RUNNING HEAD: CREATIVE DESTRUCTION APPROACH TO REPLICATION

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

Warren Tierney	INSEAD
Jay Hardy III	Oregon State University
Charles R. Ebersole	University of Virginia
Domenico Viganola	The World Bank
Elena Giulia Clemente	Stockholm School of Economics
Michael Gordon	Massey University
Suzanne Hoogeveen	University of Amsterdam
Julia Haaf	University of Amsterdam
Anna Dreber	Stockholm School of Economics, University of Innsbruck
Magnus Johannesson	Stockholm School of Economics
Thomas Pfeiffer	Massey University
Jason L. Huang	Michigan State University
Leigh Ann Vaughn	Ithaca College
Kenneth G. DeMarree	University at Buffalo, The State University of New York
Eric R. Igou	University of Limerick
Hanah Chapman	Brooklyn College CUNY
Ana Gantman	Brooklyn College CUNY
Matthew Vanaman	Brooklyn College CUNY
Jordan Wylie	Queens College CUNY
Justin Storbeck	Queens College CUNY
Michael R. Andreychik	Fairfield University
Jon McPhetres	Durham University
Culture & Work Morality Forecasting Collaboration	Many Institutions
Eric Luis Uhlmann	INSEAD

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

Author Contributions 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, JS, 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 Bayesian multiverse analysis of the results. DV, EC, MG, AD, MJ, & TP designed, ran, analyzed, and wrote the report of the forecasting study. J.Huang designed, carried out, and wrote the supplement 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.

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

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.

KEYWORDS: replication; theory testing; falsification; implicit social cognition; priming; work values; culture

The present initiative aimed to assess the robustness, generality, and cultural boundedness of prior findings on Implicit Puritanism, an 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, 1840/1990; 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 pre-specify 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 nonconscious. 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.

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, & Tweneya, 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, Schlange & 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; Mellers, Hertwig, & Kahneman, 2001; Mayo, 2018; 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 & Tikochinsky, 2001) in order to reduce the dense theoretical landscape of the sciences (Hambrick, 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" (Vandenberg & Grelle, 2008).

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, 1990; 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 meritoriously 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, deliberatively 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, Seibt, & 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. Although many faiths draw an association between sexual

restraint and divine purity, Protestantism is distinct in also placing work in the realm of the divine. Via the principle of cognitive balance (Greenwald et al., 2002; Heider, 1958), their mutual link with divine salvation forges a unique connection between Puritan sex values and the Protestant work ethic in the minds of Americans. As a result, thoughts or judgments related to hard work activate inferences and values related to sexuality, and vice versa.

Implicit Puritanism theory thus seeks to bridge prior cultural analyses of the United States (de Tocqueville, 1840/1990; Lipset, 1996) with theoretical and empirical work on implicit social cognition as applied to unconscious cultural stereotyping (Greenwald & Banaji, 1995) and principles of cognitive balance (Greenwald et al., 2002). Research in the social cognitive tradition suggests that because cultural stereotypes are ingrained and operate unconsciously, they often affect the judgements and behaviors of consciously egalitarian and consciously inegalitarian individuals to similar degrees. Critically to Implicit Puritanism theory, because the effects of the Puritan-Protestant heritage of the U.S. are held to be pervasive and unconsciously transmitted, demographic differences based on consciously endorsed religion (i.e., whether the person is a Protestant or not) and explicit religiosity (i.e., devout faith vs. atheism) should not emerge. All that should matter when it comes to exhibiting the predicted effects, for instance of subtly priming concepts related to religion (Poehlman, 2007; Uhlmann et al., 2011), is whether the person is an American or not. The absence of any moderating effects of self-reported religion or religiosity in past empirical studies thus goes hand in hand with a lack of evidence of conscious awareness (e.g., on probe questions), in supporting the original theorizing (Poehlman, 2007; Uhlmann et al., 2009, 2011). Such null effects are also broadly consistent with research on social tuning (Sinclair, Dunn, & Lowery, 2005; Sinclair, Lowery, Hardin, & Colangelo, 2005)

and cultural transmission (Boyd, Richerson, & Henrich**,** 2011), which highlight the automatic and unreflective processes via which beliefs can become pervasive in a community.

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), 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).

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").

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 et al., 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., 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").

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.

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.

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 (van't Veer & 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, Shapira, & Landy, 2017; Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013; Klein et al., 2014; McCarthy et al., 2018; O'Donnell 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 et al., 2018, 2014; McCarthy et al., 2018; Olsson-Collentine, et al., in press).

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 deliberatively 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 (Koenig & Büssing, 2010) and collected thousands rather than hundreds of participants to allow for more confident conclusions.

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 plantation culture of the South, Quakers shaped the industrial culture of the Midwest, and Scotch-Irish migration contributed to the ranch 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 primary in the New England states, the U.S. region most influenced by Puritan migration.

In the original research (Poehlman, 2007; Uhlmann 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; Uhlmann 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; Uhlmann 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.

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 and family-oriented (Stephens, Fryberg, & Markus, 2011; Stephens, Fryberg, Markus, Johnson, & Covarrubias, 2012; Snibbe & Markus, 2005; Varnum, Na, Murata, & Kitayama, 2012). Such demographic differences have been observed not only within the United States, but also other cultures, among these Italy, 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.

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.

Explicit American Exceptionalism

Another distinct possibility is that the originally hypothesized cultural differences in work and sex values (Poehlman, 2007; Uhlmann 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 judgements 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 (Tacit 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; Shafer, 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.

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.

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 explicable 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 et al., 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.[1] 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/687h5/). 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.

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 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 fulfilment-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 et al., in press; Schwarz & Strack, 2014; Stroebe & Strack, 2014). However, in the original studies the theoretical underlying processes

are nonconscious 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.

Methods

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.

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.

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 more than 1000 hits approved. Finally, we excluded participants with less than 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 (DUREL; 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.

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.

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, $F(1, 2029) = 4.43$, $p = 0.04$, $d = -0.093$, main effect of work status, $F(1, 2032) = 220.53$, $p < .001$, $d = 0.65$, and two-way interaction between age and work status, $F(1, 2027.3) = 4.596$, $p = 0.03$, $d = 0.095$ (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 for the replication was the main effect of target age (23 years or 46 years) within the target works condition. Contrary to the original research (Poehlman, 2007; Uhlmann et al., 2009) the young target who continued to

work did not receive more favorable moral evaluations than an older target who continued to work, $F(1, 1013.74) = 0.035$, $p = 0.851$, $d = -0.012$. Instead, the two-way interaction was driven by the effect of target age within the retires condition, such that the younger retiree was rated more negatively than the older retiree, $F(1, 1009.91) = 8.871$, $p = 0.003$, $d = -0.187$.

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, $F(1, 1018) = 0.268$, $p = 0.605$, $d = -0.032$, 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, $F(1, 492.32) = 0.058$, $p = 0.81$, $d = 0.022$, or USA sample, $F(1, 523) = 0.3$, $p = 0.584$, $d = -0.048$. New England region likewise failed to moderate the effect of target age within the works condition, $F(1, 1018) = 0.678$, $p = 0.411$, $d = 0.052$.

Finally, we examined theoretically relevant individual differences moderators. Neither the single item measure of religiosity, $F(1, 999) = 0.001$, $p = 0.979$, $d = -0.002$, nor the DUREL religiosity scale $F(1, 1018) = 0.251$, $p = 0.616$, $d = 0.031$, nor participant education level $F(1, 985.95) = 1.716$, $p = 0.191$, $d = -0.083$, nor the Protestant Work Ethic $F(1, 1012.15) = 0.167$, $p = 0.683$, $d = 0.026$, nor self-reported religion (Protestant or not) $F(1, 1016.62) = 3.4$, $p = 0.065$, $d = 0.116$, moderated moral judgments of a target who works based on her age.

Needless work study: PureProfile sample. A 2 (target age) x 2 (work status) ANOVA revealed a nonsignificant main effect of target age, $F(1, 4079) = 3.50$, $p = 0.06$, $d = -0.056$, 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 * 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, $F(1, 2052.56) = 1.887$, $p = 0.17$, $d = 0.061$. 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), $F(1, 2061) = 0.303$, $p = 0.582$, $d = 0.024$, the USA vs. Australia, $F(1, 1547) = 0.299$, $p = 0.585$, $d = 0.028$, or the USA vs U.K., $F(1, 1572) = 0.123$, $p = 0.725$, $d = 0.018$. Further, the target age and needless work effect was not significant within the USA sample, $F(1, 1055.87) = 1.959$, $p = 0.162$, $d = 0.086$, Australia sample, $F(1, 487) = 0.086$, $p = 0.77$, $d = 0.027$, or UK sample, $F(1, 514) = 0.266$, $p = 0.606$, $d = 0.046$. New England region again failed to emerge as a moderator $F(1, 2045.35) = 0.002$, $p = 0.97$, $d = 0.001$. The individual differences measures likewise failed to moderate, among these the single item measure of religiosity, $F(1, 2048.17) = 0.482$, $p = 0.488$, $d = 0.031$, DUREL religiosity scale, $F(1, 2056.41) = 0.308$, $p = 0.579$, $d = 0.025$, Protestant religion, $F(1, 2048.9) = 1.067$, $p = 0.302$, $d = 0.046$, education level, $F(1, 1938.1) = 0.436$, $p = 0.509$, $d = -0.03$, and PWE scores, $F(1, 2054.24) = 3.486$, $p = 0.062$, $d = 0.082$.

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 effect $F(1, 2033.89) = 27.38$, $p < 0.001$, $d = 0.232$. 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 (USA vs. India) and intuitive vs. rational responses emerged, $F(1, 2031.84) = 45.027$, $p < 0.001$, $d = 0.2977$, such that the intuitive mindset effect was stronger among American participants than Indian participants (Figure 1). The difference between intuitive and rational responses was clearly observed in the USA sample, $F(1, 1033.77) = 76.019$, $p < 0.001$, $d = 0.543$, but not the India sample, $F(1, 998) = 1.105$, $p = 0.293$, $d = -0.067$. New England region did not moderate the results, $F(1, 2033.61) = 2.009$, $p = 0.156$, $d = -0.0623$.

Self-identified religion (Protestant or not), $F(1, 2029.61) = 0.263$, $p = 0.608$, $d = 0.023$ did not moderate the effect. However education level, $F(1, 1975.39) = 5.006$, $p = 0.025$, $d = -0.101$ 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, $F(1, 1994.13) = 22.807$, $p < 0.001$, $d = -0.214$ and DUREL scale, $F(1, 2031.75) = 24.758$, $p < 0.001$, $d = -0.221$, 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 *negatively* predicted exhibiting the intuitive mindset effect, $F(1, 2033.71) = 10.17$, $p = 0.001$, $d = -0.141$. As discussed below, the moderating effects of education, religiosity and PWE endorsement in the MTurk sample did not replicate in the PureProfile sample.

Intuitive mindset study: PureProfile sample. A significant intuitive mindset effect again emerged in the PureProfile sample, $F(1, 4085.04) = 72.542$, $p < 0.001$, $d = 0.267$. However, as

seen in Figure 1, country (USA vs. UK or Australia) did not moderate the effect, $F(1, 4083.99) = 0.322$, $p = 0.57$, $d = 0.018$. Further, examining each country separately, an intuitive mindset led to more favorable judgments of a target who continued to work not only in the US, $F(1, 2117.49) = 40.965$, $p < 0.001$, $d = 0.278$, but also in the UK, $F(1, 956.66) = 7.338$, $p = 0.007$, $d = 0.175$, and Australia, $F(1, 1010) = 27.352$, $p < 0.001$, $d = 0.329$. New England region again failed to moderate the results, $F(1, 4085.82) = 0.904$, $p = 0.342$, $d = -0.03$. The single item religiosity measure, $F(1, 4071.75) = 0.299$, $p = 0.584$, $d = -0.017$, DUREL religiosity scale, $F(1, 4085.06) = 0.147$, $p = 0.701$, $d = -0.012$, self-identification as a Protestant, $F(1, 4062.19) = 0.079$, $p = 0.778$, $d = -0.009$, and the PWE, $F(1, 4084.25) = 0.931$, $p = 0.335$, $d = -0.031$, failed to emerge as significant moderators. In contrast, education level did significantly moderate the intuitive work morality effect, $F(1, 3866.82) = 13.355$, $p < 0.001$, $d = 0.118$, such that more educated participants were more likely to exhibit a difference between their intuitive and logical judgments. Note that the direction of moderation was directly opposite to that in the MTurk sample, such that these results are extremely mixed and equivocal, providing no overall support for the social class perspective.

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, $F(1, 2029.13) = 89.11$, $p < 0.001$, $d = 0.42$. Further, a significant interaction with country emerged, such that this tacit inferences effect was stronger among American participants than Indian participants, $F(1, 2027.21) = 24.882$, $p < 0.001$, $d = 0.222$ (Figure 2). Although there was a significant between-country difference, the tacit inferences effect was statistically significant not only in the USA, $F(1, 1031.8) = 103.8$, $p < 0.001$, $d = 0.632$, but also India, $F(1, 997.03) = 10.02$, $p = 0.002$, $d =$

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, $F(1, 2023.45) = 0.015$, $p = 0.902$, $d = -0.006$.

The single item measure of religiosity, $F(1, 1985.01) = 1.168$, $p = 0.28$, $d = -0.049$, and whether the participant was of the Protestant faith or not, $F(1, 2023.45) = 1.674$, $p = 0.196$, $d = 0.058$, did not moderate the tacit inferences effect in the MTurk sample. However, the DUREL religiosity scale, $F(1, 2024.49) = 5.718$, $p = 0.017$, $d = -0.106$, and Protestant Work Ethic scale, $F(1, 2024.67) = 10.143$, $p = 0.001$, $d = -0.142$, 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.

Tacit inferences study: PureProfile sample. An overall condition difference supporting the tacit inferences effect again emerged, $F(1, 4085) = 308.506$, $p < 0.001$, $d = 0.550$. Comparing the USA vs. both other countries combined (UK and Australia) did not reveal a significant difference, $F(1, 4071.27) = 0.961$, $p = 0.327$, $d = 0.031$. More fine-grained comparisons between the USA and UK, $F(1, 3078) = 0.012$, $p = 0.911$, $d = 0.034$, and USA and Australia $F(1, 3130) = 2.137$, $p = 0.144$, $d = 0.053$, were also not statistically significant. The tacit inferences effect was significant within the USA, $F(1, 2121) = 181.655$, $p < 0.001$, $d = 0.585$, Australia, $F(1, 1007) = 53.227$, $p < 0.001$, $d = 0.46$, and UK, $F(1, 951.6) = 78.326$, $p < 0.001$, $d = 0.575$, when the samples were tested separately (Figure 2). New England region was not a significant moderator of false memories consistent with an implicit link between work and sex morality, $F(1, 4069.72) = 0.069$, $p = 0.793$, $d = 0.008$.

The individual differences measures, including the single item measure of religiosity, $F(1, 4067) = 0.393$, $p = 0.531$, $d = 0.020$, the DUREL scale, $F(1, 4081) = 0.29$, $p = 0.59$, $d = 0.017$, Protestant religion, $F(1, 4058.1) = 1.193$, $p = 0.167$, $d = 0.044$, and the PWE scale, $F(1, 4079.51) = 3.102$, $p = 0.078$, $d = -0.0552$, 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.

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 their 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 Figure 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.

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).

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.

Design

The study employed a 2 (priming condition: salvation prime or neutral prime) x participant nationality between-subjects design.

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 "coupons 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 & Büssing, 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 unscrambling 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.

Results

PureProfile sample. Overall, no significant differences emerged in anagram performance between the salvation prime and neutral prime conditions, $F(1, 1120.58) = 0.034$, $p = 0.854$, $d = -0.011$. Also unlike in the original research, the priming manipulation did not interact with country: USA vs other nation (UK & Australia) $F(1, 1119.92) = 0.01$, $p = 0.989$, $d = 0.001$, USA vs UK, $F(1, 820.98) = 0.68$, $p = 0.41$, $d = -0.0576$, or USA vs Australia, $F(1, 804.37) = 0.682$, $p = 0.409$, $d = 0.058$. The salvation prime effect on task performance further failed to emerge in any of the individual countries, including the United States, $F(1, 507.73) = 0.018$, $p = 0.892$, $d = -0.012$, Australia, $F(1, 298) = 0.908$, $p = 0.341$, $d = -0.111$, and the United Kingdom, $F(1, 312) = 0.838$, $p = 0.361$, $d = 0.1036$. New England region also did not moderate the results, $F(1, 1124) = 0.019$, $p = 0.89$, $d = -0.0079$.

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, $F(1, 1112.72) = 0.24$, $p = 0.625$, $d = 0.029$, the single item measure of religiosity did not significantly interact with prime condition, $F(1, 1119.59) = 3.553$, $p = 0.06$, $d = -0.1127$, scores on the DUREL religiosity scale significantly interacted with prime condition, $F(1, 1119.95) = 6.64$, $p = 0.01$, $d = -0.154$, and scores on the PWE scale significantly interacted with prime condition, $F(1, 1117.55) = 4.202$, $p = 0.041$, $d = -0.123$. 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, $F(1, 227) = 3.043$, $p = 0.082$, $d = -0.232$, with the least religious participants exhibiting

directionally but not significantly better work performance in the salvation prime condition, $F(1, 265.86) = 1.722$, $p = 0.191$, $d = 0.161$. 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, $F(1, 177) = 0.923$, $p = 0.338$, $d = -0.144$, whereas low-PWE participants worked directionally but nonsignificantly harder in response to the primes, $F(1, 167.94) = 0.059$, $p = 0.809$, $d = 0.037$.

Laboratory data collections. In the laboratory data collections, there was again no main effect of the priming manipulation on work performance, $F(1, 728.58) = 0.269$, $p = 0.604$, $d = 0.038$, or interaction between nation of data collection and the experimental manipulation, USA vs. Republic of Ireland $F(1, 637.15) = 0.045$, $p = 0.831$, $d = -0.017$, USA vs. Canada $F(1, 648.16) = 0.25$, $p = 0.617$, $d = 0.0393$. The salvation prime effect did not emerge when the USA sample, $F(1, 649.36) = 0.165$, $p = 0.685$, $d = 0.051$, Republic of Ireland sample, $F(1, 78) = 0.166$, $p = 0.685$, $d = 0.093$, and Canadian sample, $F(1, 89) = 0.06$, $p = 0.807$, $d = -0.0525$, 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, $F(1, 721.64) = 2.375$, $p = 0.124$, $d = 0.115$, DUREL, $F(1, 727.19) = 3.423$, $p = 0.065$, $d = 0.137$, and PWE scale, $F(1, 727.91) = 0.012$, $p = 0.912$, $d = -0.008$ 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, $F(1, 711.55) = 5.764$, $p = 0.017$, $d = -0.18$. 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, $F(1, 72.75) = 5.08$, $p = 0.027$, $d = -0.5285$, whereas non-Protestants worked directionally but nonsignificantly harder when primed with salvation, $F(1, 636.78) = 1.62$, $p = 0.204$, $d = 0.1009$.

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.

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, $\beta = 0.157$, $p = 0.0005$. The Pearson correlation between the mean predicted effect size of each of the 48 effects replicated and the observed effect sizes was likewise significant, $r = 0.704$, $p < 0.0001$. 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, $p < 0.0001$, 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, $p = 0.0055$. 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, $p < 0.0001$, 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, $p = 0.9181$. 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, $p = 0.954$.

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 effect and target age and needless work effect, 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.

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 explicit 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 the 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.

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, Berkhout, 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 Figure 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.

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, & Fediuk, 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 convenience 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 crowdsource 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 for 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.

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., 2014, 2018; Open Science Collaboration, 2015), 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.

References

Adigun, I. (1997). Orientations to work: A cross-cultural approach. *Journal of Cross-Cultural Psychology, 28,* 352– 355.

Aquino, K., Freeman, D., Reed II, A., Lim, V. K., & Felps, W. (2009). Testing a social-cognitive model of moral behavior: the interactive influence of situations and moral identity centrality. *Journal of Personality and Social Psychology, 97*(1), 123-141.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423-1440.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of organizational research methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods, 11,* 9-34.

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

Arkes, H.R., Hirshleifer, D., Jiang, D.L., & Lim, S.S. (2010). A cross-cultural study of reference point adaptation: Evidence from China, Korea, and the US. *Organizational Behavior and Human Decision Processes, 112*(2), 99-111.

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

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roedeger, III, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder.* (pp.117-150). Washington, DC: American Psychological Association.

Bargh, J. A. (2014). Our unconscious mind. *Scientific American, 30,* 30-37.

Banaji, M. R., Baron, A. S., Dunham, Y., & Olson, K. (2008). The development of intergroup

social cognition: Early emergence, implicit nature, and sensitivity to group status. In S. R. Levy & M. Killen (Eds.), *Intergroup attitudes and relations in childhood through adulthood* (pp. 197-236). Oxford, UK: Oxford University Press.

Bargh, J.A., & Chartrand, T.L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H.T. Reis & C.M. Judd (Eds.), *Handbook of research methods in social and personality psychology*, Second Edition. New York: Cambridge University Press.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230-244.

Barrett, J. L., & Keil, F.C. (1996). Conceptualizing a non-natural entity: Anthropomorphism in God concepts. *Cognitive Psychology, 31*, 219-247.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science, 17*(1), 53-58.

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

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

Berman, J. S., & Reich, C. M. (2010). Investigator allegiance and the evaluation of psychotherapy outcome research. *European Journal of Psychotherapy and Counselling 12*, 11–21.

Billingsley, J., Gomes, C., & McCullough, M. (2018). Implicit and explicit influences of

religious cognition on dictator game transfers. *Royal Society Open Science, 5*(170238). https://doi.org/10.1098/rsos.170238

Boyd, R., Richerson, P.J., & Henrich**,** J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences, 108*, 10918-10925.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist, 27*, 774-775.

Brainerd, C. J., & Reyna, V. F. (2018). Replication, registration, and scientific creativity. *Perspectives on Psychological Science, 13,* 428–432. doi:10.1177/1745691617739421

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3-5.

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

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

Caruso, E. M., Shapira, O., & Landy, J. F. (2017). Show me the money: A systematic exploration of manipulations, moderators, and mechanisms of priming effects. *Psychological Science, 28,* 1148-1159.

Chartrand, T.L., Dalton, A., & Fitzsimons, G.J. (2007). Nonconscious relationship reactance: When significant others prime opposing goals. *Journal of Experimental Social Psychology, 43,* 719-726.

Cohen, A. B., & Varnum, M. E. W. (2016). Beyond east vs. west: Social class, region, and religion as forms of culture. *Current Opinion in Psychology, 8,* 5-9.

Corney, W.J., & Richards, C.H. (2005). A comparative analysis of the desirability of work characteristics: Chile versus the United States. *International Journal of Management, 22,* 159–165.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83,* 1314-1329.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., et al. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1-36.

Cuccolo, K., Irgens, M.S., Zlokovich, M.S., Grahe, J., & Edlund, J.E. (in press). What crowdsourcing can offer to cross-cultural psychological science. *Cross-Cultural Research.*

Darwin, C. (1872). *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. (6th ed.)

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

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

Dorfman, P., Hanges, P. J., & Brodbeck, F. C. (2004). Leadership and cultural variation: The identification of culturally endorsed leadership profiles. In R. J. House, P. J. Hanges, M. Javidan, P. Dorfman, & V. Gupta (Eds.), *Leadership, culture, and organizations: The GLOBE study of 62 societies* (pp. 667–718). Thousand Oaks, CA: Sage Publications.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PLoS One,* 7, e29081

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

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

Dunham, Y., Baron, A. S., & Banaji, M. R. (2006). From American city to Japanese village: The omnipresence of implicit race attitudes. *Child Development, 77,* 1268-1281.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Science, 12*(7), 248-253.

Dunham, Y., Baron, A.S., & Banaji, M. R. (2016). The development of implicit gender attitudes. *Developmental Science, 19*(5), 781–789.

Epstein, S. (1998). Cognitive-experiential self-theory: A dual process personality theory with implications for diagnosis and psychotherapy. In R.F. Bornstein and J. M. Masling (Eds.), *Empirical research on the psychoanalytic unconscious* (Vol. 7, pp. 99-140). Washington, D.C.: American Psychological Association.

Fabrigar, L.R., Wegener, D.R., & Petty, R.E. (in press). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*.

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

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

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

Foucault, M. (1978). *The history of sexuality, vol. 1*; *An introduction,* tr. Robert Hurley. New York: Pantheon.

Furnham, A. (1982). The Protestant work ethic and attitudes towards unemployment. *Journal of Occupational Psychology, 55,* 277-86.

Furnham, A. (1989). *The Protestant work ethic: The psychology of work related beliefs and behaviours.* London, New York: Routledge.

Furnham, A., Bond, M.H., Heaven, P., Hilton, D., Lobel, T., et al. (1993). A comparison of Protestant work ethic beliefs in thirteen nations. *Journal of Social Psychology, 133,* 185–197.

Gawronski, B., & Bodenhausen, G.V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731.

Gelfand, M. J., Higgins, M., Nishii, L. H., Raver, J. L., Dominguez, A., Murakami, F., Yamaguchi, S., & Toyama, M. (2002). Culture and egocentric perceptions of fairness in conflict and negotiation. *Journal of Applied Psychology, 87*(5), 833–845.

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

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

Greenwald, A. G., Oakes, M. A., & Hoffman, H. G. (2003). Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology*, *39*(4), 399-405.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*, 216-229.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1-20.

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

Gruen, L., & Panichas, G.E., eds. (1997). *Sex, morality, and the law.* London: Routledge.

Haaf, J.M., Hoogeveen, S., Berkhout, S., Gronau, Q.F., & Wagenmakers, E.J. (2020). *A Bayesian multiverse analysis of Many Labs 4: Quantifying the evidence against mortality salience.* Unpublished manuscript.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*, 779–798.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814-834.

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

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

Harris, C.R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-

performance-goal priming effects. *PLoS ONE,* 8, e72467.

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral & Brain Sciences, 33,* 61–83.

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

Hubbard, R., & Armstrong, J. S. (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing, 11,* 233-248.

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

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

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

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

Jussim, L., Coleman, L., & Lerch, L. (1987). The nature of stereotypes: A comparison and integration of three theories. *Journal of Personality and Social Psychology, 52,* 536-546.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515-526.

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational

and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology, 55,* 893–905.

King, R.C., & Bu, N. (2005). Perceptions of the mutual obligations between employees and employers: A comparative study of new generation IT professionals in China and the United States. *International Journal of Human Resource Management, 16*, 46–64

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

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

Klein, R.A., Vianello, M., Hasselman, F., et al., & Nosek, B.A. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science, 1*(4), 443-490.

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

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

Kohn, M. L. (1969). *Class and conformity: A study in values*. Chicago: University of Chicago Press.

Kohn, M. L., Naoi, A., Schoenbach, C., Schooler, C., & Slomczynski, K. M. (1990). Position in

the class structure and psychological functioning in the United States, Japan, and Poland. *American Journal of Sociology, 95*(4), 964–1008.

Kohn, M. L., Zaborowski, W., Janicka, K., Khmelko, V., Mach, B. W., Paniotto, V., et al., (2002). Structural location and personality during the transformation of Poland and Ukraine. *Social Psychology Quarterly, 65*(4), 364–385.

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

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, Musgrave eds. *Criticism and the Growth of Knowledge*. Cambridge University Press. pp. 91-195.

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

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

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

Leong, F. T. L., Huang, J. L., & Mak, S. (2014). Protestant work ethic, Confucian values, and work-related attitudes in Singapore. *Journal of Career Assessment, 22*, 304-316.

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

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science, 7*, 537–542.

Manzoli, L., Flacco, M.E., D'Addario, M., Capasso, L., DeVito, C., Marzuillo, C., et al.

(2014). Non-publication and delayed publication of randomized trials on vaccines: survey. *British Medical Journal, 348,* g3058.

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

Mayr, E. (1942). *Systematics and the origin of species.* New York, NY: Columbia University Press.

Mayr, E. (1954). Change of genetic environment and evolution. In J. Huxley, A. C. Hardy, and E. B. Ford (Eds.) *Evolution as a process.* (pp. 157–180). London: Allen & Unwin.

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., et al. (2018). Registered Replication Report: Srull & Wyer (1979). *Advances in Methods and Practices in Psychological Science, 1,* 321-336.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12,* 269–275.

Milfont, T. L.**,** & Klein, R. A. (2018). Replication and reproducibility in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 49,* 735-750.

Mirels, H., & Garrett, J. (1971). Protestant ethic as a personality variable. *Journal of Consulting and Clinical Psychology, 36,* 40-44.

Moon, J. W., Krems, J. A., & Cohen, A. B. (2018). Religious targets are trusted because they are viewed as slow life-history strategists. *Psychological Science, 29*(6), 947 –960.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., et al.

(2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1*(4), 501-515.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond WEIRD psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science, 31*(6), 678-701.

Mynatta, C.R., Dohertya, M.E., & Tweneya, R.D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*, 85–95.

Nelson, L., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*, 511-534.

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

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*, 291-310.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*, 565-584.

O'Donnell, M., et al. (2018). Registered Replication Report: Dijksterhuis & van Knippenberg (1998). *Perspectives on Psychological Science, 13*(2), 268-294.

Olsson-Collentine, A., Wicherts, J.M., & van Assen, M.A.L.M. (in press). Heterogeneity in

direct replications in psychology and its association with effect size. *Psychological Bulletin*.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). DOI: 10.1126/science.aac4716

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin, 128*(1), 3-72.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411-419.

Pashler, H., Coburn, N., & Harris, C. (2012). Priming of social distance? Failure to replicate effects on social and food judgements. *PloS ONE*, *7*(8), e42510

Pashler, H., Rohrer, D. & Harris, C. (2013). Can the goal of honesty be primed? *Journal of Experimental Social Psychology, 49,* 959-964.

Peiss, K., Simmons, C., & Padgug, R.A., eds. (1989). *Passion and power: Sexuality in history.* Philadelphia: Temple University Press.

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

Platt, J. R. (1964). Strong inference – Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*, 347-353.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*, 737-756.

Popper, K. (1959/2002). *The logic of scientific discovery*. Routledge.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on

published data on potential drug targets? *Nature Reviews Drug Discovery, 10,* 712.

Rohrer, D., Pashler, H., & Harris, C.R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General*, 144, e73-e85.

Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A Bayesian approach to finding constraints in meta-analysis. *Psychological Methods, 24,* 606–621.

Sanchez-Burks, J. (2002). Protestant Relational Ideology and (in) attention to relational cues in work settings. *Journal of Personality and Social Psychology, 83,* 919-929.

Sanchez-Burks, J. (2005). Protestant Relational Ideology: The cognitive underpinnings and organizational implications of an American anomaly. *Research in Organizational Behavior, 26,* 265-305.

Sanchez-Burks, J., & Lee, F. (2007). Culture and workways. In S. Kitayama & D. Cohen (Eds.). *Handbook of Cultural Psychology* (Vol 1, pp. 346-369). New York: Guilford.

Schafer, B.E. (1991). *Is America different? A new look at American exceptionalism.* New York, NY: Oxford University press.

Shariff, A. F., Willard, A. K., Andersen, T., & Norenzayan, A. (2016). Religious priming: A meta-analysis with a focus on prosociality. *Personality and Social Psychology Review, 20*(1), 27–48.

Schafer, B.E. (1991). *Is America different? A new look at American exceptionalism*. New York, NY: Oxford University press.

Schein, E.H. (1990). Organizational culture. *American Psychologist, 45,* 109-119.

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

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

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression.* New York, NY: Cambridge University Press.

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

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

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123-1128.

Simonsohn, U. (2014). No-way interactions. *Data Colada*. Available at: http://datacolada.org/17

Sinclair, S., Dunn, E., & Lowery, B. S. (2005). The relationship between parental racial attitudes and children's automatic prejudice. *Journal of Experimental Social Psychology, 3*, 283-289.

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). The social tuning of automatic ethnic attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology, 89*, 583-592.

Smith, R. A., Levine, T. R., Lachlan, K. A., & Fediuk, T. A. (2002). The high cost of complexity in experimental design and data analysis: Type I and type II error rates in multiway ANOVA. *Human Communication Research, 28*(4), 515-530.

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

Snir, R., & Harpaz, I. (2009). Cross-cultural differences concerning heavy work investment. *Cross Cultural Research, 43*(4), 309–319.

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

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*, 1325-1346.

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

Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S., & Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology, 102*(6), 1178–1197.

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

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

Thompson, E.A. (1978). Ancestral inference. II. The founders of Tristan da Cunha. *Annals of Human Genetics, 42*, 239-253.

Tierney, W., Hardy, J., Ebersole, C., Leavitt, K., Viganola, D., Clemente, E., Gordon, M., Dreber, A.A., Johannesson, M., Pfeiffer, T., Hiring Decisions Forecasting Collaboration, & Uhlmann, E.L. (in press). Creative destruction in science. *Organizational Behavior and Human Decision Processes*.

Uhlmann, E.L., Heaphy, E., Ashford, S.J., Zhu, L., & Sanchez-Burks, J. (2013). Acting professional: An exploration of culturally bounded norms against non-work role referencing. *Journal of Organizational Behavior, 34,* 866-886.

Uhlmann, E.L., Poehlman, T.A., & Bargh, J.A. (2008). Implicit theism. In R. Sorrentino & S. Yamaguchi (Eds.) *Handbook of motivation and cognition within and across cultures.* (pp. 71-94). St. Louis, MO: Elsevier/ Academic Press.

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

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

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

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

van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J.

(2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology, 6*, 1365.

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

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

van de Vijver, F. J. R., & Leung, K. (2010). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 15-45). Cambridge, UK: Cambridge University Press.

Varnum, M. E., Na, J., Murata, A., & Kitayama, S. (2012). Social class differences in N400 indicate differences in spontaneous trait inference. *Journal of Experimental Psychology: General*, *141*(3), 518–526.

Voss, K. (1993). *The making of American exceptionalism.* Ithaca, NY: Cornell University press.

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

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

Weeden, J., & Kurzban, R. (2013). What predicts religiosity? A multinational analysis of reproductive and cooperative morals. *Evolution and Human Behavior, 340*(6), 440-445.

Weeks, J. (2004). *Culture and Leadership at IBM*. INSEAD case.

Weingarten, E., Hepler, J., Chen, Q., McAdams, M., Yi, & Albarracín, D. (2016). On priming

action: A meta-analysis of the behavioral effects of incidentally presented stimuli. *Psychological Bulletin, 142,* 5, 472–497.

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

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, ITEC, 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 Brigden, 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. Cicerali, 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 Sciences, Kozminski University, Warsaw, Poland

Terence D. Dores Cruz, Vrije Universiteit Amsterdam

Christilene du Plessis, Singapore Management University

Dmitrii Dubrov, National Research University Higher School of Economics

Esha Dwibedi, 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 Firat, University of Alberta

Zoë Francis, University of the Fraser Valley

Yoav Ganzach, Ariel University and Tel Aviv University

Richa 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 EGuadagno, Center for International Security and Cooperation, Stanford University

Andrew C. Hafenbrack, Michael G. Foster School of Business, University of Washington

Sebastian Hafenbrädl, IESE Business School

Linda Hagen, University of Southern Califonia

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 Holzmeister, 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

Chitranjan 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, Palacky University Olomouc, Olomouc, Czech Republic

Pavol Kačmár, Department of Psychology, Faculty of Arts, Pavol Jozef 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

Yeun Joon Kim, University of Cambridge
Minjae J. Kim, Boston College

Mikael Knutsson, Linköping University

Olga Kombeiz, Loughborough University

Marta Kowal, Institute of Psychology, University of Wrocław, Wrocław, 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 Lüke, 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 Martončik, 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 Matsuo, Tokai Gakuen University

Jens Mazei, TU Dortmund University

Randy J. McCarthy, Northern Illinois University

Smriti Mehta, UC Berkeley

Chanel Meyers, Whitman College

Ewa AureliaMiendlarzewska, 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. Nissenbaum, Raindrop Games, PBC

Paweł Niszczota, Poznań University of Economics and Business

Nurit Nobel, Stockholm School of Economics

Stephan Oelhafen, Bern University of Applied Sciences

Aoife 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 Roczniewska

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

Lukas Röseler, 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 Schaerer, 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 Shtudiner, Ariel University

Deborah Shulman

Victoria Song, Fordham University

Tabea Springstein, 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

Fransisca 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 Academic College of Engineering, Ashdod, Israel

Marek A. Vranka, Charles University

Lauren E.T. Wakabayashi, Loma Linda University

Hanne M. Watkins, UMass Amherst

Erin C. Westgate, University of Florida

Margaux N. A. Wienk, 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

Footnote

[1] The 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.

Table 1. Moral judgments of a lottery winner who works vs. retires and is relatively young or older.

	India MTurk		USA MTurk		USA PP*		Australia PP		UK PP	
	Young	Older	Young	Older	Young	Older	Young	Older	Young	Older
Works	5.86	5.84	5.68	5.73	5.96	5.86	5.67	5.64	5.62	5.56
	(0.08)	(0.08)	(0.09)	(0.09)	(0.07)	(0.07)	(0.08)	(0.08)	(0.07)	(0.07)
Retires	4.90	5.08	4.84	5.14	5.03	5.33	4.65	4.81	4.75	4.90
	(0.08)	(0.08)	(0.09)	(0.09)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)

Note: *PP denotes PureProfile sample. Numbers in parentheses represent standard errors.

Figures

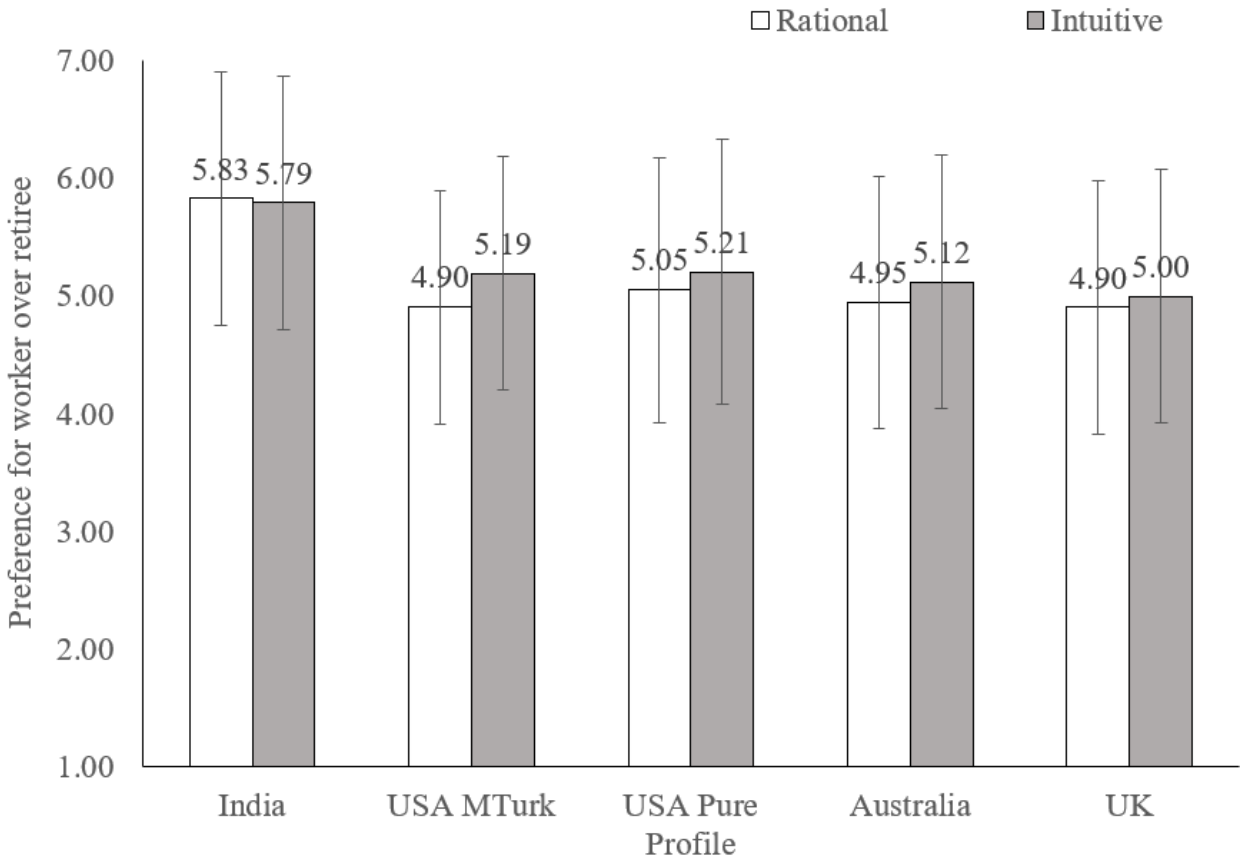


Figure 1. Intuitive vs. rational evaluations across samples. Higher numbers reflect more favorable moral judgments of a lottery winner who continues working rather than retiring. As seen in the figure, the intuitive mindset effect is present in all samples except for the Indian sample, where intuitive and rational evaluations are similar. Error bars represent standard errors.

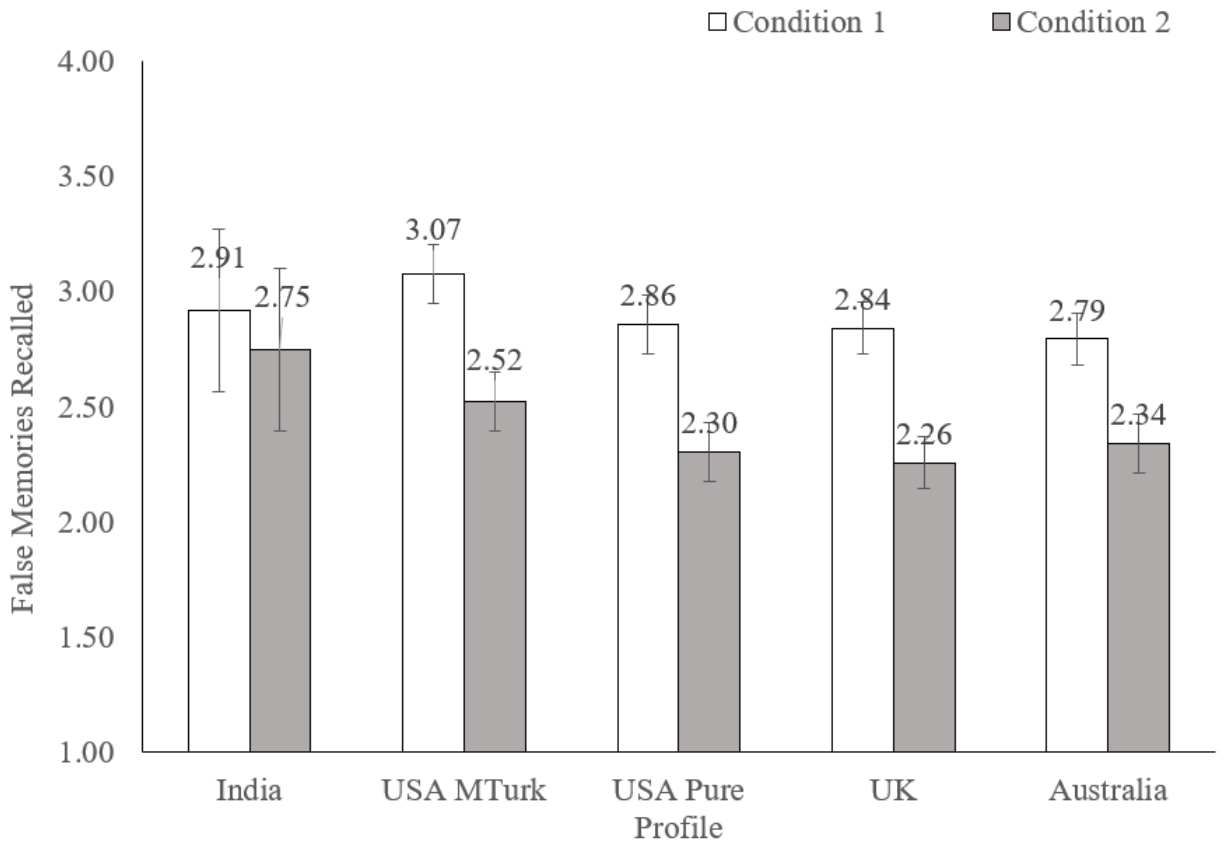


Figure 2. Tacit inferences across cultures. Higher means in Condition 1 than Condition 2 reflect false memories consistent with linking traditional work and sex morality. As seen in the figure, participants from all samples made such tacit inferences. Error bars represent standard errors.

Supplements: Table of Contents

Supplement 1: Study 1 Materials

THANKS FOR HELPING US OUT!

THIS SURVEY TAKES 15 MINUTES TO COMPLETE.

YOU WILL FIRST READ STORIES, THEN ANSWER SOME QUESTIONS ABOUT WHAT YOU REMEMBER ABOUT THE CONTENT OF THE STORIES.

PLEASE TRY AND COMPLETE THE SURVEY
AS PRIVATELY AS POSSIBLE.

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

CONSENT STATEMENT:

I understand that my responses to this survey are completely anonymous, and that my participation is strictly voluntary.
I am free to skip any questions I prefer not to answer.

LOTTERY WINNER STUDY (4 CONDITIONS)

CONDITION 1

Sarah is a 23 year old woman from Milwaukee. Sarah has worked for three years at the local post office where she is loved by her co-workers. Each week for the last 3 years she has played the same numbers in the state lottery. Last year she won 10 million dollars. After using $12,000 to pay bills and debts, she decided what she really wanted was to stay working at the post office even though she doesn't need the money anymore.

Please answer the following question about Sarah.

Is Sarah a good person?

Very Bad						Very Good
1	2	3	4	5	6	7

CONDITION 2

Sarah is a 23 year old woman from Milwaukee. Sarah has worked for three years at the local post office where she is loved by her co-workers. Each week for the last 3 years she has played the same numbers in the state lottery. Last year she won 10 million dollars. After using $12,000 to pay bills and debts, she decided what she really wanted was to never work another day. Ever since winning, she's taken it easy at home and ordered a lot of take-out food and at 23 considers herself "retired".

Please answer the following question about Sarah.

Is Sarah a good person?

Very Bad						Very Good
1	2	3	4	5	6	7

CONDITION 3

Sarah is a 46 year old woman from Milwaukee. Sarah has worked for three years at the local post office where she is loved by her co-workers. Each week for the last 3 years she has played the same numbers in the state lottery. Last year she won 10 million dollars. After using $12,000 to pay bills and debts, she decided what she really wanted was to stay working at the post office even though she doesn't need the money anymore.

Please answer the following question about Sarah.

Is Sarah a good person?

Very Bad						Very Good
1	2	3	4	5	6	7

CONDITION 4

Sarah is a 46 year old woman from Milwaukee. Sarah has worked for three years at the local post office where she is loved by her co-workers. Each week for the last 3 years she has played the same numbers in the state lottery. Last year she won 10 million dollars. After using $12,000 to pay bills and debts, she decided what she really wanted was to never work another day. Ever since winning, she's taken it easy at home and ordered a lot of take-out food and at 46 considers herself "retired".

Please answer the following question about Sarah.

Is Sarah a good person?

Very Bad						Very Good
1	2	3	4	5	6	7

TACIT INFERENCES STUDY: CONDITION 1

Mary is an attorney at a mid-sized law firm in a big city. Mary's boyfriend Matt is a doctor. Mary and Matt met in a class in college. They have been dating seriously for quite some time, but they have never had sex because Mary does not believe in premarital sex. Matt is not a virgin but Mary is, and she insists on waiting to have sex until the two of them are married. They have a strong, happy relationship. The two of them enjoy sailing and hearing about one another's careers. Matt is very supportive of Mary's law career. Because Matt works in the emergency room of a busy hospital, his work hours are often scattered throughout the week and weekend. Mary's work schedule is much more predictable. The weekends before big trials, she likes to be in bed all weekend.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

a. _____ Mary went to an all-women's college.
b. _____ Mary and Matt get along well.
c. _____ Matt is a virgin.
d. _____ Matt's dentistry career causes him to have an unpredictable schedule.
e. _____ Both Mary and Matt like to sail.
f. _____ Mary stays in bed on weekends before big trials in order to get extra work done.
g. _____ Matt wishes Mary were a doctor.
h. _____ Mary and Matt live in a sizable city.

After seven years in college, Julia graduated with a very low grade point average. She has been unemployed for the last four years and her parents are supporting her financially. Julia is not making any effort to find a job and spends a lot of time watching television. Last week, Julia was invited to a party at a guy's house not too far from her own. She ended up staying at the guy's house that night.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

a. _____ Julia did not graduate from college.
b. _____ Julia likes to watch television.
c. _____ Julia is actively looking for a job.
d. _____ Julia has friends.
e. _____ Julia's parents have to support her financially.
f. _____ Julia slept with the host of last week's party.
g. _____ Julia has gained weight since graduation.
h. _____ Julia has a low-paying job.

Ann is strongly in favor of sex and is known around her school for being promiscuous. Each week, she tells her friends about men she slept with the weekend before. She constantly flirts with the guys at her school, and she admits to having had sex with many of them. One day in history class, Ann took had to take a quiz on the Civil War's chronology. She found the quiz questions to be very hard. The next day, her teacher told Ann that she did poorly on the quiz.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

a. _____ Ann takes a history class.
b. _____ Ann goes to an all-girls school.
c. _____ Ann has a serious boyfriend.
d. _____ Ann lives in an apartment building.
e. _____ No one likes Ann at her school.
f. _____ Ann did not study hard for the Civil War chronology quiz.
g. _____ Ann is open about her sexual history.

Carl recently worked as a waiter at a restaurant where his boss praised him for his quick service and for always being on time. Carl then got a job at a library in his neighborhood. His job was to reshelf books that people returned. During the first week of his new job, he was on time every single day, worked very hard, and never took a break. At the end of one workday, as he was finishing reshelving books, he found an envelope labeled "Nude model photos" left inside a book someone had returned. Eventually, Carl put the envelope in the Lost and Found bin.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

a. _____ Carl is sometimes late.
b. _____ Before working at a library, Carl was a waiter at a French restaurant.
c. _____ His boss at the restaurant complained about Carl' service.
d. _____ A fellow waiter warned Carl about needing to improve his efficiency.
e. _____ Carl only has to reshelf mystery novels.
f. _____ Carl looked at the photos he found.
g. _____ The library has a Lost and Found bin.

TACIT INFERENCES STUDY: CONDITION 2

Mary is an attorney at a mid-sized law firm in a big city. Mary tends to date several men at the same time, preferring men in prestigious professions like doctors and lawyers. Mary meets most of her dates at bars. She has been dating around for quite some time, and tries to have as much sex as possible because she thinks she has to have fun while her body is young and virile. Mary has many fun, happy relationships where her and her lover enjoy activities like sailing and hearing about one another's careers. Mary tends to be very supportive of her boyfriends' careers even if they are wild. Mary's work schedule is much more predictable. The weekends before big trials, she likes to be in bed all weekend.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

a. _____ Mary went to an all-women's college.
b. _____ Mary and her boyfriends get along well.
c. _____ Mary is a virgin.
d. _____ Mary's career causes her to have an unpredictable schedule.
e. _____ Both Mary and her boyfriends like to sail.
f. _____ Mary stays in bed on weekends before big trials in order to get extra work done.
g. _____ Mary wishes all her boyfriends were doctors.
h. _____ Mary lives in a sizable city.

After only three years in college, Julia graduated with honors. She has held an excellent job for the last four years and is financially independent. Julia is working very hard at her job and spends hardly any time watching television. Last week, Julia was invited to a party at a guy's house not too far from her own. She ended up staying at the guy's house that night.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

a. _____ Julia did not graduate from college.
b. _____ Julia likes to watch television.
c. _____ Julia is actively looking for a job.
d. _____ Julia has friends.
e. _____ Julia's parents have to support her financially.
f. _____ Julia slept with the host of last week's party.
g. _____ Julia has gained weight since graduation.
h. _____ Julia has a low-paying job.

Ann is strongly against sex before marriage and is known around her school for being a prude. Each week, she tells her friends about her Teen Abstinence meeting the weekend before. She never flirts with the guys at her school, and has not had sex with any of them. One day in history class, Ann took had to take a quiz on the Civil War's chronology. She found the quiz questions to be very hard. The next day, her teacher told Ann that she did poorly on the quiz.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

a. _____ Ann takes a history class.
b. _____ Ann goes to an all-girls school.
c. _____ Ann has a serious boyfriend.
d. _____ Ann lives in an apartment building.
e. _____ No one likes Ann at her school.
f. _____ Ann did not study hard for the Civil War chronology quiz.
g. _____ Ann is open about her sexual history.

Carl recently worked as a waiter at a restaurant where his boss complained that his service was slow and he was always late. Carl then got a job at a library in his neighborhood. His job was to reshelf books that people returned. During the first week of his new job, he was late every single day, barely worked, and took lots of breaks. At the end of one workday, as he was finishing reshelving books, he found an envelope labeled "Nude model photos" left inside a book someone had returned. Eventually, Carl put the envelope in the Lost and Found bin.

Please click to continue to the next page. Do not refer back to this page.

- Who or what do you think was the main character in the story?

- Do you think the author of the story was female or male?

- Do you think the author of the story was older than 30 or younger than 30?

TRUE OR FALSE

h. _____ Carl is sometimes late.
i. _____ Before working at a library, Carl was a waiter at a French restaurant.
j. _____ His boss at the restaurant complained about Carl' service.
k. _____ A fellow waiter warned Carl about needing to improve his efficiency.
l. _____ Carl only has to reshelf mystery novels.
m. _____ Carl looked at the photos he found.
n. _____ The library has a Lost and Found bin.

INTUITIVE MINDSET STUDY

INSTRUCTIONS: PLEASE READ THE PARAGRAPH AND RESPOND TO THE QUESTIONS BELOW

John and Robert are two 23-year old friends who used to work together as potato peelers. Each week for 3 years they bought lotto tickets together. Last year they won 10 million dollars.

Robert decided right away to never work another day. Robert quit his job and now spends all day at home watching TV and at 23 considers himself "retired."

John decided what he really wanted was to stay working as a potato peeler even though he didn't need the money anymore. John feels that an honest day's work is its own reward.

My most rational, objective judgment is that:

Robert is a much better person than John						John is a much better person than Robert
1	2	3	4	5	6	7

My intuitive, gut feeling is that:

Robert is a much better person than John						John is a much better person than Robert
1	2	3	4	5	6	7

DUKE UNIVERSITY RELIGION INDEX (DUREL)

How often do you attend church or other religious meetings?
1 – Never
2 - Once a year or less
3 - A few times a year
4 - A few times a month
5 - Once a week
6 - More than once/week

How often do you spend time in private religious activities, such as prayer, meditation or Bible study?
1 - Rarely or never
2 - A few times a month
3 - Once a week
4 - Two or more times/week
5 – Daily
6 - More than once a day

The following section contains 3 statements about religious belief or experience. Please mark the extent to which each statement is true or not true for you.

In my life, I experience the presence of the Divine (*i.e.*, God)
1 - Definitely *not* true
2 - Tends *not* to be true
3 – Unsure
4 - Tends to be true
5 - Definitely true of me

My religious beliefs are what really lie behind my whole approach to life
1 - Definitely *not* true
2 - Tends *not* to be true
3 – Unsure
4 - Tends to be true
5 - Definitely true of me

I try hard to carry my religion over into all other dealings in life
1 - Definitely *not* true
2 - Tends *not* to be true
3 – Unsure
4 - Tends to be true
5 - Definitely true of me

PROTESTANT WORK ETHIC (PWE) SCALE

Please indicate whether you disagree or agree with the items below using the following scale:

Strongly Disagree Strongly Agree
1 2 3 4 5 6

Most people spend too much time in unprofitable amusements.

I feel uneasy when there is little work for me to do.

A distaste for hard work usually reflects a weakness of character.

People who fail at a job have usually not tried hard enough.

Anyone who is willing and able to work hard has a good chance of succeeding.

If people work hard enough they are likely to make a good life for themselves.

Most people who don't succeed in life are just plain lazy.

The person who can approach an unpleasant task with enthusiasm is the person who gets ahead.

Our society would have fewer problems if people had less leisure time.

Money acquired easily is usually spent unwisely.

Life would have very little meaning if we never had to suffer.

DEMOGRAPHIC ITEMS

My religion is:

- Protestant
- Catholic
- Islam
- Judaism
- Buddhism
- Atheist
- Agnostic
- Other (please indicate) [free response text box]

If you selected "Protestant" above, please select a denomination

- Adventist
- Anabaptist
- Anglican
- Baptist
- Calvinist (Reformed)
- Lutheran
- Methodist
- Pentecostal
- Other (please indicate) [free response text box]

If relevant, what is the name of a place of worship (e.g., church, mosque, synagogue) you attended growing up? [free response text box]

I consider myself to be:

Not at all Religious						Very Religious
1	2	3	4	5	6	7

Politically, I am *(please select one)*

- Very Progressive/Left-wing
- Progressive/Left-wing
- Somewhat Progressive/Left-wing
- Moderate/Centrist
- Somewhat Conservative/Right-wing
- Conservative/Right-wing
- Very Conservative/Right-wing

What political party do you identify with? [free response text box]

My gender is *(please select one)*:

- Male
- Female
- Other (please specify): [free response text box]

My age in years is:

My ethnicity is *(please select one)*:

- White
- Asian
- Latino
- Black
- Indigenous or native group (please specify) [free response text box]
- Other (please indicate): [free response text box]

Which of the following countries are you currently based primarily in?

- United States
- United Kingdom
- Australia
- India
- Other (please indicate): [free response text box]

If you selected the United States, which U.S. state are you primarily based in? [dropdown menu with all 50 U.S. states, Other]

If you selected the United Kingdom, which constituent country of the U.K. are you primarily based in?

- England
- Scotland
- Wales
- Northern Ireland
- Other (please indicate): [free response text box]

If you selected Australia, which Australian state are you primarily based in?

- New South Wales (including Australian Capital Territory)
- Victoria
- Queensland
- Western Australia
- South Australia
- Tasmania
- Other (please indicate): [free response text box]

If you selected India, which region of India are you primarily based in?

- South India
- Hindustan
- North-east
- Other (please indicate): [free response text box]

What country/region were you born in?

Of what nation/region are you a citizen?

How many years have you lived in the United States?

If you grew up in the United States, what U.S. state/territory did you grow up in?

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

My educational level is:

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

Are you currently a student at a university?

- Yes
- No

My occupation is: [free response text box]

My yearly household income level is:

1= Less than $10,000 United States dollars (USD) a year, or less than $13,400 Australian dollars (AUD) a year, or less than £7,387 British Pounds (GBP) a year, or less than ₹672,900 Indian Rupees (INR) a year

2= USD $10,000-$20,000, or AU$13,400- AU$26,734, or GBP £7,387-£14,781, or INR ₹672,900-₹1,345,700

3= USD $20,000-$40,000, or AU$26,734- AU$53,454, or GBP £14,781-£29,561, or INR ₹1,345,700-₹2,691,400

4= USD $40,000-$60,000, or AU$53,454- AU$80,202, or GBP £29,561-£44,342, or INR ₹2,691,400-₹4,038,000

5= USD $60,000-$80,000, or AU$80,202- AU$106,936, or GBP £44,342-£59,114, or INR ₹4,038,000-₹5,384,000

6= USD $80,000-$100,000, or AU$106,936- AU$133,670, or GBP £59,114-£73,893, or INR ₹5,384,000-₹6,730,000

7= USD $100,000 a year or more, or AU$133,670 a year or more, or GBP £73,942 a year or more, or INR ₹6,730,000 or more

What is the education level of your most educated parent?

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

AWARENESS PROBE

What do you think this survey was about? [free response text box]

ATTENTION CHECK

Please select "strongly disagree" on the scale below:

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

Supplement 2: Study 2a and 2b Materials

COVER PAGE

WORD PUZZLE STUDY!

THANKS FOR HELPING US OUT!

THIS SURVEY TAKES ABOUT 15 MINUTES TO COMPLETE

IT CONTAINS TWO WORD PUZZLES.

PLEASE TRY AND COMPLETE THE SURVEY
AS PRIVATELY AS POSSIBLE.

You must be at least 18 years old to participate in this study. If you are 17 years or younger, please tell the experimenter; you will still be compensated for your time.

CONSENT STATEMENT:

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

SALVATION PRIME

INSTRUCTIONS: For each set of words below, **one word does not belong**. Please *remove* that word and make a grammatical **FOUR-WORD** sentence. Write it down in the space provided.

Ex: flew eagle the ~~rock~~ around — The eagle flew around.

1. commander ball almighty the was

2. coupons here phone redeem your

3. face angelic paper is her

4. drink topography water gallons of

5. they righteous moisturizer women were

6. the was composition light forest

7. cough in God control is

8. the blue literature is curtain

9. grace he plays well notes

10. legacy eternal bell their is

11. her them check salvation for

12. the brown clown chair is

NEUTRAL PRIME

INSTRUCTIONS: For each set of words below, **one word does not belong**. Please *remove* that word and make a grammatical **FOUR-WORD** sentence. Write it down in the space provided.

Ex: flew eagle the ~~rock~~ around — The eagle flew around.

1. is rainbow east the ground

2. is my comfortable blue bed

3. love cup ice cream I

4. drink topography water gallons of

5. growing time now are flowers

6. the was composition light forest

7. happiness cheese envelope comes from

8. the blue literature is curtain

9. pencil children ponies rode the

10. supersede these things are hot

11. was amazing jumping the opera

12. the brown clown chair is

DEPENDENT MEASURE

Here is a word task for you to work on. Please complete as many of the anagrams as you can. To make an anagram, use the letters in the original word to make a new word. Using each provided word, please form as many different English language words FOUR or more letters in length as you can. Proper nouns (e.g. names), plurals, and tense changes (past tense, future tense) are acceptable.

Ex. P R I N C E S S: NICE (**acceptable**) RIP (**unacceptable**)

1. B I M O D A L :

2. I G N E O U S :

3. A N S W E R :

4. C U R R I E D :

MODERATOR SCALES

DUREL and PWE scales, as in Study 1.

DEMOGRAPHICS

Same as in Study 1.

AWARENESS PROBE

Did the sentence unscrambling task influence your performance on the anagram task in any way?

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

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

_____

_____

_____

ATTENTION CHECK

Please select "strongly disagree" on the scale below:

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

Supplement 3: Pre-Registered Analysis Plan for Replication Project

The studies targeted for replication are described in Uhlmann, Poehlman, Tannenbaum, and Bargh (2011) and Uhlmann, Poehlman, and Bargh (2009). Beyond examining whether the original Implicit Puritanism effects replicate, another goal of the project is to illustrate a "creative destruction" approach to replication in which competing theoretical predictions (not just the original theory vs. the null hypothesis) are put to the empirical test (see also Brainerd & Reyna, 2018). To that end, we intend to test not only the predictions of the theory of Implicit Puritanism but also competing theories drawn from the literatures on regional folkways, religious differences, explicit cultural differences, and general moralization of work. Below we outline our analytic plan, specifying key measures, the statistical analyses that will be run, and empirical predictions based on the theory of Implicit Puritanism as well as for competing theories of work and sex morality.

OVERVIEW OF DATA COLLECTIONS

There will be two major waves of data collection containing a total of 4 original experiments to be replicated.

Data collection wave 1 will contain the Tacit inferences, Lottery winner, and Intuitive mindset experiments and be run on Mechanical Turk using USA and Indian participants (1000 from each sample, 2000 participants in total), and on Pure Profile using USA New England, USA non-New England, UK, and Australian participants (1000 of each group, 4000 participants in total).

Data collection wave 2 (Salvation study) will be run online with help from the survey firm Pure Profile using participants from the U.S., U.K., and Australia (N = 1000 in total). Paper-pencil versions of the study materials will further be administered at the State University of New York, Fairfield University, University of Rochester, Ithaca College, and either Brooklyn College or Queens College in the U.S., and the University of Limerick in Ireland. Some moderator measures (e.g. DUREL multi-item religiosity scale, Protestant work ethic scale) may not be administered at some specific universities due to subject pool and time constraints.

For the Pure Profile data collections, we will temporarily stop data collection after 10% of subjects have been collected to check the online survey is working properly. So long as the survey is collecting data properly, we will then run the remaining 90% of participants regardless of whether the initial results support the predictions of the theory of Implicit Puritanism or not.

INCLUSION AND EXLUSION CRITERIA

Inclusion criteria. For our primary analyses, we will group subjects into cultural categories based on the objective location of the data collection (e.g., USA, India, UK, Australia).

Exclusion criteria. All participants who indicate they have less than 5 years' experience speaking English will be excluded from the analyses. The relevant self-report item is "How many years of experience do you have with the English language?" To further maintain the integrity of the data, we will record and screen out duplicate GPS coordinates for the online data collections. Finally,

for the MTurk data collections we will recruit only participants with a 99% acceptance rate and more than 1000 hits approved.

CONCEPTUAL OVERVIEW OF PREDICTIONS OF COMPETING THEORIES

The key statistical tests for each individual study (Tacit inferences, Lottery winner, Intuitive mindset, Salvation prime) will be carried out as outlined later in the tables. However, the theoretical conclusions will depend on the pattern of results across these four key studies. Below we describe, at a conceptual level, the overall pattern of results across the four studies that would support each respective theory of work and sex morality.

Predictions of original theory: "Implicit Puritanism"
American but not non-American participants should: 1) Prefer a lottery winner who continues to work as opposed to retiring, especially if the target person is young (Lottery winner study) and when responses are made intuitively rather than deliberatively (Intuitive mindset study); 2) falsely infer a sexually promiscuous person is lazy and vice versa (Tacit inferences study); and 3) respond to the implicit priming of concepts related to divine salvation by working harder on an unrelated task (Salvation study). Across all studies, group differences should manifest themselves at the level of national culture (USA vs. other countries), rather than regions (New England vs. not), personal religion (Protestant or not), social class, and individual differences in religiosity or explicit endorsement of the Protestant Work Ethic.

Predictions of competing theory: "False positives"
Postulates that the original findings are spurious due to the relatively small sample sizes and low statistical power of the original studies, combined with a publication filter in favor of significant results. Thus, the condition differences predicted by the theory of Implicit Puritanism will not emerge for the Lottery winner, Tacit inferences, Intuitive mindset, or Salvation studies. These effects should not emerge reliably among either Americans or members of the comparison cultures, and likewise fail to emerge for the theoretically relevant subgroups (e.g., Protestant and religious individuals, those who endorse the PWE, high SES individuals). Further, if these effects are truly null, then variability across sites (countries, regions, replication laboratories) should be relatively low (e.g., Klein et al., 2014; 2018). We will test for heterogeneity using Cochran's *Q*, generated from a random effects meta-analysis of each effect (Cochran, 1954). We will also estimate the proportion of variance due to heterogeneity using I^2 and *Tau* (Higgins, Thompson, Deeks, & Altman, 2003; Borenstein et al., 2009; Borenstein, Hedges, Higgins, & Rothstein, 2010; Borenstein, Higgins, Hedges, & Rothstein, 2017).

Predictions of competing theory: "Explicit American exceptionalism"
Expects that the work and sex morality effects predicted by the theory of Implicit Puritanism in the Lottery winner, Tacit inferences and Intuitive mindset studies will emerge in the United States but not the comparison countries (e.g., India, UK, Australia). In addition, PWE scores should moderate the effects, such that individuals who explicitly endorse the Protestant Work Ethic are significantly more likely to exhibit the Lottery winner, Tacit inferences and Intuitive mindset effects. However, this theoretical perspective predicts that the priming effect stipulated by the theory of Implicit Puritanism will not emerge in the Salvation study, since it is postulated

that the work and sex morality effects are relatively more conscious than nonconscious (in other words, intuitive rather than truly implicit or unconscious).

Unlike the theory of Implicit Puritanism, the explicit American exceptionalism perspective can easily incorporate the possibility that a preference for needless work is logically and deliberatively endorsed by Americans without qualification. If so, Americans should not exhibit any difference between their intuitive and logical preferences (i.e., no Intuitive Mindset effect), yet should still express both an intuitive and logical preference for a person who continues working rather than retires (as reflected in scores significantly above the neutral scale midpoint of 4 on both dependent measures). Further, Americans may be indifferent to the age of the target (23 or 46), and straightforwardly prefer the worker over the retiree (main effect of work status in the Lottery Winner study, with no age*work status interaction).

The remaining five theoretical perspectives can likewise incorporate the possibility that work is moralized not only intuitively but also at a logical, deliberative level. Modified versions of the Intuitive Mindset and Lottery Winner effects with a straightforward effect of work status, such that a worker is morally praised relative to an early retiree (regardless of mindset or target age), would also support these theories so long as the other patterns they predict (e.g., regional, religious, and national differences or the lack thereof) likewise hold.

Predictions of competing theory: "Regional Folkways"
Expects that the work and sex morality effects (Lottery winner, Tacit inferences, Intuitive mindset, Salvation prime) will be stronger in the New England region (Maine, Vermont, New Hampshire, Massachusetts, Rhode Island, Connecticut) than in the rest of the United States (all other USA states combined) or in the comparison cultures (India, UK, Australia, etc.).

Predictions of competing theory: "Religious Differences"
Expects that the work and sex morality effects (Lottery winner, Tacit inferences, Intuitive mindset, Salvation prime) will emerge, but significantly more strongly among 1) more religious participants, and 2) Protestant (relative to non-Protestant) participants.

Predictions of competing theory: "General moralization of work and sex"
Expects that the key work and sex morality effects (Lottery winner, Tacit inferences, Intuitive mindset, Salvation prime) will emerge in not only U.S. samples, but also in the comparison cultures (e.g., India, UK, and Australia).

Two final theories make firm predictions primarily about a subset of the effects focused on moral judgments related to work (Lottery winner and Intuitive mindset studies).

Predictions of competing theory "Social Class Differences"
Since low socioeconomic status (SES) individuals tend to perceive work as a job and means to an end (making a living), they should be less likely to moralize work than high-SES participants, who tend to see work as an end unto itself and part of a career. This theory predicts that across cultures, a higher educational and income level should be associated with exhibiting the Lottery winner and Intuitive mindset effects. 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.

Predictions of competing theory "Self-Expression Values"
Cross-national data from the World Values Survey suggests two main dimensions of culture: 1) Survival vs. Self-Expression values and 2) Traditional vs. Secular-Rational Values. Across nations, self-expression values tend to be associated with "work devotion," in other words perceiving work as an end unto itself, whereas survival values are linked to seeing work as a means of earning a living. Based on their national scores on the self-expression dimension, this perspective predicts that participants from the U.K., Australia and U.S. will exhibit the Lottery winner and Intuitive mindset effects whereas Indian participants will not. This alterative account of cultural differences makes no strong predictions for the Tacit inferences or Salvation prime effects. However, the strongest version of the theory (in which its predictions hold to the exclusion of all others), anticipates null findings. This cultural framework provides no basis to hypothesize an implicit link between work and sex values, or an automatic association between work and divine salvation.

KEY MODERATOR MEASURES

Religion and religiosity:

The potential moderator of religion will be measured using the following self-report survey item from the original studies, which will be the same in all the replications. We will categorically divide participants into Protestants and non-Protestants using this item.

My religion is:

- Protestant
- Catholic
- Islam
- Judaism
- Buddhism
- Atheist
- Agnostic
- Other (please indicate)

Below is the single item measure of religiosity, used in the original research and included in all of the present replications:

I consider myself to be:

Not at all Religious						Very Religious
1	2	3	4	5	6	7

The multi-item measure of religiosity is the five-item DUREL scale (Koenig & Büssing, 2010). This is 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*). The scale is calculated by calculating the average of all the responses, and no items are recoded.

Social class

Following on prior research (e.g., Snibbe & Markus, 2005), social class will be assessed principally using the item asking "My educational level is:"

Protestant work ethic (PWE)

The data collections will include the Katz and Hass (1988) Protestant Work Ethic (PWE) scale. The PWE measure 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*). The scale total score is calculated by taking the average of all the responses, and no items are recoded.

OVERALL ANALYSES

Consistent with past replication initiatives (Klein et al., 2014; Open Science Collaboration, 2015), for each target study one simple effect will be selected as the key comparison of interest. Whether that simple effect comparison is significant and in the expected direction, as well as moderation by national/regional and individual differences, will be used to adjudicate between the competing theories. To test the effects of the original hypothesis and moderators, linear mixed effects models with a centering within cluster approach will be used. The effect sizes will be converted into Cohen's d, bootstrapped and a meta-analysis will be conducted for each of the four effects. A test for heterogeneity using Cochran's *Q*, generated from a random effects meta-analysis, will be conducted (Cochran, 1954). Also, we will estimate the proportion of variance due to heterogeneity using I^2 and *Tau* (Higgins, Thompson, Deeks, & Altman, 2003; Borenstein et al., 2009; Borenstein, Hedges, Higgins, & Rothstein, 2010; Borenstein, Higgins, Hedges, & Rothstein, 2017).

The basic fixed effects model is represented by:

$$Y \sim A + B + C + A:B:C$$

Y = DV A = Main Effect 1 B = Main Effect 2 C = Moderator

A colon ("**:**") indicates an interaction effect

A tilde ("~") separates the dependent measure ("Y") from the main effects and interactions ("A", "B", and "C")

A plus sign ("+") indicates a new effect added to the model

All lower order interactions will be included in this model. For ease of understanding the model has been simplified. In this example the model would be Y ~ A + B + C + A:B + A:C + B:C + A:B:C

Data will be collected within regions around the world. To account for heterogeneity between regions, we will model all effects as mixed-effects models. Specifically, each model will contain a random intercept of region (defined by where in the world the data are collected from, see notes in Table S3-1 Row 1) as well as the primary effect of interest for each study modelled as random slopes. If the models fail to converge, we will begin removing random slopes terms until we achieve a convergent model. The simplest model will contain just a random intercept of the dependent variable nested within region. For the PureProfile data collections for Studies 1-2 we will be able to recruit respondents from specific subregions, allowing us to better balance our sample across different geographic areas and maximize statistical power. However, for the Mechanical Turk data collections for Study 1 we will be unable to do this and may for instance have too small a sample from one of the nine U.S. census districts. If we fail to collect a large enough sample within a given subregion to meaningfully estimate its effect, we will combine that subregion with a nearby subregion. If necessary, the decision to combine regions to achieve reasonable sample sizes will be made after the data has been collected, but prior to carrying out the key analyses testing the competing theories (Tables S3-1 through S3-4).

STUDY 1: TACIT INFERENCES, LOTTERY WINNER, AND INTUITIVE MINDSET EXPERIMENTS

The first data collection will contain three experiments that appear in counterbalanced order: the lottery winner study, tacit inferences study, and intuitive mindset study.

Lottery winner study

The dependent measure will be the item “Is Sarah a good person? (1=Very Bad, 7= Very Good)”

Analyses will consist of testing the hypothesized 3-way and 2-way interactions, tests of the main effects of age and work status, and moderator tests (single-item religiosity measure; multi-item DUREL religiosity scale; religion item with categorical division into Protestant or not; education; Protestant Work Ethic scale).

Overview of analyses and predictions for the theory of Implicit Puritanism:
-- Key interactions: Significant three-way interaction between target age (23 or 46) x work status (works or retires) x culture (for MTurk data collection, USA vs. India; for PureProfile data collection, USA vs. Australia and UK). Significant 2-way interaction for American participants only, such that younger target age is associated with more positive moral judgments, but only in the “works” condition. Note that to test the competing regional folkways thesis, New Englanders are contrasted to other Americans as well as members of the comparison national cultures.
-- Simple effect of age is tested across and within each work status condition (works vs. retires), separately for each nationality/culture.

-- Simple effect of work status (works vs. retires) is tested across and within age categories and nationalities. Main effect of work status, such that worker receives more praise than the retiree, should emerge.
-- Key difference and non-differences: Americans should rate the 23-year-old who continues to work more positively than they rate the 46-year-old who continues to work, but members of the comparison cultures should not.
-- No moderation by participant religion (Protestant or not), religiosity, education, or explicit PWE endorsement.

Table S3-1 below outlines the critical statistical tests for each competing theory more formally.

Table S3-1: Key Analyses for Lottery Winner Study

#	Theory Test	Description of Analysis	Model and Code
1	Overall analysis of study	2 (target age: 23 vs. 46) x 2 (work status: works vs. retires) x Country (US vs Other), with moral judgments as the outcome measure. This DV is responses to the question "Is Sarah a good person?" (1= Very Bad, 7= Very Good).	lmer(DV* ~ Age** + Country*** + Work_Status**** + Age:Work_Status:Country + (1 + Age:Work_Status|Region*****), data = mydata) * Responses to is Sarah a good person? (1= Very Bad, 7= Very Good) ** Age condition (target age 23 or 46) *** Here and for all of Tables S3 1-4, "Country" refers to a categorical variable where US = 0 and all the other countries will be 1. **** Work status condition (works or retires) ***** Region refers to smaller geographic regions within each world area. For the United States, it refers to census codes (nine census divisions within the U.S.). For the UK, it refers to the constituent countries in the United Kingdom, specifically England, Scotland, Wales, and Northern Ireland. For Australia, it refers to the states of New South Wales, Victoria, Queensland, Western Australia, South Australia, and Tasmania. For India, the three major regions are South India, Hindustan and North-east.
2	Key effect	The key effect for the lottery winner study is a simple effect of age (23 vs. 46) within the work condition, such that the younger target is seen more positively. Theoretically, this difference reflects the moralization of work in the absence of material need. As detailed below, the competing theories have different predictions regarding this focal effect.	lmer(DV ~ Age:Work_Status* + (1 + Age:Work_Status|Region), data = mydata) *Note only the continues to work condition will be kept in the Work_Status variable.
3	Implicit Puritanism	American, but not non-American participants, should evaluate a young person who continues working after winning the lottery more positively than an older person who continues working after winning the lottery. The critical statistical test is an interaction between the key simple effect and USA vs. other country.	The model starts with the key effect in Table S3-1, Row 2, in other words the simple effect of target age within the "target works" condition. This is then interacted with participant country, coded as USA (1) vs. Other (0). lmer(DV ~ Age:Work_Status*:Country** + (1+Age:Work_Status|Region), data = mydata) * Note only the continue to work condition will be kept in the Work_Status variable

			**Country coded as USA (1) vs. Other (0)
4	Religious Differences	Key effect (Table S3-1 Row 2) greater for participants who rate higher on religiosity (as assessed using the single item measure and DUREL), and who are Protestants rather than non-Protestants	lmer(DV ~ Age:Work_Status*:Rel** + (1 + Age:Work_Status|Region), data = mydata) lmer(DV ~ Age:Work_Status:DUREL**+ (1 + Age:Work_Status|Region), data = mydata) lmer(DV ~ Age:Work_Status:REL_Id***+ (1 + Age:Work_Status|Region), data = mydata) *Note only the continue to work condition will be kept in the Work_Status variable **Rel denotes answers to the question "I consider myself to be: Not at all Religious – Very Religious". ***DUREL denotes computed average of the answers to the DUREL scale (Koenig & Büssing, 2010) ***Rel_Id denotes the answers to "My religion is:", with Protestant = 0 and Other religions = 1
5	Regional Differences	Key effect (Table S3-1 Row 2) greater for New England participants than for non-New England participants	lmer(DV ~ Age:Work_Status*:New_Eng**+ (1 + Age:Work_Status|Region), data = mydata) *Note only the continue to work condition will be kept in the Work_Status variable **New_Eng variable will be a categorical variable whereby from the New England region will be coded as 0, while other locations will be coded as 1.
6	Social Class	Key effect (Table S3-1 Row 2) greater for high-SES participants than low-SES participants	lmer(DV ~ Age:Work_Status*:Edu**+ (1 + Age:Work_Status|Region), data = mydata) *Note only the continue to work condition will be kept in the Work_Status variable **Edu denotes answers to the question "my educational level is"
7	Explicit American Exceptionalism	Key effect (Table S3-1 Row 2) greater for Americans than non-Americans, and also greater for participants who endorse the Protestant Work Ethic (PWE) more strongly.	lmer(DV ~ Age: Work_Status*:Country**+ (1 + Age:Work_Status|Region), data = mydata) lmer(DV ~ Age:Country:Work_Status: PWE ***+ (1 + Age:Work_Status|Region), data = mydata)

			*Note only the continue to work condition will be kept in the Work_Status variable ** A categorical variable whereby US = 0 and all the other countries will be 1. *** PWE denotes answers to the Protestant Work Ethic scale (PWE; Katz & Hass, 1988)
8	General Moralization of Work	Key effect (Table S3-1 Row 2) present both for Americans and non-Americans	lmer(DV ~ Age: Work_Status*:Country**+ (1 + Age:Work_Status|Region), data = mydata) *Note only the continue to work condition will be kept in the Work_Status variable ** A categorical variable whereby US = 0 and all the other countries will be 1.
9	Self-Expression Values	Key effect (Table S3-1 Row 2) less present in India than in the United States. Note this alternative theory is tested only in the MTurk sample comparing the responses of Indians and Americans.	lmer(DV ~ Age: Work_Status*:Country**+ (1 + Age:Work_Status|Region), data = mydata) *Note only the continue to work condition will be kept in the Work_Status variable ** A categorical variable whereby US = 0 and India will be 1, for the MTurk sample.
10	False Positives	Key effect (Table S3-1 Row 2) not present for any country or relevant sub-region (e.g., New England)	

Tacit inferences study:

Dependent measures:
Scenario 1: True/False response on the item "Mary stays in bed on weekends before big trials in order to get extra work done."
Scenario 2: True/False response on the item "Julia slept with the host of last week's party."
Scenario 3: True/False response on the item "Ann did not study hard for the Civil War chronology quiz."
Scenario 4: True/False response on the item "Carl looked at the photos he found."

We will test both the tacit inferences condition x culture interactions and then, within each culture, compare the two tacit inferences conditions across the four vignettes. Finally, we will test potential moderators (single-item religiosity measure; multi-item DUREL religiosity scale; religion item with categorical division into Protestant or not; education; Protestant Work Ethic scale).

Overview of analyses and predictions for the theory of Implicit Puritanism:
-- Key interaction: Significant condition (tacit inferences condition 1 vs. tacit inferences condition 2) x culture (USA vs. other) interaction across the 4 vignettes.
-- Key differences: For Americans, significant condition differences across the four vignettes. Scenario 1: Mary is sexually abstinent in condition 1, which should lead to more "true" responses on the item: "Mary stays in bed on weekends before big trials in order to get extra work done."
Scenario 2: Julia does not work hard in school or at her job in condition 1, which should lead to more "true" responses on the item: "Julia slept with the host of last week's party."
Scenario 3: Anne is sexually active in condition 1, which should lead to more "true" responses on the item "Ann did not study hard for the Civil War chronology quiz."
Scenario 4: Carl works less hard in in condition 2, which should lead to more "true" responses on the item "Carl looked at the photos he found."
-- Key non-differences: For the comparison cultures (India, Australia, UK), no significant differences in tacit inferences between conditions 1 and 2.
-- No moderation by participant religion (Protestant or not), religiosity, education, or explicit PWE endorsement.

Table S3-2 below outlines the critical statistical tests for each competing theory more formally.

Table S3-2: Key Analyses for Tacit Inferences Study

#	Theory Test	Description of Analysis	Model and Code
1	Overall analysis of study	Tacit inferences condition (Condition 1 vs. Condition 2) x Country (US vs Other), with false memories as the outcome measure.	lmer(DV* ~ Condition**:Country*** + (1 + Condition |Region****), data = mydata) *Count of True responses to key questions across the four vignettes (1 per vignette) **Condition is a between-subjects factor manipulating whether vignette targets uphold or violate traditional morality ***Here and for all of Tables S3 1-4, "Country" refers to a categorical variable whereby US = 0 and all the other countries will be 1. ****Region refers to smaller geographic regions within each world area. For the United States, it refers to census codes (nine census divisions within the U.S.). For the UK, it refers to the constituent countries in the United Kingdom specifically England, Scotland, Wales, and Northern Ireland. For Australia, it refers to the states of New South Wales, Victoria, Queensland, Western Australia, South Australia, and Tasmania. For India, the three major regions are South India, Hindustan and North-east.
2	Key effect	The key effect is memory differences between the two conditions manipulating whether targets uphold or violate traditional morality. Specifically, individuals who violate work morality should be misremembered as also violating sexual morality, and vice versa. Such false memories reflect an implicit link between work and sex morality. As detailed below, the competing theories have different predictions regarding this focal effect.	lmer(DV ~ Condition + (1 + Condition|Region), data = mydata)
3	Implicit Puritanism	American, but not non-American participants, should exhibit an implicit link between work and sex morality. The critical statistical test is	lmer(DV ~ Condition:Country + (1 + Condition|Region), data = mydata)

		an interaction between the key simple effect and USA vs. other nationality.	
4	Religious Differences	Key effect (Table S3-2 Row 2) greater for participants who rate higher on religiosity (as assessed using the single item measure and DUREL) and who are Protestants rather than non-Protestants	lmer(DV ~ Condition:Rel* + (1 + Condition|Region), data = mydata) lmer(DV ~ Condition:DUREL** + (1 + Condition|Region), data = mydata) lmer(DV ~ Condition:REL_Id*** + (1 + Condition|Region), data = mydata) *Rel denotes answers to the question "I consider myself to be: Not at all Religious – Very Religious". **DUREL denotes computed average of the answers to the DUREL scale (Koenig & Büssing, 2010) ***Rel_Id denotes the answers to "My religion is:", with Protestant = 0 and Other religions = 1
5	Regional Differences	Key effect (Table S3-2 Row 2) greater for New England participants than for non-New England participants	lmer(DV ~ Condition:New_Eng* + (1 + Condition|Region), data = mydata) *New_Eng variable will be a categorical variable whereby from the New England region will be coded as 0, while other locations will be coded as 1.
6	Explicit American Exceptionalism	Key effect (Table S3-2 Row 2) greater for Americans than non-Americans, and also greater for participants who endorse the Protestant Work Ethic (PWE) more strongly.	lmer(DV ~ Condition:Country* + (1 + Condition|Region), data = mydata) lmer(DV ~ Condition:PWE * + (1 + Condition|Region), data = mydata) * A categorical variable whereby US = 0 and all the other countries will be 1. **PWE denotes answers to the Protestant Work Ethic scale (PWE; Katz & Hass, 1988)
7	General Moralization of Work	Key effect (Table S3-2 Row 2) present both for Americans and non-Americans	lmer(DV ~ Condition:Country*+ (1 + Condition|Region), data = mydata) * A categorical variable whereby US = 0 and all the other countries will be 1.
8	False Positives	Key effect (Table S3-2 Row 2) not present for any country or relevant sub-region (e.g., New England)	

Intuitive mindsets study

Dependent measures are the rational and intuitive items below:

My most rational, objective judgment is that:

Robert is a much better person than John						John is a much better person than Robert
1	2	3	4	5	6	7

My intuitive, gut feeling is that:

Robert is a much better person than John						John is a much better person than Robert
1	2	3	4	5	6	7

We will test the hypothesized question type x culture interaction, tests of effects of question type (rational vs. intuitive comparison) within each culture, test responses to each item separately against the neutral scale midpoint of 4, and finally carry out moderator tests (single-item religiosity measure; multi-item DUREL religiosity scale; religion item with categorical division into Protestant or not; education; Protestant Work Ethic scale).

Overview of analyses and predictions for the theory of Implicit Puritanism:
-- Key interaction: Significant mindset condition (intuitive item vs. rational item) X culture (USA vs. other) interaction.
-- Key difference: For the USA participants, the mean on the intuitive judgments item should be significantly greater than for the rational judgment item. In other words, Americans intuitively feel a lottery winner who continues to work (John) is a better person than a lottery winner who retires (Robert), but at the same time acknowledge this is not fully rational.
-- Key non-difference: Responses should be similar on the intuitive and rational item for the other countries.
-- For Americans, the mean on the intuitive judgment item should be significantly above the neutral scale midpoint of 4.
-- For Americans, the mean on the rational judgment item should <u>not</u> be significantly different from the neutral scale midpoint of 4 (reflecting indifference between the two targets).
-- For non-Americans, responses on both the intuitive and rational item should <u>not</u> be significantly different from neutral scale midpoint of 4 (again reflecting indifference between the two targets).
-- Americans should score significantly higher on the intuitive judgments item than members of the comparison cultures. In other words, Americans should be more likely than members of other cultures to intuitively prefer a lottery winner who continues to work at a menial job.
-- No moderation by participant religion (Protestant or not), religiosity, education, or explicit PWE endorsement.

Table S3-3 below outlines the critical statistical tests for each competing theory more formally.

Table S3-3: Key Analyses for Intuitive Mindsets Study

#	Theory Test	Description of Analysis	Model and Code
1	Overall analysis of study	Mindset (intuitive item vs. rational item) x Country (US vs. Non-US), with moral judgments as the outcome measure. Mindset is a within-subjects factor and country a between-subjects factor.	lmer(DV* ~ Condition**:Country*** + (1 + Condition|subj) **** + (1 + Condition|Region*****), data = mydata) * Answers to the question 7 = John is a much better person than Robert ** Condition is a within-subjects factor manipulating whether participants are asked for their intuitive or rational response. *** Here and for all of Tables S3 1-4, "Country" refers to a categorical variable whereby US = 0 and all the other countries will be 1. **** Identifying number of participants once the data is restructured ***** Region refers to smaller geographic regions within each world area. For the United States, it refers to census codes (nine census divisions within the U.S.). For the UK, it refers to the constituent countries in the United Kingdom specifically England, Scotland, Wales, and Northern Ireland. For Australia, it refers to the states of New South Wales, Victoria, Queensland, Western Australia, South Australia, and Tasmania. For India, the three major regions are South India, Hindustan and North-east.
2	Key effect	A stronger preference for the individual who upholds work morality (John) on the intuitive mindset item than on the rational mindset item. This simple within-subjects comparison reflects the intuitive moralization of work. As detailed below, the competing theories have different predictions regarding this focal effect.	lmer(DV ~ Condition + (1 + Condition|subj) + (1 + Condition|Region), data = mydata)
3	Implicit Puritanism	American, but not non-American participants, should uphold traditional work morality especially strongly in an intuitive mindset. The relevant statistical test is an interaction between the key simple effect and USA vs. other country.	lmer(DV ~ Condition:Country + (1 + Condition|subj) + (1 + Condition|Region), data = mydata)

4	Religious Differences	Key effect (Table S3-3 Row 2) greater for participants who rate higher on religiosity (as assessed using the single item measure and DUREL) and who are Protestants rather than non-Protestants	lmer(DV ~ Condition:Rel + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) lmer(DV ~ Condition:DUREL + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) lmer(DV ~ Condition:REL_Id*** + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) *Rel denotes answers to the question "I consider myself to be: Not at all Religious – Very Religious". **DUREL denotes computed average of the answers to the DUREL scale (Koenig & Büssing, 2010) ***Rel_Id denotes the answers to "My religion is:", with Protestant = 0 and Other religions = 1
5	Regional Differences	Key effect (Table S3-3 Row 2) greater for New England participants than for non-New England participants	lmer(DV ~ Condition:New_Eng* + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) *New_Eng variable will be a categorical variable whereby from the New England region will be coded as 0, while other locations will be coded as 1.
6	Social Class	Key effect (Table S3-3 Row 2) greater for high-SES participants than low-SES participants	lmer(DV ~ Condition:Edu * + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) *Edu denotes answers to the question "my educational level is?"
7	Explicit American Exceptionalism	Key effect (Table S3-3 Row 2) greater for Americans than non-Americans, and also greater for participants who endorse the Protestant Work Ethic (PWE) more strongly.	lmer(DV ~ Condition:Country * + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) lmer(DV ~ Condition:PWE* + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) * A categorical variable whereby US = 0 and all the other countries will be 1. **PWE denotes answers to the Protestant Work Ethic scale (PWE; Katz & Hass, 1988)

8	General Moralization of Work	Key effect (Table S3-3 Row 2) present both for Americans and non-Americans	lmer(DV ~ Condition:Country * + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) * A categorical variable whereby US = 0 and all the other countries will be 1.
9	Self-Expression Values	Key effect (Table S3-3 Row 2) less present in India than in the United States. Note this alternative theory is tested only in the MTurk sample comparing the responses of Indians and Americans.	lmer(DV ~ Condition:Country* + (1 + Condition|subj) + (1 + Condition|Region), data = mydata) * A categorical variable whereby US = 0 and India will be 1, for the MTurk sample.
10	False Positives	Key effect (Table S3-3 Row 2) not present for any country or relevant sub-region (e.g., New England)	

STUDY 2: SALVATION STUDY

Salvation study

Dependent measure: Total number of anagrams solved, adding up across the 4 DV stem items (bimodal, igneous, answer, curried). As noted in the instructions, only English language words four or more letters in length will be counted towards the total. Proper nouns (e.g. names), plurals, and tense changes (past tense, future tense) will be acceptable as a novel solution.

We will test the hypothesized priming x culture interaction, then tests of simple effects of priming condition within each culture, and carry out moderator tests (single-item religiosity measure; multi-item DUREL religiosity scale; religion item with categorical division into Protestant or not; education; Protestant Work Ethic scale).

Predictions of the theory of Implicit Puritanism:
-- Key interaction: Significant prime (salvation vs. neutral) x country (USA vs. other countries) interaction
-- Key difference: For Americans, significantly better anagram performance (total solutions generated) in the salvation prime condition than in the neutral prime condition
-- Key non-difference: No priming effect in comparison (non-USA) cultures
-- No moderation by participant religion (Protestant or not), religiosity, education, or explicit PWE endorsement

Table S3-4 below outlines the critical statistical tests for each competing theory more formally.

Table S3-4: Key Analyses for Salvation Prime Study

#	Theory Test	Description of Analysis	Model and Code
1	Overall analysis of study	2 prime condition (religious vs. neutral concepts) x Country (USA vs. non-USA), with anagram performance as the outcome measure.	lmer(DV* ~ Rel_Prime**+ Country *** + Rel_Prime:Country + (1 + Rel_Prime|Region****), data = mydata) *Total number of anagrams solved **Rel_Prime is a variable that indicates whether the participant was in the religion or neutral condition ***Here and for all of Tables S3 1-4, "Country" refers to a categorical variable whereby US = 0 and all the other countries will be 1. ****Region refers to smaller geographic regions within each world area. For the United States, it refers to census codes (nine census divisions within the U.S.). For the UK, it refers to the constituent countries in the United Kingdom specifically England, Scotland, Wales, and Northern Ireland. For Australia, it refers to the states of New South Wales, Victoria, Queensland, Western Australia, South Australia, and Tasmania.
2	Key effect	Improved work performance (i.e., greater number of anagrams solved) after being primed with religious concepts. This difference theoretically reflects an implicit association between work and the divine. As detailed below, the competing theories have different predictions regarding this focal effect.	lmer(DV ~ Rel_Prime + (1 + Rel_Prime |Region), data = mydata)
3	Implicit Puritanism	An interaction between the key simple effect (Table S3-4 Row 2) and Country (USA vs. other nationality). Americans, but not non-Americans, should respond to a religion prime with improved performance on a work task.	lmer(DV ~ Rel_Prime:Country + (1 + Rel_Prime|Region), data = mydata)

4	Religious Differences	Key effect (Table S3-4 Row 2) greater for participants who rate higher on religiosity (as assessed using the single item measure and DUREL) and who are Protestants rather than non-Protestants	lmer(DV ~ Rel_Prime:Rel* + (1 + Rel_Prime|Region), data = mydata) lmer(DV ~ Rel_Prime:DUREL** + (1 + Rel_Prime|Region), data = mydata) lmer(DV ~ Rel_Prime:REL_Id*** + (1 + Rel_Prime|Region), data = mydata) * Rel denotes answers to the question "I consider myself to be: Not at all Religious – Very Religious". ** DUREL denotes computed average of the answers to the DUREL scale (Koenig & Büssing, 2010) *** Rel_Id denotes the answers to "My religion is:", with Protestant = 0 and Other religions = 1
5	Regional Differences	Key effect (Table S3-4 Row 2) greater for New England participants than for non-New England participants	lmer(DV ~ Rel_Prime: New_Eng** + (1 + Rel_Prime|Region), data = mydata) * New_Eng variable will be a categorical variable whereby from the New England region will be coded as 0, while other locations will be coded as 1.
6	Explicit American Exceptionalism	Key effect (Table S3-4 Row 2) not present, and not moderated by the Protestant work ethic (PWE)	lmer(DV ~ Rel_Prime:Country* + (1 + Rel_Prime|Region), data = mydata) lmer(DV ~ Rel_Prime:PWE * + (1 + Rel_Prime|Region), data = mydata) * A categorical variable whereby US = 0 and all the other countries will be 1. ** PWE denotes answers to the Protestant Work Ethic scale (PWE; Katz & Hass, 1988)
7	General Moralization of Work	Key effect (Table S3-4 Row 2) present both for Americans and non-Americans	lmer(DV ~ Rel_Prime:Country* + (1 + Rel_Prime|Region), data = mydata) * A categorical variable whereby US = 0 and all the other countries will be 1.
8	False Positives	Key effect (Table S3-4 Row 2) not present for any country or relevant sub-region (e.g., New England)	

DETERMINING WHICH THEORY BEST FITS THE DATA

Below in Table S3-5 we use simplified scores to heuristically capture the contrasting predictions about mean differences made by several of the different theories, following on Jussim et al. (1987).

Again, as described in Tables S3 1-4, our four key pre-specified effects for the purposes of comparing the effectiveness of the different theories are:

1. Lottery winner study: Higher mean scores for moral character ratings in the 23 year old who continues to work condition than in the 46 year old who continues to work condition.
Note: Every theory except for Implicit Puritanism and False Positives can also incorporate a modified lottery winner effect in which the worker is preferred over the retiree regardless of age, so long as the theory's other predictions about variability across populations hold. This modified lottery winner effect is captured by an effect of work status (worker ratings > retiree ratings) in the specified population (e.g., Americans, Protestants, New Englanders), no work status x target age interaction, plus the specific theory's predictions about group differences (e.g., Americans vs. non-Americans, Protestants vs. non-Protestants, New England vs. other regions).

2. Intuitive mindset study: Higher scores on the intuitive item than on the rational item.
Note: Every theory except for Implicit Puritanism and False Positives can also incorporate a modified lottery winner effect in which the worker is preferred over the retiree in both an intuitive and deliberative mindset. This modified prediction is scores above the neutral scale midpoint of 4 on *both* the intuitive and rational items in the expected population (e.g., Americans, Protestants), plus each theory's specific predictions about variability across populations (e.g., regional or religious differences).

3. Tacit inferences study: True/false responses overall across each of the four vignettes consistent with linking work and sex morality. Specifically:
Scenario 1: In condition 1, more "true" responses on the item: "Mary stays in bed on weekends before big trials in order to get extra work done."
Scenario 2: In condition 1, more "true" responses on the item: "Julia slept with the host of last week's party."
Scenario 3: In condition 1, more "true" responses on the item "Ann did not study hard for the Civil War chronology quiz."
Scenario 4: In condition 2, more "true" responses on the item "Carl looked at the photos he found."

4. Salvation study: More total anagrams solved in the religion priming condition than in the neutral prime condition

In Table S3-5 below, +1 means the relevant theory predicts the effect in question, and a zero (0) means it does not. For example, the theory of Implicit Puritanism predicts a tacit inferences effect for Americans (coded as 1 in Table S3-5), but not for non-Americans (coded as 0 in Table S3-5). "Scenario" refers to the studies relying on vignettes, specifically the Lottery Winner, Tacit

Inferences, and Intuitive Mindset effects (#1-3 above). "Priming" refers to the study relying on a scrambled-sentences manipulation, specifically the Salvation Prime effect (#4 above). We will separately examine the outcomes of the scenario studies and priming study given recent difficulties in replicating priming studies (e.g., Doyens et al., 2012; Harris et al., 2013; Klein et al., 2014; Open Science Collaboration, 2015), as well as theoretical interest in distinguishing between relatively unconscious and consciously accessible aspects of work and sex morality.

Notably, certain subsets of effects are particularly relevant to different claims from the theory of Implicit Puritanism. To assess the intuitive moralization of work hypothesis, we will focus on the replications for the Lottery Winner and Intuitive Mindset effects. For the implicit link between work and sex morality, our focus is on the Tacit Inferences replication. The Salvation Prime replication tests the hypothesized implicit link between work and divine salvation.

We will separately examine the predictions of each theory for each of the four effects, since distinct theories may best explain each.

Table S3-5. Predictions of the strong versions of each of the 6 main theories

	USA sample				**Other national culture**	
	New England		**Other USA states**			
THEORY	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**
Implicit Puritanism	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0
Religious differences	Scenario: +1 Priming: +1	Scenario: 0 Priming: 0	Scenario: +1 Priming: +1	Scenario: 0 Priming: 0	Scenario: +1 Priming: +1	Scenario: 0 Priming: 0
Regional folkways	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0
Explicit American exceptionalism	Scenario: +1 Priming: +0	Scenario: +1 Priming: +0	Scenario: +1 Priming: +0	Scenario: +1 Priming: +0	Scenario: +0 Priming: +0	Scenario: +0 Priming: +0
General moralization of work and sex	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1
False positives	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0	Scenario: 0 Priming: 0

To explain Table S3-5 above in plainer language:

- Implicit Puritanism theory predicts the key effects (lottery winner, potato peeler, memory effect, and salvation prime) will emerge among Americans, but not non-Americans. Again, an effect emerging is heuristically indicated as "1" in the table, and an effect not emerging is indicated as a "0" in the table.
- The religious differences perspective predicts the key effects will hold for Protestants and religious participants ("1" in the relevant cells), but not non-Protestants or less religious participants ("0" in the relevant cells).

- Regional folkways: New England participants will exhibit the key effects ("1" in the relevant cells), but not Americans outside New England or individuals from other countries ("0" in the relevant cells).
- Explicit American Exceptionalism: Americans will exhibit the scenario effects (lottery winner, potato peeler, and memory effects) whereas non-Americans will not ("1" for Americans for scenario studies; "0" for non-Americans). Both U.S. and non-U.S. participants will fail to exhibit the salvation prime effect (reflected in a consistent "0" in every cell for the priming effect).
- General moralization of work and sex: Both Americans and non-Americans will exhibit both the scenario and priming effects (reflected in a consistent "1" in every cell).
- False positives: none of the key effects will emerge for either Americans or non-Americans (reflected in a consistent "0" in every cell).

Following on Jussim et al. (1987) scores in Table S3-5 are based on the strong version of each theory in which it provides an "exhaustive and mutually exclusive" account of work and sex morality. In other words, the table above captures only each theory's individual predictions in isolation, to the exclusion of all the other theories. However, one can readily imagine scenarios where multiple theories are additively true (as Jussim et al., 1987, found for three major theories of racial stereotyping). Consider for instance the possibility that Implicit Puritanism effects will replicate robustly among Americans in general, but at the same time the effects are shown particularly strongly by Protestant and/or religious Americans. This potential outcome is captured in Table S3-6 below.

Table S3-6. Pattern of results if both the Implicit Puritanism and Religious Differences perspectives are correct, and predictions are combined additively.

	USA sample				**Other national culture**	
	New England		**Other USA states**			
THEORY	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**
Implicit Puritanism + Religious Differences predictions combined additively	Scenario: +2 Priming: +2	Scenario: +1 Priming: +1	Scenario: +2 Priming: +2	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: 0 Priming: 0

The pattern in Table S3-6 above reflects the key effects emerging for Americans more so than non-Americans, as predicted by the theory of Implicit Puritanism— and at the same time emerging more strongly among religious Protestants, as expected by the religious differences perspective. In this potential outcome, the lottery winner, potato peeler, memory effect, and salvation prime effect emerge especially strongly for religious and Protestant Americans, as indicated by a "2" in the table.

Alternatively, consider the additive case in which Implicit Puritanism effects are true of Americans in general, but especially true of those living in the New England states (Table S3-7).

Table S3-7. Pattern of results if Implicit Puritanism and Regional Folkways perspectives are both correct, and their predictions are combined additively.

	USA sample				Other national culture	
	New England		Other USA states			
THEORY	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**
Implicit Puritanism + Regional Folkways combined additively	Scenario: +2 Priming: +2	Scenario: +2 Priming: +2	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1	Scenario: +0 Priming: +0	Scenario: 0 Priming: 0

The pattern in Table S3-7 above reflects the key effects emerging for Americans more so than non-Americans, as predicted by the theory of Implicit Puritanism, and at the same time emerging most strongly for New Englanders, as expected by the regional folkways perspective. In this potential outcome, the lottery winner, potato peeler, memory effect, and salvation prime effect emerge especially strongly for Americans from the New England states, as indicated by a "2" in the table.

As illustrated in Table S3-8 below, a candidate theory can be correct for scenario effects but not priming or vice versa. For instance, the false positive perspective might be most supported for priming (i.e., the Salvation Priming effect failing to emerge across all replication samples), whereas the general moralization of work is most supported for the scenario studies (Lottery winner, Tacit inferences, and Intuitive mindset effects replicate not only in the United States but also in India, Australia, and the United Kingdom).

Table S3-8. Outcomes if general moralization of work and sex is supported for scenario studies and false positives for the priming experiment.

USA sample				Other national culture	
New England		Other USA states			
Protestant/ Religious	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**
Scenario: +1 Priming: 0	Scenario: +1 Priming: 0	Scenario: +1 Priming: 0	Scenario: +1 Priming: 0	Scenario: +1 Priming: 0	Scenario: +1 Priming: 0

The pattern in Table S3-8 above reflects the lottery winner, potato peeler, and memory effect emerging for both Americans and non-Americans (as indicated by a "1" in the table), and the salvation prime replications returning null effects for both Americans and non-Americans (as indicated by a "0" in the table).

As illustrated in Table S3-9 below, toned-down versions of a theory that makes weaker claims might also emerge from these large-scale data collections. For instance, Implicit Puritanism effects (Lottery winner, Tacit inferences, Intuitive mindset, Salvation prime) might replicate in all cultures, but be twice as strong among Americans (e.g., $d = .40$ for Americans, $d = .20$ for Australia, India, and the UK). The conclusion then would be that work and sex are moralized

across cultures, but relatively more so in the United States. This would be consistent with the additive predictions from the general moralization of work and sex and Implicit Puritanism perspectives.

Table S3-9. Implicit Puritanism and General Moralization of Work and Sex predictions combined additively

	USA sample				Other national culture	
	New England		Other USA states			
THEORY	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**	**Protestant/ Religious**	**Non-Protestant/ Less Religious**
Implicit Puritanism + General Moralization of Work and Sex combined additively	Scenario: +2 Priming: +2	Scenario: +2 Priming: +2	Scenario: +2 Priming: +2	Scenario: +2 Priming: +2	Scenario: +1 Priming: +1	Scenario: +1 Priming: +1

In Table S3-9 above, the key effects are present in all cultures (as indicated by a value of at least a "1" in each cell, rather than "0"), but the effects are strongest in the United States (as indicated by a "2" in the table).

In terms of the ultimate theoretical conclusions regardless underlying processes, we specify in advance that the case for the "implicit" (Bargh, 2014; Bargh et al., 1996; Greenwald & Banaji, 1995) nature of Puritanism effects hinges on the Salvation Priming replications (Study 2). In contrast, the scenario studies (Intuitive mindset, Tacit inferences, and Lottery winner studies) capture moral intuitions, in other words responses that may implicate automatic or unreasoned processing to some extent, but whose outputs are conscious and introspectively accessible (Haidt, 2001). There is no "Implicit Puritanism" without the priming effects, but there may still be an "Intuitive Puritanism" if the other theoretically predicted effects and cultural differences emerge (e.g., Americans intuitively lauding a young lottery winner who continues working at her job while members of comparison cultures do not). Indeed, "Intuitive Puritanism" is specifically anticipated by the Explicit American Exceptionalism perspective, which postulates that American values are deeply influenced by a Puritan-Protestant heritage, but that the resulting intuitive moral values are conscious and reportable.

We will use model-fitting tests to help assess which theory best accounts for the pattern of results. To test which theories fit the data, we will construct a series of linear mixed-effects models for each paradigm. After fitting an unconditional model to examine site independence assumptions, we will fit a base model predicting a paradigm's given DV from experimental condition (as a fixed effect) and a random intercept and random slope of experimental condition nested within site. If the model fails to converge, we will remove the random slope. We will then add additional fixed effects to this base model to test each individual theory. For most theories, this will involve adding an additional fixed effect term, and then an interaction term between the added fixed effect and experimental condition (e.g., for Implicit Puritanism, the second step model would include fixed main effects of culture [US vs. non-US], region [New England vs. non-New England], and religion [Protestant vs. Non-protestant], and the third step would contain

the interaction of culture and experimental condition). We will examine parameter estimates and conduct a likelihood ratio test to determine whether the addition of each theory's additional fixed effect improves the model.

A base model for Implicit Puritanism would look like the following:

Level 1 variables: DV, religion, experimental condition
Level 2 variables: culture, region, site

Step 0
DV =
random intercept = (site)

Step 1
DV = experimental condition
random intercept= experimental condition, nested within site; random slope = experimental condition, nested within site

Step 2
DV = experimental condition + culture + region + religion
random intercept= experimental condition, nested within site; random slope = experimental condition, nested within site

Step 3
*Testing implicit puritanism
DV = experimental condition + culture + region + religion +
experimental condition * culture
random intercept= experimental condition, nested within site; random slope = experimental condition, nested within site

*Testing religious differences
DV = experimental condition + culture + region + religion +
experimental condition * religion
random intercept= experimental condition, nested within site; random slope = experimental condition, nested within site

*Testing regional folkways differences
DV = experimental condition + culture + region + religion +
experimental condition * region
random intercept= experimental condition, nested within site; random slope = experimental condition, nested within site

*Full theory tests (in case multiple theories come out)
DV = experimental condition + culture + region + religion +
experimental condition * culture +
experimental condition * religion +

experimental condition * region
random intercept= experimental condition, nested within site; random slope = experimental condition, nested within site

As per the above, in addition to testing the strong predictions of each theory (Table S3-5) we will test various "additive models" (e.g., Tables S3-6 through S3-9) in which the predictions of different theories are combined together to see if this best accounts for the pattern of results.

STATISTICAL POWER

Without much inter-site heterogeneity, some of the theories tested will not be supported. Our model comparison process should help us to test for that possibility. We ran several power simulations in order to better understand the likelihood of detecting effects between some of the high-level variables in our models, specifically, comparing the strength of effects in the United States to all other countries. We simulated data using our planned sample sizes and using effect sizes from previous studies on Implicit Puritanism as the basis for effect sizes in the United States, and assumed, as the theory would predict, a null effect in other countries. With little intra-region heterogeneity (d +/- .1), our sample should be well powered (~100%) to detect even the smallest previously observed Implicit Puritanism effect (d = .37). However, with added intra-region heterogeneity (d +/- .4) our power falls to ~37% to detect the smallest previously estimated Implicit Puritanism effect. Breaking our sample into smaller sub-samples (e.g., breaking each sample into five sub-samples based around state or province boundaries) increases power substantially (~72%, again based on the smallest previously estimated Implicit Puritanism effect). As such, instead of collecting a few larger samples, we will collect a large number of participants, but in smaller groups. This increase in number of sites will help us to increase power if we observe high intra-region heterogeneity and should not cost us power if the regions are fairly homogenous.

DATA-DEPENDENT VS. DATA-INDEPENDENT DECISIONS

The resulting dataset will provide a rich opportunity for further supplementary analyses beyond the preregistered ones. For example, various demographic variables, either individually or in combination with one another, may explain the results. There may also be specific subgroups of participants, not identified previously, for whom Implicit Puritanism effects hold especially strongly (e.g., White American Protestants who are highly religious and lifelong residents of New England).

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

For instance, if we indeed find that White American Protestant men who are highly religious and lifelong residents of New England most strongly exhibit Implicit Puritanism effects in the initial

test sample, we will pre-register a formal test of that hypothesis and run it on the data-independent holdout sample.

Below we outline at a conceptual level some analyses we anticipate conducting on the initial test sample. However, we will do not pre-specify each and every statistical test, since ultimately any promising analyses from the test sample will be pre-registered and applied to the holdout sample.

Further checks on data quality

We will repeat analyses not only selecting out participants with not only less than 5 years of experience with the English language (our pre-registered exclusion criteria), but also less than 10 years of prior experience (key demographic item: "How many years of experience do you have with the English language?"). This is to avoid effect sizes being artificially reduced due to participants misunderstanding the study materials. In addition, we will re-run the primary analyses selecting only participants who correctly responded "strongly disagree" on the attention check.

To further check on the integrity of the data, we will blindly code written responses to the free response awareness probe ("What do you think this survey was about?") for nonsensical and incoherent written comments, and likewise screen out duplicate written comments (e.g., two supposedly different participants write word-for-word identical free responses to the same open-ended query).

Prior studies find that less engaged participants who speed through the survey actually exhibit priming effects more strongly (Huang, 2014). More generally, dual process models suggest that faster responding should increase the influence of intuitive and implicit mental processes (Gawronski & Bodenhausen, 2006). Therefore, for the online data collections, we will assess whether completion speeds moderate whether the hypothesized Implicit Puritanism effects replicate, following the approach used by Huang (2014). Specifically, low response effort will be operationalized as below-the-median survey duration times. To prevent insufficient effort responding from clouding the results, participants whose page completion times would require reading a non-believable 675 words per minute will be removed from the sample.

Examining order of measures

For Study 1, which contains the tacit inferences, lottery winner, and intuitive mindset studies, we will test whether it matters if the experiment in question came first, second, or third in terms of order of administration. This is to address the participant fatigue issue, and potential interference effects from running multiple studies together. If the predicted work and sex morality effects are stronger when a given study is administered first, we will repeat our analyses using only participants for which that study came first and report those results separately, in addition to the overall findings collapsing across study order.

Tests of conscious awareness of being primed

We will repeat our analyses for the salvation priming experiment (Study 2) honing in on those participants most naïve to the purpose of the priming manipulation, and therefore theoretically most likely to exhibit implicit social cognition effects. We will test whether responses to the item "Did the sentence unscrambling task influence your performance on the anagram task in any way?" (*1 = no, 9 = yes*) moderate the hypothesized priming effect. We will then repeat our analyses above selecting only participants who score a 5 or below on this item. Finally, we will blindly code free responses to the items reading "If yes, please explain how and why it influenced you in your own words", flag subjects who may have suspected the true purpose of the study, and re-run the analyses excluding these participants.

Alterative means of sorting participants into cultural groups

As outlined earlier under "inclusion criteria," following on the approach used in the original studies (Uhlmann et al., 2009, 2011), for Study 1 (Tacit inferences, Lottery winner, and Intuitive mindset) and Study 2 (Salvation study) we will primarily group subjects into cultural categories based on the objective location of the data collection (e.g., USA, India, UK, Australia), rather than selecting participants based on their immigrant status, time spent in the United States, citizenship, etc. Likewise, initial regional classifications (New England) will be based on the state in which the data is collected (e.g., for New England, Maine, Vermont, New Hampshire, Massachusetts, Rhode Island, and Connecticut).

However, as a supplementary strategy, we will also repeat our analyses using participants' self-reports regarding the country and state/sub-region they are based primarily in (relevant items include: "Which of the following countries are you currently based primarily in?" and "If you selected the United States, which U.S. state are you primarily based in?"). We will also use self-reported nationality to group participants as Americans, Indians, UK, Australian, or others (the relevant demographic item is "Of what nation are you a citizen?"), and self-reported home state to separate New Englanders from other Americans (the relevant demographic item is "If you grew up in the United States, what U.S. state/territory did you grow up in?"). We will further repeat all analyses using self-reported nation of birth ("What country/region were you born in?) to group participants by nationality.

In addition, we will repeat the analyses using years spent in the U.S. as a continuous measure of U.S. cultural exposure ("How many years have you lived in the United States?"). We will also repeat these analyses taking into account participant age to avoid confounding cultural exposure with being chronologically older. Participants with greater levels of U.S. cultural exposure, regardless of nation of residence or origin, may exhibit the hypothesized Implicit Puritanism effects more strongly.

Alternative measures of social class

As an additional supplementary measure of social class we will assess self-reported income with the item "My yearly household income level is:"

As another supplementary measure of social class, we will assess parental education level using the question: "What is the education level of your most educated parent?"

QUALITATIVE ASSESSMENT OF REPLICATION RESULTS

The last author will provide a qualitative assessment of scientific status of the theory of Implicit Puritanism, which is contingent on the replicability of the four critical effects in question. A "failure to replicate" could take one of two forms. First, the original effect size might not emerge among Americans. Second, a pattern of cross-cultural differences might be obtained that is contrary to that in the original studies. In other words, each effect (Lottery Winner, Intuitive Mindset, Tacit Inferences, Salvation prime) needs to emerge among Americans, *and* more strongly than in members of the comparison cultures, for the original Implicit Puritanism predictions to be truly supported.

Non-replication of the salvation prime effect undermines the claim that at an implicit level, work morality is not fully secular in America and retains a cognitive residue of its religious roots. However, failed replications of the Salvation prime effect still leave room for a modified theory of "Intuitive Puritanism," most compatible with the Explicit Moral Exceptionalism perspective—so long as the other effects (Lottery winner, Intuitive mindset, Tacit inferences) replicate among Americans but not non-Americans. Non-replication of both the Lottery Winner and Intuitive Mindset effects fails to support the intuitive moralization of work hypothesis, core to the theory. A systematic failure to replicate the Tacit Inferences effect calls into question the hypothesized implicit work-sex link that is one of Implicit Puritanism's core predictions.

Failed replications of any combination of 3 of the key original findings (i.e., no effect in U.S. samples *or* no predicted cultural difference for at least 3/4 of the original effects), represent a breach of the theory's theoretical core (Lakatos, 1970). In the case of a core breach the theory of Implicit Puritanism should most likely be abandoned, and either replaced by one or more of the alternative theories considered here, or an entirely new account of work and sex morality.

So long as the basic pattern of experimental effects and cultural differences predicted by Implicit Puritanism holds, some individual and group differences could be straightforwardly be incorporated into the theory. Specifically, Protestant Americans may exhibit the effects more strongly than non-Protestants, religious Americans more so than the non-religious, New Englanders comparatively more so than individuals from other U.S. states, and high-SES Americans more so than low-SES Americans. Although relative differences between these subgroups are reconcilable with the original theorizing, Implicit Puritanism theory does make the strong prediction that the four original effects should still be significant among non-Protestant, less religious, and low-SES Americans not from the New England states. This again based on the thesis that Puritan-Protestant values implicitly permeate U.S. culture.

References for Supplement 3 (Not cited in main manuscript)

Borenstein, M., et al. (2009). *Introduction to meta-analysis*. Chichester (UK): Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97-111.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5-18.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101-129.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), 557.

Supplement 4: Pre-Registered Analysis Plan for the Forecasting Survey

CULTURE AND WORK REPLICATION PROJECT: PRE-ANALYSIS PLAN FOR THE FORECASTING SURVEY

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

Summary: In this survey we will examine whether researchers can predict the extent to which experimental findings regarding work morality replicate in data collections in different cultures and populations around the world. Of particular interest is the tendency to morally praise individuals for working in the absence of material need to work (the "needless work" effect), as well as linking work to other forms of traditional morality and divine salvation (Poehlman, 2007; Uhlmann, Poehlman, & Bargh, 2008, 2009; Uhlmann, Poehlman, Tannenbaum, & Bargh, 2011). The data for the replications are collected in the United States (differentiating the New England states from the rest of the country), United Kingdom, Australia, and India.

We are targeting researchers with training in judgment and decision making/social psychology research to participate in the forecasting survey, with no exclusion based on seniority or any other demographic characteristic.

Each participant (also referred to as forecaster in the rest of this pre-analysis plan) makes a total of $p = 48$ predictions. These will focus on five different work morality effects:

1. Needless work effect - 6 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects
2. Target age effect - 6 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects
3. Intuitive work morality effect - 6 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects
4. Tacit inferences effect - 6 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects
5. Salvation primes and work behavior - 4 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects

The data for these direct and conceptual replications are collected in the USA as a whole (MTurk sample), USA New England states (PureProfile sample), USA non-New England states (PureProfile sample), UK (PureProfile sample), Australia (PureProfile sample), and India (MTurk sample) for effects #1-4. For the fifth effect, no MTurk data was collected, hence the

predictions are for USA New England States, USA non-New-England states, Australia, and UK, all sampled via the professional survey firm PureProfile. In addition to making these predictions, the participants are asked to answer a set of demographic questions.

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

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

Primary Hypotheses

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

Individual-level regression to test whether forecasters' beliefs are significantly related to the realized effect sizes:

$$RES_{hc} = \beta_0 + \beta_1 PES_{ihc} + \varepsilon_{ihc} \quad (1)$$

where:

- RES_{hc} is a continuous variable indicating the realized effect size of the hypothesis h object of the prediction in population c;
- PES_{ihc} is a continuous variable indicating the predicted effect size of the effect of hypothesis h in population c by forecaster i;

In equation (1) we plan to cluster standard errors at the individual level (number of clusters determined by the number of forecasters with $N = 48$ observations per cluster), since doing so allows us to take into account the fact that the predictions elicited from the same forecaster might be correlated.

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

Robustness test of Hypothesis 1: we will estimate regression (1) separately for the two sets of predictions - predictions regarding simple effects and regarding moderator effects. Moreover, we will also carry out a robustness test where we estimate the Pearson correlation between the two vectors ($N = 48$ each) with the mean predicted effect size (PES_{hc}) of each of the 48 effects

replicated and the realized effect sizes RES_{hc}. Finally, we will estimate the Pearson correlation separately for the predictions regarding simple effects and the predictions regarding the moderator effects.

Primary Hypothesis 2: Forecasts regarding simple effect sizes are more accurate than forecasts regarding moderator effect sizes

Can participants predict complex experimental results, such as those associated with each candidate moderator, with the same accuracy achieved in predictions of simple effect sizes? To answer this question, first we compute the *accuracy* achieved in forecast hc by each survey-taker i in terms of squared prediction error (Brier score), according to the formula:

$$BS_{ihc} = (PES_{ihc} - RES_{hc})^2$$

where RES_{hc} and PES_{ihc} should be interpreted as specified above. Then, we regress the variable BS_{ihc} on a dummy variable identifying the forecasts regarding moderators (MES_{ihc})and on the individual fixed effects FE_i, clustering the standard errors at the individual level in line with model (1).

$$(2) \qquad BS_{ihc} = \beta_0 + \beta_1 MES_{ihc} + FE_i + \varepsilon_{ihc}$$

Tests: *t*-test on coefficient $\beta_1$in regression equation (2). Under the assumption that the forecasts regarding the moderators effects are more demanding, we expect $\beta_1$to be positive.

Secondary Hypothesis

Secondary hypothesis: Forecasted effect sizes are not significantly different from the realized effect sizes.

Hypothesis 1 tests the correlation between forecasts and realized effect sizes, but is not informative about the difference between the realized effects and their forecasted counterparts. To investigate whether the forecasted effect sizes are significantly different from the realized ones, we plan to apply the following procedure. First, for each of the 5 key effects we estimate the meta-analytic mean effect size $PES^m{}_h$, h ranging between 1 and 5, by pooling the effect sizes across the different cultures and populations (namely, across 6 populations for key effects 1 to 4 and across 4 populations for effect 5, as specified above) in a random effects meta-analysis. Then, we estimate the average at the individual level of the effect size of each key effect across the different populations for each participant (PES_{ih}). Finally, for each of the five key effects we implement a z-test comparing the meta-analyzed effect size $PES^m{}_h$ to the mean of PES_{ih}.

Exploratory Hypotheses

Do participants predict experimental results across different populations with different degrees of accuracy? To answer this question we plan to estimate equation (3):

$$(3) \qquad BS_{ihc} = \beta_0 + \beta_1 USNE_c + \beta_2 USNNE_c + \beta_3 US_c + \beta_4 AUS_c + \beta_5 IN_c + FE_i + \varepsilon_{ihc}$$

where BS_{ihc} and FE_i should be interpreted as above and $USNE_c$, $USNNE_c$,US_c, AUS_c and IN_c are dummy variables identifying forecasts on New England states in US (data collected via PureProfile), non-New England states in US (PureProfile), US (MTurk), Australia (PureProfile), and India (MTurk) respectively (United Kingdom being the baseline population). In line with previous regressions, in equation (3) the standard errors are clustered at the individual level.

Tests: separate *t*-test on coefficients $\beta_1$to β_5 in regression equation (3); Wald test on coefficients β_i being different from β_j for $i, j \in (1,2,3,4,5)$.

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

- Predictions regarding the five key work morality effect sizes
- Predictions regarding the four moderator effects

Are the forecasters' years of academic experience related to higher accuracy? To answer this question, we plan to regress BS_{ihc} on the variable SEN_i which represents the year from when the PhD was awarded (this variable takes value zero if a PhD title is not awarded yet). We will again cluster the standard errors at the individual level to take into account potential correlations across forecasts made by the same forecaster.

$$(4) \qquad BS_{ihc} = \beta_0 + \beta_1 SEN_i + \varepsilon_{ihc}$$

Test: *t*-tests on coefficient $\beta_1$in regression equation (4).

As a robustness check for hypothesis 3, we will analyze the accuracy of predictions on simple effects and on moderators effects separately. We will also use a different proxy of seniority, namely, academic job rank.

Incentives scheme

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

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

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

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

where $\underline{Sq.Error}$ is the average of the squared errors for all the 48 forecasts of the 'Work and Culture Forecasting Study' made by the forecasters.

References for Supplement 4

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ...Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10. doi:10.1038/s41562-017-0189-z

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

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

Uhlmann, E.L., Poehlman, T.A., & Bargh, J.A. (2008). Implicit theism. In R. Sorrentino & S. Yamaguchi (Eds.) *Handbook of Motivation and Cognition Across Cultures*. (pp. 71-94). St. Louis, MO: Elsevier/Academic Press.

Uhlmann, E.L., Poehlman, T.A., & Bargh, J.A. (2009). American moral exceptionalism. In J.T. Jost, A.C. Kay, & H. Thorisdottir (Eds.) *Social and Psychological Bases of Ideology and System Justification.* (pp. 27-52). New York, NY: Oxford University Press.

Supplement 5: Forecasting Survey

WORK MORALITY ACROSS CULTURES: FORECASTING SURVEY

We are scientists at the Stockholm School of Economics, University of Limerick, and INSEAD conducting an investigation of forecasting accuracy. We are interested in whether researchers can predict the extent to which experimental findings regarding work morality replicate in data collections in different cultures around the world. We are recruiting researchers with training in judgment and decision making/social psychology research to participate in this study. All levels of expertise are welcome, from graduate students to senior professors. In addition to providing your forecasts, you will also complete a brief demographic questionnaire.

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

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

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

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

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

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

[Page break here; do not include page breaks unless directly indicated]

Your Contact Information

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

Then click "next" to complete the survey.

[Page break here]

Forecasting Survey: Work Morality Across Cultures

About the initiative

The culture-and-work-morality project tested eight competing theories of work morality across cultures against one another, by directly and conceptually replicating previously observed effects across multiple countries and measuring a number of theoretically important individual differences moderators. Of particular interest is the tendency to morally praise individuals for working in the absence of material need to work (the "needless work" effect), as well as linking work to other forms of traditional morality and divine salvation (Poehlman, 2007; Uhlmann, Poehlman, & Bargh, 2008, 2009; Uhlmann, Poehlman, Tannenbaum, & Bargh, 2011; Uhlmann & Sanchez-Burks, 2014).

The eight competing theories are the following:

Implicit Puritanism perspective: Americans intuitively and automatically moralize work, due the legacy of the Puritan-Protestant founding in American culture.

Religious differences perspective: Religious individuals moralize work more than non-religious individuals, and Protestants moralize work more than those who follow other faiths.

Regional folkways perspective: New Englanders moralize work, due to the founding legacy of the Puritan-Protestants in this region of what became the United States.

Explicit American exceptionalism perspective: Americans consciously moralize work, to the extent that they consciously endorse traditional beliefs such as the Protestant Work Ethic.

General moralization of work perspective: People across cultures moralize work. In other words, work morality effects should emerge in all cultures, not just the United States.

False positives perspective: The original findings are spurious due to small sample sizes and researcher degrees of freedom in data analysis.

Self-expression values perspective: Individuals from wealthy nations (e.g., U.S., U.K., Australia) moralize work more than individuals from less wealthy nations (e.g., India). This is because national wealth is associated with self-expression values linking work to personal fulfillment and a sense of meaning, not just instrumental goals such as earning money to pay for necessities.

Social class perspective: High socioeconomic status (high SES) persons moralize work. Better educated individuals moralize work because they tend to view work as a source of meaning and fulfillment. In contrast, less educated individuals view work in instrumental terms, as a means of making a living, and should not moralize needless labor.

References:

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

Uhlmann, E.L., Poehlman, T.A., Tannenbaum, D., & Bargh, J.A. (2011). Implicit Puritanism in American moral cognition. *Journal of Experimental Social Psychology, 47,* 312-320.
Full text: socialjudgments.com/docs/Uhlmann.Poehlman.Tannebaum.Bargh.2011.pdf

Uhlmann, E.L., Poehlman, T.A., & Bargh, J.A. (2009). American moral exceptionalism. In J.T. Jost, A.C. Kay, & H. Thorisdottir (Eds.) *Social and Psychological Bases of Ideology and System Justification.* (pp. 27-52). New York, NY: Oxford University Press.
Full text: socialjudgments.com/docs/AME%20CHAPTER.POSTING.pdf

Uhlmann, E.L., Poehlman, T.A., & Bargh, J.A. (2008). Implicit theism. In R. Sorrentino & S. Yamaguchi (Eds.) *Handbook of Motivation and Cognition Across Cultures*. (pp. 71-94). St. Louis, MO: Elsevier/Academic Press.
Full text: http://socialjudgments.com/docs/Final%20Theism%20Chapter.pdf

Uhlmann, E.L., & Sanchez-Burks, J. (2014). The implicit legacy of American Protestantism. *Journal of Cross-Cultural Psychology, 45,* 991-1005.
Full text: socialjudgments.com/docs/Uhlmann.SanchezBurks.LegacyPaper.JCCP.pdf

[Page break here]

Format of predictions

We will ask you to make specific predictions about the primary effect size for each original effect replicated in each targeted population, and also for your forecasts regarding potential moderators of work morality effects. We will ask you about the expected effect sizes in terms of Cohen's d (Cohen, 1988; Sawilowsky, 2009). For more on Cohen's d please see this link: https://en.wikipedia.org/wiki/Effect_size#Cohen.27s_d

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

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

Incentives for accuracy

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

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

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

Please click the “forward” button to begin reviewing each study targeted for replication and provide your forecasts. You will make predictions about 5 work morality effects across up to 6 populations and 4 moderators, for a total of just under 50 predictions. Data for effects #1-4 were collected together (i.e., the studies were packaged together in the same survey), whereas effect #5 was a separate data collection.

Please note:

-- Your answers are saved in real time, so you can complete the survey in more than one session. To do this simply click on the survey link: the survey will automatically continue where you stopped at the end of your previous session

-- The "back button" on the bottom right allows you to go back and update the answers that you submitted previously

-- Please complete this survey on a sufficiently large screen

-- Please do not clear cookies or browsing history of your browser, especially if you are planning to complete the survey in multiple sittings

-- Please do not complete the survey in private/incognito mode on your browser, as your progress will not be saved then

[Page break here. Do not include page breaks unless otherwise indicated. Of particular importance, descriptions of the effect should be on the same page as the predictions regarding the effect]

EFFECT #1 OF 5: NEEDLESS WORK EFFECT

Conceptual description of effect: A person with a low-paying job who continues to work after winning the lottery is perceived as a morally good person.

Design: 2 (target works vs. retires) x 2 (target age: 23 years or 46 years) between-subjects design.

Sample study materials: Sarah is a [Young condition: 23 year old; Older condition: 46 year old] woman from Milwaukee. Sarah has worked for three years at the local post office where she is loved by her co-workers. Each week for the last 3 years she has played the same numbers in the state lottery. Last year she won 10 million dollars. [Works condition: After using $12,000 to pay bills and debts, she decided what she really wanted was to stay working at the post office even though she doesn't need the money anymore.] [Retires condition: Ever since winning, she's taken it easy at home and ordered a lot of take-out food and at 23 considers herself "retired".]
Is Sarah a good person?

Very Bad						Very Good
1	2	3	4	5	6	7

Key statistical comparison: Main effect of whether the target works vs. retires (regardless of target age) with ratings of moral goodness as the dependent measure. The needless work effect is supported if a lottery winner who continues to work receives higher evaluations than a lottery winner who retires.

Methods and planned analyses for replication:
https://www.dropbox.com/s/q482uzxdotwbv18/Methods%20and%20Planned%20Analyses.Culture%20and%20Work%20Replications.docx?dl=0

Complete study materials for replication:
https://www.dropbox.com/s/3arydioj11rvr2y/Study%20Materials.Culture%20and%20Work%20Replications.docx?dl=0

Predictions about the replication effect size for each population:

We will ask for your effect size prediction separately for each replication sample. In each case, we ask about the effect size in terms of Cohen's d. In the social sciences, a Cohen's d of 0.20 is considered to be a small effect, 0.50 is considered to be a medium effect, and 0.80 is considered to be a large effect. Please put a negative symbol (-) in front of your predicted effect size if you expect it to be in the opposite direction from the original hypothesis.

1. What do you predict will be the effect size for the needless work effect in the United States as a whole? The replication sampled 1036 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

2. What do you predict will be the effect size for the needless work effect in the New England states of the U.S.? In other words, the states in U.S. census district 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. The replication sampled 1012 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

3. What do you predict will be the effect size for the needless work effect in the non-New-England states of the U.S.? In other words, the states in U.S. census districts 2-9. The replication sampled 991 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

4. What do you predict will be the effect size for the needless work effect in Australia? The replication sampled 1011 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

5. What do you predict will be the effect size for the needless work effect in the United Kingdom? The replication sampled 960 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

6. What do you predict will be the effect size for the needless work effect in India? The replication sampled 1000 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

Predictions about moderators of the needless work effect:

Below, please estimate the effect size associated with each candidate moderator, aggregating across all of the data collection locations included in the present project.

7. Aggregating across all the replication sites, do you think Protestants will exhibit the needless work effect more than non-Protestants? Here we ask about the effect size in terms of Cohen's d. The relevant demographic item reads: *"My religion is: Protestant, Catholic, Islam, Judaism, Buddhism, Atheist, Agnostic, Other."* Please put a negative symbol (-) in front of your predicted effect size if you expect Protestants to exhibit the needless work effect *less* than non-Protestants. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

8. Aggregating across all the replication sites, do you think religious participants will exhibit the needless work effect more than non-religious participants? Here we ask about the effect size in terms of Cohen's d. A representative item from the DUREL religiosity scale used in this research is *"My religious beliefs are what really lie behind my whole approach to life."* Please put a negative symbol (-) in front of your predicted effect size if you expect religious persons to exhibit the needless work effect *less* than non-religious persons. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

9. Aggregating across all the replication sites, do you think participants who endorse the Protestant Work Ethic (PWE) will exhibit the needless work effect more than participants who do not endorse the PWE? Here we ask about the effect size in terms of Cohen's d. A representative PWE item is *"Most people who don't succeed in life are just plain lazy."* Please put a negative symbol (-) in front of your predicted effect size if you expect high-PWE individuals to exhibit the needless work effect *less* than low-PWE individuals. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

10. Aggregating across all the replication sites, do you think formally educated participants will exhibit the needless work effect more than less-educated participants? Here we ask about the effect size in terms of Cohen's d. The education item read as follows: *"My educational level is: Some high school/secondary school, High school degree/completed secondary school, Some university, University degree, Some graduate/postgraduate education, Graduate/postgraduate degree (e.g., doctoral degree)."* Please put a negative symbol (-) in front of your predicted effect size if you expect highly educated individuals to exhibit the needless work effect *less* than individuals who do not have as much advanced education. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

[Page break here]

EFFECT #2 OF 5: TARGET AGE EFFECT

Conceptual description of effect: Praise for needless work effect is greater when the target is young rather than older.

Design: 2 (target works vs. retires) x 2 (target age: 23 years or 46 years) between-subjects design.
Note: This effect is tested within the same research design and data collection as the needless work effect (Effect #1 of 5), but comparing different conditions.

Sample study materials: Sarah is a [Young condition: 23 year old] [Older condition: 46 year old] woman from Milwaukee. Sarah has worked for three years at the local post office where she is loved by her co-workers. Each week for the last 3 years she has played the same numbers in the state lottery. Last year she won 10 million dollars. [Works condition: After using $12,000 to pay bills and debts, she decided what she really wanted was to stay working at the post office even though she doesn't need the money anymore.] Is Sarah a good person?

Very Bad						Very Good
1	2	3	4	5	6	7

Key statistical comparison: Selecting only the condition in which the target works, the effect of target age (23 years old vs. 46 years old) is tested, with moral goodness ratings as the dependent measure. The target age effect is supported if a 23 year old lottery winner who continues to work receives higher moral goodness ratings than a 46 year old lottery winner who continues to work.

Methods and planned analyses for replication:
https://www.dropbox.com/s/q482uzxdotwbv18/Methods%20and%20Planned%20Analyses.Culture%20and%20Work%20Replications.docx?dl=0

Complete study materials for replication:
https://www.dropbox.com/s/3arydioj11rvr2y/Study%20Materials.Culture%20and%20Work%20Replications.docx?dl=0

Predictions about the replication effect size for each population:

We will ask for your effect size prediction separately for each replication sample. In each case, we ask about the effect size in terms of Cohen's d. In the social sciences, a Cohen's d of 0.20 is considered to be a small effect, 0.50 is considered to be a medium effect, and 0.80 is considered to be a large effect. Please put a negative symbol (-) in front of your predicted effect size if you expect it to be in the opposite direction from the original hypothesis.

1. What do you predict will be the effect size for the target age effect in the United States as a whole? The replication sampled 523 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

2. What do you predict will be the effect size for the target age effect in the New England states of the U.S.? In other words, the states in U.S. census district 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. The replication sampled 520 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

3. What do you predict will be the effect size for the target age effect in the non-New-England states of the U.S.? In other words, the states in U.S. census districts 2-9. The replication sampled 523 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

4. What do you predict will be the effect size for the target age effect in Australia? The replication sampled 489 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

5. What do you predict will be the effect size for the target age effect in the United Kingdom? The replication sampled 514 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

6. What do you predict will be the effect size for the target age effect in India? The replication sampled 495 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

Predictions about moderators of the target age effect:

Below, please estimate the effect size associated with each candidate moderator, aggregating across all of the data collection locations included in the present project.

7. Aggregating across all the replication sites, do you think <u>Protestants</u> will exhibit the target age effect more than non-Protestants? Here we ask about the effect size in terms of Cohen's d. The relevant demographic item reads: *"My religion is: Protestant, Catholic, Islam, Judaism, Buddhism, Atheist, Agnostic, Other."* Please put a negative symbol (-) in front of your predicted effect size if you expect Protestants to exhibit the target age effect *less* than non-Protestants. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

8. Aggregating across all the replication sites, do you think <u>religious participants</u> will exhibit the target age effect more than non-religious participants? Here we ask about the effect size in terms of Cohen's d. A representative item from the DUREL religiosity scale used in this research is *"My religious beliefs are what really lie behind my whole approach to life."* Please put a negative symbol (-) in front of your predicted effect size if you expect religious persons to exhibit the target age effect *less* than non-religious persons. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

9. Aggregating across all the replication sites, do you think participants who <u>endorse the Protestant Work Ethic</u> (PWE) will exhibit the target age effect more than participants who do not endorse the PWE? Here we ask about the effect size in terms of Cohen's d. A representative PWE item is *"Most people who don't succeed in life are just plain lazy."* Please put a negative symbol (-) in front of your predicted effect size if you expect high-PWE individuals to exhibit the target age effect *less* than low-PWE individuals. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

10. Aggregating across all the replication sites, do you think formally <u>educated</u> participants will exhibit the target age effect more than less-educated participants? Here we ask about the effect size in terms of Cohen's d. The education item read as follows: *"My educational level is: Some high school/secondary school, High school degree/completed secondary school, Some university, University degree, Some graduate/postgraduate education, Graduate/postgraduate degree (e.g., doctoral degree)."* Please put a negative symbol (-) in front of your predicted effect size if you expect highly educated individuals to exhibit the target age effect *less* than individuals who do not have as much advanced education. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

[Page break here]

EFFECT #3 OF 5: INTUITIVE WORK MORALITY EFFECT

Conceptual description of effect: The tendency to morally praise needless work is greater in an intuitive mindset than in a deliberative mindset

Design: Within-subjects comparison of intuitive vs. deliberative preference for worker over retiree.

Sample study materials: John and Robert are two 23-year old friends who used to work together as potato peelers. Each week for 3 years they bought lotto tickets together. Last year they won 10 million dollars. Robert decided right away to never work another day. Robert quit his job and now spends all day at home watching TV and at 23 considers himself "retired." John decided what he really wanted was to stay working as a potato peeler even though he didn't need the money anymore. John feels that an honest day's work is its own reward.

My most rational, objective judgment is that:

Robert is a much better person than John						John is a much better person than Robert
1	2	3	4	5	6	7

My intuitive, gut feeling is that:

Robert is a much better person than John						John is a much better person than Robert
1	2	3	4	5	6	7

Key statistical comparison: Within-subjects comparison between responses on the intuitive and deliberative preferences items. The intuitive work morality effect is supported if mean responses are higher on the intuitive item than on the deliberative item, reflecting greater intuitive than deliberative approval for a lottery winner who continues to work.

Methods and planned analyses for replication:
https://www.dropbox.com/s/q482uzxdotwbv18/Methods%20and%20Planned%20Analyses.Culture%20and%20Work%20Replications.docx?dl=0

Complete study materials for replication:
https://www.dropbox.com/s/3arydioj11rvr2y/Study%20Materials.Culture%20and%20Work%20Replications.docx?dl=0

Predictions about the replication effect size for each population:

We will ask for your effect size prediction separately for each replication sample. In each case, we ask about the effect size in terms of Cohen's d. In the social sciences, a Cohen's d of 0.20 is considered to be a small effect, 0.50 is considered to be a medium effect, and 0.80 is considered to be a large effect. Please put a negative symbol (-) in front of your predicted effect size if you expect it to be in the opposite direction from the original hypothesis.

1. What do you predict will be the effect size for the intuitive work morality effect in the United States as a whole? The replication obtained 2072 responses from 1036 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

2. What do you predict will be the effect size for the intuitive work morality effect in the New England states of the U.S.? In other words, the states in U.S. census district 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. The replication obtained 2024 responses from 1012 participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

3. What do you predict will be the effect size for the intuitive work morality effect in the non-New-England states of the U.S.? In other words, the states in U.S. census districts 2-9. The replication obtained 1982 responses from 991 participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

4. What do you predict will be the effect size for the intuitive work morality effect in Australia? The replication obtained 2022 responses from 1011 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

5. What do you predict will be the effect size for the intuitive work morality effect in the United Kingdom? The replication obtained 1920 responses from 960 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

6. What do you predict will be the effect size for the intuitive work morality effect in India? The replication obtained 2000 responses from 1000 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

Predictions about moderators of the intuitive work morality effect:

Below, please estimate the effect size associated with each candidate moderator, aggregating across all of the data collection locations included in the present project.

7. Aggregating across all the replication sites, do you think Protestants will exhibit the intuitive work morality effect more than non-Protestants? Here we ask about the effect size in terms of Cohen's d. The relevant demographic item reads: *"My religion is: Protestant, Catholic, Islam, Judaism, Buddhism, Atheist, Agnostic, Other."* Please put a negative symbol (-) in front of your predicted effect size if you expect Protestants to exhibit the intuitive work morality effect *less* than non-Protestants. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

8. Aggregating across all the replication sites, do you think religious participants will exhibit the intuitive work morality effect more than non-religious participants? Here we ask about the effect size in terms of Cohen's d. A representative item from the DUREL religiosity scale used in this research is *"My religious beliefs are what really lie behind my whole approach to life."* Please put a negative symbol (-) in front of your predicted effect size if you expect religious persons to exhibit the intuitive work morality effect *less* than non-religious persons. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

9. Aggregating across all the replication sites, do you think participants who endorse the Protestant Work Ethic (PWE) will exhibit the intuitive work morality effect more than participants who do not endorse the PWE? Here we ask about the effect size in terms of Cohen's d. A representative PWE item is *"Most people who don't succeed in life are just plain lazy."* Please put a negative symbol (-) in front of your predicted effect size if you expect high-PWE individuals to exhibit the intuitive work morality effect *less* than low-PWE individuals. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

10. Aggregating across all the replication sites, do you think formally educated participants will exhibit the intuitive work morality effect more than less-educated participants? Here we ask about the effect size in terms of Cohen's d. The education item read as follows: *"My educational level is: Some high school/secondary school, High school degree/completed secondary school, Some university, University degree, Some graduate/postgraduate education, Graduate/postgraduate degree (e.g., doctoral degree)."* Please put a negative symbol (-) in front of your predicted effect size if you expect highly educated individuals to exhibit the intuitive work morality effect *less* than individuals who do not have as much advanced education. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

[Page break here]

EFFECT #4 OF 5: TACIT INFERENCES EFFECT

Conceptual description of effect: Target persons who fail to uphold traditional work morality are misremembered as violating traditional sex morality, and vice versa.

Design: 2-celled between-subjects design manipulating whether targets uphold or violate traditional moral values.

Sample study materials: Below is one example scenario.

[Violates work morality: After seven years in college, Julia graduated with a very low grade point average. She has been unemployed for the last four years and her parents are supporting her financially. Julia is not making any effort to find a job and spends a lot of time watching television.]

[Upholds work morality: After only three years in college, Julia graduated with honors. She has held an excellent job for the last four years and is financially independent. Julia is working very hard at her job and spends hardly any time watching television.]

Last week, Julia was invited to a party at a guy's house not too far from her own. She ended up staying at the guy's house that night.

True or False: Julia slept with the host of last week's party.

Key statistical comparison: Comparison of false memories regarding individuals who uphold vs. violate traditional morality across two between-subject conditions. Aggregating across the four scenarios, participants should misremember individuals who violated work morality as also having violated sexual morality, and vice versa.

Methods and planned analyses for replication:
https://www.dropbox.com/s/q482uzxdotwbv18/Methods%20and%20Planned%20Analyses.Culture%20and%20Work%20Replications.docx?dl=0

Complete study materials for replication:
https://www.dropbox.com/s/3arydioj11rvr2y/Study%20Materials.Culture%20and%20Work%20Replications.docx?dl=0

Predictions about the replication effect size for each population:

We will ask for your effect size prediction separately for each replication sample. In each case, we ask about the effect size in terms of Cohen's d. In the social sciences, a Cohen's d of 0.20 is considered to be a small effect, 0.50 is considered to be a medium effect, and 0.80 is considered to be a large effect. Please put a negative symbol (-) in front of your predicted effect size if you expect it to be in the opposite direction from the original hypothesis.

1. What do you predict will be the effect size for the tacit inferences effect in the United States as a whole? The replication sampled 1036 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

2. What do you predict will be the effect size for the tacit inferences effect in the New England states of the U.S.? In other words, the states in U.S. census district 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. The replication sampled 1012 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

3. What do you predict will be the effect size for the tacit inferences effect in the non-New-England states of the U.S.? In other words, the states in U.S. census districts 2-9. The replication sampled 1015 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

4. What do you predict will be the effect size for the tacit inferences effect in <u>Australia</u>? The replication sampled 1011 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

5. What do you predict will be the effect size for the tacit inferences effect in <u>the United Kingdom</u>? The replication sampled 960 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

6. What do you predict will be the effect size for the tacit inferences effect in <u>India</u>? The replication sampled 1000 participants from this country using Amazon Mechanical Turk. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

Predictions about moderators of the tacit inferences effect:

Below, please estimate the effect size associated with each candidate moderator, aggregating across all of the data collection locations included in the present project.

7. Aggregating across all the replication sites, do you think <u>Protestants</u> will exhibit the tacit inferences effect more than non-Protestants? Here we ask about the effect size in terms of Cohen's d. The relevant demographic item reads: *"My religion is: Protestant, Catholic, Islam, Judaism, Buddhism, Atheist, Agnostic, Other."* Please put a negative symbol (-) in front of your predicted effect size if you expect Protestants to exhibit the tacit inferences effect *less* than non-Protestants. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

8. Aggregating across all the replication sites, do you think <u>religious participants</u> will exhibit the tacit inferences effect more than non-religious participants? Here we ask about the effect size in terms of Cohen's d. A representative item from the DUREL religiosity scale used in this research is *"My religious beliefs are what really lie behind my whole approach to life."* Please put a negative symbol (-) in front of your predicted effect size if you expect religious persons to exhibit the tacit inferences effect *less* than non-religious persons. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

9. Aggregating across all the replication sites, do you think participants who <u>endorse the Protestant Work Ethic</u> (PWE) will exhibit the tacit inferences effect more than participants who do not endorse the PWE? Here we ask about the effect size in terms of Cohen's d. A representative PWE item is *"Most people who don't succeed in life are just plain lazy."* Please put a negative symbol (-) in front of your predicted effect size if you expect high-PWE individuals to exhibit the tacit inferences effect *less* than low-PWE individuals. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

10. Aggregating across all the replication sites, do you think formally <u>educated</u> participants will exhibit the tacit inferences effect more than less-educated participants? Here we ask about the effect size in terms of Cohen's d. The education item read as follows:

"My educational level is: Some high school/secondary school, High school degree/completed secondary school, Some university, University degree, Some graduate/postgraduate education, Graduate/postgraduate degree (e.g., doctoral degree)." Please put a negative symbol (-) in front of your predicted effect size if you expect highly educated individuals to exhibit the tacit inferences effect *less* than individuals who do not have as much advanced education. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

[Page break here]

EFFECT #5 OF 5: SALVATION PRIMES AND WORK BEHAVIOR

Conceptual description of effect: Activating religious, as opposed to neutral concepts, improves performance on a subsequent work task (solving anagrams).

Design: 2-celled between-subjects design, religion prime vs. neutral primes

Sample study materials:

Salvation prime condition: Participants unscrambled sentences like "her them check salvation for"

Neutral prime condition: Participants unscrambled sentences like "the brown clown chair is"

Dependent measure: Please complete as many of the anagrams as you can. To make an anagram, use the letters in the original word to make a new word. Anagrams: BIMODAL, IGNEOUS, ANSWER, CURRIED.

Key statistical comparison: Comparison of anagram performance across two between-subjects conditions: religion prime and neutral prime. The salvation prime effect is supported if participants in the religion prime condition complete more anagrams than participants in the neutral prime condition.

Methods and planned analyses for replication:
https://www.dropbox.com/s/q482uzxdotwbv18/Methods%20and%20Planned%20Analyses.Culture%20and%20Work%20Replications.docx?dl=0

Complete study materials for replication:
https://www.dropbox.com/s/3arydioj11rvr2y/Study%20Materials.Culture%20and%20Work%20Replications.docx?dl=0

Predictions about the replication effect size for each population:

We will ask for your effect size prediction separately for each replication sample. In each case, we ask about the effect size in terms of Cohen's d. In the social sciences, a Cohen's d of 0.20 is considered to be a small effect, 0.50 is considered to be a medium effect, and 0.80 is considered

to be a large effect. Please put a negative symbol (-) in front of your predicted effect size if you expect it to be in the opposite direction from the original hypothesis.

1. What do you predict will be the effect size for the salvation primes in the New England states of the U.S.? In other words, the states in U.S. census district 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. The replication sampled 271 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

2. What do you predict will be the effect size for the salvation primes in the non-New-England states of the U.S.? In other words, the states in U.S. census districts 2-9. The replication sampled 245 adult participants from this region using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

3. What do you predict will be the effect size for the salvation primes in Australia? The replication sampled 300 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

4. What do you predict will be the effect size for the salvation primes in the United Kingdom? The replication sampled 314 participants from this country using the survey firm PureProfile. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

Predictions about moderators of the salvation prime effect:

Below, please estimate the effect size associated with each candidate moderator, aggregating across all of the data collection locations included in the present project.

5. Aggregating across all the replication sites, do you think Protestants will exhibit the salvation prime effect more than non-Protestants? Here we ask about the effect size in terms of Cohen's d. The relevant demographic item reads: *"My religion is: Protestant, Catholic, Islam, Judaism, Buddhism, Atheist, Agnostic, Other."* Please put a negative symbol (-) in front of your predicted effect size if you expect Protestants to exhibit the salvation prime effect *less* than non-Protestants. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

6. Aggregating across all the replication sites, do you think religious participants will exhibit the salvation prime effect more than non-religious participants? Here we ask about the effect size in terms of Cohen's d. A representative item from the DUREL religiosity scale used in this research is *"My religious beliefs are what really lie behind my whole approach to life."* Please put a negative symbol (-) in front of your predicted effect size if you expect religious persons to exhibit the salvation prime effect *less* than non-religious persons. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

7. Aggregating across all the replication sites, do you think participants who endorse the Protestant Work Ethic (PWE) will exhibit the salvation prime effect more than

participants who do not endorse the PWE? Here we ask about the effect size in terms of Cohen's d. A representative PWE item is *"Most people who don't succeed in life are just plain lazy."* Please put a negative symbol (-) in front of your predicted effect size if you expect high-PWE individuals to exhibit the salvation prime effect *less* than low-PWE individuals. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

8. Aggregating across all the replication sites, do you think formally <u>educated</u> participants will exhibit the salvation prime effect more than less-educated participants? Here we ask about the effect size in terms of Cohen's d. The education item read as follows: *"My educational level is: Some high school/secondary school, High school degree/completed secondary school, Some university, University degree, Some graduate/postgraduate education, Graduate/postgraduate degree (e.g., doctoral degree)."* Please put a negative symbol (-) in front of your predicted effect size if you expect highly educated individuals to exhibit the salvation prime effect *less* than individuals who do not have as much advanced education. [Free response bounded between -3 and 3 with a pop-up message if the bound is exceeded].

[Page break here]

Demographic Questions

What is your age? [Free response]

What is your gender?
1= Male
2= Female
3= Other: [Free response text box]
4= Prefer not to disclose

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

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

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

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

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

What is your job rank? (please select one)

- ○ Research assistant (1)
- ○ Graduate student (2)
- ○ Postdoctoral researcher (3)
- ○ Assistant Professor (4)
- ○ Associate Professor (5)
- ○ Full Professor (6)
- ○ Other (please indicate) (7)

Other job rank, please indicate: [Free response]

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

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

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

- ○ Amazon US voucher (2)
- ○ Amazon UK voucher (3)
- ○ Amazon DE voucher (4)
- ○ Paypal account (1)

Consortium Co-authorship

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

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

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

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

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

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

Feedback

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

Supplement 6: Detailed Report of the Forecasting Results

Methodological details

Materials. Respondents (forecasters) to the forecasting survey were asked to each make a total of 48 predictions regarding five different work morality effects ('key effects') in terms of effect sizes (Cohen's d) and the direction of the effect. Twenty-eight predictions were regarding effect sizes in different populations and 20 predictions regarding moderator effects. Effect sizes were bounded between -3 and 3. Forecasters were also asked to answer a set of demographic questions including their PhD year and job rank.

Forecasters could access all the relevant study materials. These included detailed information about the sample sizes, sample characteristics, study design and materials, including links to the original articles and the complete study materials and pre-analysis plans for the replication.

Recruiting forecasters. As in our other forecasting projects, we targeted researchers with training in judgment and decision making and/or social psychology research to participate in the forecasting survey. We excluded no respondents based on e.g. seniority or any other demographic characteristic. The link to a signup page for the forecasting project was posted on various academic websites, platforms, and Facebook pages aimed at researchers in psychology, judgment and decision making, and research methodology (e.g. Psych Map, Psych Methods Discussion Group, Judgment and Decision Making list). Colleagues with large followings on Twitter were also asked to post the link to the signup page. After having signed up, respondents received an individualized link to the forecasting survey, which allowed them to complete the survey in multiple sittings if they wished. Respondents received at least two reminders to finish the survey.

We incentivized participation in two ways. First, forecasters were offered coauthorship on the manuscript through a consortium credit ('Culture & Work Forecasting Collaboration'). Second, two forecasters were randomly selected and monetarily rewarded based on the accuracy of their forecasts using the following scoring rule:

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

where $\underline{Sq.Error}$ is the average of the squared errors for all the 48 forecasts of the 'Culture & Work Forecasting Study' made by the forecasters.

Initially, 429 individuals signed up for the forecasting survey, out of which 222 completed the survey. One hundred and fifty of the individuals who had initially signed up started but did not ultimately complete the survey, and 57 signed up but never started their forecasts. One forecaster

was removed from the sample for a technical issue that rendered her/his data unusable. Therefore, the final set of forecasters includes 221 respondents. This final sample size is comparable to past academic forecasting surveys (e.g., Landy et al., 2020; Tierney et al., 2020). In terms of gender, 38.9% of the forecasters reported that they were women, 59.7% that they were men, 0.005% chose ´other´ and 0.01% chose ´prefer not to tell.´ Forecasters reported 48 countries of birth and 38 countries of residence. Out of 221 forecasters, 72% of them were born in either Europe (100 forecasters) or North America (58 forecasters), and 80% of them currently reside in Europe (96 forecasters) or North America (80 forecasters). The most represented countries of birth were the United States with 50 forecasters, Germany with 20 and the United Kingdom with 10, while the most represented countries of residence were the United States with 72 forecasters, the United Kingdom with 20, and the Netherlands with 13. The average number of years since PhD was four (*SD* = 5.6). Given the nature of the recruitment method (social media), sample size (and thus statistical power) as well as the sample composition were not under our full control. We simply tried to recruit as many forecasters as we could within the pre-registered time frame for data collection.

Results

Hypothesis tests. The planned analysis is reported in our pre-analysis plan on https://osf.io/7uhcg/ and in Supplement 4. We follow the pre-analysis plan unless otherwise specified.

Our primary hypothesis 1 was that there would be a positive association between the predictions (beliefs) of the forecasters and the observed effect sizes. As expected, the individual-level regression and t-test show a positive association between the predictions of the forecasters and the observed replication effect sizes, $\beta_1 = 0.157$, $p = 0.0008$. See Table S6-1 for the individual-level regression estimates.

Table S6-1. Association between forecasted and observed effect sizes.

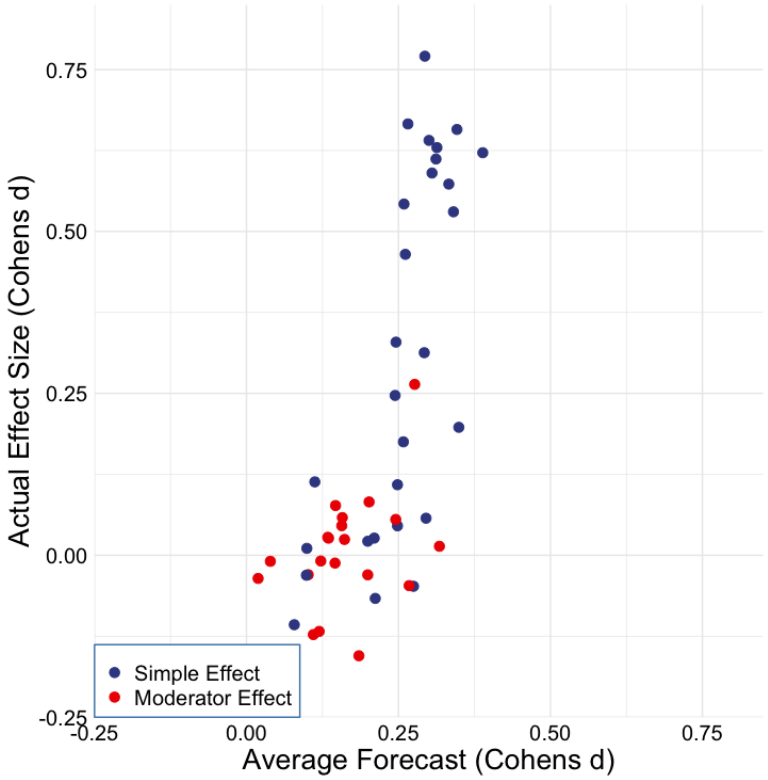
	Dependent variable:
	Realized effect size
Forecasted effect size	0.157**
	(0.045)
Observations	10608
R^2	0.037

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at individual level.

As a robustness test, we estimate hypothesis 1 separately for the two sets of predictions (simple effects and moderator effects). Focusing on simple effects only, there is a positive association (β = 0.164, p = 0.002). For the moderators alone, the association between predictions and effect sizes is significant using the traditional p-value cutoff of .05, but not the stricter .005 significance threshold proposed by Benjamin et al. (2018) for which it represents suggestive evidence (β = 0.019, p = 0.04).

In another robustness test, we estimate the Pearson correlation between the mean predicted effect size of each of the 48 effects replicated and the observed effect sizes. Figure S6-1 displays the correlation (r = 0.704, p < 0.0001) between the average predicted effect sizes and the observed effect size. We also estimate the Pearson correlation separately for the predictions regarding simple effects (r = 0.688, p < 0.0001) and the predictions regarding the moderator effects (r = 0.375, p = 0.104). The correlations for all effects combined and the simple effects separately are large and significant, but the correlation for moderator effects separately is not found to be statistically significant. This suggests that forecasters are for the most part able to anticipate the realized effect sizes, but their accuracy is not perfect. Further research is needed to establish whether or not forecasters are able to accurately predict the moderators of replication effect sizes.

Figure S6-1. Actual effect size vs average forecast (Cohen's d). Correlation between forecasted and actual effects for both simple and moderator effects (differentiated by the colors blue and red).



Our primary hypothesis 2 was that forecasters would be able to predict simple effect sizes more accurately than moderator effect sizes. For this we compute the *accuracy* achieved in each prediction by each forecaster in terms of squared prediction error (Brier score). In the regression of the Brier score we find no evidence for a relationship between effect type and accuracy (see Table S6-2). The coefficient for the variable identifying the forecasts regarding moderator effects is $\beta = 0.008$, $p = 0.5221$. Thus, we cannot conclude that forecasters are significantly better at predicting simple effects than moderator effects.

Table S6-2: Forecasts of moderator effects relative to simple effects in terms of squared prediction error (Brier score).

	Dependent variable:
	Brier Score
Forecasts for moderator effects	0.008
	(0.013)
Observations	10608
R^2	0.0001

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at the individual level.

While our primary hypothesis 1 tests the correlation between forecasts and realized effect sizes, it does not take into account the absolute difference between them. Our secondary hypothesis was that forecasted effect sizes would not be statistically significantly different from the realized effect sizes. We compare the meta-analyzed effect size for each of the five key effects (by pooling the effect sizes across the different cultures and populations) to the mean at the individual level of the effect size of each key effect (across the different populations for each participant). We then implement a z-test comparing whether these are statistically significantly different. The results are summarized in Table S6-3.

Table S6-3: Summary of the differences between meta-analyzed effect sizes and forecasts (standard errors in parenthesis)

Effect	Meta-analyzed effect	Mean of the forecasts	Difference	P-value
Needless work main effect (works vs. retires)	0.652 (0.031)	0.323 (0.014)	0.329	<0.0001
Target age and needless work effect	0.032 (0.041)	0.246 (0.017)	0.214	<0.0001
Intuitive work morality effect	0.257 (0.082)	0.252 (0.015)	0.005	0.954
Tacit inferences effect	0.505 (0.068)	0.311 (0.017)	0.1939	0.0055
Salvation prime and work behavior	0.010 (0.844)	0.097 (0.012)	0.087	0.9181

For two key effect sizes out of five, the main effect of needless work (works vs. retires) and target age and needless work effect, the mean of the forecasts and the meta analyzed effects are statistically significantly different from each other at the .005 level (Benjamin et al., 2018), with forecasts underestimating the former effect and overestimating the latter one. For the tacit inferences effect forecasters significantly underestimate the effect using the traditional .05 significance criterion, but not the more conservative .005 criterion proposed by Benjamin et al. (2018). The z-tests for salvation prime and work behavior and the intuitive work morality effect fail to reject the null hypothesis that the means of the forecasts and the meta-analyzed effects are not statistically different.

Additional analyses. We prespecified further analyses we regard as exploratory given the number of statistical tests involved and lack of strong theoretical predictions. First, we test whether forecasters can predict experimental results across different populations with different degrees of accuracy. In a regression we have binary variables for the New England states in the USA (data collected via PureProfile), non-New England states in the USA (PureProfile), USA (MTurk), Australia (PureProfile), and India (MTurk) respectively, with United Kingdom being the baseline population. We do separate t-tests on the coefficients for these binary variables ($\beta 1$

to β5) and a set of Wald tests on whether these coefficients are pairwise statistically significantly different. As Table S6-4 shows, we find that accuracy varies statistically significantly across some locations compared to the United Kingdom baseline population (β_1 = -0.008, p =0.351; β_2 = -0.023, p = 0.188; β_3 = 0.083, p < 0.0001; β_4 = 0.016, p = 0.0003; β_5 = 0.042, p < 0.0001). The set of pairwise Wald tests summarized in Table S6-5 indicate that we cannot reject the null hypothesis that the coefficients are the same for two pairs of populations among these pairwise tests: New England states/non-New England states in the US and USA/India.

Table S6-4: Regression estimates of accuracy on country indicators.

	Dependent variable:
	Brier Score
USA – New England States (PureProfile)	-0.008
	(0.008)
USA – Non-New England States (PureProfile)	-0.023
	(0.017)
USA (MTurk)	0.083**
	(0.021)
Australia (PureProfile)	0.016**
	(0.004)
India (MTurk)	0.042*
	(0.009)
Observations	6188
R^2	0.009

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at individual level.

Table S6-5: P-values resulting from pairwise Wald tests on country coefficients shown in Table S6-4 being different from each other.

	USA – New England States (PureProfile)	USA – Non-New England States (PureProfile)	USA (MTurk)	Australia (PureProfile)	India (MTurk)
USA – New England States (PureProfile)	-	0.1704	<0.0001	0.0016	<0.0001
USA – Non-New England States (PureProfile)	-	-	0.0007	0.0100	<0.0001
USA (MTurk)	-	-	-	0.0026	0.1251
Australia (PureProfile)	-	-	-	-	0.0075
India (MTurk)	-	-	-	-	-

Finally, in an exploratory vein, we test whether the forecasters' years of academic experience (i.e., years since PhD) are related to higher accuracy. The results from the regression and the t-test on the seniority coefficient indicate that years since PhD is not statistically significant correlated with accuracy ($\beta_1 = 0.00024$, $p = 0.96$). As a robustness check for this exploratory hypothesis, we analyze the accuracy of predictions on simple effects and on moderator effects separately. We find a similar result for simple effects ($\beta_1 = 0.001$, $p = 0.724$) and moderator effects ($\beta_1 = -0.001$, $p = 0.843$). Also, as a robustness check, we use academic job rank as a different proxy of seniority. We find that none of the academic ranks has a statistically significant correlation with accuracy relative to the reference group, i.e. those who selected "other" as job rank (see Table S6-6).

Table S6-6: Regression estimating the effects of academic seniority on forecasting accuracy.

	Dependent variable: Brier Score		
	(1) Full Sample	(2) Simple effects	(3) Moderators effects
Full Professor	-0.360	-0.308	-0.432
	(0.234)	(0.225)	(0.248)
Associate Professor	-0.316	-0.250	-0.409
	(0.234)	(0.225)	(0.248)
Assistant Professor	-0.313	-0.266	-0.379
	(0.234)	(0.225)	(0.248)
Postdoctoral researcher	-0.325	-0.273	-0.398
	(0.234)	(0.225)	(0.248)
Graduate student	-0.240	-0.218	-0.271
	(0.242)	(0.231)	(0.262)
Research Assistant	-0.291	-0.266	-0.327
	(0.234)	(0.225)	(0.250)
Constant	0.408*	0.368*	0.466
	(0.234)	(0.225)	(0.248)
Observations	10680	6188	4420
R^2	0.026	0.021	0.034

Note: $*p < 0.05$; $**p < 0.005$. Standard errors clustered at individual level.

Deviation from the pre-analysis plan for the forecasting survey

Below we list three deviations from the pre-registered plan with regard to our forecasting analyses and descriptions of these results.

Testing forecasts about moderators separately by country. As a robustness check for the exploratory hypothesis 'Do participants predict experimental results across different populations with different degrees of accuracy?' (estimates presented in Table S6-4), we pre-specified that we would analyze the accuracy of predictions regarding simple effects and regarding moderator effects separately. However, we were mistaken in planning this as the test is in fact impossible. Based on the design of the forecasting survey, the predictions regarding moderators do not vary across countries, since the participants were asked about the effect sizes of the moderators *'aggregating across all the replication sites*.' Since we have no variation in the effect of moderators within different populations, we simply could not run this analysis.

Comparing significance levels of forecasted and replicated effect sizes. We did not pre-register that we would compare whether the forecasted and realized effect sizes for each original effect targeted for replication would respectively differ from zero. However, it can be readily inferred from Table S6-3 where we report the effect sizes and their associated standard errors. It is clear from Table S6-3 that forecasters predicted that all five key effects would be observed (the mean of the forecasts is statistically significantly higher than zero, $p < 0.005$, for all the five key effects). This differs from the realized effect sizes where there was statistically significant support for three of the five key effects: the needles work main effect (works vs. retires comparison), the intuitive work morality effect, and the tacit inferences effect. In contrast, the null hypothesis of no observed effect could not be rejected for the target age and needles work effect and the salvation prime and work behavior effect (see Table S6-3 for the realized effect sizes and their standard errors).

Splitting forecasted and realized effect sizes by sample. In Table S6-7 below, we report the forecasted and realized effect sizes separately for each sample. This is done for descriptive purposes, without any statistical tests for differences. Note that estimates for "All USA" and "India" are based on Amazon Mechanical Turk (MTurk) samples, whereas the subregions of the USA (New England U.S. states vs. other U.S. states), Australia, and the UK are PureProfile (PP) samples. Data for the salvation prime replication was not collected on MTurk, therefore those entries are blank.

Table S6-7: Forecasted and realized effect sizes separately for each major sample of participants.

		Needless work main effect (works vs. retires)	Target age and needless work effect	Intuitive work morality effect	Tacit inferences effect	Salvation primes and work behavior
New England U.S. States (PP)	Mean Forecast	0.389	0.295	0.292	0.340	0.099
	Actual Effect Size	0.622	0.057	0.313	0.530	0.011
Non New England U.S. States (PP)	Mean Forecast	0.333	0.248	0.244	0.312	0.098
	Actual Effect Size	0.573	0.109	0.247	0.612	-0.031
All USA (MTurk)	Mean Forecast	0.346	0.275	0.259	0.300	-
	Actual Effect Size	0.658	-0.048	0.542	0.641	-
Australia (PP)	Mean Forecast	0.293	0.210	0.246	0.261	0.079
	Actual Effect Size	0.771	0.027	0.329	0.465	-0.107
India (MTurk)	Mean Forecast	0.266	0.200	0.212	0.349	-
	Actual Effect Size	0.666	0.022	-0.067	0.198	-
UK (PP)	Mean Forecast	0.313	0.248	0.258	0.305	0.112
	Actual Effect Size	0.630	0.045	0.175	0.590	0.113

Supplement 7: Further analyses of the replication results

Below we report some additional analyses of the replication results that were not reported in the main manuscript, either due to space constraints or because they represent secondary or ancillary analyses.

Aggregated effect sizes and tests of heterogeneity

Below we aggregate the effect size estimates across all samples for each pre-registered key effect of interest. Further included are tests for heterogeneity using Cochran's *Q*, I^2, and *Tau* (Borenstein et al., 2009; Borenstein, Hedges, Higgins, & Rothstein, 2010; Borenstein, Higgins, Hedges, & Rothstein, 2017; Cochran, 1954; Higgins, Thompson, Deeks, & Altman, 2003).

Key effect	**Aggregated effect size**	**Cochran's *Q***	***I*²**	***Tau***
Target age and needless work	0.0335	$Q(df=6) = 2.0552$, $p = 0.9146$	0.00	0.00
Intuitive work morality	0.2513	$Q(df=6) = 172.7514$, $p < .0001$	97.7806	0.1780
Tacit inferences	0.4893	$Q(df=6) = 49.3908$, $p < .0001$	89.7965	0.1362
Salvation prime	0.0384	$Q(df=7) = 0.0142$, $p = 1.0000$	0.00	0.00

There is statistically significant and substantial cross-sample variability for the intuitive work morality and tacit inferences effects, but not for the salvation prime effect or target age and needless work effect. This follows the general pattern in replication initiatives, such that effects that replicate successfully are associated with cross-site heterogeneity, whereas null findings tend to fail to replicate consistently across populations (Olsson-Collentine, Wicherts, & van Assen, in press).

Further analyses of the "Sarah" lottery winner study

The analyses of this study design in the main manuscript focus on our key effect of interest, specifically the simple effect of target age within the works condition ("target age and needless work effect"). A secondary pre-registered effect of interest is the main effect of working vs. retiring on moral judgments of the target, also reported in the main manuscript. Below, we carry out further analyses of the complete study design, as per our pre-registered analysis plan (Supplement 3).

The three-way interaction between target age (23 vs. 46) x work status (works vs. retires) x culture (USA vs. other) did not emerge for either the MTurk sample, $F(1, 2033) = 0.035$, $p = 0.852$, $d = 0.008$, or PureProfile sample, $F(1, 4083) = 1.92$, $p = 0.166$, $d = 0.043$.

The three-way interaction between target age (23 vs. 46) x work status (works vs. retires) x region (New England vs. other) did not emerge for either the MTurk sample, $F(1, 2033) = 0.255$, $p = 0.613$, $d = -0.0225$, or PureProfile sample, $F(1, 4069.06) = 0.692$, $p = 0.405$, $d = -0.0261$.

The main effect of works vs. retires is statistically significant separately examining the USA MTurk sample, $F(1, 1033) = 111.649$, $p < 0.001$, $d = 0.6575$, USA PureProfile sample, $F(1, 2121.93) = 189.142$, $p < 0.001$, $d = 0.597$, India sample, $F(1, 996) = 110.456$, $p < 0.001$, $d = 0.666$, Australia sample, $F(1, 1007) = 149.507$, $p < 0.001$, $d = 0.771$, and UK sample, $F(1, 956) = 94.727$, $p < 0.001$, $d = 0.630$. In each country examined, the worker is consistently preferred over the retiree, supporting the pre-registered predictions of the General Moralization of Work account.

Further analyses of the rational-intuitive mindset study

The analyses of this study in the main manuscript focus on our key effect of interest, specifically the difference between intuitive and rational evaluations of targets who retire vs. work after winning the lottery ("intuitive mindset effect"). A secondary pre-registered effect of interest is the overall preference for the target who works vs. retires. This is tested below by comparing preferences on the 1-7 scale to the neutral scale midpoint of 4, with scores above 4 indicating positive reactions to needless work.

Averaging together the intuitive and rational item and testing against the scale midpoint of 4, the worker is preferred over the retiree across all samples (see Figure 1 of the main manuscript). A series of one sample t-tests indicated that scores on the 2-item composite measure are significantly above the scale midpoint separately examining the USA MTurk sample, $t(1022) = 29.72$, $p < 0.001$, $d = .93$, USA PureProfile sample, $t(2121)= 41.75$, $p < 0.001$, $d = .91$, India sample, $t(996) = 46.95$, $p < 0.001$, $d = 1.46$, Australia sample, $t(1008) = 27.41$, $p < 0.001$, $d = .86$, and UK sample, $t(957) = 24.35$, $p < 0.001$, $d = .79$.

Examining the rational mindset item alone, the worker is preferred over the retiree across all samples. Scores on the rational mindset item are significantly above the scale midpoint separately examining the USA MTurk sample, $t(1022) = 23.77$, $p < 0.001$, $d = .74.$, USA PureProfile sample, $t(2117) = 35.53$, $p < 0.001$, $d = .77$, India sample, $t(996) = 43.48$, $p < 0.001$, $d = 1.37$, Australia sample, $t(1008) = 23.00$, $p < 0.001$, $d = .72$, and UK sample, $t(954) = 21.42$, $p < 0.001$, $d = .69$.

Examining the intuitive mindset item alone, the worker is again preferred over the retiree across all samples. Scores on the intuitive mindset item are significantly above the scale midpoint separately examining the USA MTurk sample, $t(1021) = 29.92$, $p < 0.001$, $d = .93$, USA PureProfile sample, $t(2118) = 40.69$, $p < 0.001$, $d = .69$, India sample, $t(996) = 41.86$, $p < .0001$, $d = 1.33$, Australia sample, $t(1008) = 27.10$, $p < 0.001$, $d = .85$ and UK sample, $t(957) = 23.13$, $p < 0.001$, $d = .74$.

Contrary to the predictions of the Implicit Puritanism account, Americans do not score higher on the intuitive preference item than members of the comparison cultures (see Figure 1 of the main

manuscript). Indeed, the highest mean is observed in the India MTurk sample (M = 5.78, SD = 1.35), which is further significantly above that in the USA MTurk sample (M = 5.18, SD = 1.26), $t(2001) = 10.39$, $p < 0.001$, $d = .46$. Indians were more likely than Americans to intuitively prefer a lottery winner who continues to work at a low-wage job.

In sum, for each country sampled and for both the logical and intuitive mindset item, the worker is preferred over the retiree. This supports the pre-registered predictions of the General Moralization of Work account. Further, cross-national differences in intuitive evaluations sharply contradict the Implicit Puritanism account, with Indians intuitively moralizing work significantly more than Americans. That Indian participants exhibited the highest means on both the logical and intuitive mindset items also undermines the conclusion that the cross-national replications of the intuitive mindset effect support the Self-Expression account. Recall that the intuitive mindset effect is based on the difference score between intuitive and logical evaluations, which was indeed sharply reduced in India relative to the other samples, as demonstrated by the analyses in the main manuscript. However, the unexpected reason for this appears to be that Indian participants strongly moralize work at both an intuitive and logical level, and hence do not exhibit any mindset differences. As such, the Self-Expression account, which posits that members of survival cultures (e.g., India) intuitively moralize work less than members of self-expression cultures (e.g., USA, UK, Australia), is not supported by a more fine-grained examination of the results.

References for Supplement 7 (not cited in main manuscript)

Borenstein, M., et al. (2009). *Introduction to meta-analysis*. Chichester (UK): Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97-111.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5-18.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101-129.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal, 327*(7414), 557.

Supplement 8: Pre-Registered Plan for Bayesian Multiverse Analysis

Preregistration: Bayesian Multiverse Analysis for the Culture & Work Morality Project

Suzanne Hoogeveen & Julia Haaf

8/21/2020

Overview

We outline a Bayesian multiverse analysis for the 6 key effects in the culture and work morality project: the four primary effects that are the main focus of the article, as well as two preregistered effects of further theoretical interest. For each effect, we will construct various hierarchical models that reflect the predictions from the proposed theories. The evidence for each of these different theories will be quantified by Bayes factor model comparison, following the approach by Haaf & Rouder (2017) and Rouder, Haaf, Davis-Stober, & Hilgard (2019). In addition, we will adopt a multiverse approach in which we assess evidence for various a priori specified alternative analysis paths (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

Six key effects

1. **[Moralization of work vs. retirement 1]:** In a between-subjects design, is needless work, relative to retirement, specifically and exclusively praised by (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
2. **[Target age and needless work]:** Is needless work by a young rather than old target specifically and exclusively praised by (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
3. **[Moralization of work vs. retirement 2]:** In a within-subjects design, is a person who continues to work needlessly preferred over a person who retires, specifically and exclusively by (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
4. **[Intuitive moralization of work]:** Is needless work more intuitively rather than rationally specifically and exclusively praised by (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work

ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.

5. **[Link between work & sex morality]** Are inferences about work and sex morality specifically and exclusively implicitly linked by (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
6. **[Link between salvation and work]:** Does subtle activation of the concept of divine salvation induce enhanced work performance specifically and exclusively for (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) individuals who endorse the Protestant work ethic, or (6) everyone.

We will construct hierarchical Bayesian regression models that reflect the predictions from the 9 theories + the false positives model, for each of the 6 key effects. For each effect, the predictive adequacy of these models as well as the unconstrained model will be compared using Bayes factor model comparison. In addition, we conduct a multiverse analysis varying operationalizations of cultural groupings, variables, data exclusions, and analytic approaches.

Model Adequacy

With this many models and analyses in the multiverse it is common that variables overlap to a high degree, or that sample sizes are too low after data exclusions to learn anything. We therefore propose two criteria for the inclusion of variables and alternative operationalisations/exclusions in the multiverse analysis.

7. Multicollinearity: we suggest to use a criterion of $r < .7$, where two predictors cannot be both included in the models if they are correlated more than .7. For instance, if $r > .7$ for age and religiosity, the variable age will be removed from the models. We will try to include the variable that is most theoretically relevant.
8. Correlation between alternative operationalizations and exclusions: we suggest to use a correlation check for variables that use an alternative operationalisation, change assignment, score or change exclusion criteria. If $r > .9$ we will not analyze the level as a separate path in the multiverse. To assess whether we want to do an exclusion or no exclusion at all (which means we cannot calculate a correlation) we suggest to only analyze the level if more than 5% of the sample are excluded. We assume that otherwise the operationalisation differences will not have a meaningful effect on the results.

We will dichotomize the continuous predictors for *religiosity* and *Protestant work ethic (PWE)*, and categorize participants into ‘religious’ vs. ‘nonreligious’ and ‘endorse PWE’ vs. ‘reject PWE’. By using dichotomized variables we can both (a) more directly test the predictions of the relevant theories (e.g., effect for Protestants, no effect for non-Protestants), and (b) make a fairer comparison between the various models derived from theories that relate to categorical variables and continuous variables. For the religiosity measure, we will use the DUREL scale in the primary analyses, and include the single-item measure in the multiverse. We will split the scales as follows: The DUREL scale has two 6-point items and three 5-point, so the minimum score is 5, the maximum is 27. We will use a cutoff of 16, with 16 or higher is religious, and 15 or lower

is nonreligious. For the single item (7-point scale) 4 or higher is religious and 3 or lower is nonreligious. The PWE scale has 11 items on a 6-point scale. We will categorize a score of 39 and higher as endorsing PWE and 38 and lower as rejecting PWE.

The SES measure will be a composite score of (1) personal level of education, (2) parental level of education, and (3) yearly income (personal for online samples / household for university student samples). For the US online samples will we use the following criteria to select high SES participants: (1) and (2) should be a score of 3 or higher (some college), and (3) should be a score of 4 of higher (> USD 40,000 as USD 34,000 is the median US personal income). For the US labs samples: (1) and (2) should be a score of 3 or higher (some college), and (3) should be a score of 5 of higher (> USD 60,000 as USD 64,000 is the median US household income).For the UK: (1) and (2) should be a score of 3 or higher (some college), and (3) should be a score of 4 of higher (> GBP £29,561 as GBP 24,000 is the median UK personal income). For Australia: (1) and (2) should be a score of 3 or higher (some college), and (3) should be a score of 4 of higher (> AU dollars 53,454 as AU dollars 48,000 is the median Australian personal income). For India: (1) and (2) should be a score of 3 or higher (some college), and (3) should be a score of 2 of higher (> INR 672,900 as INR 150,000 is the median Indian personal income). For Ireland: (1) and (2) should be a score of 3 or higher (some college), and (3) should be a score of 5 of higher (> €50,540 as €45,256 is the median Irish household income). Note that we use personal income for the online samples, and household income for the lab samples, as we assume that these household income are more representative for students' SES than their current personal incomes.[1]

Multiverse Analysis

The multiverse approach allows us to assess the evidence across multiple potential sets of sample compositions, without a priori committing to any particular set of exclusion criteria or variable specifications (Steegen et al., 2016). Nevertheless, for the primary analysis, we propose to include everyone who completed the relevant measures. Although for individual variables there are often exclusion criteria that might be preferred by researchers in the field for theoretical reasons (e.g., excluding subjects with less English experience or who fail to fully complete the scrambled sentences in the priming task), in practice, when many simultaneous exclusions are made the sample size can drop dramatically. In this case, effect size estimates may be inaccurate, it may not be possible to make any inference because of low sensitivity, and concerns are raised about differential attrition across conditions potentially confounding the results. Thus there is a principled case to be made on behalf of an intent-to-treat approach, in which few to no observations or participants are excluded (Gupta, 2011; McCoy, 2017). We consider the intent-

[1] In case the scores are very skewed using these specified criteria, we may reconsider the categorization and choose different criteria for the split between high/low on these measures. However, we prefer to stick to these theoretically-driven decisions rather than using data-driven criteria such as median splits, as we believe the categorization of an individual's wealth/religiosity/Protestant work ethic should not be dependent on their relative position (i.e., based on the other people in the sample), but rather their absolute ratings, at least when possible given the measure. We believe the measures used here are sufficiently informative to use absolute scores. If this eventually renders the analysis impossible because there are too few people in either category, we will reassess and decide on a new theoretically motivated split criterion.

to-treat principle a conservative but valuable approach that can be one key part of a multiverse analysis.

Variables to include in the multiverse for all 6 key effects

Modeling-related dimensions

- Modeling random effects vs. common effects.

9. Use a common effect model
10. Use a varying effects model (random effects across regions)

While the random effects models are much more informative they are also more complex and it is more difficult to find evidence for them when the experimental effect is small. Since we don't know the data well enough and we don't know whether we will have the resolution for complex models we will use both simple common-effect models and random-effects model.

Variable operationalizations

- Religiosity.

11. Use the DUREL scale (validated scale not used in the original studies)
12. Use the single-item religiosity measure (direct replication of original approach)

Exclusion criteria

- Attention check.

13. Exclude no one based on the attention check.
14. Exclude participants who missed the instructional manipulation check (i.e., did not select "strongly disagree" on the item telling them to choose "strongly disagree")

- English fluency.

15. Include everyone regardless of years of English experience
16. Exclude participants with less than 5 years of English experience

- Cultural grouping.

17. Use current objective location of data collection
18. Use self-reported nation of birth

- Regional grouping.

19. Use current objective location of data collection
20. Use self-reported region "grew up in" (US only)

In total, there will be $2 \times 2 \times 2 \times 2 \times 2 \times 2 = 64$ conditions for the multiverse analysis of the first 5 key effects.

Priming-related dimensions (key effect 6)

For the salvation priming study (key effect 6), we additionally evaluate the following dimensions:

- Awareness measure.

At the end of the priming study, participants indicated on a 1-9 scale whether they thought they were influenced by the scrambling task.

21. Exclude no one based on the awareness item.
22. Select only participants who score a 4 or below on numeric awareness rating item (claim not to be influenced).

- Anagram dependent measure.

Participants were instructed to create words with 4 or more letters. Of interest conceptually is work effort in addition to the ability to follow instructions. However, 3-letters solutions may be considerably easier and hence result in more solved anagrams. We will therefore include the following in the multiverse analysis:

23. Any solution counts regardless of number of letters.
24. Only 4+ letter solutions.

- Scrambled sentence completion.

Some participants may not have completed all scrambled sentences, and hence be less strongly primed than participants who unscrambled all the sentences.

25. Exclude no one based on completing the priming task.
26. Select only participants who completed all scrambled sentences.

This results in a total of $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 512$ conditions for the multiverse analysis of the priming study.

Superselect sample and antiselect sample

In addition to the main analyses comparing the predictions from all theories, we will create sub-samples in which the effect would theoretically be most likely to emerge ("superselect samples") and comparison groups in which the effect would be least likely to emerge ("antiselect samples").

Combining a number of the theories, 1) Americans (currently located and born) 2) from New England 3) who are self-identified Protestants, 4) endorse the PWE, and 5) are religious should be most likely to exhibit the original effects. We can contrast this "select group" to a "mirror image" group of non-Americans who are not Protestant, reject the PWE, and are not religious.

For the first 5 key effects, we will compare the evidence for an effect for religious Protestant Americans from New England who endorse the PWE (superselect) to nonreligious, non-Protestant Indians who reject the PWE (antiselect). In case the selection results in fewer than 50 participants per condition in either of these groups, we will sequentially remove layers starting from 5), i.e., we will first remove the religiosity criterion, then PWE and so on.

For the salvation prime study we will only use the lab samples, as we assume that a priming effect might be stronger in the lab compared to online. Specifically, we will compare the presence of an effect for religious Protestant Americans from New England who endorse the PWE (superselect) to nonreligious, non-Protestant Irish participants who reject the PWE

(antiselect). Again, in case we have fewer than 50 participants per condition, we will remove layers starting from 5. In addition, we will conduct a separete "priming methodology superselect" analysis, focusing on factors that are believed to facilitate priming effects. Specifically: 1) score of 1-4 on the 1-9 awareness of influence scale, and 2) completion all sentence scrambles.

For the superselect and antiselect group analyses, we will simply conduct Bayesian t-tests for the presence of an experimental effect. We expect evidence for an effect in the superselect group and evidence against the effect in the antiselect group.

Further considerations

Prior settings

We think small effects in the predicted direction may still be meaningful. We therefore propose to use a scale of 0.25 for the overall effect in each of the six original findings targeted for replication. A scale of 0.25 assumes an size effect that is 25% of the sampling noise (standard deviation), which is generally considered a small effect. For the variation between regions/labs we use a setting of 60% of the overall size of the effect, which means a scale of 0.15.

Online vs. lab samples

In case the random intercepts and random slopes indicate systematic differences between the online and lab samples, we may investigate to what extent sampling method affects the priming effect (if present) by including a sampling method (online vs. lab) indicator. However, such differences should be accounted for in the multilevel structure of the models, and we will therefore not include the sampling method indicator in the primary models.

Additional unconstrained models

The specified unconstrained models include all predictors and parameters of each of the separate models. Based on the estimates of these huge models, we may assess more specific models that include a combination of predictors that appear to best explain the data. For instance, we may create an additive model that includes US vs. non-US and religiosity, but without SES and New England region. In addition, the unconstrained model captures the possibility that observed effects may be directionally opposite to the predictions of any of the competing theories.

Model Specification and Predictions

General Model Structure

We will use Bayesian hierarchical modeling with participants nested in regions/labs. For each key effect, we will first construct an unconstrained model that includes all individual parameters from the separate theories, which are free to vary in size and direction. We will then construct up to 10 additional models that incorporate the predictions from each theory (see below). Bayes factor model comparison will be used to compare the models and determine what theory best

predicts the empirical data. For key effects 1, 2, 5 and 6, the critical theoretical predictions relate to interaction effects between experimental condition and the moderator of interest (e.g., country, religiosity, region). Here, we include main effects of these moderators in all models. This way, we isolate the critical interaction effect from any main effect of the moderator and eliminate the possibility that the preference for any specific model is driven by a moderator main effect instead of the hypothesized interaction. The main effects included in these base models are: Protestantism, religiosity, PWE, and SES.[2]. Data for key effect 6 (salvation prime) is collected both online and in university laboratories. Since the online and lab samples vary greatly in the age composition, we include age as a common main effect to the models for the salvation prime study. Key effects 3 and 4 concern within-subjects designs that are modeled as intercept effects (key effect 3) or as a difference score (key effect 4). The theoretical predictions are main effects, rather than interactions and hence do not require the inclusion of additional predictors.

Key effect 1: Moralization of Work vs. Retirement 1

Effect of working vs. retiring on moral judgments in the "Sarah" lottery winner scenario (between subjects). Participants should give higher ratings for "Is Sarah a good person?" (1-7) in the condition where Sarah continues to work after winning the lottery compared to when Sarah retires after winning the lottery.

The base model for the moralization of work vs. retirement 1 effect is an unconstrained model that includes all parameters proposed by the 9 different theories. Let Y_{ijk} be the rating for the ith region, the jth participant, and the kth condition. Then

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + x_k\eta_i + u_{ik}\theta_{i1} + v_{ik}\theta_2 + l_{jk}\theta_{i3} + m_{jk}\theta_{i4} + c_{jk}\theta_{i5} + d_{jk}\theta_{i6} + t_{ik}\theta_{i7}, \sigma^2),$$

where,

- α_i is the baseline rating for ith region
- β_i is the Protestantism effect for ith region
- γ_i is the religiosity effect for ith region
- ζ_i is the protestant work ethic effect for ith region
- ω_i is the SES effect for ith region
- η_i is the general experimental effect for ith region
- θ_{i1} is the ith region's experimental effect for US participants
- θ_2 is the experimental effect for New England participants
- θ_{i3} is the ith region's experimental effect for Protestant participants
- θ_{i4} is the ith region's experimental effect for highly religious participants
- θ_{i5} is the ith region's experimental effect for participants who endorse PWE
- θ_{i6} is the ith region's experimental effect for high SES participants
- θ_{i7} is the ith region's experimental effect for participants from self-expression cultures

[2] SES is not included for key effect 6 (salvation priming), because there is no prediction regarding SES for this study.

The variable p_j is the indicator that is 0.5 if a participant is protestant and -0.5 otherwise. r_j is the indicator for the dichotomized religiosity score (low vs. high religiosity) for each person where 0.5 is high religiosity and -0.5 is low religiosity. w_j is the indicator for the dichotomized Protestant work ethic (PWE) scale, that is 0.5 if a participant endorses the PWE and -0.5 if a participant rejects the PWE. l_j is the indicator for the dichotomized SES measure, that is 0.5 for high SES participants and -0.5 for low SES participants.

Additionally, the variable $x_k = -0.5,0.5$ if $k = 1,2$ respectively, with $k = 1$ when condition is 'retires' and $k = 2$ when condition is 'continues to work'. For the specific predictors in each theory, we used contrast coding to allow for interaction effects in the absence of a main effect of experimental condition. That is:

- u_{ik} is the indicator that is 3 for US regions in the experimental condition, and -1 otherwise
- v_{ik} is the indicator that is 3 for New England regions in the experimental condition, and -1 otherwise
- l_{jk} is the indicator that is 3 for Protestant participants in the experimental condition, and -1 otherwise
- m_{jk} is the indicator that is 3 for highly religious participants in the experimental condition, and -1 otherwise
- c_{jk} is the indicator that is 3 for participants who endorse PWE in the experimental condition, and -1 otherwise
- d_{jk} is the indicator that is 3 for high SES participants in the experimental condition, and -1 otherwise
- t_{ik} is the indicator that is 3 for self-expression cultures regions (US, UK, Australia) in the experimental condition, and -1 otherwise

Theoretical Predictions:

27. False Positives: Sarah is rated equally positively when she continues to work or retires (i.e., no work vs. retires effect 1 present).

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i, \sigma^2),$$

2. Implicit Puritanism: works vs. retires effect 1 is present for participants from the US but not for participants from the UK, Australia, and India.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + u_{ik}\theta_1, \sigma^2),$$

with $\theta_1 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + u_{ik}\theta_{i1}, \sigma^2),$$

with all $\theta_{i1} > 0$

3. Regional Folkways: works vs. retires effect 1 is present for participants from New England but not for participants from other regions and nations.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + v_{ik}\theta_2, \sigma^2),$$

with $\theta_2 > 0$

4. Religious Differences A: works vs. retires effect 1 is present for Protestant participants but not for non-Protestant participants.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + l_{jk}\theta_3, \sigma^2),$$

with $\theta_3 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + l_{jk}\theta_{i3}, \sigma^2),$$

with all $\theta_{i3} > 0$

5. Religious Differences B: works vs. retires effect 1 is present for high religiosity participants but not for low religiosity participants.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + m_{jk}\theta_4, \sigma^2),$$

where $\theta_4 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + m_{jk}\theta_{i4}, \sigma^2),$$

where all $\theta_{i4} > 0$

6. Explicit American Exceptionalism: works vs. retires effect 1 is present for participants from the US but not for participants from the UK, Australia, and India and more so for US participants who endorse the PWE rather than reject the PWE. For this model, we add a new parameter and predictor that captures an experimental effect for US participants who endorse PWE; θ_{i8} = ith region's experimental effect for US participants who endorse PWE, o_{ijk} is the indicator that is 7 for Americans in the 'worker' condition who endorse PWE, and -1 otherwise.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + u_{ik}\theta_1 + o_{ijk}\theta_8, \sigma^2),$$

with $\theta_1 > 0$ and $\theta_8 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + u_{ik}\theta_{i1} + o_{ijk}\theta_{i8}, \sigma^2),$$

with all $\theta_{i1} > 0$ and $\theta_{i8} > 0$

7. Protestant Work Ethic: works vs. retires effect 1 is present for participants who endorse the PWE but not for participants who reject the PWE.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + c_{jk}\theta_5, \sigma^2),$$

where $\theta_5 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + c_{jk}\theta_{i5}, \sigma^2),$$

where all $\theta_{i5} > 0$

8. Generalized Moralization: works vs. retires effect 1 is present for all samples of participants (USA, UK, Australia, India).

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + x_k\eta_i, \sigma^2),$$

where $\eta > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + x_k\eta_i, \sigma^2),$$

where all $\eta_i > 0$

9. Social Class Differences: works vs. retires effect 1 is present for high SES participants but not for low SES participants.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + d_{jk}\theta_6, \sigma^2),$$

where $\theta_6 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + d_{jk}\theta_{i6}, \sigma^2),$$

where all $\theta_{i6} > 0$

10. Self Expression Values: works vs. retires effect 1 is present for participants from the US, UK, and Australia, but not for participants from India.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + t_{ik}\theta_7, \sigma^2),$$

where $\theta_7 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + t_{ik}\theta_{i7}, \sigma^2),$$

where all $\theta_{i7} > 0$

Key effect 2: Target Age and Needless Work Effect

Target age (young vs. old) effect, selecting the 'continues working' condition, on moral judgments in the "Sarah" lottery winner scenario. Targets who are young and continue to work are rated more positively than targets who are old and continue to work. Participants should give higher ratings for "Is Sarah a good person?" (1-7) in the condition where Sarah is young and continues to work than when Sarah is relatively older and continues to work after winning the lottery. Note: this effect only uses half of the original data, as we only focus on the 'continues to work' condition, not the 'retires' condition.

Theoretical Predictions:

Models are the same as those for key effect 1

28. False Positives: No effect of target age.
29. Implicit Puritanism: target age effect is present for participants from the US but not for participants from the UK, Australia, and India.
30. Regional Folkways: target age effect is present for participants from New England but not for participants from other regions.
31. Religious Differences A: target age effect is present for Protestant participants but not for non-Protestant participants.
32. Religious Differences B: target age effect is present for high religiosity participants but not for low religiosity participants.
33. Explicit American Exceptionalism: target age effect is present for participants from the US but not for participants from the UK, Australia, and India, and more so for US participants who endorse the PWE rather than reject the PWE.
34. Protestant Work Ethic: target age effect is present for participants who endorse the PWE but not for participants who reject the PWE.
35. Generalized Moralization: target age effect is present for all samples of participants (USA, UK, Australia, India).
36. Social Class Differences: target age effect is present for high SES participants but not for low SES participants.
37. Self Expression Values: target age effect is present for participants from the US, UK and Australia but not for participants from India.

Key effect 3: Moralization of Work vs. Retirement 1

Main effect of preference for a worker over a retiree in the potato peelers scenario (within subjects). In a choice on a 1-7 scale with 4 as the neutral midpoint, workers are preferred overall over retirees (averaged over the intuitive preference and rational preference items), as indicated by an average score > 4. We will transform the scale to range from -3 to 3 and test if the intercept is larger than 0.

$$Y_{ij} \sim N(\alpha_i + u_i\theta_{i1} + v_i\theta_2 + p_j\theta_{i3} + r_j\theta_{i4} + w_j\theta_{i5} + l_j\theta_{i6} + t_i\theta_{i7}, \sigma^2),$$

with:

- θ_{i1} is the ith region's main effect of US vs. non-US on preference.
- θ_2 is the main effect of New England vs. non-New England .
- θ_{i3} is the ith region's main effect of Protestant vs. non-Protestant.
- θ_{i4} is the ith region's main effect of being religious vs. non-religious.
- θ_{i5} is the ith region's main effect of endorsing PWE vs. rejecting PWE.
- θ_{i6} is the ith region's main effect of having high SES vs. low SES.
- θ_{i7} is the ith region's main effect of being from a self-expression culture (US, UK, Australia) vs. from a survival culture (India).

We will use the following indicators to test the predictions from each theory:

- u_i is the indicator that is 1 if region is in the US, and 0 otherwise
- v_i is the indicator that is 1 if the region is New England, and 0 otherwise
- p_j is the indicator that is 1 if the participant is Protestant, and 0 otherwise.
- r_j is the indicator that is 1 if the participant is religious, and 0 otherwise.
- w_j is the indicator that is 1 if the participants endorses the PWE, and 0 otherwise.
- l_j is the indicator that is 1 for high SES participants and 0 otherwise.
- t_i is the indicator that is 1 for self-expression cultures regions (US, UK, Australia), and 0 otherwise (India).

Theoretical Predictions:

38. False Positives: No preference for worker over retiree in the potato peeler scenario.

$$Y_{ij} \sim N(0, \sigma^2),$$

3. Implicit Puritanism: work status effect is present for participants from the US but not for participants from the UK, Australia, and India.

Common:

$$Y_{ij} \sim N(u_i\theta_1, \sigma^2),$$

with: $\theta_1 > 0$

Random:

$$Y_{ij} \sim N(u_i\theta_{i1}, \sigma^2),$$

with all: $\theta_{i1} > 0$

4. Regional Folkways: work status effect is present for participants from New England but not for participants from other regions.

Common:

$$Y_{ij} \sim N(v_i\theta_2, \sigma^2),$$

with: $\theta_2 > 0$

5. Religious Differences A: work status effect is present for Protestant participants but not for non-Protestant participants.

Common:

$$Y_{ij} \sim N(p_j\theta_3, \sigma^2),$$

with: $\theta_3 > 0$

Random:

$$Y_{ij} \sim N(p_j\theta_{i3}, \sigma^2),$$

with all: $\theta_{i3} > 0$

6. Religious Differences B: work status effect is present for high religiosity participants but not for low religiosity participants.

Common:

$$Y_{ij} \sim N(r_j\theta_4, \sigma^2),$$

with: $\theta_4 > 0$

Random:

$$Y_{ij} \sim N(r_j\theta_{i4}, \sigma^2),$$

with all: $\theta_{i4} > 0$

7. Explicit American Exceptionalism: work status effect is present for participants from the US but not for participants from the UK, Australia, and India, and more so for US participants who endorse the PWE rather than reject the PWE. For this model, we add a new parameter and predictor that captures an effect for US participants who endorse PWE; θ_{i8} is the *i*th region's effect for US participants who endorse PWE, o_{ij} is the indicator that is 1 for Americans who endorse PWE, and 0 otherwise.

Common:

$$Y_{ij} \sim N(u_i\theta_1 + o_{ij}\theta_8, \sigma^2),$$

with: $\theta_1 > 0$ and $\theta_8 > 0$

Random:

$$Y_{ij} \sim N(u_i\theta_{i1} + o_{ij}\theta_{i8}, \sigma^2),$$

with all: $\theta_{i1} > 0$ and $\theta_{i8} > 0$

8. Protestant Work Ethic: work status effect is present for participants who endorse the PWE but not for participants who reject the PWE.

Common:

$$Y_{ij} \sim N(w_j\theta_5, \sigma^2),$$

with: $\theta_5 > 0$

Random:

$$Y_{ij} \sim N(w_j\theta_{i5}, \sigma^2),$$

with all: $\theta_{i5} > 0$

9. Generalized Moralization: work status effect is present for all samples of participants (USA, UK, Australia, India).

Common:

$$Y_{ij} \sim N(\alpha_i, \sigma^2),$$

with: $\alpha_i = \mu_\alpha, \quad \mu_\alpha > 0$

Random:

$$Y_{ij} \sim N(\alpha_i, \sigma^2),$$

with all: $\alpha_i > 0$

10. Social Class Differences: work status effect is present for high SES participants but not for low SES participants.

Common:

$$Y_{ij} \sim N(l_j\theta_6, \sigma^2),$$

with: $\theta_6 > 0$

Random:

$$Y_{ij} \sim N(l_j\theta_{i6}, \sigma^2),$$

with all: $\theta_{i6} > 0$

11. Self Expression Values: work status effect is present for participants from the US, UK, and Australia, but not for participants from India.

Common:

$$Y_{ij} \sim N(t_i\theta_7, \sigma^2),$$

with: $\theta_7 > 0$

Random:

$$Y_{ij} \sim N(t_i\theta_{i7}, \sigma^2),$$

with all: $\theta_{i7} > 0$

Key effect 4: Intuitive Mindset Effect

Main effect of mindset (intuitive vs. rational) on moral judgments in the potato peelers scenario (within subject comparison). Participants should show a stronger intuitive preference than a rational preference for the target who continues to work after winning the lottery. The intuitive - rational preference (1-7 scales) difference score per participant serves as the dependent variable.

Theoretical Predictions:

Models are the same as those for key effect 3 (moralization of work vs. retirement 2)

39. False Positives: no effect of mindset, i.e., difference score (intercept) is zero.
40. Implicit Puritanism: mindset effect is present for participants from the US but not for participants from the UK, Australia, and India.
41. Regional Folkways: mindset effect is present for participants from New England but not for participants from other regions.
42. Religious Differences A: mindset effect is present for Protestant but not for non-Protestant participants.
43. Religious Differences B: mindset effect is present for high religiosity but not for low religiosity participants
44. Explicit American Exceptionalism: mindset effect is present for participants from the US but not for participants from the UK, Australia, and India, and more so for US participants who endorse the PWE rather than reject the PWE.
45. Protestant Work Ethic: mindset effect is present for participants who endorse the PWE but not for participants who reject the PWE.
46. Generalized Moralization: mindset effect is present for all samples of participants (USA, UK, Australia, India) (i.e., intercept > 0)
47. Social Class Differences: mindset effect is present for high SES but not for low SES participants.
48. Self Expression Values: mindset effect is present for participants from the US, UK and Australia but not for participants from India.

Key effect 5: Tacit Inferences Effect

Main effect of condition (sexually promiscuous or lazy vs. sexually abstinent or hard working) on false memories for linked violations of work or sex related morality. Participants should misremember a promiscuous person as lazy and vice versa, and an abstinent person as hard working and vice versa, compared to misremembering a promiscuous person as hardworking and vice versa, and an abstinent person as lazy and vice versa. Number of false memories serves as the dependent variable (0-4), with one scenario recoded such that higher scores always reflect false memories consistent with an intuitive work-sex link.

Theoretical Predictions:

Models are the same as those for key effect 1 and 2

49. False Positives: no tacit inferences effect consistent with an implicit link between work and sex values.
50. Implicit Puritanism: tacit inferences effect present for participants from the US but not for participants from the UK, Australia, and India.
51. Regional Folkways: tacit inferences effect present for participants from New England but not for participants from other regions.
52. Religious Differences A: tacit inferences effect present for Protestant but not for non-Protestant participants.
53. Religious Differences B: tacit inferences effect present for high religiosity but not for low religiosity participants.
54. Explicit American Exceptionalism: tacit inferences effect present for participants from the US but not for participants from the UK, Australia, and India, and more so for US participants who endorse the PWE than for participants who reject PWE.
55. Protestant Work Ethic: tacit inferences effect present for participants who endorse the PWE but not for participants who reject the PWE.
56. Generalized Moralization: tacit inferences effect present for all samples of participants (US, UK, Australia, India).
57. Social Class Differences: tacit inference effect present for high SES but not for low SES participants.
58. Self Expression Values: tacit inferences effect present for participants from the US, UK and Australia but not for participants from India.

Key effect 6: Salvation Prime Effect

Main effect of prime condition (religious prime vs. neutral prime) on anagram solving as a measure of work effort. Participants should solve more anagrams when they are primed with religiosity compared to a neutral prime.

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + x_k\eta_i + u_{ik}\theta_{i1} + v_{ik}\theta_2 + l_{jk}\theta_{i3} + m_{jk}\theta_{i4} + c_{jk}\theta_{i5}, \sigma^2),$$

This model is similar to the unconstrained model for key effect 1, 2 and 5, but without the main effect of SES and the condition-by-SES and condition-by-culture (self-expression vs. survival)

interactions, because there are no theoretical predictions for SES and the self-expression vs. survival culture dimensions. Since data was collected both online and in laboratories at universities, and the age compositions vary greatly between these two sources, we included participant age, δ, as an additional common main effect in the models. The indicator q_j gives the centered participant age (in decades). A separate intercept (α_i) will be modeled for the different labs (in addition to the different regions).

Theoretical Predictions:

59. False Positives: no salvation prime effect on solving anagrams.

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i, \sigma^2),$$

4. Implicit Puritanism: salvation prime effect present for participants from the US but not for participants from the UK, Australia, Ireland, and Canada.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + u_{ik}\theta_1, \sigma^2),$$

with $\theta_1 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + u_{ik}\theta_{i1}, \sigma^2),$$

with all $\theta_{i1} > 0$

5. Regional Folkways: salvation prime effect present for participants from New England but not for participants from other regions.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + v_{ik}\theta_2, \sigma^2),$$

with $\theta_2 > 0$

6. Religious Differences A: salvation prime effect present for Protestant but not for non-Protestant participants

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + l_{jk}\theta_3, \sigma^2),$$

with $\theta_3 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + l_{jk}\theta_{i3}, \sigma^2),$$

with all $\theta_{i3} > 0$

7. Religious Differences B: salvation prime effect present for high religiosity but not for low religiosity participants

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + m_{jk}\theta_4, \sigma^2),$$

with $\theta_4 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + m_{jk}\theta_{i4}, \sigma^2),$$

with all $\theta_{i4} > 0$

8. Explicit American Exceptionalism: no salvation prime effect present for anyone (same as false positives)

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i, \sigma^2),$$

9. Protestant Work Ethic: salvation prime effect present for participants who endorse the PWE but not for participants who reject the PWE.

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + c_{jk}\theta_5, \sigma^2),$$

with $\theta_5 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + c_{jk}\theta_{i5}, \sigma^2),$$

with all $\theta_{i5} > 0$

10. Generalized Moralization: salvation prime effect present for all samples of participants (US, UK, Australia, Ireland, Canada).

Common:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + x_k\eta, \sigma^2),$$

with $\eta > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + q_j\delta + w_j\zeta_i + x_k\eta_i, \sigma^2),$$

with all $\eta_i > 0$

11. Social Class Differences: no prediction

12. Self Expression Values: no prediction

References for Supplement 8

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, *2,* 109–112. doi:*10.4103/2229-3485.83221*

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22,* 779–798. doi:*10.31234/osf.io/ktjnq*

McCoy, C. E. (2017). Understanding the Intention-to-treat Principle in Randomized Controlled Trials. *Western Journal of Emergency Medicine*, *18,* 1075–1078. doi:*10.5811/westjem.2017.8.35985*

Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A Bayesian approach to finding constraints in meta-analysis. *Psychological Methods*, *24,* 606–621. doi:*10.1037/met0000216*

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11,* 702–712. doi:*10.1177/1745691616658637*

Supplement 9: Bayesian Multiverse Analysis of Project Results

Bayesian Multiverse Analysis Culture & Work Morality Project

Suzanne Hoogeveen & Julia Haaf

8/30/2020

This document contains the results of the Bayesian hierarchical multiverse analysis for the 6 key effects in the culture and work morality project: the four primary effects that are the main focus of the article, as well as two preregistered effects of further theoretical interest. For each effect, we constructed various hierarchical models that reflect the predictions from the proposed theories. The evidence for each of these different theories is quantified by Bayes factor model comparison, following the approach by Haaf & Rouder (2017) and Rouder, Haaf, Davis-Stober, & Hilgard (2019). In addition, we applied a multiverse approach in which we assessed evidence for various a priori specified alternative analysis paths (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). The preregistration for the analysis can be found at https://osf.io/pgfm8.

1 Six key effects

Here, we outline the key six effects that are targeted in the two studies.

1. **[Moralization of work vs. retirement 1]:** In a between-subjects design, a target who wins the lottery is more positively evaluated when she continues to work than when she retires. This effect specifically and exclusively occurs for (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
2. **[Target age and needless work]:** In a between-subjects design, a target who continues to work after winning the lottery is more positively evaluated when she is relatively younger than when she is older. This effect specifically and exclusively occurs for (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
3. **[Moralization of work vs. retirement 2]:** In a within-subjects design (1 item), a target who continues to work after winning the lottery is preferred over a person who retires. This effect specifically and exclusively occurs for (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.

4. **[Intuitive moralization of work]:** In a within-subjects design (2 items), a target who continues to work after winning the lottery is more strongly intuitively than rationally preferred over a target who retires. This effect specifically and exclusively occurs for (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
5. **[Link between work & sex morality]** In a between-subjects design, people have more false memories of a target violating (upholding) traditional norms in one domain (sex / work) when the target was described as violating (upholding) norms in the other domain (work /sex), compared to violating or upholding in one domain and upholding (violating) in the other domain. This effect specifically and exclusively occurs for (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) Americans who endorse the Protestant work ethic, (6) individuals who endorse the Protestant work ethic (7) everyone, (8) individuals with high socio-economic status, or (9) individuals from self-expression cultures.
6. **[Link between salvation and work]:** In a between-subjects design, subtle activation of the concept of divine salvation induce enhanced work performance, compared to subtle activation of general positivity. This effect specifically and exclusively occurs for (1) Americans, (2) Americans from New England, (3) Protestants, (4) religious individuals, (5) individuals who endorse the Protestant work ethic, or (6) everyone.

We constructed hierarchical Bayesian regression models that reflect the predictions from the 9 substantive theories, and the false positives model, for each of the 6 key effects. For each effect, the relative predictive adequacy of these models as well as the unconstrained model was compared using Bayes factors. In addition, we assessed the robustness of the findings to somewhat arbitrary analysis decisions by conducting a multiverse analysis: We varied operationalizations of cultural groupings and other variables, used different data exclusions, and specified more and less complex models for analysis.

2 Method

In this section we apply the preregistered model adequacy criteria, assess whether the applied dichotomization of variables is sensible, briefly summarize the multiverse paths specified, and note the necessary deviations from the preregistration.

2.1 Model adequacy

In the preregistration we specified the following criteria for variables to be simultaneously included in the models and for alternative operationalizations and groupings to be included as viable separate paths in the multiverse analysis:

1. Multicollinearity: two predictors cannot be both included in the models if they are correlated more than 0.7.

2. Multiverse paths: (a) an alternative operationalization or grouping variable will not be analyzed as a separate path if they are correlated more than 0.9, and (b) data exclusions will not be analyzed as a separate path when less than 5% of the sample are excluded.

Tables S9-1 and S9-2 verify that the condition (1) is met for both Study 1 and Study 2. Additionally, condition (2a) is met for all variables as well. This means that all preregistered variables can be simultaneously included in the models and that the alternative operationalization of the religiosity measure, and the alternative cultural and regional groupings are retained as separate paths in the multiverse analysis. Note that we did not include all data exclusions specified in the preregistration (see section "Deviations from preregistration").

Table S9-1: Correlation Matrix Model Predictors Study 1

	American	New England	Religious	Protestant	PWE	SES	Self-expression culture
American	1	0.44	0.01	0.11	-0.07	0.16	-0.45
New England	0.44	1	-0.05	-0.04	-0.05	0.03	-0.2
Religious	0.01	-0.05	1	0.16	0.24	0.07	0.38
Protestant	0.11	-0.04	0.16	1	0.05	-0.02	-0.16
PWE	-0.07	-0.05	0.24	0.05	1	0	0.24
SES	0.16	0.03	0.07	-0.02	0	1	0.01
Self-expression culture	-0.45	-0.2	0.38	-0.16	0.24	0.01	1

Note: Correlation (Cramer's V) between dichotomized covariates.

Table S9-2: Correlation Matrix Model Predictors Study 2

	American	New England	Religious	Protestant	PWE	SES
American	1	0.42	0.2	-0.04	0.1	-0.22
New England	0.42	1	0.09	-0.03	0	0.05
Religious	0.2	0.09	1	0.15	0.18	0.08
Protestant	-0.04	-0.03	0.15	1	0.04	0.29
PWE	0.1	0	0.18	0.04	1	0.05
Age	-0.22	0.05	0.08	0.29	0.05	1

Note: Correlation (Cramer's V) between dichotomized covariates.

We dichotomized the continuous predictors for *religiosity* and *Protestant work ethic (PWE)*, and categorized participants into 'religious' vs. 'nonreligious' and 'endorse PWE' vs. 'reject PWE'.

By using dichotomized variables we could both (a) more directly test the predictions of the relevant theories (e.g., effect for Protestants, no effect for non-Protestants), and (b) make a fairer comparison between the various models derived from theories that relate to categorical variables and continuous variables. We constructed a composite measure for socio-economic status based on personal education, parental education, and personal (online samples) or household income (laboratory samples of university students). See the preregistration for details on how we determined the composite score and the dichotomization. Participants with missing values on any of the relevant measures were excluded from the analysis. For the SES measure, however, we retained data from participants who completed 2 out of the 3 components, because 264 participants did not fill out their personal level of education and would otherwise have been removed from the sample. The distribution of participants across all theoretically relevant predictor categories are given in the table below.

Table S9-3: Participant demographics.

	Percentage	
Category	Study 1	Study 2
US	52	58.2
New England	17.1	19.9
Religious	41.4	34.3
Protestant	22.6	19
Endorse PWE	65	53.8
High SES	36.2	–
Self-Expression Culture	16	–

Note: The socio-economic status and self-expression vs. survival culture dimensions are not theoretically relevant in Study 2.

2.2 Multiverse paths

The eventual multiverse dimensions consisted of:

1. Modeling strategy: Common vs. random effects
2. Religiosity: the DUREL scale vs. the single-item religiosity measure
3. Cultural grouping (US or other): current objective location vs. self-reported nation of birth
4. Regional grouping (New England or other): current objective location vs. self-reported region "grew up in".

This resulted in $2 \times 2 \times 2 \times 2 = 16$ conditions for the multiverse analysis of the first 5 key effects. For the salvation prime study (key effect 6), the following dimensions were added:

5. Awareness measure: no exclusions vs. exclude participants who suspected they may have been influenced by the prime (56.2% excluded)

6. Prime task completion: no exclusions vs. exclude participants who did not complete all scrambled sentences (10.8% excluded).

This resulted in 2 × 2 × 2 × 2 × 2 × 2 = 64 conditions for the multiverse analysis of key effect 6.

2.3 Deviations from preregistration

Before presenting the results of the multiverse analysis we first want to highlight deviations from the pre-registration that occurred during the analysis.

The first deviation concerns data exclusions: We stated that we would assess exclusions based on attention check failure and based on having less than 5 years of English experience. However, for the online samples, these participants were prevented from completing the study or already screened out. Therefore, in Study 1, none of the included participants failed the attention check or had less than 5 years of English experience. For Study 2, which included an online sample as well as lab-based samples, 4.3% of participants failed the attention check, and 0.8% of participants had less than 5 years of English experience. This does not surpass the 5% that was specified as the minimum difference in sample size between two paths. Based on these considerations, we did not include the attention check and English fluency dimensions in the multiverse analysis. Note that in order to stay consistent with the analyses for Study 1 (key effects 1-5), the frequentist analysis, and the online sample in Study 2, we decided to also exclude attention check failures and participants with less than 5 years of English experience for the main analysis in Study 2.

The second deviation is that we did not include how the anagram dependent measure for key effect 6 was scored as a choice point in the multiverse. Participants in the salvation prime study were instructed to produce anagram solutions of four letters or more, and thus only 4+ letter solutions were counted as measures of work productivity. We initially intended to create an alternative scoring for this DV counting any anagram solution, regardless of number of letters, as a measure of productivity. However we belatedly discovered we did not have this data for most of the crowd laboratory sites, which are critical in light of arguments that priming effects should be more likely to emerge in controlled laboratory environments. Due to the ongoing COVID-19 pandemic, retrieving and scanning all of the original paper-and-pencil questionnaires at each replication site in order to retrieve the necessary fine-grained data was not feasible. Therefore, we removed the choice point of anagram scoring from the analyses.

The third deviation concerns the analyzed models for the within-subjects designs: For key effects 3 and 4, we planned to use a within-subjects comparison using a single item score (key effect 3) or difference score (key effect 4) and test whether the intercept = 0 and the effect of the predictor of interest > 0. However, this is not easily possible in the software we use. For key effect 4, we therefore decided to run a mixed model with a *mindset* within-subjects factor (intuitive vs. rational), and test the crucial interaction between mindset condition and theoretical moderator (e.g., American, Protestant etc.). In this approach, the models for key effect 4 are equivalent to those for key effect 1, 2, and 5 (i.e., assuming a pattern such as American-intuitive > American-

rational = non-American-intuitive = non-American-rational). For key effect 3, this was not possible, as we only had 1 (averaged) score per participant and no experimental factor. Given the fact that the main effect of preference for the worker over the retiree in the potato peeler scenario is clearly robust, we now tested the following: in addition to an overall effect, do any of the predictors add anything, e.g., do Americans *more strongly* morally praise needless work, compared to the other groups.

3 Results

Here, we briefly describe the results of the multiverse analysis. For interested researchers the analysis code is provided at https://github.com/jstbcs/multi-destruction.

3.1 Summary

1. We find evidence for key effects 1 and 3 across nationalities and cultures. The analyses indicate strong evidence for the basic moralization of work effects: Targets who decide to continue working after winning the lottery are considered morally superior to targets who retire. We find evidence against most of the moderators considered here. The most supported moderator is the Protestant work ethic; endorsement of the Protestant work ethic appears to enhance the basic moralization of work effect.
2. The false positive model is supported for key effect 2. That is, we find evidence against a target age effect on the moralization of needless work.
3. Although less strong, we find evidence for key effect 4 across all countries and cultures. Lottery winners who continue working are especially morally praised on an intuitive level, more than on a rational level. Although the effect does not emerge in all regions separately, there is substantial evidence for an overall effect.
4. We find strong evidence for key effect 5, an implicit link between work and sex morality. We find evidence against all moderators under consideration. Therefore, this effect seems to be general rather than nationally or culturally specific even though we do find evidence for heterogeneity across geographic regions.
5. The false positive model is supported for key effect 6. That is, there is substantial evidence *against* an effect of priming the concept of salvation on work performance.
6. Across all key effects the analyses for different multiverse paths seemingly converge to a large degree. Except for key effect 2 the winning model is preferred for all analysis paths, and the patterns of evidence are very similar.

Table S9-4: Best model per key effect

Effect	Best Model	BF_{w0} Winning Model	BF_{10} Implicit Puritanism	BF_{w1} Winning Model vs. Implicit Puritanism
1	Generalized Moralization (common)	10^{131}	10^{55}	10^{75}
2	False Positives	1	0.12	8.38
3	Unconstrained Model	10^{138}	58.42	10^{136}
4	Generalized Moralization (common)	10^{7}	10^{6}	2.39
5	Generalized Moralization (random)	10^{80}	10^{54}	10^{25}
6	False Positives/Explicit American Exceptionalism	1	0.05	21.87

Note: The BF_{w0} gives the Bayes factor for the winning model vs. the null model per key effect. The BF_{10} Implicit Puritanism gives the Bayes factor for central Implicit Puritanism-based model versus the null-model for each key effect. The BF_{w1} gives the Bayes factor for the winning model vs. the Implicit Puritanism model for each key effect.

3.2 Bayes factor analysis

Table S9-4 and Figure S9-1 provide an overview over the results from the Bayes factor analysis for all six key effects. Table S9-4 shows which model is preferred for which key effect, and how this model compares to the false positive model (first column) and the target model that specifies the implicit puritanism hypothesis (i.e., only Americans' judgments are affected by Puritan-Protestant values; third column). As can be seen, the target implicit puritanism model performs worse than the winning model for all key effects. The Bayes factors range from 1-to-2.4 against the implicit puritanism model for key effect 4, all the way up to 1-to-10^{136} against the implicit puritanism model for key effect 3. Column 2 additionally shows how the implicit puritanism model performs in comparison to the false positive model. For this comparison, the implicit puritanism model is preferred for 4/6 of the key effects. Note, however, that the false positive model is actually the winning model for two of the six key effects.

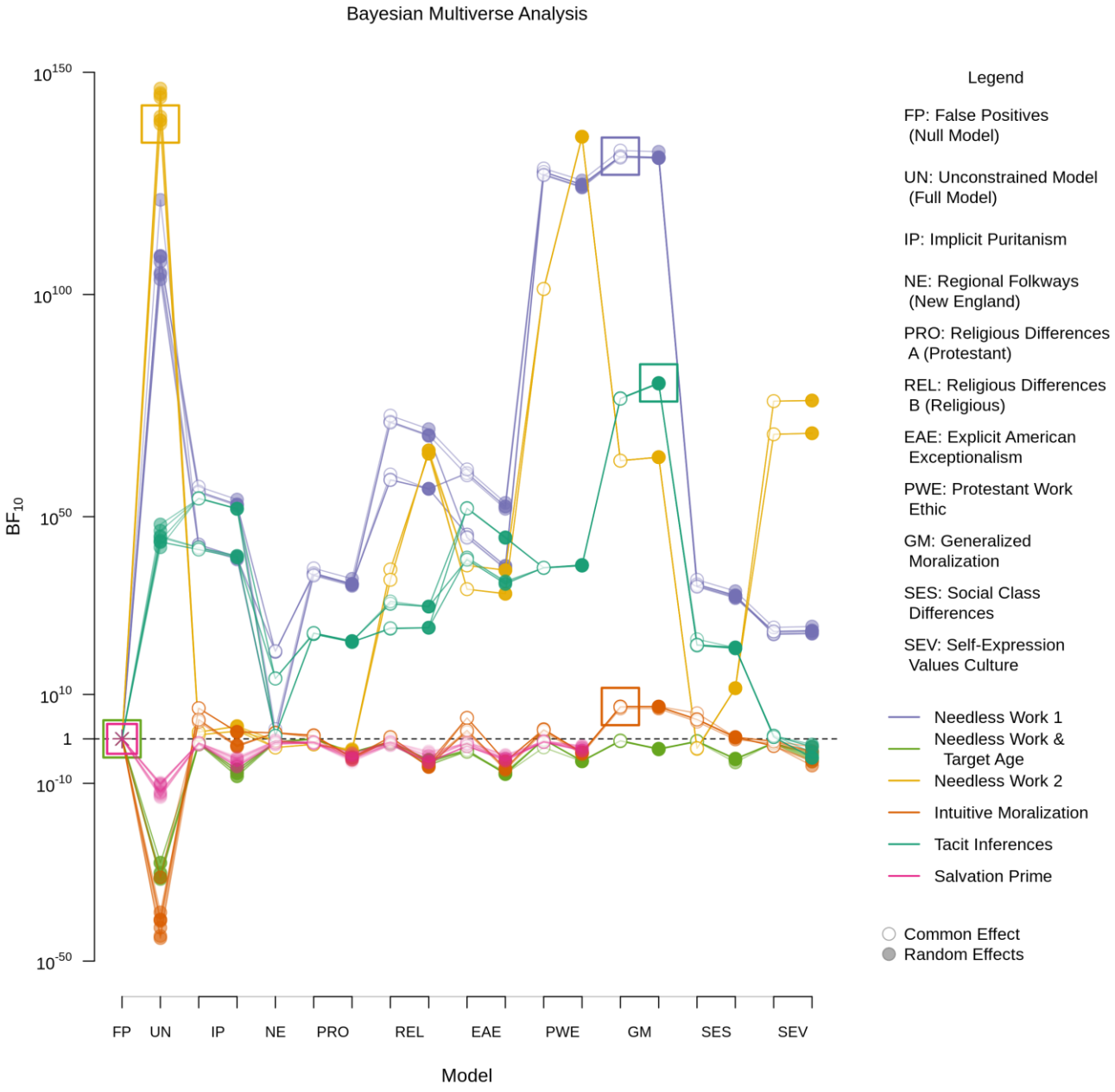


Figure S9-1: Summary of the Bayesian multiverse analysis for all 6 key effects. Bayes factors for each of the theory-based models versus the null model. The colors denote the 6 key effects and the multiple lines and points per color per model reflect the different multiverse paths. Open circles denote common effect models and solid circles denote random effects models. The squared box gives the preferred model for each key effect.

Figures S9-5-S9-10 show Bayes factors for all models relative to the false positive model for the six key effects (as separate figures) and all analysis paths of the multiverse (as separate lines). Each model is implemented as a common-effect model (open points) and as a random-effect model (filled points). The winning models are highlighted with a red box. Across all key effects the analyses for different multiverse paths seemingly converge to a large degree. Except for key effect 2 the winning model is preferred for all analysis paths, and the patterns of evidence are very similar (for key effect 2, the common Protestantism model is preferred in 2/8 paths, but with a maximum of 1.26-to-1 versus the null-model). One reason for this strong convergence is that the different variable specifications in the multiverse do not lead to data exclusions in Study 1 (key effects 1-5). Therefore, all the data are used for all analyses in the multiverse, but they are somewhat differently distributed across levels of moderators. These slight changes in the allocation of data affect the results much less than excluding large proportions of data, or adding different moderators in the analyses. Another reason for the seeming convergence is the scale of the Bayes factors depicted on the y-axis of the figures. For example, depending on the path taken in the multiverse for key effect 1, the false positive model is preferred over the implicit puritanism model by $4x10^{40}$-to-1 or by $5x10^{55}$-to-1. Even though both Bayes factors provide overwhelming evidence against the implicit puritanism model, the amount of overwhelming evidence is still quite variable. This variability is even more relevant for the evidence against the regional folkways model that specifies that the moralization of work effect should only occur for participants from New England. For key effect 1 and different paths in the multiverse the Bayes factors range from 7-to-1 to 10^{19}-to-1 in favor of the false positive model. Even though the false positive model is preferred for all paths in some cases the evidence could be considered much more modest than in others.

3.3 Superselect and Antiselect Samples

In addition to the main analyses comparing the predictions from all theories on all the data, we created sub-samples in which, according to the implicit puritanism hypothesis, the effects should most likely emerge. We call these samples the *superselect samples*. In addition to these superselect groups, we also specified comparison groups in which the effect would be least likely to emerge and call these the *antiselect samples*. We expect to find evidence in favor of the key effects in the superselect samples and evidence against the key effects in the antiselect samples.

We constructed the superselect samples by combining a number of the theories, 1) Americans (currently located and born) 2) from New England 3) who are self-identified Protestants, 4) endorse the PWE, and 5) are religious. These participants should be most likely to exhibit the original effects. We contrast this superselect group of participants with the mirror image of non-Americans who are not Protestant, reject the PWE, and are not religious. These participants serve as our antiselect sample. For key effect 6 we additionally specified a superselect group theoretically more likely to exhibit priming effects (e.g., were not suspicious about the manipulation) as well as a comparison antiselect group (see the preregistration for more details).

As specified in the preregistration, in case these criteria would result in fewer than 50 participants per condition, we would sequentially remove layers starting from 5. Since Bayes

factors are sensitive to sample size, we tried to match the sizes of the superselect and antiselect samples for each effect. That is, when the mirrored antiselect sample was substantially larger than the superselect sample, we randomly selected a subset of the antiselect group that was equal in size to the superselect group. This happened for key effects 1 and 5.

These selection rules resulted in the following samples for the select groups: For the between-subjects key effect 1 and 5, the superselect group consisted of US-born Protestants from New England ($n = 166$) and the antiselect group consisted of non-Protestant Indians ($n = 166$). For the between-subjects key effect 2, which uses only half of the data (the 'work' condition), the superselect group consisted of US-born New Englanders ($n = 452$) and the antiselect group consisted of non-Protestant Indians ($n = 424$). For the within-subjects key effects 3 and 4, the superselect group consisted of US-born Protestants from New England who endorse the PWE ($n = 108$) and the antiselect group consisted of non-Protestant Indians who reject the PWE ($n = 76$). The demographics-based superselect group for key effect 6 consisted of US-born Protestants from New England who completed the study in the lab ($n = 92$) and the antiselect group consisted of Irish individuals who completed the study in the lab ($n = 70$). Note that we could not reach 50 participants per condition for this group analysis, as we only had less than 100 Irish lab participants. Finally, the priming-based superselect group for key effect 6 consisted of participants who were unaware of the effect of the prime, completed all scrambled sentences in the priming task, and completed the study in the lab ($n = 212$) and the antiselect group consisted of individuals who were unsure or aware of the prime effect, did not complete all scrambled sentences, and completed the study either online or in the lab ($n = 121$).

Table S9-5: Bayes Factors for Selective Samples

	BF_{10}		*N*	
Effect	Superselect	Antiselect	Superselect	Antiselect
1	15581	10^7	166	166
2	0.13	0.10	452	424
3	10^{21}	10^6	108	76
4	0.17	0.34	108	76
5	36.37	0.43	166	166
6a	0.16	0.34	92	70
6b	0.20	0.29	212	121

Note: All Bayes factors reflect the evidence for the respective experimental effect vs. no effect per group.

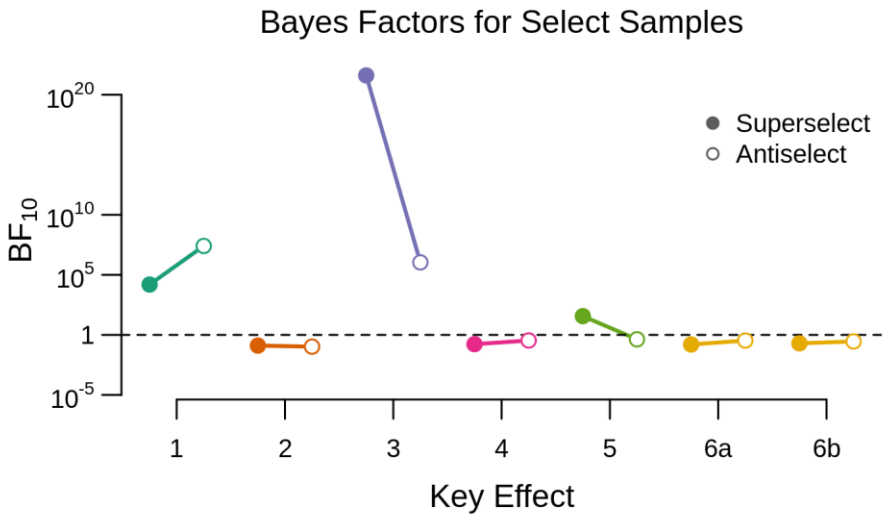


Figure S9-2: Bayes factors for each key effect for the superselect and antiselect samples. The expected pattern is Bayes factor greater than 1 for the superselect samples and Bayes factors less than 1 for the antiselect samples. This pattern only emerged for key effect 5.

Table S9-5 and Figure S9-2 show the results of this analysis. Remember that we predicted Bayes factors greater than one for the superselect samples and Bayes factors smaller than one for the antiselect samples. As can be seen in the figure, for most key effects the evidence is quite consistent across the two samples. Key effect 3 is a notable exception; for key effect 3 we see that there is much more evidence in favor of the moralization of work effect for the superselect sample, yet, there is strong support for the effect for both samples. Also note that in both the superselect and antiselect samples we get evidence in favor of the null for key effect 4, while we obtain evidence for a general effect in the overall sample. This suggests that the difference in intuitive vs. rational praise for needless work is rather small, and large samples are needed to obtain convincing evidence for the effect.

3.4 Key effect 3: Moralization of Work vs. Retirement 2

The pattern of evidence is quite clear except for the results for key effect 3. Here, the unconstrained model is preferred over all other models indicating that none of the theoretically motivated models was appropriate on its own. Additionally, as highlighted in the section "Deviation from preregistration" we were not able to conduct the analysis as planned. The models compared here all include an intercept corresponding to an overall moralization of work effect across countries, religion and culture. These two issues, that none of the theoretical predictions was adequate and that our modeling approach was altered, warrant a more thorough analysis of key effect 3.

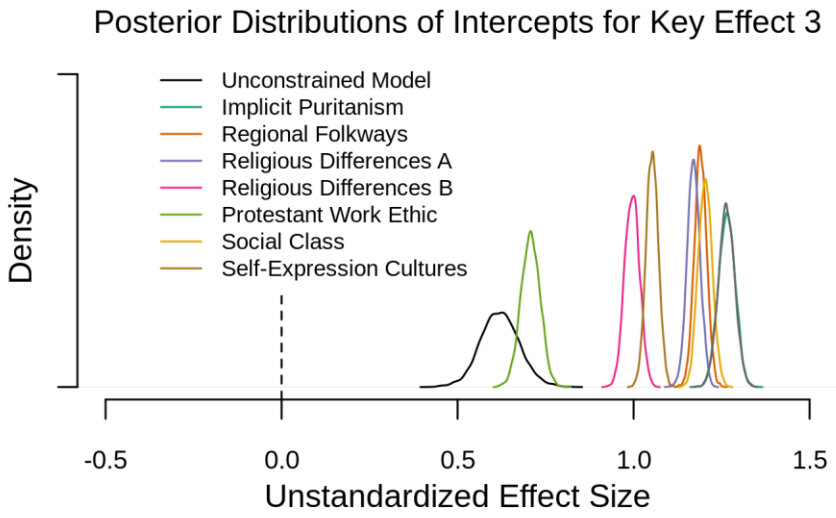


Figure S9-3: Posterior distributions of the intercepts for all models specified for key effect 3. Except for the generalized moralization hypothesis, all substantive theories predict the intercept to be zero.

First, we show that the intercept is far from zero for all models specified (see Figure S9-3). Except for the generalized moralization hypothesis, all substantive theories predict the intercept to be zero, and this is clearly not the case. The two models with the smallest intercept are the protestant work ethic model and the unconstrained model highlighting that they capture at least some of the overall effect. Based on these results and the Bayes factor model comparison results in Figure S9-7 we decided to posthoc construct an exploratory model in an attempt to capture the variability in the effect. In this exploratory model we included a random effect of region, the effects of protestant work ethic and religiosity (either as common effects or random effects).

Figure S9-4 again illustrates the Bayes factors for the key effect 3, now including the additional exploratory model. For 3 out of the 8 multiverse paths the exploratory model with common effects for religiosity and protestant work ethic is preferred over the unconstrained model (Bayes factors between 6-to-1 and 65-to-1 in favor of the exploratory model); for one path the models are equivalent (Bayes factor of 1); and for the other four paths the unconstrained model is still preferred over the exploratory model (Bayes factors between 9279-to-1 and 10^6-to-1 in favor of the unconstrained model). This variability of the results dependent on the multiverse paths suggest that there might be some interactions between several covariates in the data, and some more subtle effects that are captured by the unconstrained model. To fully understand key effect 3 a more thorough exploratory investigation of the data is needed, and, if new hypotheses about the data pattern emerge, an additional replication in an independent sample is warranted.

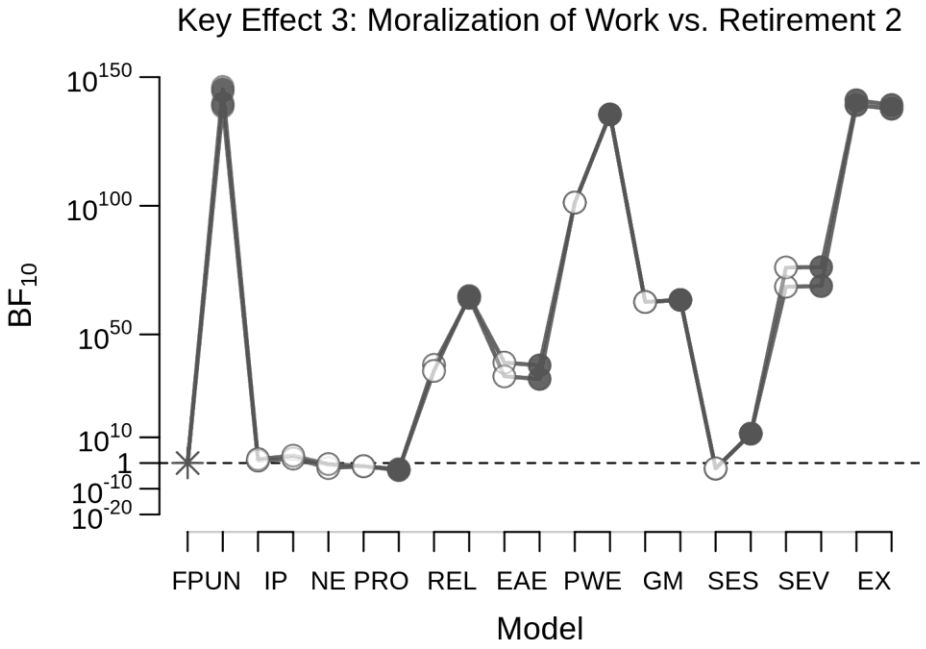


Figure S9-4: Bayes factors for key effect 3 including an additional exploratory model. The new model is preferred for four out of eight multiverse paths.

3.5 Supplemental figures

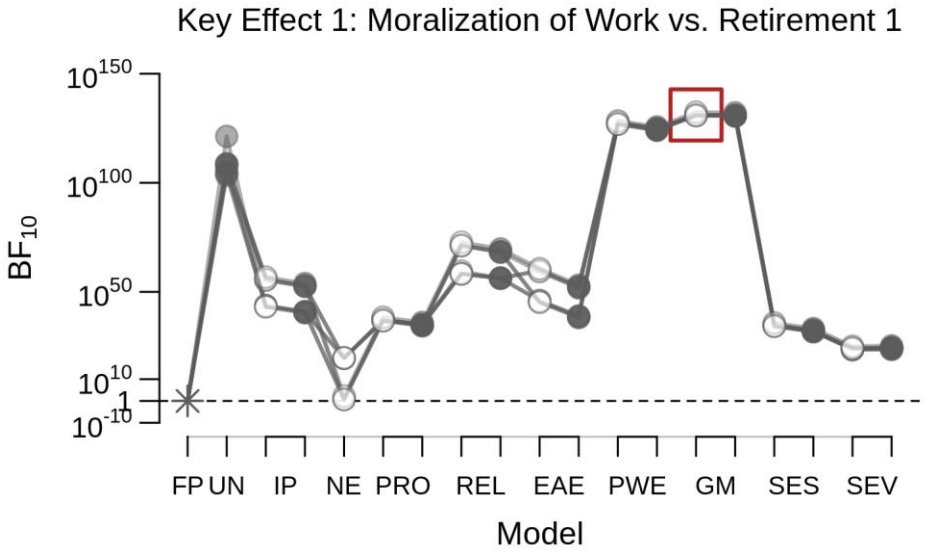


Figure S9-5: Bayes factors for all models (x-axis) compared to the false positive model (starred) for all multiverse analysis paths (different lines) for key effect 1. The preferred model is highlighted by the red square. The model with a general effect of moralization of work is preferred over all others.

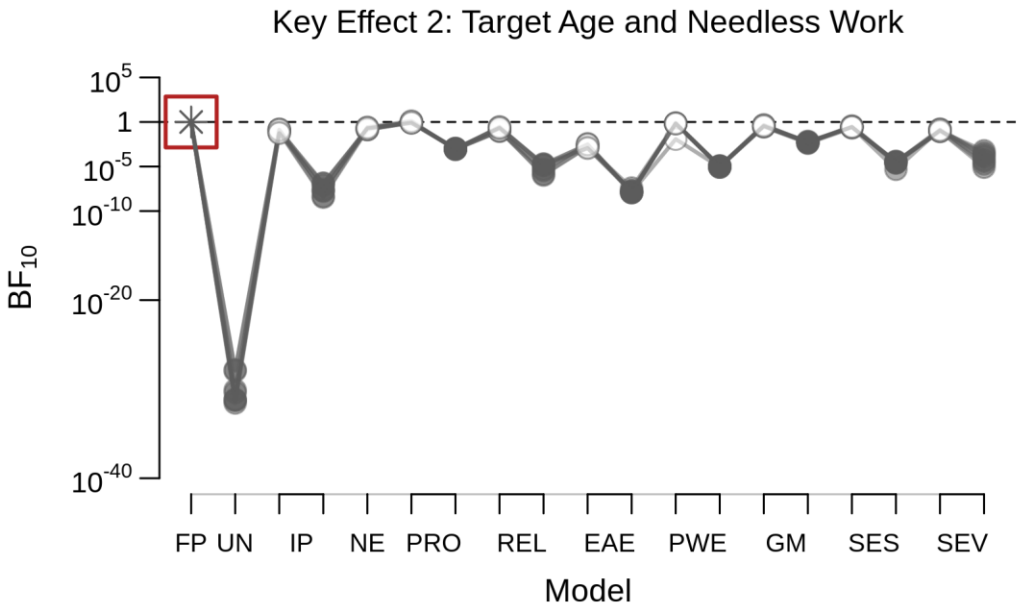


Figure S9-6: Bayes factors for all models (x-axis) compared to the false positive model (starred) for all multiverse analysis paths (different lines) for key effect 2. The false positive model is preferred over all others.

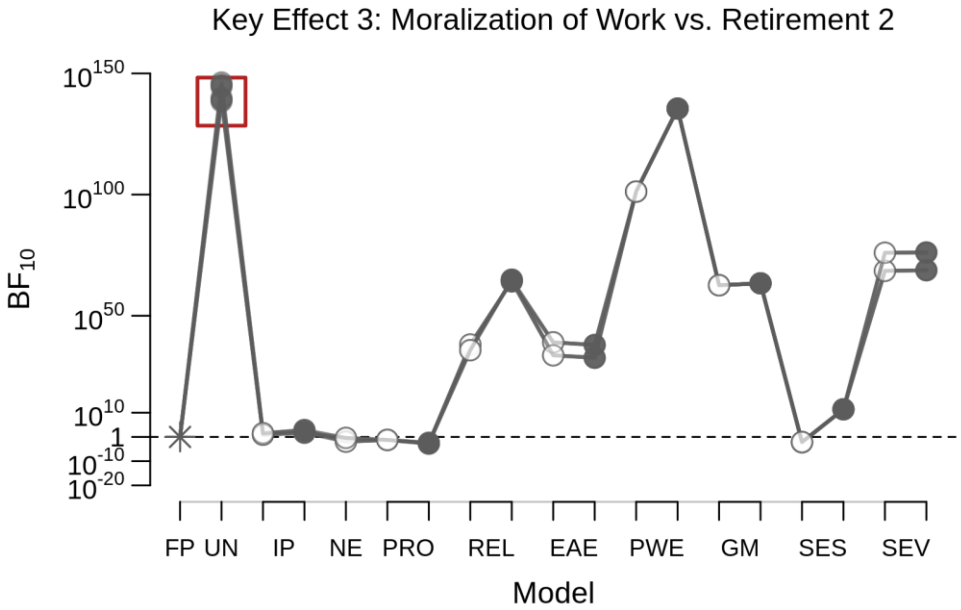


Figure S9-7: Bayes factors for all models (x-axis) compared to the false positive model (starred) for all multiverse analysis paths (different lines) for key effect 3. The preferred model is highlighted by the red square. The unconstrained model including all covariates and a general moralization of work effect is preferred over all others.

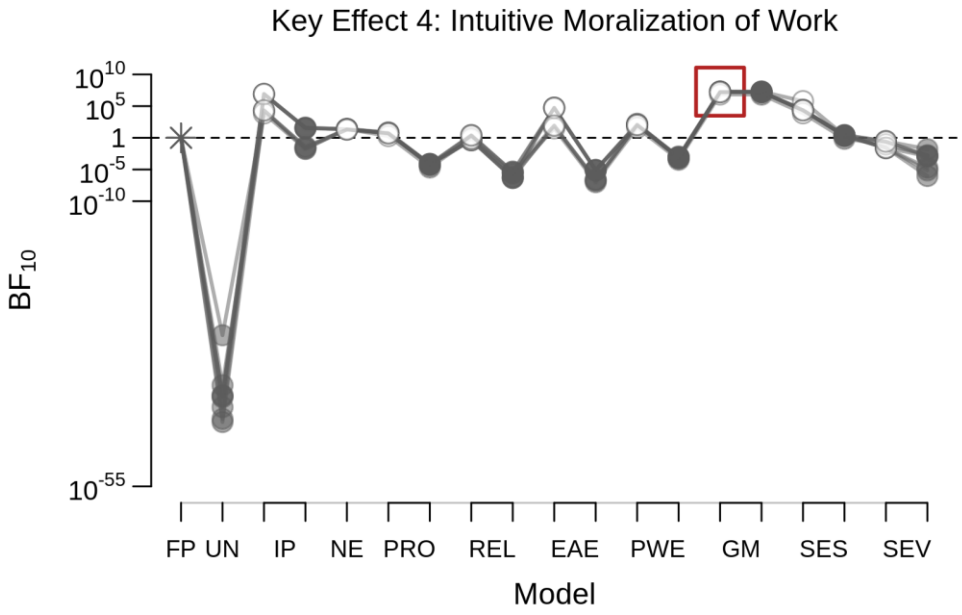


Figure S9-8: Bayes factors for all models (x-axis) compared to the false positive model (starred) for all multiverse analysis paths (different lines) for key effect 4. The model with a general intuitive mindset effect is preferred over all others.

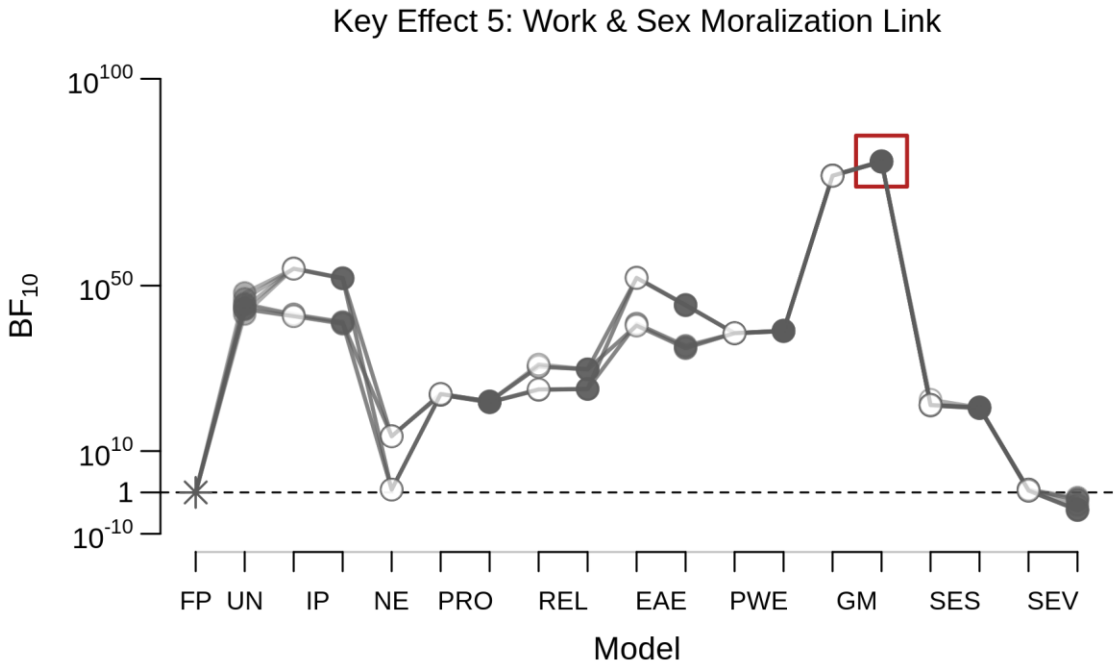


Figure S9-9: Bayes factors for all models (x-axis) compared to the false positive model (starred) for all multiverse analysis paths (different lines) for key effect 5. The model with a general tacit inferences effect is preferred over all others.

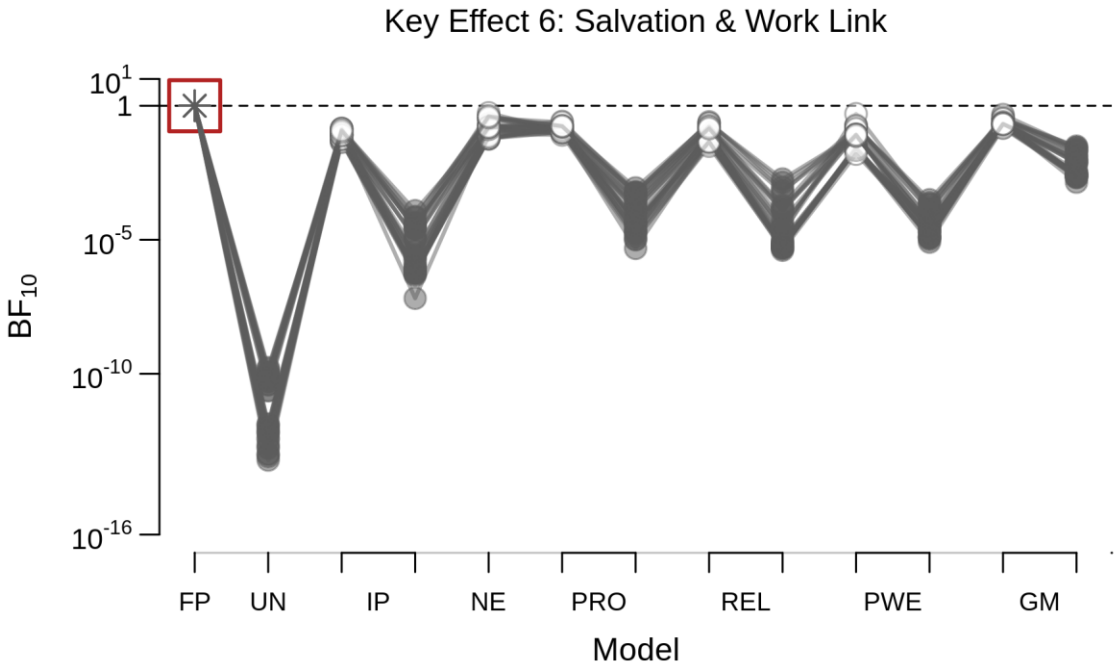


Figure S9-10: Bayes factors for all models (x-axis) compared to the false positive model (starred) for all multiverse analysis paths (different lines) for key effect 6. The false positive model is preferred over all others.

3.6 Effect size estimates

In order to better understand the cross-national patterns, we visualized the estimates of the 6 experimental effects per key effect, per region. These effect size estimates are taken from the random generalized moralization models, which assume a general experimental effect for everyone and estimate this effect separately per region. The patterns again show that in most cases, the respective effect is either present across all regions (key effect 1, 3, and 5) or in none of the regions (key effect 2 and 6). Except for key effect 3 and 5, there does not seem to be much variability between regions. Notably, for key effect 3 (moralization of work 2), it appears that Indians in fact moralize work *more* instead of less than people from the US, UK, and Australia (see also Supplement 7).

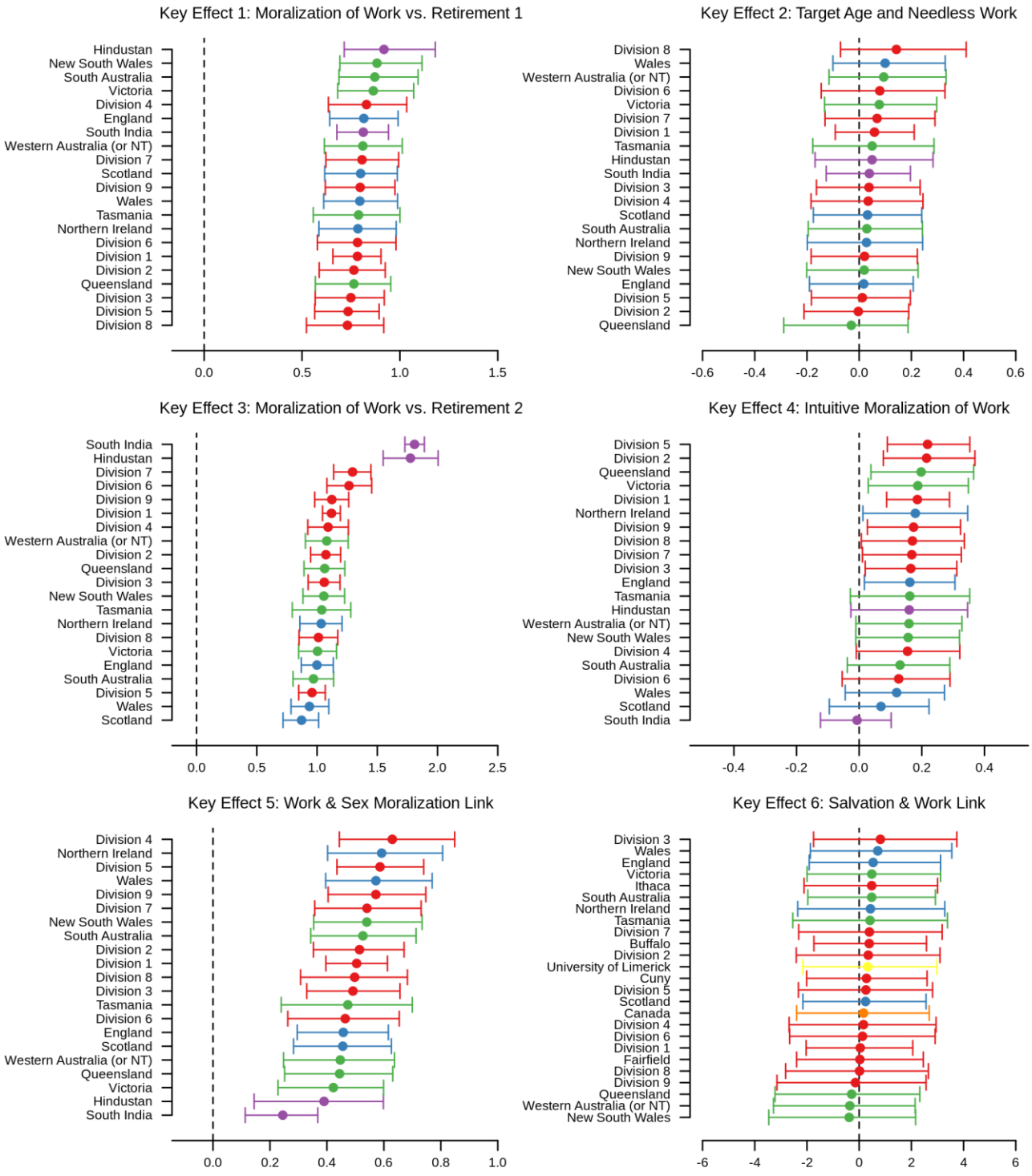


Figure S9-11: Posterior estimates of the experimental effect per region for each key effect. Estimates are taken from the random generalized moralization models (i.e., assuming an experimental effect for every person in each region). The colors denote the different countries of the regions; red is US, blue is UK, green is Australia, purple is India, orange is Canada and yellow is Ireland. Effects are unstandardized.

References for Supplement 9

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*, 779–798. doi:10.31234/osf.io/ktjnq

Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A Bayesian approach to finding constraints in meta-analysis. *Psychological Methods*, *24*, 606–621. doi:10.1037/met0000216

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*, 702–712. doi:10.1177/1745691616658637

Supplement 10: Departures from Pre-Registered Replication Plan

Below we list ways in which our analyses, presentation of the results, and theoretical conclusions drawn deviated from our pre-registered plan.

***Model fitting tests*.** We originally pre-registered that we would carry out model-fitting tests following on the primary specification and participant inclusion/exclusion criteria reported in the main text. After the project was already in progress, we recruited a sub-team of data blind experts to conduct a Bayesian multiverse analysis with more comprehensive model fitting tests capturing numerous defensible specifications (see General Discussion and Supplements 8 and 9). The Bayesian multiverse subsumes the originally planned model fitting tests, which were therefore not conducted separately.

Exclusions based on attention check. We included an instructional attention check item asking participants to select "strongly disagree" on a Likert scale. The survey firm PureProfile offered to exclude *a priori* and not charge for respondents who failed the attention check, thus these individuals were not in the PureProfile samples at all. For consistency, and in the interests of data quality, all participants who failed the instructional attention check were excluded from all analyses, not just PureProfile samples but also the MTurk sample and the crowdsourced laboratory data collections. This decision was made prior to running the main analyses and without knowing the implications for the replication results.

Exclusions based on English experience. We pre-registered that we would exclude participants who indicated they had less than 5 years of experience with the English language. However, this left ambiguous what to do with participants who did not respond to the English experience item (i.e., left it blank). In the primary analyses reported in the main manuscript we included non-respondents in the sample if they were located in English speaking countries (USA, UK, Australia), and excluded them if they were not (India). This decision was made prior to running the replication analyses, without knowledge of the implications for the replication results.

Test-holdout approach. We initially planned to publicly distribute half the replication data online for a crowdsourced exploratory analysis involving numerous colleagues. Following on this, the remaining half of the data would be provided to the crowd of analysts for pre-registered analyses confirming or disconfirming their initial findings. However, we instead recruited a sub-team of data-blind experts to carry out all defensible specifications in the context of a pre-registered Bayesian multiverse analysis. This achieves the same goal, namely testing the research questions using as many alternative means as possible while avoiding false positive findings.

Income item in University of Limerick sample. For all samples but one, we converted the income item into local currency. For the University of Limerick sample in the Republic of Ireland the survey stated the conversion rate above the question "note that 1 dollar is approximately 0.90 euro" and presented income brackets in U.S. dollars. Note that our primary measure of socioeconomic status was education level, not income, thus this does not affect the primary tests of the social class account in the main manuscript. However, it could potentially affect the Bayesian multiverse analyses of the salvation prime effect using income as an

alternative measure of social class, since 80 participants from the University of Limerick were included in the crowdsourced laboratories sample (see Table S14-2).

Regional partitioning in India. We pre-registered that we would divide India into the following regions: South India, Hindustan, and North-East. However, due to uneven sampling from throughout the country we instead ended up separating South India from other, effectively creating two regions. Over 80% of the Indian sample was located within the South of India, thus this seemed the most reasonable method of separating the sample. This change in how Indian regions were partitioned occurred before we conducted the main analyses, with the research team unaware of the implications for the research results.

Meta-analysis. For conducted the meta-analysis, we pre-registered that we would bootstrap the Cohen's d. However given the random effect in the mixed method model, it was considered unwise to bootstrap the Cohen's d values since the random effect would not be taken into consideration in the bootstrapping.

Second MTurk sample. Although not a part of our original pre-registration (see Supplement 3), after discovering New England residents composed only 4.3% of our MTurk sample, we planned to conduct a second MTurk data collection for Study 1 oversampling the New England U.S. states. Unfortunately, we were unable to realize this goal due a lack of research funds in light of of the COVID-19 pandemic. Notably however we were able to recruit over 1000 New England (U.S. census district 1) residents using the survey firm PureProfile, providing an adequate sample to test the regional folkways account of culture and work morality.

Sample size in Ireland. We originally intended to collect 300 participants in a paper-pencil replication of the salvation prime effect at the University of Limerick. However, the data collection was cut short due to the COVID-19 pandemic, and ultimately only 80 Irish participants completed the paper-pencil questionnaire version of the study. Having already collected 312 adult participants from the United Kingdom online, we felt that collecting data from additional University of Limerick students online to meet the numeric sample size target would add limited value. Doing so would also have introduced the potential confound of pandemic/lockdown conditions, and defeated the purpose of complementing online participants with data collected under controlled laboratory conditions.

Data collections at CUNY. A second planned data collection wave at the City University of New York for Spring 2020 was canceled due to the pandemic. However, we were able to collect 161 participants from CUNY in Fall 2019, making for a respectable overall sample size.

Data collection in Canada. One replicator for the crowdsourced data collections for the salvation prime study (J. McPhetres) changed academic institutions during the course of the project, moving from the University of Rochester in New York State to the University of Regina in Canada. Departing from our original plan, we therefore included non-USA samples from not only the Republic of Ireland but also Canada. This had the benefit of more directly replicating the original experiment, which compared the responses of Americans and Canadians to religion primes. It also increased our non-USA sample for the crowdsourced laboratory data collections to 171, totaling across the Republic of Ireland and Canadian samples.

Theoretical conclusions drawn from mindset differences. A final deviation involves our theoretical conclusions drawn from the study in which two potato peelers win the lottery and decide to either continue working or retire, and participants are asked for their intuitive and logical evaluations of the two characters. We pre-registered that we would interpret a smaller difference score between intuitive and logical judgments in the survival culture (India) compared to the self-expression cultures (USA, UK, Australia) as support for the Self Expression account of culture and work morality. As described in the main manuscript, we indeed found that USA, UK, and Australian participants exhibited an intuitive mindset effect (i.e., statistically significant difference score), whereas Indians did not. However, pre-registered follow up analyses indicated that Indians scored unexpectedly higher in the tendency to intuitively moralize work than Americans (Supplement 7). Inspection of the means in Figure 1 clarifies that the lack of an intuitive-logical mindset difference score in the India sample is attributable to their greater logical moralization of work, not reduced intuitive moralization of work, relative to the other samples. In light of this, we cannot interpret the empirical patterns regarding intuitive and logical evaluations of needless work as supporting the Self-Expression account.

Supplement 11: Further discussion of the priming failed replication

In light of recent controversies regarding the reproducibility of priming effects (Bargh, 2012, 2014; Dijksterhuis, 2018; Doyen et al., 2012; Harris et al., 2013; O'Donnell et al., 2018; Pashler et al., 2012, 2013; Rohrer et al., 2015; Schnall, 2014; Weingarten et al., 2016), we elaborate below and in Supplement 12 on the failed religion prime replication in the present Study 2.

Lack of a mediator measure

We cannot rule out the possibility that the sentence-unscrambling manipulation did not impact the theoretically hypothesized mediator of the accessibility of religious concepts (Fabrigar et al., in press; Schwarz & Strack, 2014; Stroebe & Strack, 2014). This mediator was also not measured in the original research, since most measures of construct accessibility (e.g., lexical decision tasks) would inadvertently prime the target construct, interfering with the priming manipulation. We concur with other scholars that replications should not be held to higher standards than the original research investigations (Zwaan et al., 2018), which frequently lacked thorough process checks and construct validation evidence (Ejelöv & Luke, 2020; Flake, Pek, & Hehman, 2017). We further suggest that the argument that replicators should demonstrate an effect of the independent variable on a manipulation check or mediator, and then a null effect on the dependent measure, misses the point of the strong version of the false positives account. According to the most skeptical false positives account, the original study was underpowered, lacked constraints against the use of researcher degrees of freedom, and was subject to perverse publication incentives. As a result, all of the claimed effects of the experimental manipulation—not only on the dependent variable but also hypothesized mediating processes—are unreliable and potentially spurious. There is no need to demonstrate an effect of the manipulation on the mediator, but not the DV, when that is not necessarily the claim being tested.

Suspicion and awareness

A potential interfering factor for Study 2's salvation prime effect is awareness of a potential influence attempt. On the numeric rating item from the funneled debriefing (Poehlman, 2007; Uhlmann et al., 2011), a full 56.2% of participants indicated they were either unsure or believed that the sentence-unscrambling task had influenced their subsequent responses (Supplement 9). This figure is far higher than the suspicion level of 5% or less recommended by Bargh and Chartrand (2000). As highlighted by Dijksterhuis (2018), a similar pattern occurred in the crowdsourced replication of the professor priming and intellectual performance effect, with 65% of participants suspicious or aware (O'Donnell et al., 2018). This raises the possibility that widespread publicity, for example in best-selling books like *Blink* (Gladwell, 2004), could be one culprit for the recent reproducibility tribulations of priming research. Strack (2016) raises a similar concern regarding the use of undergraduate students in replications of findings featured in introductory textbooks and lectures. Tierney et al. (in press) found that participants who reporting having been in a similar experiment before exhibited favoritism towards female job candidates, perhaps to avoid appearing biased or sexist, reversing the original pattern of results from Uhlmann and Cohen (2005, 2007). In the present replication, selecting participants who indicated on the numeric awareness probe that they did *not* believe they were influenced by the sentence-unscrambling task still revealed no evidence of a salvation prime effect (Supplement 9).

However, the high baseline level of awareness remains of concern not only for replicating priming effects, but any experimental finding subject to significant media attention and inclusion in educational curriculums. An ongoing replication ring (Kahneman, 2012) for prime-to-behavior effects will systematically assess participants' previous degree of experience with research studies, enrollment in psychology courses, and suspicions about the study hypothesis (Schweinsberg, Tierney, et al., 2020). Moving forward, we recommend that replication initiatives routinely include not only funneled debriefings about the specific effect in question (Bargh & Chartrand, 2000) but also general indices of study-savviness.

Replicator expertise

A number of the Study 2 data collections were carried out by experts on implicit social cognition and prime-to-behavior effects, ruling out the common counter-explanation that the effects did not emerge reliably due to a lack of replicator expertise (Bargh, 2012; Baumeister, 2016; Schnall, 2014; Weingarten et al., 2016; Wilson, 2014). This adds further weight to prior failed replications of religion priming (Billingsley, Gomes, & McCullough, 2018; Gomes & McCullough, 2015; Miyatake & Higuchi, 2017) and behavioral priming more generally (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, 2013; Rohrer et al., 2015).

References for Supplement 11 (not cited in main manuscript)

Bargh, J.A. (2012). Priming effects replicate just fine, thanks. *Psychology Today.* Retrieved April 11, 2017 at: https://www.psychologytoday.com/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66,* 153-158.

Dijksterhuis, A. (2018). Reflection on the Professor-Priming Replication Report. *Perspectives on Psychological Science, 13*(2), 295-296.

Ejelöv, E., & Luke, T.J. (2020). Rarely safe to assume: Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology, 87,* 103937.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8,* 370-378.

Gomes, C. M., & McCullough, M. E. (2015). The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. *Journal of Experimental Psychology: General, 144*(6), e94.

Gladwell, M. (2004). *Blink: The power of thinking without thinking*. New York: Little, Brown.

Kahneman, D. (2012). *A proposal to deal with questions about priming effects.* Retrieved at: http://www.decisionsciencenews.com/2012/10/05/kahneman-on-the-storm-of-doubts-surrounding-social-priming-research/ (August 4, 2018).

Miyatake, S., & Higuchi, M. (2017). Does religious priming increase the prosocial behaviour

of a Japanese sample in an anonymous economic game? *Asian Journal of Social Psychology, 20*(1), 54–59.

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

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

Schweinsberg, M., Tierney, W.T., Viganola, D., Ebersole, C., Hardy, J., Gordon, M., Dreber, A., Johannesson, M., Pfeiffer, T., et al., & Uhlmann, E. L. (2020). *Replication ring for priming effects on judgments and actions.* Registered report proposal.

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

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

Wilson, T. D. (2014). Is there a crisis of false negatives in psychology? Available at: https://timwilsonredirect.wordpress.com/

Zwaan, R. A., Etz, A., Lucas, R L., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, e120.

Supplement 12: Post-hoc analysis of response effort as a moderator of priming

We performed additional post-hoc analyses to evaluate the possibility that response effort may moderate the Salvation Prime Effect in Study 2 (see Huang, 2014). The analyses were performed using the PureProfile sample where time data were available to serve as proxy for response effort. Because no a priori cutoffs were set for participants' response effort, the following analysis is exploratory in nature. It should be noted that cutoffs for response effort below were determined prior to conducting relevant moderating analysis.

*Removal of participants suspected of insufficient effort responding***.** Although an instructional attention check item was employed to screen participants, this single item may fail to detect participants who failed to pay sufficient attention in the relevant study section. We adopted three rationally developed screening criteria to remove participants who likely engaged in insufficient effort responding (IER; Huang, Keeney, Curran, Poposki, & DeShon, 2012). These cutoffs include: (a) leaving blank more than half of the priming puzzles (i.e., 6 out of 12); (b) spending less than 36 seconds on the priming task (i.e., 3 seconds per sentence unscrambling puzzle); and (c) spending less than 40 seconds on the anagram task (i.e., 10 seconds per anagram). These cutoffs were quite lenient and should only remove the most egregious form of IER. The three sequential screening criteria removed 20, 0, and 6 respondents (26 in total; 2.30%). Subsequent analyses on response effort were based on the remaining respondents (N = 1104).

*Operationalization of response effort***.** We used the amount of time each participant spent on the priming task as a proxy to assess response effort. We focused on time on the priming task as opposed to time for the entire study (see Huang, 2014) because the salvation prime might influence participants' effort in subsequent part of the study. We performed a median split on time on the priming task to classify participants into low versus high response effort groups. Due to the lack of a pre-test to determine an appropriate cutoff, a median split allowed us to reasonably capture the overall difference between participants with lower versus higher response effort.

Response effort as a moderator. We examined whether response effort (low vs. high) moderated the salvation priming effect on anagram performance. The moderating effects of response effort reported below mirror the order of results reported on the PureProfile sample in Study 2.

Overall, response effort did not moderate the salvation versus neutral prime conditions to influence anagram performance, $F(1, 1089.46) = 2.155$, $p = 0.142$, $d = 0.089$. The three-way interaction involving response effort, priming manipulation, and country was nonsignificant comparing USA vs other nation (UK & Australia) $F(1, 1087.39) = 2.387$, $p = 0.123$, $d = 0.094$, and USA vs Australia, $F(1, 777.17) = 0.368$, $p = 0.544$, $d = 0.044$, but was significant comparing USA vs UK, $F(1, 798.22) = 3.988$, $p = 0.046$, $d = 0.141$. Although response effort interacted with the salvation prime to influence task performance in the USA, $F(1, 493.3) = 5.838$, $p = 0.016$, $d = 0.218$, the salvation prime effect was nonsignificant in either response effort condition: in the low response effort condition, $F(1, 242.9) = 2.663$, $p = 0.104$, $d = -0.209$; whereas in the high response effort condition, $F(1, 251) = 3.089$, $p = 0.080$, $d = 0.222$. In contrast, response effort failed to moderate the salvation prime effect on task performance in the United Kingdom, $F(1, 310) = 0.530$, $p = 0.467$, $d = -0.082$; or in Australia, $F(1, 289) = 0.505$, p

= 0.478, d = 0.084. The three-way interaction involving response effort, New England region, and prime condition was nonsignificant, $F(1, 1079.68) = 0.003$, $p = 0.959$, $d = -0.003$.

Next, we examined the three-way interactions between response effort, prime condition, and moderator measures. None of the three-way interactions was significant, including response effort by prime condition by Protestant faith, $F(1, 1077.49) = 0.458$, $p = 0.498$, $d = -0.041$; response effort by prime condition by the single item measure of religiosity, $F(1, 1087.33) = 0.170$, $p = 0.681$, $d = 0.025$; response effort by prime condition by DUREL religiosity scale, $F(1, 1086.61) = 1.111$, $p = 0.292$, $d = 0.064$; and response effort by prime condition by PWE, $F(1, 1087.62) = 0.980$, $p = 0.322$, $d = 0.060$.

Reference for Supplement 12

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

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Supplement 13: Summaries of other creative destruction projects

Below we describe several other completed creative destruction replication initiatives as well as their results. Similar to the present project, these initiatives failed to find support for the original theory, instead supporting one or more of the competing theoretical accounts (see Tierney et al., in press, for a more in-depth review).

Motivated reasoning and child care decisions

The first project pitted the motivated reasoning account of how people process scientific evidence against the cognitive confirmation bias account and accuracy-driven reasoning accounts (Ebersole, 2019; Tierney et al., in press). We attempted to replicate the "wishful thinking" effect that desired outcomes are more important than cognitive beliefs in driving how people reason about scientific evidence (Bastardi, Uhlmann, & Ross, 2011). Relying on the biased assimilation paradigm (Lord, Ross, & Lepper, 1979) we manipulated the methodology and conclusions of ostensive research studies examining the downstream consequences of home care vs. day care for children's development. Of interest was whether intended parents who planned to use day care for their future children, but believed home care was superior, would prefer the methodology of studies finding day care was okay for kids vs. detrimental. Expanding beyond the original sample (Bastardi et al., 2011) we recruited not only individuals planning to become parents but also those who were already parents at the time of the study, many of whom had already made their child care decisions. According to theories of motivated reasoning, such as Festinger's (1962) theory of cognitive dissonance, individuals should be more prone to rationalize their actions in high-stakes as opposed to hypothetical situations. The results of the replication initiative resoundingly supported the cognitive confirmation bias account: only prior factual beliefs, not desired conclusions or parental status, drove the processing of evidence.

Motivated gender discrimination

Another investigation (Tierney et al., in press) sought to replicate our earlier findings that decision makers construct criteria biased against female job candidates, especially if led to believe they are an objective and rational person (Uhlmann & Cohen, 2005, 2007). The original theory posits that evaluators engage in motivated rationalizations for discrimination against women, and further suffer from an illusion of objectivity regarding their biases. The new study repeated the original hiring paradigm, but included additional conditions and measures, among these an affirmation-threat manipulation (Steele, 1988) and individual-differences measures of exposure to the #MeToo movement and endorsement of gender ideologies (McCormick-Huhn & Shields, 2019). The motivated discrimination account was pitted against the cognitive schema account in which group stereotypes influence perceptions of candidate characteristics, motivated liberalism account in which feminist ideologies lead to reverse discrimination against male candidates, and study-savviness account positing participants are suspicious the study is about gender bias and overcorrect their judgments to avoid appearing sexist. The empirical results of the replication study were a mirror-image reversal of the findings originally reported by Uhlmann and Cohen (2005, 2007). Specifically, male evaluators shifted their hiring criteria in favor of *female* candidates, and were also more likely to select women than men for the job. These reverse gender biases were exacerbated when evaluators were led to feel objective.

Rejection of societal sexism and prior experience with research studies predicted favoritism towards female candidates, supporting the motivated liberalism and study-savviness accounts. Providing some partial support for the motivated discrimination account, a self-threat (relatively to a self-affirmation) caused male evaluators' hiring evaluations of female candidates to become less positive.

References for Supplement 13 (not cited in main manuscript)

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

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

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

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

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

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

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

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

Supplement 14: Demographic details regarding study samples

On the following pages, please find summary demographic tables for the replication samples for Studies 1 and 2 (Tables S14-1 and S14-2).

Table S14-1. Demographics for the replication samples for the needless work, tacit inferences, and intuitive work morality studies.

		MTurk India	MTurk USA	PureProfile USA	PureProfile Australia	PureProfile UK
Sample	N	1000	1036	2127	1011	960
Age	Mean	30.51	39.48	49.62	46.24	48.05
	Standard deviation	7.12	12.36	16.61	16.35	15.25
Gender	Male	718 (71.8%)	494 (47.68%)	566 (26.61%)	331 (32.74%)	389 (40.52%)
	Female	281 (28.1%)	536 (51.74%)	1552 (72.97%)	679 (67.16%)	564 (58.75%)
	Other	0 (0%)	4 (0.39%)	4 (0.19%)	1 (0.10%)	2 (0.21%)
	No response	1 (0.01%)	2 (0.19%)	5 (0.24%)	0 (0%)	5 (0.52%)
Religion	Agnostic	5 (0.50%)	230 (22.2%)	188 (8.84%)	126 (12.46%)	105 (10.94%)
	Atheist	44 (4.40%)	206 (19.88%)	129 (6.06%)	257 (25.42%)	240 (25%)
	Buddhism	50 (5.00%)	10 (0.97%)	29 (1.36%)	26 (2.57%)	7 (0.73%)
	Catholic	290 (29.00%)	164 (15.83%)	619 (29.1%)	187 (18.5%)	133 (13.85%)
	Islam	71 (7.10%)	10 (0.97%)	16 (0.75%)	18 (1.78%)	11 (1.15%)
	Judaism	10 (1.00%)	11 (1.06%)	57 (2.68%)	7 (0.69%)	4 (0.42%)
	Other (please indicate)	456 (45.60%)	120 (11.58%)	522 (24.54%)	225 (22.26%)	150 (15.63%)
	Protestant	72 (7.20%)	285 (27.51%)	562 (26.42%)	153 (15.13%)	305 (31.77%)
	No response	2 (0.20%)	0 (0%)	5 (0.24%)	12 (1.19%)	5 (0.52%)
Income	Less than $10,000 USD a year	413 (41.30%)	43 (4.15%)	183 (8.60%)	61 (6.03%)	91 (9.48%)
	USD $20,000-$40,000	218 (21.80%)	111 (10.71%)	222 (10.44%)	112 (11.08%)	142 (14.79%)
	USD $40,000-$60,000	163 (16.30%)	287 (27.7%)	500 (23.51%)	219 (21.66%)	298 (31.04%)
	USD $60,000-$80,000	98 (9.80%)	208 (20.08%)	417 (19.61%)	188 (18.6%)	195 (20.31%)
	USD $80,000-$100,000	43 (4.30%)	162 (15.64%)	297 (13.96%)	157 (15.53%)	105 (10.94%)
	USD $100,000 a year or more	17 (1.70%)	97 (9.36%)	182 (8.56%)	102 (10.09%)	50 (5.21%)
	No response	48 (4.80%)	128 (12.36%)	326 (15.33%)	172 (17.01%)	79 (8.23%)
Political Views	Very Progressive/Left-wing	74 (7.40%)	165 (15.93%)	159 (7.48%)	57 (5.64%)	55 (5.73%)
	Progressive/Left-wing	66 (6.60%)	226 (21.81%)	224 (10.53%)	114 (11.28%)	95 (9.9%)
	Somewhat Progressive/Left-wing	105 (10.50%)	158 (15.25%)	179 (8.42%)	104 (10.29%)	148 (15.42%)

	Moderate/Centrist	325 (32.50%)	215 (20.75%)	803 (37.75%)	486 (48.07%)	419 (43.65%)
	Somewhat Conservative/Right-wing	165 (16.50%)	105 (10.14%)	290 (13.63%)	131 (12.96%)	135 (14.06%)
	Conservative/Right-wing	146 (14.60%)	112 (10.81%)	232 (10.91%)	67 (6.63%)	56 (5.83%)
	Very Conservative/Right-wing	70 (7.00%)	52 (5.02%)	194 (9.12%)	17 (1.68%)	17 (1.77%)
	No response	49 (4.90%)	3 (0.29%)	46 (2.16%)	35 (3.46%)	35 (3.65%)
Education	Some high school/secondary school	25 (2.50%)	8 (0.77%)	54 (2.54%)	108 (10.68%)	72 (7.50%)
	High school degree/completed secondary school	15 (1.50%)	141 (13.61%)	513 (24.12%)	272 (26.9%)	347 (36.15%)
	Some university	71 (7.10%)	277 (26.74%)	575 (27.03%)	127 (12.56%)	98 (10.21%)
	University degree	308 (30.80%)	407 (39.29%)	539 (25.34%)	221 (21.86%)	228 (23.75%)
	Some graduate/postgraduate education	318 (31.80%)	66 (6.37%)	129 (6.06%)	126 (12.46%)	52 (5.42%)
	Graduate/postgraduate degree (e.g., doctoral degree)	238 (23.80%)	106 (10.23%)	241 (11.33%)	93 (9.20%)	88 (9.17%)
	No response	25 (2.50%)	31 (2.99%)	76 (3.57%)	64 (6.33%)	75 (7.81%)

Table S14-2. Demographics for replications of the salvation prime effect.

	PureProfile USA	PureProfile Australia	PureProfile UK	Crowdsourced Labs USA	Crowdsourced Lab Canada	Crowdsourced Lab Ireland
N	514	298	312	563	91	80
Mean	47.21	44.53	47.66	19.54	21.09	20.23
Standard deviation	16.46	17.19	15.75	3.15	2.73	5.18
Male	131 (25.49%)	94 (31.54%)	103 (33.01%)	220 (39.08%)	27 (29.67%)	22 (27.5%)
Female	378 (73.54%)	203 (68.12%)	206 (66.03%)	338 (60.04%)	63 (69.23%)	58 (72.5%)
Other	1 (0.19%)	0 (0%)	2 (0.64%)	4 (0.71%)	1 (1.1%)	0 (0%)
No response	4 (0.78%)	1 (0.34%)	1 (0.32%)	1 (0.18%)	(0%)	0 (0%)
Agnostic	51 (9.92%)	28 (9.4%)	38 (12.18%)	41 (7.28%)	7 (7.69%)	4 (5%)
Atheist	25 (4.86%)	78 (26.17%)	69 (22.12%)	50 (8.88%)	16 (17.58%)	14 (17.5%)
Buddhism	4 (0.78%)	16 (5.37%)	0 (0%)	11 (1.95%)	3 (3.3%)	1 (1.25%)
Catholic	167 (32.49%)	43 (14.43%)	38 (12.18%)	207 (36.77%)	13 (14.29%)	54 (67.5%)
Islam	4 (0.78%)	4 (1.34%)	4 (1.28%)	34 (6.04%)	11 (12.09%)	1 (1.25%)
Judaism	17 (3.31%)	2 (0.67%)	0 (0%)	62 (11.01%)	1 (1.1%)	0 (0%)
Other	114 (22.18%)	86 (28.86%)	56 (17.95%)	87 (15.45%)	19 (20.88%)	4 (5%)
Protestant	131 (25.49%)	39 (13.09%)	103 (33.01%)	57 (10.12%)	18 (19.78%)	2 (2.5%)
No response	1 (0.19%)	2 (0.67%)	4 (1.28%)	14 (2.49%)	3 (3.3%)	0 (0%)
Less than $10,000 USD a year	45 (8.75%)	19 (6.38%)	37 (11.86%)	64 (11.37%)	15 (16.48%)	28 (35%)
USD $20,000-$40,000	114 (22.18%)	66 (22.15%)	112 (35.9%)	66 (11.72%)	10 (10.99%)	11 (13.75%)
USD $40,000-$60,000	91 (17.7%)	54 (18.12%)	60 (19.23%)	67 (11.9%)	20 (21.98%)	21 (26.25%)
USD $60,000-$80,000	73 (14.2%)	56 (18.79%)	27 (8.65%)	59 (10.48%)	6 (6.59%)	4 (5%)
USD $80,000-$100,000	39 (7.59%)	27 (9.06%)	16 (5.13%)	68 (12.08%)	11 (12.09%)	6 (7.5%)
USD $100,000 a year or more	76 (14.79%)	39 (13.09%)	13 (4.17%)	157 (27.89%)	15 (16.48%)	2 (2.5%)
No response	4 (0.78%)	5 (1.68%)	6 (1.92%)	38 (6.75%)	4 (4.4%)	2 (2.5%)
Very Progressive/Left-wing	43 (8.37%)	12 (4.03%)	17 (5.45%)	18 (3.2%)	7 (7.69%)	3 (3.75%)
Progressive/Left-wing	47 (9.14%)	34 (11.41%)	42 (13.46%)	123 (21.85%)	15 (16.48%)	1 (1.25%)

Somewhat Progressive/Left-wing	41 (7.98%)	36 (12.08%)	53 (16.99%)	65 (11.55%)	14 (15.38%)	15 (18.75%)
Moderate/Centrist	217 (42.22%)	166 (55.7%)	141 (45.19%)	215 (38.19%)	25 (27.47%)	36 (45%)
Somewhat Conservative/Right-wing	54 (10.51%)	33 (11.07%)	39 (12.5%)	66 (11.72%)	9 (9.89%)	5 (6.25%)
Conservative/Right-wing	59 (11.48%)	13 (4.36%)	16 (5.13%)	35 (6.22%)	12 (13.19%)	4 (5%)
Very Conservative/Right-wing	52 (10.12%)	4 (1.34%)	3 (0.96%)	7 (1.24%)	3 (3.3%)	0 (0%)
No response	1 (0.19%)	(0%)	1 (0.32%)	34 (6.04%)	9 (9.89%)	16 (20%)
Some high school/secondary school	17 (3.31%)	34 (11.41%)	25 (8.01%)	8 (1.42%)	3 (3.3%)	0 (0%)
High school degree/completed secondary school	115 (22.37%)	85 (28.52%)	115 (36.86%)	161 (28.6%)	12 (13.19%)	24 (30%)
Some university	140 (27.24%)	50 (16.78%)	32 (10.26%)	363 (64.48%)	59 (64.84%)	39 (48.75%)
University degree	129 (25.1%)	67 (22.48%)	87 (27.88%)	24 (4.26%)	13 (14.29%)	13 (16.25%)
Some graduate/postgraduate education	35 (6.81%)	40 (13.42%)	21 (6.73%)	5 (0.89%)	3 (3.3%)	3 (3.75%)
Graduate/postgraduate degree (e.g., doctoral degree)	73 (14.2%)	20 (6.71%)	26 (8.33%)	2 (0.36%)	1 (1.1%)	0 (0%)
No response	5 (0.97%)	2 (0.67%)	6 (1.92%)	(0%)	(0%)	1 (1.25%)