiation in human societies provides no basis to hypothesize an implicit link between work and sex values, or an automatic association between work and divine salvation.

4.5. Self-expression values

Cross-national data from the World Values Survey identifies two primary dimensions of culture: 1) traditional vs. secular-rational values, and 2) survival vs. self-expression values (Inglehart, 1997; Inglehart & Welzel, 2005). Traditional societies emphasize the importance of religious faith and absolute standards for morality, and people tend to be opposed to divorce, euthanasia, and abortion; in secular societies, fewer people self-identify as devoutly religious and such practices are more socially acceptable. In cultures high in self-expression values, individuals pursue their own individual happiness and personal fulfillment, whereas in survival cultures economic security is the overriding goal.

High national scores on self-expression values tend to be associated with “work devotion,” in other words perceiving work to be an enjoyable pursuit above and beyond money, whereas survival values are linked to “work investment,” or seeing work as a means of earning a living (Snir & Harpaz, 2009). There are no major differences between the United States and other nations in the English-speaking cultural cluster in terms of self-expression values (Inglehart & Welzel, 2005). This leads to a predicted pattern of cross-national similarities and differences in results that deviates sharply from the Implicit Puritanism perspective. Based on their scores on self-expression values, participants from the United States, United Kingdom, and Australia should all intuitively moralize work, and to similar degrees. In contrast, participants from survival-oriented societies, such as India, should view work arrangements as instrumental and therefore not valorize needless work. The Inglehart and Welzel (2005) cultural framework provides no reason to expect the Tacit Inferences or Salvation Prime effects to emerge in any culture.

4.6. Explicit American Exceptionalism

Another distinct possibility is that the originally hypothesized cultural differences in work and sex values (Poehlman, 2007; Uhllmann et al., 2009, 2011) are in fact more explicit than implicit. Such deep-seated cultural beliefs may have a strong intuitive component, in that associated judgments appear suddenly in consciousness without much subjective experience of deliberation (Haidt, 2001). However, they could still be introspectively accessible and consciously reportable. As noted earlier, the results of cross-national surveys such as the World Values Survey (Inglehart & Welzel, 2005), Hofstede’s classic study of IBM employees (Hofstede, 2001), and GLOBE survey (Dorfman, Hanges, & Brodbeck, 2004), already capture the strikingly religious and traditional values of the United States. Comparisons of societal institutions and work practices provide converging evidence of American exceptionalism (Baker, 2005; Landes, 1998; Lipset, 1996). The valorization of long work hours in America, and conservative views on sexuality, may be reflected in emotional gut responses that are fully verbalizable and conscious.

Notably, many Americans explicitly endorse the Protestant work ethic (PWE) on self-report scales, agreeing to items like “Most people who don’t succeed in life are just plain lazy” (Furnham, 1989; Katz & Hass, 1988; Mirels & Garrett, 1971). The PWE correlates with attitudes towards social groups such as the unemployed, Black Americans, and the obese; as well as views on policies such as affirmative action and
welfare (Furnham, 1982, 1989; Katz & Hass, 1988; Sidanius & Pratto, 1999). However, this prior scholarship does not directly predict that such complex ideologies will operate unconsciously in the manner suggested by research on implicit social cognition (Bargh, 2014; Bargh et al., 1996). Americans are perhaps exceptional in intuitively lauding individuals who engage in needless work (Target Age and Needless Work effect and Intuitive Mindset effect), and may intuitively infer that hard-working individuals are sexually chaste and vice versa (Tact Inferences effect), all judgments flowing from their explicit endorsement of the Protestant work ethic. However, merely priming words related to religion will not necessarily have the same impact on downstream judgments and behaviors (e.g., Salvation Prime effect).

Importantly, prior scholarship in fields such as sociology, political science, and cultural history identifies consciously self-reported cultural differences in values, but is largely silent on whether or not traditional American values further operate unconsciously. The Explicit American Exceptionalism alternative theory tested here, in which traditional work and sex values are observable in consciously self-reported judgments, but not on implicit indicators, is suggested by the recent wave of replication failures for nonconscious priming effects (Caruso et al., 2017; Doyen et al., 2012; Harris et al., 2013; Klein et al., 2014; McCarthy et al., 2018; O’Donnell et al., 2018; Olsson-Collentine, et al., in press; Pashler et al., 2012; Pashler et al., 2013; Rohrer et al., 2015). In other words, the Explicit American Exceptionalism account places great stock in earlier multi-disciplinary work on U.S. cultural mores, which relied heavily on high powered cross-national surveys (e.g., Baker, 2005; Lipset, 1996; Schafer, 1991), and has little faith in small sample experiments on implicit priming (Bargh, 2014; Bargh et al., 1996; Poehlman, 2007; Uhlmann et al., 2011). However, that religious and work values may be prime-able in experimental settings and exert unconscious influences on judgments and behaviors does not challenge the work of Lipset (1996), Baker (2005), and other scholars of U.S. exceptionalism in fields outside of psychology.

4.7. General moralization of work and sex

A final possibility is that the key experimental effects outlined earlier (Poehlman, 2007; Uhlmann et al., 2009, 2001) may be exhibited not only by Americans, but members of other cultures as well. Historically, moralization and regulation of sexual behavior is characteristic of most religious faiths and societies (Foucault, 1978; Gruen & Panichas, 1997; Peiss, Simmons, & Padgug, 1989). A general distaste for individuals who under-contribute to work tasks is suggested by research on costly punishment of defectors and free riders (Dreber, Rand, Fudenberg, & Nowak, 2008; Jordan, Hoffman, Bloom, & Rand, 2016), and may have evolutionary roots. The original Implicit Puritanism studies provide preliminary evidence of cross-cultural differences, but with samples too small to draw strong conclusions. Higher powered tests may be necessary to detect the implicit moralization of work and sex across human societies.

Notably, neither the original studies nor the present replication initiative examined whether moral intuitions related to work and sexuality are potentially useful in identifying social targets with strong moral identities (Aquino, Freeman, Reed II, Lim, & Felps, 2009; Aquino & Reed II, 2002). Sexually restricted and hard-working individuals may or may not actually be more “moral” on other dimensions— such as empathy, generosity, fairness, or trustworthiness— and the strength of such relationships could also vary by culture (Weeden & Kurzban, 2013). Even if there is an ecological relationship between traditional Puritan morality and ethical behavior more generally, it is likely to be far from perfect, and also imperfectly aligned with social inferences and perceptions (Moon, Krems, & Cohen, 2018). The original Implicit Puritanism studies dealt with social judgments, not social reality. The present replications sought to reproduce the original results, and also test for alternative patterns in social judgments predicted by competing theories. The potential general moralization of work and sexuality across cultures is one of these alternative possibilities. The validity or rationality of such inferences is a fascinating question that will have to be left to follow-up research.

5. Overview of the present investigations

These novel data collections used the creative destruction approach to replication to further our theoretical understanding of moral values related to work and sexuality. A set of key effects originally predicted by the theory of Implicit Puritanism, but potentially replicable under other frameworks, were systematically re-examined. The replications occurred across six nations (United States, United Kingdom, Australia, Republic of Ireland, Canada, and India), oversampling the particularly relevant New England region of the United States. As in the original research (Poehlman, 2007; Uhlmann et al., 2011), data were collected both online and in research laboratories.

The original Implicit Puritanism studies adhered to pre-2011 standards for experimental research, in that studies were not pre-registered and sample sizes were moderate (Nelson et al., 2018). Indeed, historically only 8% of studies in the field of psychology have achieved 80% power to detect the reported effects (Stanley, Carter, & Doucouliagos, 2018). In the replication initiative, planned sample sizes totaled many times those of the original experiments, allowing for more precise effect size estimates as well as better powered tests of potential moderators— such as regional variation within the United States, as well as individual differences in religion and religiosity. This allowed us to empirically adjudicate between the Implicit Puritanism, false positives, religious differences, regional variability, social class, self-expression values, explicit American moral exceptionalism, and general moralization accounts of work and sex values. We considered both the strong version of each theory, in which its predictions hold to the exclusion of all others, as well as whether multiple theories in combination best explained the results. All measures and manipulations in this research are disclosed, and sample sizes were determined in advance. The complete study materials are provided in Supplements 1–2, the preregistered analysis plan in Supplement 3 and https://osf.io/xwu4v/, and the datafiles at (Study 1: https://osf.io/k236g/, Study 2: https://osf.io/687h/). Our hope is that this initiative will not only shed novel light on cultural values, but also serve as a model for future efforts to assess the replicability of published findings and explanatory power of competing theories.

6. Study 1

This large-scale online data collection attempted to replicate the target age and needless work effect, intuitive mindset effect, and tacit inferences effect (Poehlman, 2007; Uhlmann et al., 2009) across four nations. A professional survey firm, PureProfile, was used to recruit large samples from the United States, United Kingdom, and Australia, while sampling as evenly as feasible from the constituent regions of each country with the exception of oversampling from the theoretically important New England region of the United States. Amazon’s Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010) was used to collect data from further

1The ultimate origins of cultural values related to work and sexuality are difficult to test empirically. Adaptive pressures may have led human groups to regulate sexual behavior, engage in costly punishment of free riders, and confer status on over-contributors to group efforts. Such morally charged reactions could also reflect more proximal influences such as a society's history of economic activity (Talhelm et al., 2014) or religious migrations (Fisher, 1989; Lipset, 1996). Far more tractable is assessing what values predominate in a society, explicitly and implicitly, and whether they can be situationally activated or primed. These individual-level outputs, predicted based on the expected influence of past events on present day social cognition, are the focus of the present research.
groups of Indian and USA participants (see also Uhlmann, Heaphy, Ashford, Zhu, & Sanchez-Burks, 2013). This online microwork website provided an efficient means of recruiting English speakers from both a survival-oriented society (India) and personal fulfillment-oriented society (U.S.) in order to test the self-expression values hypothesis.

Notably, we held methods and materials constant across these populations to allow for direct replication (Simons, 2014). One can also make iterative modifications to the materials across research sites, assessing mediating states each time, in an effort to achieve psychological rather than methodological equivalence (Fabrigar, Wegener, & Petty, in press; Schwarz & Strack, 2014; Stroebe & Strack, 2014). However, in the original studies the theoretical underlying processes are non-conscious and were inferred rather than measured (Poehlman, 2007; Uhlmann et al., 2009, 2011), seriously complicating such an approach. As the original studies sampled some of the same populations (e.g., USA, UK, and Canadian participants) without modifications across sites, the present replication initiative did the same. Future research using a creative destruction approach to replication may prioritize either methodological or psychological equivalence.

6.1. Methods

6.1.1. Participants

PureProfile sample. The professional survey firm PureProfile was used to recruit participants (total N = 4098) from Australia (24.67%), the United Kingdom (23.43%), and the United States (51.90%) while oversampling the New England states (Maine, Vermont, New Hampshire, Massachusetts, Rhode Island, and Connecticut; 47.58% of the USA sample). Thus, the PureProfile sample was split more or less equally between Australia, the U.K., USA New England states, and USA non-New-England states.

Amazon Mechanical Turk sample. MTurk was used to collect data from a further 2036 Indian (49%) and USA (51%) participants. The MTurk data collection in the USA had a smaller percentage of respondents from the New England region (only 4.3%), limiting our ability to test regional variability.

Demographic information for each major sample for Study 1 is summarized in Table S14-1 in Supplement 14.

6.1.2. Design

The three experiments appeared in counterbalanced order, with assignment to condition within each study randomized. The Lottery Winner study featured a 2 (work status: retired or continues working) x 2 (age: 23 years or 46 years) x participant nationality between-subjects design. The Intuitive Mindset study included a within-subjects factor comparing participants’ preferences in the intuitive framing and logical framing conditions, with participant nationality a between-subjects factor. The Tacit Inferences study had two between-subjects conditions manipulating whether targets uphold or violate traditional morality, with participant nationality again serving as the second between-subjects factor. At the end of the study, after exposure to the manipulations and completing the dependent measures, all participants filled out individual differences and demographic measures.

6.1.3. Materials and procedure

In all of the present data collections, we employed a variety of safeguards to maintain data quality. The cover page for all our online experiments included a captcha item to avoid contamination by bots, and we further screened out participants with duplicate GPS coordinates. For the MTurk data collections for Study 1 we recruited only participants with a 99% acceptance rate and > 1000 hits approved. Finally, we excluded participants with < 5 years of English experience or who failed an instructional manipulation check from all analyses (see Supplements 3 and 10).

Lottery winner study. Participants read a vignette about Sarah, a postal worker who wins the lottery and either decides to retire immediately or to continue in her job. Depending on the experimental condition, she was either 23 or 46 years of age. Participants provided their assessment of Sarah’s moral character (1 = very bad, 7 = very good).

Intuitive mindset study. Participants were presented with a scenario about Robert and John, two potato peelers who shared a winning lottery ticket. Robert immediately chose to retire young, whereas John continued working peeling potatoes. In the intuitive mindset frame, participants were asked for their “intuitive, gut feeling” as to whom is the better person (1 = Robert is a much better person than John, 7 = John is a much better person than Robert). In the logical mindset frame, they were asked for their “rational, objective judgment” to the same question (Epstein, 1998).

Tacit inferences study. In this experiment, participants were first informed that “You will first read stories, then answer some questions about what you remember about the content of the stories” (Barrett & Keil, 1996). They then read four vignettes, each featuring a target person who either conforms to or transgresses traditional work or sex morality. The two scenarios in which the protagonist clearly violates or upholds traditional sex norms also contain ambiguous information about the person’s work ethic. At the same time, the two scenarios in which the protagonists clearly exhibit a strong work ethic or not further contain ambiguous information about her or his sexual behavior.

For example, participants read about a character name Julia, who either worked long hours at her job or was unemployed and not actively looking for a job. In both conditions the vignette went on to indicate that Julia attended a local party and stayed overnight. Participants were then presented with a set of distractor questions prompting open-ended written responses (e.g., “Do you think the author of the story was older than 30 or younger than 30?”). Finally, they were tested on their recall of the scenario using eight true/false questions. Embedded among seven filler true/false items was the critical target item, “Julia slept with the host of last week’s party.” An intuitive link between work and sex morality is reflected in falsely remembering the vignette as stating that Julia had sex only in the condition in which she was previously described as lazy.

The following measures were administered after the key manipulations and dependent measures.

Religiosity. Our multi-item measure of religiosity was the Duke University Religion Index (DUREI; Koenig & Büssing, 2010), a validated five-item measure widely used across fields. Example items include “My religious beliefs are what really lie behind my whole approach to life” and “In my life, I experience the presence of the Divine (i.e., God)” (1 = definitely not true, 5 = definitely true of me). Also included was the single item religiosity item from the original Implicit Puritanism studies (Poehlman, 2007; Uhlmann et al., 2019, 2011), which simply states “I consider myself to be” and provides a numeric scale ranging from 1 (not at all religious) to 7 (very religious). Responses on the numeric scale effectively complete the statement in the initial question—for instance, choosing “7” indicates “I consider myself to be... very religious.”

Protestant work ethic (PWE). The PWE scale from Katz and Hass (1988) is an 11-item questionnaire including statements such as “A distaste for hard work usually reflects a weakness of character” and “Most people who don’t succeed in life are just plain lazy” (1 = strongly disagree, 6 = strongly agree).

Demographics. Participants completed demographic measures including their religion (Protestant, Catholic, Islam, Judaism, Buddhism, atheist, agnostic, other), religious denomination within Protestantism if applicable (Adventist, Anabaptist, Anglican, Baptist, Calvinist,
Lutheran, Methodist, Pentecostal, other), place of worship if any, political orientation (1 = very progressive/left-wing, 7 = very conservative/right-wing), political party identification (free response), gender, age, ethnicity, country and state/region they are currently primarily based in, country of birth, country of citizenship, years spent in the United States, state of origin with the USA if relevant, years of experience with the English language, occupation, income, personal educational level, and education level of most highly educated parent.

Awareness probe. In contrast to the priming paradigm used in Study 2 below, participants’ level of awareness of the manipulations (e.g., target work behavior or age) should not theoretically interfere with the effects in Study 1. However, an exploratory free response item asked “What do you think this survey was about?”

Attention check. An instructional attention check told participants to “please select strongly disagree” and provided a scale ranging from 1 (strongly disagree) to 5 (strongly agree). Participants who failed this check were excluded from all analyses.

6. Results

Mixed models were conducted using the condition values as the fixed effect, while using the region as the random effect. Thereafter, F statistics were derived from the ANOVA produced by these models.

6.2. Results

6.2.1. Needless work study: MTurk sample

A 2 (target age: 23 or 46 years) x 2 (target works vs. retires) ANOVA revealed a statistically significant main effect of target age, F(1, 2029) = 4.43, p = .04, d = −0.093, main effect of work status, F(1, 2032) = 220.53, p < .001, d = 0.65, and two-way interaction between age and work status, F(1, 2027.3) = 4.596, p = .03, d = 0.095 (see Table 1). The target received more moral praise when she continued working compared to when she retired, and when she was older rather than young. Further, reactions to a lottery winner who continued working vs. retired depended on her age.

Although target age and work status interacted significantly, unpacking this interaction revealed a markedly different pattern of results than in the original Implicit Puritanism research. As per the pre-registered analysis plan, the key effect of primary interest was the replication was the main effect of target age (23 years or 46 years) within the target works condition, such that the younger retiree was rated more negatively than the older retiree, F(1, 1009.91) = 8.871, p = .003, d = −0.187.

We next examined potential moderating effects of country, focusing again on the pre-registered key effect of interest (i.e., target age effect within the target works condition). A 2 (23 or 46 years) x 2 (India vs. USA) ANOVA revealed no significant interaction, F(1, 1018) = 0.268, p = .605, d = −0.032, indicating no evidence of moderation by participant nation. Further, testing for the key effect separately by country (USA and India) revealed no effect of target age within the works condition in either the India sample, F(1, 492.32) = 0.058, p = .81, d = 0.022, or USA sample, F(1, 523) = 0.3, p = .584, d = −0.048. New England region likewise failed to moderate the effect of target age within the works condition, F(1, 1018) = 0.678, p = .411, d = 0.052.

Finally, we examined theoretically relevant individual differences moderators. Neither the single item measure of religiosity, F(1, 999) = 0.001, p = .997, d = −0.002, nor the DUREL religiosity scale F(1, 1018) = 0.251, p = .616, d = 0.031, nor participant education level F(1, 985.95) = 1.716, p = .191, d = −0.083, nor the Protestant Work Ethic F(1, 1012.15) = 0.167, p = .683, d = 0.026, nor self-reported religion (Protestant or not) F(1, 1016.62) = 3.4, p = .065, d = 0.116, moderated moral judgments of a target who works based on her age.

6.2.2. Needless work study: PureProfile sample

A 2 (target age) x 2 (work status) ANOVA revealed a nonsignificant main effect of target age, F(1, 4079) = 3.50, p = .06, d = −0.056, a statistically significant main effect of work status, F(1, 4082) = 423.24, p < .001, d = 0.367, and a significant interaction between age and work status, F(1, 4077) = 16.15, p < .001, d = 0.125. With the exception of the main effect of age not reaching statistical significance, this overall pattern paralleled the results reported above for the MTurk sample (see Table 1). Unpacking the target age x work status interaction, the young target who stayed on the job after winning the lottery received similar evaluations to the older target who continued to work, F(1, 2052.56) = 1.887, p = .17, d = 0.061. Instead, the two-way interaction was driven by a target age effect within the retires condition, with the younger retiree rated significantly less favorably than the older retiree, F(1, 2019.88) = 17.675, p < .001, d = −0.1871.

With regard to the moderating effects of nation, there was no significant difference between the USA and the other two countries (Australia & UK), F(1, 2061) = 0.303, p = .582, d = 0.024, the USA vs. Australia, F(1, 1547) = 0.299, p = .585, d = 0.028, or the USA vs U.K., F(1, 1572) = 0.123, p = .725, d = 0.018. Further, the target age and needless work effect was not significant within the USA sample, F(1, 1055.87) = 1.959, p = .162, d = 0.086, Australia sample, F(1, 487) = 0.086, p = .77, d = 0.027, or UK sample, F(1, 514) = 0.266, p = .606, d = 0.046. New England region again failed to emerge as a moderator F(1, 2045.35) = 0.002, p = .97, d = 0.001. The individual differences measures likewise failed to moderate, among these the single item measure of religiosity, F(1, 2048.17) = 0.482, p = .488, d = 0.031, DUREL religiosity scale, F(1, 2056.41) = 0.308, p = .579, d = 0.025, Protestant religion, F(1, 2048.9) = 1.067, p = .302, d = 0.046, education level, F(1, 1938.1) = 0.436, p = .509, d = −0.03, and PWE scores, F(1, 2054.24) = 3.486, p = .062, d = 0.082.

6.2.3. Intuitive mindset study: MTurk sample

A within-subjects ANOVA comparing intuitive and deliberative responses as to whom was the better person revealed a significant overall

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moral judgments of a lottery winner who works vs. retires and is relatively young or older.</td>
</tr>
<tr>
<td>India MTurk</td>
</tr>
<tr>
<td>Young</td>
</tr>
<tr>
<td>Works</td>
</tr>
<tr>
<td>(0.08)</td>
</tr>
<tr>
<td>Retires</td>
</tr>
<tr>
<td>(0.08)</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses represent standard errors.

* PP denotes PureProfile sample.
effect $F(1, 2033.89) = 27.38$, $p < .001$, $d = 0.232$. Specifically, participants expressed a preference for the worker over the retiree that was stronger on the intuitive mindset item than on the rational mindset item.

A significant interaction between country (US vs. India) and intuitive vs. rational responses emerged, $F(1, 2031.84) = 45.027$, $p < .001$, $d = 0.2977$, such that the intuitive mindset effect was stronger among American participants than Indian participants (Fig. 1). The difference between intuitive and rational responses was clearly observed in the USA sample, $F(1, 1033.77) = 76.019$, $p < .001$, $d = 0.543$, but not the India sample, $F(1, 998) = 1.105$, $p = .293$, $d = −0.067$. New England region did not moderate the results, $F(1, 2033.61) = 2.009$, $p = .156$, $d = −0.0623$.

Self-identified religion (Protestant or not), $F(1, 2029.61) = 0.263$, $p = .608$, $d = 0.023$ did not moderate the effect. However education level, $F(1, 1975.39) = 5.006$, $p = .025$, $d = −0.106$ did significantly moderate the results, such that less educated participants were more likely to demonstrate the intuitive mindset effect, directionally contrary to the expectations of the social class perspective. Highly religious individuals, as assessed by both the single-item measure, $F(1, 1994.13) = 22.807$, $p < .001$, $d = −0.214$ and DUREL scale, $F(1, 2031.75) = 24.758$, $p < .001$, $d = −0.221$, were significantly less likely to exhibit a difference between their intuitive and rational responses, directly opposite to the predictions of the religious differences perspective. Contrary to any of the theories tested, endorsement of the PWE, $F(1, 2027.21) = 24.882$, $p < .001$, $d = 0.222$ (Fig. 2). Although there was a significant between-country difference, the tacit inferences effect was statistically significant not only in the USA, $F(1, 1031.8) = 103.8$, $p < .001$, $d = 0.632$, but also India, $F(1, 997.03) = 10.02$, $p = .002$, $d = 0.201$. In other words, the effect was present in both comparison countries, but relatively larger in one nation (US) than in the other (India). New England region did not moderate the results, $F(1, 2023.45) = 0.015$, $p = .902$, $d = −0.006$.

6.2.5. Tacit inferences study: MTurk sample

An overall condition effect emerged such that when the target upheld (violated) traditional work morality, she/he was falsely remembered as upholding (violating) traditional sexual morality, and vice versa, $F(1, 2029.13) = 89.11$, $p < .001$, $d = 0.42$. Further, a significant interaction with country emerged, such that this tacit inferences effect was stronger among American participants than Indian participants, $F(1, 2024.49) = 24.882$, $p < .001$, $d = 0.222$ (Fig. 2). Although there was a significant between-country difference, the tacit inferences effect was statistically significant not only in the USA, $F(1, 1031.8) = 103.8$, $p < .001$, $d = 0.632$, but also India, $F(1, 997.03) = 10.02$, $p = .002$, $d = 0.201$. In other words, the effect was present in both comparison countries, but relatively larger in one nation (US) than in the other (India). New England region did not moderate the results, $F(1, 2023.45) = 0.015$, $p = .902$, $d = −0.006$.

The single item measure of religiosity, $F(1, 1985.01) = 1.168$, $p = .28$, $d = −0.049$, and whether the participant was of the Protestant faith or not, $F(1, 2023.45) = 1.674$, $p = .196$, $d = 0.058$, did not moderate the tacit inferences effect in the MTurk sample. However, the DUREL religiosity scale, $F(1, 2024.49) = 5.718$, $p = .017$, $d = −0.106$, and Protestant Work Ethic scale, $F(1, 2024.67) = 10.143$, $p = .001$, $d = −0.142$, did significantly moderate the effect. Surprisingly, more religious participants on the DUREL scale, and individuals who explicitly endorsed the PWE, were significantly less likely to exhibit false memories consistent with an intuitive link between work and sex morality. These results are inconsistent with any of the theories considered here, and as noted below failed to replicate in the PureProfile sample.
Comparing the USA vs. both other countries combined (UK and Australia) did not reveal a significant difference, $F(1, 4071.27) = 0.961, p = .327$, $d = 0.031$. More fine-grained comparisons between the USA and UK, $F(1, 3078) = 0.012, p = .911$, $d = 0.034$, and USA and Australia $F(1, 3130) = 2.137, p = .144$, $d = 0.053$, were also not statistically significant. The tacit inferences effect was significant within the USA, $F(1, 2121) = 181.655, p < .001$, $d = 0.585$, Australia, $F(1, 1007) = 53.227, p < .001, d = 0.46$, and UK, $F(1, 951.6) = 78.326, p < .001, d = 0.575$, when the samples were tested separately (Fig. 2). New England region was not a significant moderator of false memories consistent with an implicit link between work and sex morality, $F(1, 4069.72) = 0.069, p = .793$, $d = 0.008$.

The individual differences measures, including the single item measure of religiosity, $F(1, 4067) = 0.393, p = .531, d = 0.020$, the DUREL scale, $F(1, 4081) = 0.29, p = .59, d = 0.017$, Protestant religion, $F(1, 4058.1) = 1.193, p = .271, d = 0.044$, and the PWE scale, $F(1, 4079.51) = 3.102, p = .078, d = −0.0552$, did not moderate the tacit inferences effect in the PureProfile sample. Notably, this fails to replicate the initial evidence of moderation by religiosity (DUREL) and PWE scores in the MTurk sample.

6.3. Discussion

The results of this first set of replications confirm a number of the original experimental effects (Poehlman, 2007; Uhlmann et al., 2009, 2011), yet at the same time depart in theoretically informative ways from the original research. One original effect, specifically the moderating role of target age in judgments of needless work, failed to replicate across four nations (India, USA, Australia, and the United Kingdom) and is identified as a likely false positive. At the same time, a pre-registered secondary effect of interest in this “lottery winner” paradigm, the simple main effect of working vs. retiring on judgments of moral goodness, emerged robustly across samples and nations (see Table 1 and Supplement 7). Although neither Americans nor members of several comparison cultures appear to be sensitive to the age of a lottery winner who decides to retire vs. continue working (contrary to the Implicit Puritanism account), people across a number of cultures do appear to morally praise needless work (consistent with the General Moralization of Work account).

Of further theoretical interest was the extent to which positive reactions to needless work are especially strong in an intuitive rather than deliberative mindset. Consistent with the original research, American participants praised needless work more strongly when asked for intuitive gut reaction rather than their more deliberative response. Inconsistent with the theory of Implicit Puritanism, however, not only Americans but also participants from the United Kingdom and Australia exhibited this intuitive work morality effect, while Indian participants did not. This cross-national pattern of results is highly inconsistent with the claim of a unique American work morality, and could reflect the greater intuitive moralization of work in self-expression cultures (USA, UK, Australia) relative to survival-oriented cultures (India). A more nuanced interpretation is that Indian participants strongly moralized work both intuitively and deliberatively, such that a difference in evaluations based on mindset was unlikely to emerge. Indeed, in a pre-registered secondary analysis, a preference for the worker over the retiree emerged robustly across mindsets and cultures (Supplement 7). Scores consistently above the neutral scale midpoint of 4, indicating a preference for needless work, support the General Moralization of Work account. Thus, larger-scale research including a greater number of societies characterized by self-expression and survival values (Inglehart, 1997; Inglehart & Welzel, 2005) will be needed before drawing strong conclusions. We also cannot rule out that the study materials were psychologically nonequivalent between the Western and Indian populations in some unintended manner, or that some other confound in measurement led to the lack of differences in intuitive and deliberative judgments in the India sample (Fabrigar et al., in press; Milfont & Klein, 2018; Poortinga, 1989; van de Vijver & Leung, 2010).

Another interesting cross-national pattern emerged with regards to the tacit inferences drawn from ambiguous scenarios. As in the original experiment, U.S. participants falsely remembered individuals who had violated work values as having also violated traditional sexual mores, and vice versa. However, contrary to the Implicit Puritanism and Explicit American Exceptionalism accounts, such false recollections likewise emerged robustly in the India, U.K., and Australia samples. The effect was statistically significant but diminished in the India sample (see Fig. 2). MTurk respondents in India are more likely to hold a university degree (86.4% of the sample, as shown in Table S14-1) than the general population, potentially artificially attenuating cultural differences. However, the presence of the tacit inferences effect across all samples is most consistent with the pre-registered predictions of the General Moralization of Work account.

Finally, no consistent evidence was found for regional differences within the USA (i.e., New England vs. other parts of the country), or the expected moderating effects of Protestantism, religiosity, and education level. In those few cases where an individual-differences factor significantly moderated the effect, the direction of moderation was more often opposite to rather than consistent with theoretical expectations. Thus, we consider the Social Class, Regional Differences, and Religious Differences accounts unsupported by this first cross-national data collection in the replication initiative.

7. Study 2: methods

Our second study included both online and crowdsourced laboratory replications of the salvation prime effect on work performance. The original salvation prime experiment was conducted with lay adults recruited from public areas in New York State in the United States and Ontario, Canada (Poehlman, 2007; Uhlmann et al., 2011). The present online data collection recruited adults from the United States, the United Kingdom, and Australia via the survey firm PureProfile. The laboratory data collections strategically oversampled populations in New York state to remain as faithful as possible to the original study in terms of region of data collection, with materials administered in paper pencil format as in the original experiment. Replication laboratories were recruited through the last author's professional network and the Study Swap platform (http://osf.io/view/StudySwap/), and relied on locally available samples of university undergraduates. Note that participant age and method of data collection are not theoretically anticipated moderators of the salvation prime effect, and that the original line of research on Implicit Puritanism featured students and lay adult participants, and both paper-pencil and online administration of priming paradigms (Poehlman, 2007; Uhlmann et al., 2009, 2011).

7.1.1. Participants

Online data was collected by the survey firm Pure Profile, and included 514 (45.73%) USA based participants, 312 (27.76%) participants from the United Kingdom, and 298 (26.51%) participants from Australia. The constituent regions of each country were sampled as evenly as feasible, with the exception of again oversampling the New England states ($N = 270$, or 52.52% of the USA sample), in order to compare their responses to participants from other USA regions ($N = 244$, or 47.48% of the USA sample).

The crowdsourced laboratory data collections in the northeastern region of the United States included 95 participants from Ithaca College, 161 participants from the City University of New York, 208 participants from the State University of New York, and 99 participants from Fairfield University. Data collections outside the U.S. included the University of Regina in Canada ($N = 91$), and the University of Limerick in Ireland ($N = 80$). See Table S14–2 in Supplement 14 for an overview of the demographics of the online and laboratory samples.
7.1.2. Design
The study employed a 2 (priming condition: salvation prime or neutral prime) x participant nationality between-subjects design.

7.1.3. Materials and procedure
Participants completed two ostensibly unrelated puzzle tasks. The first was a scrambled-sentences task (Srull & Wyer, 1979) containing either words related to salvation (e.g., redeem, divine, heaven) or similarly valanced words unrelated to religion (e.g., flowers, rainbow, happiness). For instance, in the salvation prime condition the scrambled sentence “coupes here phone redeem your” could be unscrambled to read “redeem your coupons here,” after omitting the word “phone.” Following on prior research using anagram performance as a work task (Chartrand, Dalton, & Fitzsimons, 2007), participants then completed an anagram challenge in which they attempted to derive as many words four or more letters in length as possible out of four source words (bimodal, igneous, answer, and curried).

Moderators. Subsequent to the manipulation and key dependent measures, participants completed the PWE scale (Katz & Hass, 1988) and DUREL (Koenig & Buessing, 2010), as well as the single item religiosity measure from the original experiment (Poehlman, 2007; Uhlmann et al., 2011).

Demographics. Participants fill out a set of demographic items paralleling those from Study 1.

Awareness probe. A set of questions assessed awareness of the influence of the priming manipulation (Poehlman, 2007; Uhlmann et al., 2011; adapted from Bargh & Chartrand, 2000). The numeric probe item asked “Did the sentence unscramble task influence your performance on the anagram task in any way?” (1 = no, 5 = not sure, 9 = yes). The subsequent free response item inquired “If yes, please explain how and why it influenced you in your own words.”

Attention check. Participants completed the same instructional attention check as in Study 1. All participants who failed to follow the simple instruction to “please select strongly disagree” on a Likert-type scale were excluded from the analyses.

7.2. Results
7.2.1. PureProfile sample
Overall, no significant differences emerged in anagram performance between the salvation prime and neutral prime conditions, F(1, 1120.58) = 0.034, p = .854, d = −0.011. Also unlike in the original research, the priming manipulation did not interact with country: USA vs other nation (UK & Australia) F(1, 1119.92) = 0.01, p = .989, d = 0.001, USA vs UK, F(1, 820.98) = 0.68, p = .41, d = −0.0576, or USA vs Australia, F(1, 804.37) = 0.682, p = .409, d = 0.058. The salvation prime effect on task performance further failed to emerge in any of the individual countries, including the United States, F(1, 507.73) = 0.018, p = .892, d = −0.012, Australia, F(1, 298) = 0.908, p = .341, d = −0.111, and the United Kingdom, F(1, 312) = 0.838, p = .361, d = 0.1036. New England region also did not moderate the results, F(1, 1124) = 0.019, p = .89, d = −0.0079.

Note that any significant interactions between prime condition and moderator measures must be interpreted in light of the absence of any main effect of the primes. Whether the participant was of Protestant faith did not interact with the priming manipulation to predict anagram performance, F(1, 1112.72) = 0.24, p = .625, d = 0.029, the single item measure of religiosity did not significantly interact with prime condition, F(1, 1119.59) = 3.553, p = .06, d = −0.1127, scores on the DUREL religiosity scale significantly interacted with prime condition, F(1, 1119.95) = 6.64, p = .01, d = −0.154, and scores on the PWE scale significantly interacted with prime condition, F(1, 1117.55) = 4.202, p = .041, d = −0.123. The directions of these latter two interactions were, however, contrary to any of the present theories of work morality. Specifically, participants high in religiosity (DUREL) exhibited directionally but non-significantly worse work performance in the salvation prime condition relative to the neutral primes, F(1, 227) = 3.043, p = .082, d = −0.232, with the least religious participants exhibiting directionally but not significantly better work performance in the salvation prime condition, F(1, 265.86) = 1.722, p = .191, d = 0.161. Similarly, participants who endorsed the Protestant Work Ethic performed directionally but not significantly worse on a subsequent work task after being primed with salvation relative to neutral concepts, F(1, 177) = 0.923, p = .338, d = −0.144, whereas low-PWE participants worked directionally but nonsignificantly harder in response to the primes, F(1, 167.94) = 0.059, p = .809, d = 0.037.

7.2.2. Laboratory data collections
In the laboratory data collections, there was again no main effect of the priming manipulation on work performance, F(1, 728.58) = 0.269, p = .604, d = 0.038, or interaction between nation of data collection and the experimental manipulation, USA vs. Republic of Ireland F(1, 637.15) = 0.045, p = .831, d = −0.017, USA vs. Canada F(1, 648.16) = 0.25, p = .617, d = 0.0393. The salvation prime effect did not emerge when the USA sample, F(1, 649.36) = 0.165, p = .685, d = 0.051, Republic of Ireland sample, F(1, 778) = 0.166, p = .685, d = 0.093, and Canadian sample, F(1, 89) = 0.06, p = .807, d = −0.0525, were analyzed separately. Regional differences (New England vs. other) were not tested since USA laboratory data collections intentionally focused on the northeastern United States (i.e., New York State and Connecticut).

The single-item measure of religiosity, F(1, 721.64) = 2.375, p = .124, d = 0.115, DUREL, F(1, 727.19) = 3.423, p = .065, d = 0.137, and PWE scale, F(1, 727.91) = 0.012, p = .912, d = −0.008 did not moderate the results of the crowdsourced data collection in partner laboratories. Unlike in the PureProfile sample, in the laboratory data collections Protestant religious faith interacted with the priming manipulation, F(1, 711.55) = 5.764, p = .017, d = −0.18. The pattern of the interaction was directly contrary to the religious differences account, such that Protestants performed significantly worse on the work task in the salvation prime condition relative to the neutral prime condition, F(1, 72.75) = 5.08, p = .027, d = 0.5285, whereas non-Protestants worked directionally but nonsignificantly harder when primed with salvation, F(1, 636.78) = 1.62, p = .204, d = 0.1009.

7.3. Discussion
In contrast to the complex pattern of experimental and cross-national results from Study 1, the priming replication (Study 2) returned null effects and little to no reliable evidence of moderation. Whether the experimental paradigm was administered electronically online, or in paper-pencil format in more controlled conditions, played no apparent role in the primary outcome. Implicitly activating religious concepts such as redeem and divine had no reliable main effect on subsequent task performance, either in the United States or in the other nations examined (UK, Australia, Canada, and the Republic of Ireland).

Sharply contradicting the predictions of the religious differences account, in the online sample less religious participants were more likely than religious participants to exhibit the salvation prime effect on work performance. In the online sample, the direction of moderation from endorsement of the Protestant Work Ethic was likewise precisely opposite to what one might expect based on prior scholarship on work morality (Weber, 1904/1958). However, these individual-differences moderators failed to replicate in the laboratory data collections. Further, a recent meta-analysis concluded that participants who are more religious are more susceptible to the activation of religious concepts (Shariff, Willard, Andersen, & Norenzayan, 2016), a pattern of results
opposite to that for DUREL religiosity scores in our online investigation. Self-identification as a Protestant interacted with the priming manipulation in the crowdsourced laboratory data collection, in the direction contrary to the religious differences account, but this interaction failed to replicate in the online sample. Overall, this decidedly mixed set of results calls for further pre-registered, cross-national investigations of the role of individual religiosity and related ideologies in responses to the temporary accessibility of religion (van Elk et al., 2015). Subtly increasing the accessibility of religious concepts could potentially influence other dependent measures, such as moral judgments and actions (Shariff et al., 2016; cf. Billingsley, Gomes, & McCullough, 2018). However, despite a few caveats (see Supplements 11 and 12), the present results regarding salvation priming and work productivity are most consistent with the false positives account.

8. Forecasting survey

Given the findings from both Studies 1 and 2 are quite contrary to the original theorizing (Poehlman, 2007; Uhlmann et al., 2009, 2011), an interesting question is whether the replication results are predictable by psychologists and other scholars. In a forecasting survey accompanying the present project, independent scientists were provided with descriptions of the competing theories and asked to try to predict the replication effect sizes associated with each targeted effect. Two hundred and twenty-one colleagues made predictions about the target age and needless work effect, needless work main effect (works vs. retires) in the same “postal worker” scenario, tacit inference effect, intuitive work morality effect, and salvation prime effect, across each online sample for which data was collected (MTurk: USA and India; PureProfile: New England U.S. states, non-New-England U.S. states, Australia, and United Kingdom). For each targeted effect, we also asked forecasters to predict the aggregated effect size across samples for four key theoretical moderators: participant religious affiliation (Protestant or not), religiosity (DUREL score), Protestant work ethic endorsement, and education level.

Prior investigations demonstrate that scientists can anticipate simple condition differences based on mere examination of study abstracts or materials (Camerer et al., 2016; DellaVigna & Pope, 2018; Dreber et al., 2015; Forsell et al., 2019). We examined, for the first time, whether they can likewise accurately predict empirical outcomes when the same research paradigms are repeated in multiple cultural contexts. See https://osf.io/7uhcg/ and Supplements, 4, 5, and 6 for the forecasting survey pre-registered analysis plan, survey materials, and a detailed report of the results. Summarizing briefly, in our primary hypothesis test, we found a statistically significant positive overall association between realized and predicted effect sizes, $\beta = 0.157$, $p = .0005$. The Pearson correlation between the mean predicted effect size of each of the 48 effects replicated and the observed effect size was likewise significant, $r = 0.704$, $p < .0001$. Thus, even when the pattern of results being predicted is quite complex, the accuracy of scientific forecasters remains a robust phenomenon (Landy et al., 2020; Tierney et al., in press).

At the same time, comparing the absolute differences between the forecasted and realized effect sizes (Cohen’s $d$) for each original effect underscores that this accuracy was less than perfect. Specifically, forecasted effect sizes averaged across populations were significantly different from the realized effect sizes, aggregated for each key effect via a random effect meta-analysis, for two of the five key effects at the $p < .005$ level (Benjamin et al., 2018) and for a third effect at the traditional $p < .05$ level. For the needless work main effect (works vs. retires), mean forecasts = 0.3233, and meta analyzed realized effect size = 0.6524, with the difference between the two statistically significant, $p < .0001$, such that participants underestimated the replication effect size. Forecasters likewise believed the tacit inferences effect would be smaller than it turned out to be, mean forecasts = 0.3114, meta analyzed effect size = 0.5053, $p = .0055$. In contrast, for the target age moderating needless work effect, participants systematically overestimated the effect size, mean forecasts = 0.2461, meta analyzed realized effect size = 0.032, $p < .0001$, believing the effect would replicate when in fact it did not. Forecasters expected a small but significant overall salvation prime effect, mean forecasts = 0.0972, which did not emerge, meta analyzed effect size = 0.0104, but the difference between forecasted and realized effect sizes was not statistically significant, $p = .9181$. Finally, for the intuitive work morality effect, mean forecasts = 0.2520, were closely aligned with the meta analyzed realized effect size = 0.2568, with no significant difference between them, $p = .954$.

Overall, forecasters did quite well in anticipating the replication outcomes, although they were less accurate in predicting absolute effect sizes than their direction and relative ordering. Based on their pattern of forecasted results, these independent scientists appear to have endorsed the general moralization of work theoretical perspective, in that they forecasted all the original effects would emerge and further would do so across cultures (see Tables S6–3 and S6–7 in Supplement 6). For the most part this facilitated successful forecasts, the general moralization of work being the most empirically supported theory in this replication initiative. The major exceptions are of course the salvation prime effects and target age and needless work effects, which failed to replicate as anticipated by the false positives account. Further research should continue to examine the extent to which scientists are able to anticipate cross-cultural replication results, ideally using a larger number of cultural populations than the relatively small set sampled here, as well as effects that exhibit greater heterogeneity across societies.

9. General discussion

This large-scale creative destruction replication initiative, which involved over eight thousand participants from half a dozen nations, systematically competed theories of culture and work morality against one another. In addition to directly replicating a set of original experimental effects central to the theory of Implicit Puritanism (Poehlman, 2007; Uhlmann et al., 2009, 2011), we included new measures and populations facilitating novel conceptual tests of the predictions of the Implicit American Exceptionalism, general moralization of work, self-expression values, social class, religious differences, and regional folkways accounts of work values.

The observed pattern of experimental and cross-national differences and similarities severely undermines the original theory of Implicit Puritanism. In every instance, the targeted effect either failed to replicate entirely, or unexpectedly replicated in multiple cultures when it had been predicted to emerge only among Americans. Two original effects—specifically, the moderating effect of target age on judgments of needless work, and influence of implicit salvation primes on work behavior—failed to replicate in all populations examined and are identified as likely false positives (Poehlman, 2007; Uhlmann et al., 2011). In contrast, the main effect of moral praise for a lottery winner who continues to work, and false memories consistent with an implicit link between work and sex morality (Poehlman, 2007; Uhlmann et al., 2009), were robust across cultures (India, the United States, Australia, and United Kingdom). Finally, the effects of an intuitive mindset on moral judgments of needless work replicated across the USA, Australia, and UK samples, but not the India sample. The emergence of a number of key effects across a number of different nations sharply contradicts Implicit Puritanism’s core theoretical claim of a unique American work morality.

Rather than leaving a theoretical void in the form of reduced confidence in the original findings and the underlying ideas, these results point in new theoretical directions. Specifically, they provide initial evidence that work behavior elicits strong moral intuitions across cultures, and that the gap between intuitive and deliberative feelings about work could be larger in wealthier societies. Personal religion (e.g.,
Protestant faith), degree of religiosity, socioeconomic status, and region of the United States (e.g., historically Puritan-Protestant New England) did not moderate any of the observed experimental effects, failing to support the associated accounts of work values. More investigations involving larger samples of countries, especially societies in which survival rather than self-expression values are widely endorsed (Inglehart, 1997; Inglehart & Welzel, 2005), and with varied historic backgrounds and diverse workways (Sanchez-Burks & Lee, 2007) are needed before drawing strong conclusions (Simons, Shoda, & Lindsay, 2017). At the same time, we believe the present investigation highlights the feasibility and generative nature of the creative destruction approach to replication, in identifying the most promising theories to guide further empirical research.

9.1. A Bayesian multiverse analysis

A pre-registered (https://osf.io/pgfm8) Bayesian multiverse analysis examined the consequences of different inclusion criteria, variable operationalizations, and statistical approaches for the replication results (see Haaf, Hoogeveen, Berkhow, Gronau, & Wagenmakers, 2020; Haaf & Rouder, 2017; Rouder, Haaf, Davis-Stober, & Hilgard, 2019). Overall, the results of the Bayesian multiverse are highly consistent with the frequentist analyses reported earlier (see Supplement 9 for a more detailed report). Strong evidence emerged that the tacit inference effect and overall valorization of needless work (regardless of target age or participant mindset) are true-positives and further present across samples. Although less strongly, the data also support an overall intuitive mindset effect across all samples combined. Finally, strong evidence emerged against the target age and needless work effect, and the salvation prime effect. The latter remained unsupported even in those conditions pre-specified as most favorable for priming effects, specifically controlled laboratory studies and excluding participants suspicious of being influenced or whom had failed to complete all the scrambled sentences. The Implicit Puritanism model performed worse than the winning model for all six original effects. The General Moralization of Work and False Positives accounts were the best fitting models overall, depending on the effect in question. The Protestant work ethic was found to positively predict the main effects of needless work (i.e., preference for worker over retiree regardless of target age or participant mindset), but such judgments did not vary across cultures as predicted by the Explicit American Exceptionalism account or any of the other competing theories (see Furnham et al., 1993, and Leong, Huang, & Mak, 2014, for evidence “Protestant” work ethic beliefs are broadly applicable). Empirical estimates converged across the different universes of potential analyses (see Fig. S9–1 in Supplement 9). Effects that were not replicated in the primary analyses were not supported under any specification in the Bayesian multiverse, and replicable effects found evidentiary support across many different specifications.

9.2. False inferences in cross-cultural experiments

The present replication results highlight potential broader challenges for producing robust and reliable cross-cultural experimental research (Milfont & Klein, 2018). We define an x-cultural experiment as a study containing a manipulation (e.g., random assignment to condition A or condition B) and sampling at least two distinct cultural populations (e.g., university students in China and the United States). More broadly than the typical concerns about false positive findings (Open Science Collaboration, 2015; Simmons et al., 2011), such cross-cultural investigations are open to false inferences about patterns of experimental results across different human populations. In addition to the expected condition differences failing to emerge (e.g., salvation prime effect, target age and needless work effect), cross-cultural findings may prove over-robust, in other words emerging in societies where they were theoretically expected not to (e.g., the tacit inferences effect and intuitive work morality effect replicating outside the United States). False inferences could also involve concluding a phenomenon is culturally bounded when it is fact universal, and mis-estimating the direction or relative magnitude of an effect between two cultures, among other empirical patterns.

At least two major features of an x-cultural experiment increase the chances of drawing such false conclusions, relative to a simple two-condition experiment in a single population. First, x-cultural studies often rely on an interaction between membership in a cultural group and an experimental manipulation as the key statistical test of the hypothesized cultural difference. Between-subjects interaction tests are typically underpowered unless very large samples are recruited (Simonsohn, 2014; Smith, Levine, Lachlan, & Fiduik, 2002). The Open Science Collaboration’s Reproducibility Project: Psychology replicated 23 of 49 targeted studies (47%) whose key test was a main or simple effect, and only 8 of 37 studies (22%) when the key test was an interaction. Second, x-cultural experiments typically rely on small conveniences samples and attempt to generalize to broader cultures. For example, 100 participants per location might be recruited from universities in New Haven, USA, and Xiamen, China. Since societies are quite heterogeneous (Kitayama et al., 2006; Muthukrishna et al., 2020; Nisbett & Cohen, 1996; Talhelm et al., 2014), this approach may or may not capture central tendencies in the United States and China.

In the present replication initiative a number of the experimental condition differences emerged (i.e., tacit inferences effect, intuitive work morality effect, needless work main effect), yet none of the original condition x national culture interactions (Poehlman et al., 2007; Uhlmann et al., 2009, 2011) were obtained again. The Many Labs 2 crowd initiative likewise failed to replicate previously reported interactions between experimental manipulations and cultural populations, even some considered well-established findings (Klein et al., 2018). To guard against such problems, future cross-cultural behavioral research should seek to collect larger and more varied samples. Researchers might form a network of laboratories and crowdsourcing data collections at multiple sites in each nation (Cuccolo, Irgens, Zlokovich, Grahe, & Edlund, in press; Moshontz et al., 2018), or partner with a survey firm to systematically sample respondents from different regions of the same country, ideally achieving representative sampling.

Different cultural theories predict distinct patterns of empirical results, and some may be more subject to false inferences than others. In a presence-absence pattern, an experimental effect is hypothesized to emerge in one culture, but not in the other. Most of the original Implicit Puritanism studies predicted and found such a pattern, for example an implicit link between work and sex morality among Americans, but not members of other cultures. In a reduced pattern, the effect is in the same direction for both cultures, but diminished in some cultures relative to others (e.g., varying degrees of loss aversion among members of different nations; Arkes, Hirshleifer, Jiang, & Lim, 2010). Finally, in a reversal pattern, the effects of an experimental manipulation are expected to fully reverse between a focal culture and comparison culture. For example, Gelfand et al. (2002) predicted and found that whereas American participants were significantly more disposed to accept positive than negative feedback, Japanese participants exhibited the opposite pattern, accepting more personal responsibility for negative than positive feedback. We suggest that future theorizing on culture focus on developing such reversal predictions, which rely on better powered crossover interactions, and are less likely to be confounded by measurement challenges than presence-absence patterns or reduced patterns.

9.3. The broader utility of the creative destruction approach

The present culture and work morality project is the first of several recent initiatives applying the creative destruction approach to replication to previously published findings from our research group (see Tierney et al., in press, for a review). Adding to the recent deluge of failed replications of experimental behavioral findings (e.g., Klein et al.,
none of these replication studies succeeding in reproducing the original patterns of results. However, unlike prior replication initiatives, we were able to obtain positive evidence for alternative theoretical accounts (Supplement 13).

We believe this highlights the general utility of the creative destruction approach to replication, which seeks to combine theory pruning methods from the management literature (Leavitt et al., 2010), with best practices from the open science movement in psychology such as pre-registration (Van't Veer & Giner-Sorolla, 2016; Wagenmakers et al., 2012) to achieve critical tests (Mayo, 2018) of competing intellectual ideas. Unlike traditional replication approaches, in which the original finding is tested against the expectation of null effects, the creative destruction approach seeks to identify the strongest theory currently operating in a given intellectual space.

Of course, not all research topics and original findings are well suited for large-scale competitive theory testing. As discussed at greater length by Tierney et al. (in press), the creative destruction approach is best suited to mature research areas with substantial published evidence, common methodological approaches, and well-developed theories that make precise, bounded predictions distinct from those of other theories. In contrast, traditional replications simply repeating the original method are better suited to confirming or disconfirming potential new breakthrough findings. Scientists should carefully allocate scarce replication resources for maximum impact, leveraging the methods best suited to the situation. It is our hope the present line of research contributes to a Replication 2.0 movement, in which rather than solely probing the reliability of past findings, scientists also focus on replacing them with new and improved accounts of human behavior.

CRedit authorship contribution statement

The first three and last authors contributed equally. WT, J. Hardy, CE, & EU designed the culture and work replication studies. WT, J. Hardy, CE, LAV, KD, EI, HC, AG, MV, JW, IS, MA, JM, & EU served as replicators. WT, J. Hardy, & CE carried out the frequentist statistical analysis of the replication results. SH & J. Haaf designed, carried out, and wrote the report of the multivariate analysis of the results. DV, EC, MG, AD, MJ, & TP designed, ran, and analyzed, and wrote the report of the forecasting study. J. Huang designed, carried out, and wrote the replication reporting the response effort analyses. Members of the “Culture & Work Morality Forecasting Collaboration” lent their expertise as forecasters, and are listed with full names and affiliations in Appendix 1. All authors collaboratively edited the final project report.

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 Forskningsstiftelse), the Knut and Alice Wallenberg Foundation and the Marianne and Marcus Wallenberg Foundation (Anna Dreber is a Wallenberg Scholar), and Anna Dreber and Magnus Johansson are grateful for a grant from the Swedish Foundation for Humanities and Social Sciences.

Appendix 1. Names and affiliations for the culture & work morality forecasting collaboration

The following colleagues lent their time and expertise as forecasters: Ajay T. Abraham, Seattle University
Matus Adamkovic, Institute of social sciences, CSPS Slovak Academy of Sciences, and Institute of psychology, Faculty of Arts, University of Presov
Jais Adam-Troian, College of Arts and Sciences, American University of Sharjah, Sharjah, UAE

Elena Agadullina, National Research University Higher School of Economics
Handan Akkas, Ankara Science University, Department of Management Information Systems
Dorsa Amir, Boston College
Michele Anne, University of Nottingham Malaysia
Kelly J. Arbeau, Trinity Western University
Mads N. Arnestad, BI Norwegian Business School, Department of Leadership and Organization
John Jamir Benzon Aruta, De La Salle University
Mujeeba Ashraf, Institute of Applied Psychology, University of the Punjab, Lahore
Ofer H. Azar, Ben-Gurion University of the Negev
Bradley J. Baker, University of Massachusetts
Gabriel Baník, University of Presov
Sergio Barbosa, School of Medicine and Health Sciences, Universidad del Rosario
Ana Barbosa Mendes, IITEC, Faculty of Psychology and Educational Sciences, KU Leuven, Belgium
Ernest Baskin, Saint Joseph’s University
Christopher W. Bauman, University of California, Irvine
Jozef Bavolar, Pavol Jozef Safarik University in Kosice
Stephanie E. Beckman, The Chicago School of Professional Psychology
Theiss Bendixen, Department of the Study of Religion, Aarhus University
Aaron S. Benjamin, University of Illinois at Urbana-Champaign
Ruud M.W.J. Berkers, Max Planck Research Group: Adaptive Memory, Max Planck Institute for Human Cognitive & Brain Sciences, Leipzig, Germany
Amit Bhattacharjee, INSEAD
Samuel E. Bodily, Darden Business School, University of Virginia
Helena Bonache, Universidad de La Laguna
Vincent Bottom, Washington University School of Medicine in St. Louis
Cameron Brick, University of Amsterdam
Neil Brildenafil, Bow Valley College and University of Alberta
Stephanie E. V. Brown, Texas A&M University
Jeffrey Buckley, Faculty of Engineering and Informatics, Athlone Institute of Technology, Westmeath, Ireland and Department of Learning, KTH Royal Institute of Technology, Stockholm, Sweden
Max E. Butterfield, Point Loma Nazarene University
Neil R.Caton, The University of Queensland
Zhang Chen, Department of Experimental Psychology, Ghent University
Jessica F. Chen
Fadong Chen, School of Management, Zhejiang University
Irene Christensen, GUST University
Ensari E. Cicerald, Nisantasi University
Simon Columbus, University of Copenhagen
David J. Cox, GuideWell, Endicott College
Emiel Cracco, Department of Experimental Psychology, Ghent University
Daina Crafa, CrafaLab, Interacting Minds Centre, Aarhus University
Jamie Cummins, Ghent University
Jo Cutler, The University of Birmingham
Zech O. Dahms, UW-Milwaukee, MBA
Alexander F. Danvers, University of Arizona
Liora Daum-Avital, Ben-Gurion University of the Negev
Ian G. J. Dawson, University of Southampton
Martin V. Day, Memorial University of Newfoundland
Philippe O. Deprez, Indiana University Southeast
Erik Dietl, Loughborough University
Eugen Dimant, University of Pennsylvania
Gönül Doğan, University of Cologne
Artur Domurat, Centre for Economic Psychology and Decision
Economics, Erasmus School of Law, Erasmus University Rotterdam
Department of Economics, Leiden University
Jozef Šafárik University in Košice, Slovakia
Palacky University Olomouc, Olomouc, Czech Republic
University of Washington
Cooperation, Stanford University
Economics
Innsbruck
University of Geneva
Sciences, Kozminski University, Warsaw, Poland
Terence D. Dores Cruz, Vrije Universiteit Amsterdam
Christiene de Plessis, Singapore Management University
Dmitrii Dubrov, National Research University Higher School of Economics
Esha Dwivedi, Virginia Tech
Christian T. Elbaek, Aarhus University, Department of Management
Mahmoud M. Elsherif, University of Birmingham
Thomas R. Evans, School of Psychological, Social and Behavioural Sciences, Coventry University
Sarahanne M. Field, University of Groningen
Mustafa Fırat, University of Alberta
Zoé Francis, University of the Fraser Valley
Yoav Ganzach, Ariel University and Tel Aviv University
Rica Gautam, University of Delaware
Brian Gearin, University of Oregon
Sandra J. Geiger, University of Amsterdam
Omid Ghasemi, Macquarie University
Lorenz Graf-Vlachy, ESCP Business School
Lu Gram, Institute for Global Health, University College London
Dmitry Grigoryev, National Research University Higher School of Economics
Rosanna B'Guadagno, Center for International Security and Cooperation, Stanford University
University of Washington
Sebastian Hafenbrädl, IESE Business School
Linda Hagen, University of Southern California
David Hagmann, Harvard Kennedy School
Jonathan J. Hammersley, Western Illinois University
Hyemin Han, University of Alabama
Andree Hartanto, Singapore Management University
Renata M. Heilman, Babes-Bolyai University, Department of Psychology
Alexander P. Henkel, Open University of the Netherlands
Felix Holmeister, Department of Economics, University of Innsbruck
Qian Huang, University of Miami
Tina S.-T. Huang, University College London
Barbora Hubena, Ministerstvo zdravotnictví České republiky
Jeffrey R. Huntsinger, Loyola University Chicago
Hirotaka Imada, University of Kent
Michael J. Ingels
Tatsunori Ishii, Waseda University
Chitrankan Jain, Birla Institute of Technology and Science Pilani
Konrad Jamro, St. Bonaventure University
Kristin Jankowsky, University of Kassel
Steve M. J. Janssen, University of Nottingham Malaysia
Nilotpal Jha, Singapore Management University
Fanli Jia, Seton Hall University
Daniel Jolles, University of Essex
Bibiana Jozefiakova, Olomouc University Social Health Institute, Czech Republic
Pavel Kačmár, Department of Psychology, Faculty of Arts, Pavol Šafárik University in Košice, Slovakia
Kyriaki Kalimeri, ISI Foundation, Turin, Italy
Jaroslaw Kantorowicz, Institute of Security and Global Affairs and Department of Economics, Leiden University
Elena Kantorowicz-Reznichenko, Rotterdam Institute of Law and Economics, Erasmus School of Law, Erasmus University Rotterdam
Matthias Kasper, Tulane University and University of Vienna
Edgar E. Kausel, Pontificia Universidad Católica
Lucas Keller, Department of Psychology, University of Konstanz
Yeon Joon Kim, University of Cambridge
Minjae J. Kim, Boston College
Mikael Knutsson, Linköping University
Olga Kombez, Loughborough University
Marta Kowal, Institute of Psychology, University of Wroclaw, Wroclaw, Poland
Tei Laine
Aleksandra Lazić, University of Belgrade
Johannes Leder, University of Bamberg
Margarita Leib, University of Amsterdam
Carmel A. Levitan, Occidental College
Alex Lloyd, Royal Holloway, University of London
Ronda F. Lo, York University
Andrey Lovakov, National Research University Higher School of Economics
Timo Lükë, TU Dortmund University
Albert L. Ly, Loma Linda University
Victor S. Maas, University of Amsterdam
Zoe Magraw-Mickelson, Ludwig-Maximilians-Universität München
Elizabeth A. Mahar, University of Florida
James C. Marcus, Evidera
Melvin S. Marsh, Georgia Southern University
Abigail A. Marsh, Georgetown University
Chris C. Martin, Georgia Institute of Technology
Marcel Martončík, Institute of Psychology, Faculty of Arts, University of Presov, Slovakia
Sébastien Massoni, Université de Lorraine, Université de Strasbourg,
CNRS, BETA, Nancy, France
Theodore C. Masters-Waage, Singapore Management University
Akiko Matsu, Tokai Gaku University
Jens Mazié, TU Dortmund University
Randy J. McCarthy, Northern Illinois University
Smriti Mehta, UC Berkeley
Chanel Meyers, Whitman College
Ewa AureliaMiedlarzewska, Geneva Finance Research Institute, University of Geneva
Philip Millroth, Department of Psychology, Uppsala University, Sweden
Marina Milyavskaya, Carleton University
Talya Miron-Shatz, Ono Academic College, Israel
Pooja D. Mistry
Karina Mitropoulou
Mao Mogami, New York University
David Moreau, School of Psychology and Centre for Brain Research, The University of Auckland
Yuki Mori, Graduate School of Human-Environment Studies, Kyushu University
Annalisa Myer, University of Virginia
Philip W. S. Newall, CQUniversity
Phuong Linh L. Nguyen, University of Minnesota
Annika S. Nieper, Vrije Universiteit Amsterdam
Gustav Nilsonne, Karolinska Institutet and Stockholm University
Abigail L. Ly, Loma Linda University
Timo Lükë, TU Dortmund University
Andrey Lovakov, National Research University Higher School of Economics
Tilo Oelhafen, Bern University of Applied Sciences
Nurit Nobel, Stockholm School of Economics
Stephan Oelhafen, Bern University of Applied Sciences
Anna A. O' Mahony, Cardiff University, U.K.
Mehmet A. Orhan, PSB Paris School of Business
Flora Oswald, The Pennsylvania State University
Tobias Otterbring, University of Agder
Philipp E. Otto, European University Viadrina
Mariola Paruzel-Czachura, University of Silesia in Katowice, Institute of Psychology
Gerit Pfuhl, UiT The Arctic University of Norway
Jessica M. Plourde, Fordham University
Madeleine Pownall, School of Psychology, University of Leeds
Anushree Prashant, University of Glasgow, Scotland, UK, and GEMS
World Academy, Dubai, UAE
Marjorie L. Prokosch, Tulane University
John Protzko, University of California, Santa Barbara
Danka B. Purić, University of Belgrade, Faculty of Philosophy,
Department of Psychology and Laboratory for Research of Individual Differences
M. S. Rad, New School
Louis Raes, Tilburg University
Rima-Maria Rahal, Tilburg University
Liz Redford
Christopher M. Redker, Ferris State University
Niv Reggev, Ben-Gurion University of the Negev
Caleb J. Reynolds, Florida State University
Marta Roczniwksa
Ivan Ropovik, Charles University, Faculty of Education, Institute for Research and Development of Education & University of Presov, Faculty of Education
Lukas Rösseler, Harz University of Applied Sciences, University of Bamberg
Robert M. Ross, Macquarie University
Amanda Rotella, Department of Psychology, University of Waterloo, Canada
Raluca Rusu
Michael Schaeber, Lee Kong Chian School of Business, Singapore Management University
William M. Schiavone, University of Georgia
Landon Schnabel, Stanford University and Cornell University
Brendan A. Schuetze, The University of Texas at Austin
Irene Scopelliti, City, University of London
Zeev Shbudiner, Ariel University
Deborah Shulman
Victoria Song, Fordham University
Tabea Springer, Washington University in St. Louis
Eirik Strømland, University of Bergen
Kevin P. Sweeney, Western Kentucky University
Maria A. Terskova, National Research University Higher School of Economics
Kian Siong Tey, INSEAD
Franisca Ting, University of Illinois at Urbana-Champaign
Joshua M. Tybur, Vrije Universiteit Amsterdam
Karolina Urbanska, Department of Psychology, University of Sheffield
Paul Vanags, University of Oxford Brookes
Joseph A. Vitriol, Stony Brook University
Alisa Voslinsky, Department of Industrial Engineering and Management, Sami Shamoon Academy of Engineering, Ashdod, Israel
Marek A. Vranka, Charles University
Lauren E.T. Wakabayashi, Loma Linda University
Hanne M. Watkins, UMass Amherst
Erik C. Westgate, University of Florida
Margaux N. A. Wink, Department of Psychology, Columbia University
Jan K. Woike, University of Plymouth, UK
Conny E.Wollbrant, University of Stirling
Amanda J. Wright, Washington University in St. Louis
Qinyu Xiao, University of Hong Kong
Alon Yakter, Tel Aviv University
Yurik Yang, Fakultas Psikologi Universitas Indonesia
Zhixu Yang, Purdue University
Siu Kit Yeung, The University of Hong Kong
Onurcan Yilmaz, Kadir Has University
Meltem Yucel, University of Virginia
Cristina Zogmaister, Università degli Studi di Milano-Bicocca
Ro’i Zultan, Ben-Gurion University of the Negev

Appendix 2. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2020.104060.

References
