Supplements for “A creative destruction approach to replication: Implicit work and sex morality across cultures”
# Table of Contents

Supplement 1: Materials for Needless Work, Tacit Inferences, and Intuitive Work Morality Replications (Study 1) .................................................3  

Supplement 2: Materials for Salvation Prime Replication (Study 2) .................................................29  

Supplement 3: Pre-Registered Analysis Plan for Replication Project .................................................34  

Supplement 4: Pre-Registered Analysis Plan for the Forecasting Survey .................................................65  

Supplement 5: Forecasting Survey .................................................70  

Supplement 6: Detailed Report of the Forecasting Results .................................................89  

Supplement 7: Further Analyses of the Replication Results .................................................99  

Supplement 8: Pre-Registered Plan for Bayesian Multiverse Analysis .................................................102  

Supplement 9: Bayesian Multiverse Analysis of Project Results .................................................121  

Supplement 10: Departures from Pre-Registered Replication Plan .................................................139  

Supplement 11: Further Discussion of the Priming Failed Replication .................................................142  

Supplement 12: Post-Hoc Analysis of Response Effort as a Moderator of Priming .................................................145  

Supplement 13: Summaries of Other Creative Destruction Projects .................................................147  

Supplement 14: Demographic Details for Study Samples .................................................149
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
1 2 3 4 5 6
Very Good
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
1 2 3 4 5 6
Very Good
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?

<table>
<thead>
<tr>
<th>Very Bad</th>
<th>Very Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

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?

<table>
<thead>
<tr>
<th>Very Bad</th>
<th>Very Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
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.
- 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.
• 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.
• 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.
• 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 INFERENCE 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.
• 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.
• 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.
• 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.
• 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:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robert is a much better person than John</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John is a much better person than Robert</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

My intuitive, gut feeling is that:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robert is a much better person than John</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John is a much better person than Robert</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

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
  1 2 3 4 5 6 7
  Very Religious

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 £73,893 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
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?

<table>
<thead>
<tr>
<th>NO</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Not Sure</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>YES</th>
<th>9</th>
</tr>
</thead>
</table>

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 EXCLUSION 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,
SUPPLEMENTS: CREATIVE DESTRUCTION APPROACH TO REPLICATION

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 alternative 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:

<table>
<thead>
<tr>
<th>Not at all Religious</th>
<th>Very Religious</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

The multi-item measure of religiosity is the five-item DUREL scale (Koenig & Büsing, 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

<table>
<thead>
<tr>
<th>#</th>
<th>Theory Test</th>
<th>Description of Analysis</th>
<th>Model and Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overall analysis of</td>
<td>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).</td>
<td>lmer(DV ~ Age + Country + Work_Status + Age:Work_Status:Country + (1 + Age:Work_Status</td>
</tr>
<tr>
<td></td>
<td>study</td>
<td></td>
<td>* 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.</td>
</tr>
<tr>
<td>2</td>
<td>Key effect</td>
<td>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.</td>
<td>lmer(DV ~ Age:Work_Status* + (1 + Age:Work_Status</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Note only the continues to work condition will be kept in the Work_Status variable.</td>
</tr>
<tr>
<td>3</td>
<td>Implicit Puritanism</td>
<td>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.</td>
<td>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).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* Note only the continue to work condition will be kept in the Work_Status variable.</td>
</tr>
<tr>
<td>4</td>
<td>Religious Differences</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>lmer(DV ~ Age:Work_Status*Rel</strong> + (1 + Age:Work_Status</td>
<td>Region), data = mydata)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>lmer(DV ~ Age:Work_Status:DUREL</strong> + (1 + Age:Work_Status</td>
<td>Region), data = mydata)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>lmer(DV ~ Age:Work_Status:REL_Id</strong>* + (1 + Age:Work_Status</td>
<td>Region), data = mydata)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Note only the continue to work condition will be kept in the Work_Status variable</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rel denotes answers to the question “I consider myself to be: Not at all Religious – Very Religious”.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DUREL denotes computed average of the answers to the DUREL scale (Koenig &amp; Büssing, 2010)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rel_Id denotes the answers to “My religion is:”, with Protestant = 0 and Other religions = 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5</th>
<th>Regional Differences</th>
<th>Key effect (Table S3-1 Row 2) greater for New England participants than for non-New England participants.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>lmer(DV ~ Age:Work_Status*New_Eng</strong> + (1 + Age:Work_Status</td>
<td>Region), data = mydata)**</td>
<td></td>
</tr>
<tr>
<td><em>Note only the continue to work condition will be kept in the Work_Status variable</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>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.</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>Social Class</th>
<th>Key effect (Table S3-1 Row 2) greater for high-SES participants than low-SES participants.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>lmer(DV ~ Age:Work_Status*Edu</strong> + (1 + Age:Work_Status</td>
<td>Region), data = mydata)**</td>
<td></td>
</tr>
<tr>
<td><em>Note only the continue to work condition will be kept in the Work_Status variable</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Edu denotes answers to the question “my educational level is”</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7</th>
<th>Explicit American Exceptionalism</th>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>lmer(DV ~ Age:Work_Status*Country</strong> + (1 + Age:Work_Status</td>
<td>Region), data = mydata)**</td>
<td></td>
</tr>
<tr>
<td>**lmer(DV ~ Age:Country:Work_Status: PWE *** + (1 + Age:Work_Status</td>
<td>Region), data = mydata)**</td>
<td></td>
</tr>
</tbody>
</table>
### General Moralization of Work

Key effect (Table S3-1 Row 2) present both for Americans and non-Americans

\[
lmer(DV \sim Age: \text{Work Status}\cdot\text{Country}^* + (1 + Age: \text{Work Status}|\text{Region}), \text{data} = \text{mydata})
\]

* 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)

---

### 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.

\[
lmer(DV \sim Age: \text{Work Status}\cdot\text{Country}^* + (1 + Age: \text{Work Status}|\text{Region}), \text{data} = \text{mydata})
\]

* 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.

---

### 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 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

<table>
<thead>
<tr>
<th>#</th>
<th>Theory Test</th>
<th>Description of Analysis</th>
<th>Model and Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overall analysis of study</td>
<td>Tacit inferences condition (Condition 1 vs. Condition 2) x Country (US vs Other), with</td>
<td>lmer(DV* ~ Condition**:Country*** + (1 + Condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>false memories as the outcome measure.</td>
<td>* Count of True responses to key questions across the four vignettes (1 per vignette)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>** Condition is a between-subjects factor manipulating whether vignette targets uphold or violate traditional morality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*** 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.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>**** 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.</td>
</tr>
<tr>
<td>2</td>
<td>Key effect</td>
<td>The key effect is memory differences between the two conditions manipulating whether</td>
<td>lmer(DV ~ Condition + (1 + Condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>targets uphold or violate traditional morality. Specifically, individuals who violate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>work morality should be misremembered as also violating sexual morality, and vice versa.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Such false memories reflect an implicit link between work and sex morality. As detailed</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>below, the competing theories have different predictions regarding this focal effect.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Implicit Puritanism</td>
<td>American, but not non-American participants, should exhibit an implicit link between</td>
<td>lmer(DV ~ Condition:Country + (1 + Condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>work and sex morality. The critical statistical test is</td>
<td></td>
</tr>
</tbody>
</table>
| 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:

<table>
<thead>
<tr>
<th></th>
<th>Robert is a much better person than John</th>
<th>John is a much better person than Robert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

My intuitive, gut feeling is that:

<table>
<thead>
<tr>
<th></th>
<th>Robert is a much better person than John</th>
<th>John is a much better person than Robert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

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 not 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 not 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

<table>
<thead>
<tr>
<th>#</th>
<th>Theory Test</th>
<th>Description of Analysis</th>
<th>Model and Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overall analysis of study</td>
<td>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.</td>
<td>lmer(DV ~ Condition** : Country*** + (1 + Condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>* Answers to the question 7 = John is a much better person than Robert</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>** Condition is a within-subjects factor manipulating whether participants are asked for their intuitive or rational response.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>*** 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.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>**** Identifying number of participants once the data is restructured</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>***** 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.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Key effect</td>
<td>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.</td>
<td>lmer(DV ~ Condition + (1 + Condition</td>
</tr>
<tr>
<td>3</td>
<td>Implicit Puritanism</td>
<td>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.</td>
<td>lmer(DV ~ Condition:Country + (1 + Condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Religious Differences</td>
<td>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</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lmer(DV ~ Condition:Rel + (1 + Condition</td>
<td>subj) + (1 + Condition</td>
</tr>
<tr>
<td>5</td>
<td>Regional Differences</td>
<td>Key effect (Table S3-3 Row 2) greater for New England participants than for non-New England participants</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lmer(DV ~ Condition:New_Eng* + (1 + Condition</td>
<td>subj) + (1 + Condition</td>
</tr>
<tr>
<td>6</td>
<td>Social Class</td>
<td>Key effect (Table S3-3 Row 2) greater for high-SES participants than low-SES participants</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lmer(DV ~ Condition:Edu* + (1 + Condition</td>
<td>subj) + (1 + Condition</td>
</tr>
<tr>
<td>7</td>
<td>Explicit American Exceptionalism</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lmer(DV ~ Condition:Country* + (1 + Condition</td>
<td>subj) + (1 + Condition</td>
</tr>
</tbody>
</table>
|   | General Moralization of Work | Key effect (Table S3-3 Row 2) present both for Americans and non-Americans | \( \text{lm} \text{er(DV} ~ \text{Condition: Country}^* + (1 + \text{Condition}|\text{subj}) + (1 + \text{Condition}|\text{Region}), \text{data = mydata}) \)
|   |                             | * A categorical variable whereby US = 0 and all the other countries will be 1. |

|   | 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. | \( \text{lm} \text{er(DV} ~ \text{Condition: Country}^* + (1 + \text{Condition}|\text{subj}) + (1 + \text{Condition}|\text{Region}), \text{data = mydata}) \)
|   |                             | * A categorical variable whereby US = 0 and India will be 1, for the MTurk sample. |

|   | 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

<table>
<thead>
<tr>
<th>#</th>
<th>Theory Test</th>
<th>Description of Analysis</th>
<th>Model and Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overall analysis of study</td>
<td>2 prime condition (religious vs. neutral concepts) x Country (USA vs. non-USA), with anagram performance as the outcome measure.</td>
<td>lmer(DV* ~ Rel_Prime** + Country*** + Rel_Prime:Country + (1 + Rel_Prime</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*Total number of anagrams solved</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>** Rel_Prime is a variable that indicates whether the participant was in the religion or neutral condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*** Here and for all of Tables S3-4, “Country” refers to a categorical variable whereby US = 0 and all the other countries will be 1.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>****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.</td>
</tr>
<tr>
<td>2</td>
<td>Key effect</td>
<td>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.</td>
<td>lmer(DV ~ Rel_Prime + (1 + Rel_Prime</td>
</tr>
<tr>
<td>3</td>
<td>Implicit Puritanism</td>
<td>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.</td>
<td>lmer(DV ~ Rel_Prime:Country + (1 + Rel_Prime</td>
</tr>
</tbody>
</table>
**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.

\[
\text{lmer(DV ~ Rel\_Prime:Rel}^* + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]
\[
\text{lmer(DV ~ Rel\_Prime:DUREL}^{**} + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]
\[
\text{lmer(DV ~ Rel\_Prime:REL\_Id}^{***} + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]

*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

**Regional Differences**

Key effect (Table S3-4 Row 2) greater for New England participants than for non-New England participants.

\[
\text{lmer(DV ~ Rel\_Prime: New\_Eng}^{**} + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]

*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.

**Explicit American Exceptionalism**

Key effect (Table S3-4 Row 2) not present, and not moderated by the Protestant work ethic (PWE).

\[
\text{lmer(DV ~ Rel\_Prime:Country}^* + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]
\[
\text{lmer(DV ~ Rel\_Prime:PWE}^* + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]

*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)

**General Moralization of Work**

Key effect (Table S3-4 Row 2) present both for Americans and non-Americans.

\[
\text{lmer(DV ~ Rel\_Prime:Country}^* + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]

*A categorical variable whereby US = 0 and all the other countries will be 1.

**False Positives**

Key effect (Table S3-4 Row 2) not present for any country or relevant sub-region (e.g., New England).

\[
\text{lmer(DV ~ Rel\_Prime:Country}^* + (1 + \text{Rel\_Prime}|\text{Region}), \text{data = mydata)}
\]
DETERRMINING 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

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

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.

<table>
<thead>
<tr>
<th>THEORY</th>
<th>USA sample</th>
<th>Other national culture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New England</td>
<td>Other USA states</td>
</tr>
<tr>
<td>Implicit Puritanism + Religious Differences predictions combined additively</td>
<td>Scenario: +2 Priming: +2</td>
<td>Scenario: +2 Priming: +2</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>THEORY</th>
<th>USA sample</th>
<th>Other national culture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New England</td>
<td>Other USA states</td>
</tr>
<tr>
<td>Implicit Puritanism + Regional Folkways combined additively</td>
<td>Scenario: +2 Priming: +2</td>
<td>Scenario: +1 Priming: +1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>USA sample</th>
<th>Other national culture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New England</td>
<td>Other USA states</td>
</tr>
<tr>
<td>Protestant/ Religious</td>
<td>Scenario: +1 Priming: 0</td>
<td>Scenario: +1 Priming: 0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>THEORY</th>
<th>USA sample</th>
<th>Other national culture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New England</td>
<td>Other USA states</td>
</tr>
<tr>
<td>Protestant/ Religious</td>
<td>Non-Protestant/ Less Religious</td>
<td>Protestant/ Religious</td>
</tr>
<tr>
<td>Implicit Puritanism + General Moralization of Work and Sex combined additively</td>
<td>Scenario: +2 Priming: +2</td>
<td>Scenario: +2 Priming: +2</td>
</tr>
<tr>
<td>Scenario: +2 Priming: +2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 +
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 ($\sim100\%$) to detect even the smallest previously observed Implicit Puritanism effect ($d = .37$). However, with added intra-region heterogeneity ($d +/-.4$) our power falls to $\sim37\%$ 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 ($\sim72\%$, 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.

Alternative 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)

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

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


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:

\[
RES_{hc} = \beta_0 + \beta_1 PES_{thc} + \epsilon_{thc}
\]

where:
- \( RES_{hc} \) is a continuous variable indicating the realized effect size of the hypothesis \( h \) object of the prediction in population \( c \);
- \( PES_{thc} \) 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 \( \beta_1 \) 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).

$$BS_{ihc} = \beta_0 + \beta_1 MES_{ihc} + FE_i + \epsilon_{ihc}$$

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 $PES_{ihc}^{m}$, $h$ 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 $PES_{ihc}^{m}$ 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):

\[
BS_{ihc} = \beta_0 + \beta_1 USNE_c + \beta_2 USNNE_c + \beta_3 US_c + \beta_4 AUS_c + \beta_5 IN_c + FE_i + \varepsilon_{ihc}
\]

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 $\beta_5$ in regression equation (3); Wald test on coefficients $\beta_i$ being different from $\beta_j$ 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.

\[
BS_{ihc} = \beta_0 + \beta_1 SEN_i + \varepsilon_{ihc}
\]

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:

\[ \$200 - (S\text{q.Error} \times 200) \]

where \( \text{Sq.Error} \) 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


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.
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:


Full text: socialjudgments.com/docs/AME%20CHAPTER.POSTING.pdf

Full text: http://socialjudgments.com/docs/Final%20Theism%20Chapter.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:

\[ 200 - (\text{Sq.Error} \times 200) \]

where \(\text{Sq.Error}\) 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 \(\text{Sq.Error}\) 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:

Complete study materials for replication:

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

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:

Complete study materials for replication:

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 **Protestants** 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 **religious participants** 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 **endorse the Protestant Work Ethic** (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 educated** 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].

[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:

<table>
<thead>
<tr>
<th></th>
<th>Robert is a much better person than John</th>
<th>John is a much better person than Robert</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

My intuitive, gut feeling is that:

<table>
<thead>
<tr>
<th></th>
<th>Robert is a much better person than John</th>
<th>John is a much better person than Robert</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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:**

**Complete study materials for replication:**

**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].

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:**

**Complete study materials for replication:**

**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 **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 tacit inferences 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 tacit inferences 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 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 **Protestants** 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 **religious participants** 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 **endorse the Protestant Work Ethic** (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 **educated 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-cellled 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:*

*Complete study materials for replication:*

*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 educated 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.

- Yes, you may use my anonymized data in this research
- 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:

\[ 200 - (\text{Sq.Error} \times 200) \]

where \( \text{Sq.Error} \) 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 \(\text{SD} = 5.6\). 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/](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>
<thead>
<tr>
<th>Table S6-1. Association between forecasted and observed effect sizes.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable:</strong></td>
</tr>
<tr>
<td><strong>Realized effect size</strong></td>
</tr>
<tr>
<td>Forecasted effect size</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>(R^2)</td>
</tr>
</tbody>
</table>

*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 ($\beta = 0.164, p = 0.002$). 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 ($\beta = 0.019, p = 0.04$).

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).

<table>
<thead>
<tr>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brier Score</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forecasts for moderator effects</th>
<th>0.008</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.013)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observations</th>
<th>10608</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

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)

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

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 (β1
to $\beta_5$) 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 ($\beta_1 = -0.008, p = 0.351; \beta_2 = -0.023, p = 0.188; \beta_3 = 0.083, p < 0.0001; \beta_4 = 0.016, p = 0.0003; \beta_5 = 0.042, p < 0.0001$). 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.

<table>
<thead>
<tr>
<th>Dependent variable: Brier Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA – New England States (PureProfile)</td>
</tr>
<tr>
<td>USA – Non-New England States (PureProfile)</td>
</tr>
<tr>
<td>USA (MTurk)</td>
</tr>
<tr>
<td>Australia (PureProfile)</td>
</tr>
<tr>
<td>India (MTurk)</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>$R^2$</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>USA – New England States (PureProfile)</th>
<th>USA – Non-New England States (PureProfile)</th>
<th>USA (MTurk) (PureProfile)</th>
<th>India (MTurk)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA – New England States</td>
<td>-</td>
<td>0.1704</td>
<td>&lt;0.0001</td>
<td>0.0016</td>
</tr>
<tr>
<td>States (PureProfile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA – Non-New England</td>
<td>-</td>
<td></td>
<td>0.0007</td>
<td>0.0100</td>
</tr>
<tr>
<td>States (PureProfile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA (MTurk)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0026</td>
</tr>
<tr>
<td>(PureProfile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia (MTurk)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>( PureProfile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India (MTurk)</td>
<td>-</td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>(MTurk)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td></td>
<td>Full Sample</td>
<td>Simple effects</td>
<td>Moderators effects</td>
</tr>
<tr>
<td>Full Professor</td>
<td>-0.360</td>
<td>-0.308</td>
<td>-0.432</td>
</tr>
<tr>
<td></td>
<td>(0.234)</td>
<td>(0.225)</td>
<td>(0.248)</td>
</tr>
<tr>
<td>Associate Professor</td>
<td>-0.316</td>
<td>-0.250</td>
<td>-0.409</td>
</tr>
<tr>
<td></td>
<td>(0.234)</td>
<td>(0.225)</td>
<td>(0.248)</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td>-0.313</td>
<td>-0.266</td>
<td>-0.379</td>
</tr>
<tr>
<td></td>
<td>(0.234)</td>
<td>(0.225)</td>
<td>(0.248)</td>
</tr>
<tr>
<td>Postdoctoral researcher</td>
<td>-0.325</td>
<td>-0.273</td>
<td>-0.398</td>
</tr>
<tr>
<td></td>
<td>(0.234)</td>
<td>(0.225)</td>
<td>(0.248)</td>
</tr>
<tr>
<td>Graduate student</td>
<td>-0.240</td>
<td>-0.218</td>
<td>-0.271</td>
</tr>
<tr>
<td></td>
<td>(0.242)</td>
<td>(0.231)</td>
<td>(0.262)</td>
</tr>
<tr>
<td>Research Assistant</td>
<td>-0.291</td>
<td>-0.266</td>
<td>-0.327</td>
</tr>
<tr>
<td></td>
<td>(0.234)</td>
<td>(0.225)</td>
<td>(0.250)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.408*</td>
<td>0.368*</td>
<td>0.466</td>
</tr>
<tr>
<td></td>
<td>(0.234)</td>
<td>(0.225)</td>
<td>(0.248)</td>
</tr>
<tr>
<td>Observations</td>
<td>10680</td>
<td>6188</td>
<td>4420</td>
</tr>
<tr>
<td>R^2</td>
<td>0.026</td>
<td>0.021</td>
<td>0.034</td>
</tr>
</tbody>
</table>

*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.

<table>
<thead>
<tr>
<th></th>
<th>Needless work main effect (works vs. retires)</th>
<th>Target age and needless work effect</th>
<th>Intuitive work morality effect</th>
<th>Tacit inferences effect</th>
<th>Salvation primes and work behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New England U.S. States (PP)</strong></td>
<td>Mean Forecast: 0.389 Actual Effect Size: 0.622</td>
<td>0.295</td>
<td>0.292</td>
<td>0.340</td>
<td>0.099</td>
</tr>
<tr>
<td><strong>Non New England U.S. States (PP)</strong></td>
<td>Mean Forecast: 0.333 Actual Effect Size: 0.573</td>
<td>0.248</td>
<td>0.244</td>
<td>0.312</td>
<td>0.098</td>
</tr>
<tr>
<td><strong>All USA (MTurk)</strong></td>
<td>Mean Forecast: 0.346 Actual Effect Size: 0.658</td>
<td>0.275</td>
<td>0.259</td>
<td>0.300</td>
<td>-</td>
</tr>
<tr>
<td><strong>Australia (PP)</strong></td>
<td>Mean Forecast: 0.293 Actual Effect Size: 0.771</td>
<td>0.210</td>
<td>0.246</td>
<td>0.261</td>
<td>0.079</td>
</tr>
<tr>
<td><strong>India (MTurk)</strong></td>
<td>Mean Forecast: 0.266 Actual Effect Size: 0.666</td>
<td>0.200</td>
<td>0.212</td>
<td>0.349</td>
<td>-</td>
</tr>
<tr>
<td><strong>UK (PP)</strong></td>
<td>Mean Forecast: 0.313 Actual Effect Size: 0.630</td>
<td>0.248</td>
<td>0.258</td>
<td>0.305</td>
<td>0.112</td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th>Key effect</th>
<th>Aggregated effect size</th>
<th>Cochran’s $Q$</th>
<th>$\tau$</th>
<th>$I^2$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target age and needless work</td>
<td>0.0335</td>
<td>$Q(\text{df}=6) = 2.0552, p = 0.9146$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Intuitive work morality</td>
<td>0.2513</td>
<td>$Q(\text{df}=6) = 172.7514, p &lt; .0001$</td>
<td>97.7806</td>
<td>0.1780</td>
<td></td>
</tr>
<tr>
<td>Tacit inferences</td>
<td>0.4893</td>
<td>$Q(\text{df}=6) = 49.3908, p &lt; .0001$</td>
<td>89.7965</td>
<td>0.1362</td>
<td></td>
</tr>
<tr>
<td>Salvation prime</td>
<td>0.0384</td>
<td>$Q(\text{df}=7) = 0.0142, p = 1.0000$</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

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, $F(1, 2121.93) = 189.142$, $p < 0.001$, $d = 0.666$, India sample, $F(1, 996) = 110.456$, $p < 0.001$, $d = 0.666$, Australia sample, $F(1, 1007) = 149.507$, $p < 0.001$, $d = 0.666$, 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
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)

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.
- Use a common effect model
- 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.
- Use the DUREL scale (validated scale not used in the original studies)
- Use the single-item religiosity measure (direct replication of original approach)

**Exclusion criteria**

- Attention check.
- Exclude no one based on the attention check.
- 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.
- Include everyone regardless of years of English experience
- Exclude participants with less than 5 years of English experience
- Cultural grouping.
- Use current objective location of data collection
- Use self-reported nation of birth
- Regional grouping.
- Use current objective location of data collection
- 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.
   • 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 separate “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 $i$th region, the $j$th participant, and the $k$th condition. Then

$$Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \zeta_i + l_j \omega_i + x_k \eta_i + u_{ik} \theta_{i1} + v_{ik} \theta_{i2} + l_{jk} \theta_{i3} + m_{jk} \theta_{i4} + c_{jk} \theta_{i5} + d_{jk} \theta_{i6} + t_{ik} \theta_{i7}, \sigma^2),$$

where,

- $\alpha_i$ is the baseline rating for $i$th region
- $\beta_i$ is the Protestantism effect for $i$th region
- $\gamma_i$ is the religiosity effect for $i$th region
- $\zeta_i$ is the protestant work ethic effect for $i$th region
- $\omega_i$ is the SES effect for $i$th region
- $\eta_i$ is the general experimental effect for $i$th region
- $\theta_{i1}$ is the $i$th region’s experimental effect for US participants
- $\theta_{i2}$ is the experimental effect for New England participants
- $\theta_{i3}$ is the $i$th region’s experimental effect for Protestant participants
- $\theta_{i4}$ is the $i$th region’s experimental effect for highly religious participants
- $\theta_{i5}$ is the $i$th region’s experimental effect for participants who endorse PWE
- $\theta_{i6}$ is the $i$th region’s experimental effect for high SES participants
- $\theta_{i7}$ is the $i$th 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).

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

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:

$$Y_{ijk} \sim N(\alpha_i + p_j \beta_l + r_j \gamma_l + w_j \zeta_i + l_j \omega_i + u_{ik} \theta_1, \sigma^2),$$

with $\theta_1 > 0$

Random:

$$Y_{ijk} \sim N(\alpha_i + p_j \beta_l + r_j \gamma_l + w_j \zeta_i + l_j \omega_i + u_{ik} \theta_{11}, \sigma^2),$$

with all $\theta_{11} > 0$
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; \( \theta_{i8} = i \)th 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:
\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \xi_i + l_j \omega_i + u_{ik} \theta_{i1} + o_{ijk} \theta_{i8}, \sigma^2), \]
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:
\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \xi_i + l_j \omega_i + c_{jk} \theta_5, \sigma^2), \]
where \( \theta_5 > 0 \)

Random:
\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \xi_i + l_j \omega_i + c_{jk} \theta_{i5}, \sigma^2), \]
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:
\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \xi_i + l_j \omega_i + x_k \eta_i, \sigma^2), \]
where \( \eta > 0 \)

Random:
\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \xi_i + l_j \omega_i + x_k \eta_i, \sigma^2), \]
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:
\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \xi_i + l_j \omega_i + d_{jk} \theta_6, \sigma^2), \]
where \( \theta_6 > 0 \)

Random:
\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + w_j \xi_i + l_j \omega_i + d_{jk} \theta_{i6}, \sigma^2), \]
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.
SUPPLEMENTS: CREATIVE DESTRUCTION APPROACH TO REPLICATION

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:

\[ Y_{ijk} \sim N(\alpha_i + p_j\beta_i + r_j\gamma_i + w_j\zeta_i + l_j\omega_i + t_{ik}\theta_7, \sigma^2), \]

where all \( \theta_7 > 0 \)

**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

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:

- \( \theta_{i1} \) is the \( i \)th region’s main effect of US vs. non-US on preference.
- \( \theta_2 \) is the main effect of New England vs. non-New England.
- \( \theta_{i3} \) is the \( i \)th region’s main effect of Protestant vs. non-Protestant.
- \( \theta_{i4} \) is the \( i \)th region’s main effect of being religious vs. non-religious.
- \( \theta_{i5} \) is the \( i \)th region’s main effect of endorsing PWE vs. rejecting PWE.
- \( \theta_{i6} \) is the \( i \)th region’s main effect of having high SES vs. low SES.
- \( \theta_{i7} \) is the \( i \)th 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; \( \theta_{i8} \) 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.

\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j\gamma_i + q_k \delta + w_j \lambda_i + x_k \eta_i + u_{ik}\theta_{l1} + v_{ik}\theta_{l2} + l_{jk}\theta_{l3} + m_{jk}\theta_{l4} + c_{jk}\theta_{l5}, \sigma^2), \]

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 (\( \alpha_i \)) 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:

\[
Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + q_j \delta + w_j \zeta_i + u_{ik} \theta_{11}, \sigma^2),
\]

with all \( \theta_{11} > 0 \)

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:

\[
Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + q_j \delta + w_j \zeta_i + l_{jk} \theta_{13}, \sigma^2),
\]

with all \( \theta_{13} > 0 \)
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:

\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + q_j \delta + w_j \zeta_i + m_{jk} \theta_{t4}, \sigma^2), \]

with all \( \theta_{t4} > 0 \)

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:

\[ Y_{ijk} \sim N(\alpha_i + p_j \beta_i + r_j \gamma_i + q_j \delta + w_j \zeta_i + c_{jk} \theta_{t5}, \sigma^2), \]

with all \( \theta_{t5} > 0 \)

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


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

<table>
<thead>
<tr>
<th></th>
<th>American</th>
<th>New England</th>
<th>Religious</th>
<th>Protestant</th>
<th>PWE</th>
<th>SES</th>
<th>Self-expression culture</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>1</td>
<td>0.44</td>
<td>0.01</td>
<td>0.11</td>
<td>-0.07</td>
<td>0.16</td>
<td>-0.45</td>
</tr>
<tr>
<td>New England</td>
<td>0.44</td>
<td>1</td>
<td>-0.05</td>
<td>-0.04</td>
<td>-0.05</td>
<td>0.03</td>
<td>-0.2</td>
</tr>
<tr>
<td>Religious</td>
<td>0.01</td>
<td>-0.05</td>
<td>1</td>
<td>0.16</td>
<td>0.24</td>
<td>0.07</td>
<td>0.38</td>
</tr>
<tr>
<td>Protestant</td>
<td>0.11</td>
<td>-0.04</td>
<td>0.16</td>
<td>1</td>
<td>0.05</td>
<td>-0.02</td>
<td>-0.16</td>
</tr>
<tr>
<td>PWE</td>
<td>-0.07</td>
<td>-0.05</td>
<td>0.24</td>
<td>0.05</td>
<td>1</td>
<td>0</td>
<td>0.24</td>
</tr>
<tr>
<td>SES</td>
<td>0.16</td>
<td>0.03</td>
<td>0.07</td>
<td>-0.02</td>
<td>0</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Self-expression culture</td>
<td>-0.45</td>
<td>-0.2</td>
<td>0.38</td>
<td>-0.16</td>
<td>0.24</td>
<td>0.01</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Correlation (Cramer’s V) between dichotomized covariates.

Table S9-2: Correlation Matrix Model Predictors Study 2

<table>
<thead>
<tr>
<th></th>
<th>American</th>
<th>New England</th>
<th>Religious</th>
<th>Protestant</th>
<th>PWE</th>
<th>SES</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>1</td>
<td>0.42</td>
<td>0.2</td>
<td>-0.04</td>
<td>0.1</td>
<td>-0.22</td>
</tr>
<tr>
<td>New England</td>
<td>0.42</td>
<td>1</td>
<td>0.09</td>
<td>-0.03</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Religious</td>
<td>0.2</td>
<td>0.09</td>
<td>1</td>
<td>0.15</td>
<td>0.18</td>
<td>0.08</td>
</tr>
<tr>
<td>Protestant</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0.15</td>
<td>1</td>
<td>0.04</td>
<td>0.29</td>
</tr>
<tr>
<td>PWE</td>
<td>0.1</td>
<td>0</td>
<td>0.18</td>
<td>0.04</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>Age</td>
<td>-0.22</td>
<td>0.05</td>
<td>0.08</td>
<td>0.29</td>
<td>0.05</td>
<td>1</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Table S9-3: Participant demographics.</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>Study 1</td>
</tr>
<tr>
<td>US</td>
<td>52</td>
</tr>
<tr>
<td>New England</td>
<td>17.1</td>
</tr>
<tr>
<td>Religious</td>
<td>41.4</td>
</tr>
<tr>
<td>Protestant</td>
<td>22.6</td>
</tr>
<tr>
<td>Endorse PWE</td>
<td>65</td>
</tr>
<tr>
<td>High SES</td>
<td>36.2</td>
</tr>
<tr>
<td>Self-Expression Culture</td>
<td>16</td>
</tr>
</tbody>
</table>

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 $2 \times 2 \times 2 \times 2 \times 2 \times 2 = 64$ 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

<table>
<thead>
<tr>
<th>Effect</th>
<th>Best Model</th>
<th>BF&lt;sub&gt;W0&lt;/sub&gt;</th>
<th>BF&lt;sub&gt;10&lt;/sub&gt;</th>
<th>BF&lt;sub&gt;W1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Winning Model vs. Implicit Puritanism Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Generalized Moralization (common)</td>
<td>10&lt;sup&gt;131&lt;/sup&gt;</td>
<td>10&lt;sup&gt;55&lt;/sup&gt;</td>
<td>10&lt;sup&gt;75&lt;/sup&gt;</td>
</tr>
<tr>
<td>2</td>
<td>False Positives</td>
<td>1</td>
<td>0.12</td>
<td>8.38</td>
</tr>
<tr>
<td>3</td>
<td>Unconstrained Model</td>
<td>10&lt;sup&gt;138&lt;/sup&gt;</td>
<td>58.42</td>
<td>10&lt;sup&gt;136&lt;/sup&gt;</td>
</tr>
<tr>
<td>4</td>
<td>Generalized Moralization (common)</td>
<td>10&lt;sup&gt;7&lt;/sup&gt;</td>
<td>10&lt;sup&gt;6&lt;/sup&gt;</td>
<td>2.39</td>
</tr>
<tr>
<td>5</td>
<td>Generalized Moralization (random)</td>
<td>10&lt;sup&gt;80&lt;/sup&gt;</td>
<td>10&lt;sup&gt;54&lt;/sup&gt;</td>
<td>10&lt;sup&gt;25&lt;/sup&gt;</td>
</tr>
<tr>
<td>6</td>
<td>False Positives/Explicit American Exceptionalism</td>
<td>1</td>
<td>0.05</td>
<td>21.87</td>
</tr>
</tbody>
</table>

Note: The BF<sub>W0</sub> gives the Bayes factor for the winning model vs. the null model per key effect. The BF<sub>10</sub> Implicit Puritanism gives the Bayes factor for central Implicit Puritanism-based model versus the null-model for each key effect. The BF<sub>W1</sub> 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<sup>136</sup> 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 $4 \times 10^{40}$-to-1 or by $5 \times 10^{55}$-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

<table>
<thead>
<tr>
<th></th>
<th>BF(_{10})</th>
<th>(N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Superselect</td>
<td>Antiselect</td>
</tr>
<tr>
<td>1</td>
<td>15581</td>
<td>(10^7)</td>
</tr>
<tr>
<td>2</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>3</td>
<td>(10^{21})</td>
<td>(10^6)</td>
</tr>
<tr>
<td>4</td>
<td>0.17</td>
<td>0.34</td>
</tr>
<tr>
<td>5</td>
<td>36.37</td>
<td>0.43</td>
</tr>
<tr>
<td>6a</td>
<td>0.16</td>
<td>0.34</td>
</tr>
<tr>
<td>6b</td>
<td>0.20</td>
<td>0.29</td>
</tr>
</tbody>
</table>

*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.
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 10⁶-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.
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), 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


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 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)


Miyatake, S., & Higuchi, M. (2017). Does religious priming increase the prosocial behaviour


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, $F(1, 289) = 0.505, p$
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

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)**


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.

<table>
<thead>
<tr>
<th>Sample</th>
<th>N</th>
<th>MTurk India</th>
<th>MTurk USA</th>
<th>PureProfile USA</th>
<th>PureProfile Australia</th>
<th>PureProfile UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>30.51</td>
<td>39.48</td>
<td>49.62</td>
<td>46.24</td>
<td>48.05</td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td>7.12</td>
<td>12.36</td>
<td>16.61</td>
<td>16.35</td>
<td>15.25</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>718</td>
<td>494</td>
<td>566</td>
<td>331</td>
<td>389</td>
<td>398</td>
</tr>
<tr>
<td>Female</td>
<td>281</td>
<td>536</td>
<td>1552</td>
<td>679</td>
<td>564</td>
<td>564</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>No response</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Religion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agnostic</td>
<td>5</td>
<td>230</td>
<td>188</td>
<td>126</td>
<td>105</td>
<td>105</td>
</tr>
<tr>
<td>Atheist</td>
<td>44</td>
<td>206</td>
<td>129</td>
<td>257</td>
<td>240</td>
<td>240</td>
</tr>
<tr>
<td>Buddhism</td>
<td>50</td>
<td>10</td>
<td>29</td>
<td>26</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Catholic</td>
<td>290</td>
<td>164</td>
<td>619</td>
<td>187</td>
<td>133</td>
<td>133</td>
</tr>
<tr>
<td>Islam</td>
<td>71</td>
<td>10</td>
<td>16</td>
<td>18</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Judaism</td>
<td>10</td>
<td>11</td>
<td>57</td>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Other (please indicate)</td>
<td>456</td>
<td>120</td>
<td>522</td>
<td>225</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>Protestant</td>
<td>72</td>
<td>285</td>
<td>562</td>
<td>153</td>
<td>305</td>
<td>305</td>
</tr>
<tr>
<td>No response</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than $10,000 USD a year</td>
<td>413</td>
<td>43 (4.15%)</td>
<td>183 (8.60%)</td>
<td>61 (6.03%)</td>
<td>91 (9.48%)</td>
<td></td>
</tr>
<tr>
<td>USD $20,000-$40,000</td>
<td>218</td>
<td>111 (10.71%)</td>
<td>222 (10.44%)</td>
<td>112 (11.08%)</td>
<td>142 (14.79%)</td>
<td></td>
</tr>
<tr>
<td>USD $40,000-$60,000</td>
<td>163</td>
<td>287 (27.7%)</td>
<td>500 (23.51%)</td>
<td>219 (21.66%)</td>
<td>298 (31.04%)</td>
<td></td>
</tr>
<tr>
<td>USD $60,000-$80,000</td>
<td>98</td>
<td>208 (20.08%)</td>
<td>417 (19.61%)</td>
<td>188 (18.6%)</td>
<td>195 (20.31%)</td>
<td></td>
</tr>
<tr>
<td>USD $80,000-$100,000</td>
<td>43</td>
<td>162 (15.64%)</td>
<td>297 (13.96%)</td>
<td>157 (15.53%)</td>
<td>105 (10.94%)</td>
<td></td>
</tr>
<tr>
<td>USD $100,000 a year or more</td>
<td>17</td>
<td>97 (9.36%)</td>
<td>182 (8.56%)</td>
<td>102 (10.09%)</td>
<td>50 (5.21%)</td>
<td></td>
</tr>
<tr>
<td>No response</td>
<td>48</td>
<td>128</td>
<td>326</td>
<td>172</td>
<td>79</td>
<td>79</td>
</tr>
<tr>
<td>Political Views</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Progressive/Left-wing</td>
<td>74</td>
<td>165 (15.93%)</td>
<td>159 (7.48%)</td>
<td>57 (5.64%)</td>
<td>55 (5.73%)</td>
<td></td>
</tr>
<tr>
<td>Progressive/Left-wing</td>
<td>66</td>
<td>226 (21.81%)</td>
<td>224 (10.53%)</td>
<td>114 (11.28%)</td>
<td>95 (9.9%)</td>
<td></td>
</tr>
<tr>
<td>Somewhat Progressive/Left-wing</td>
<td>105</td>
<td>158 (15.25%)</td>
<td>179 (8.42%)</td>
<td>104 (10.29%)</td>
<td>148 (15.42%)</td>
<td></td>
</tr>
<tr>
<td>Political Orientation</td>
<td>Moderate/Centrist</td>
<td>Somewhat Conservative/Right-wing</td>
<td>Conservative/Right-wing</td>
<td>Very Conservative/Right-wing</td>
<td>No response</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-------------------</td>
<td>----------------------------------</td>
<td>--------------------------</td>
<td>----------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>325 (32.50%)</td>
<td>215 (20.75%)</td>
<td>803 (37.75%)</td>
<td>486 (48.07%)</td>
<td>419 (43.65%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>165 (16.50%)</td>
<td>105 (10.14%)</td>
<td>290 (13.63%)</td>
<td>131 (12.96%)</td>
<td>135 (14.06%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>146 (14.60%)</td>
<td>112 (10.81%)</td>
<td>232 (10.91%)</td>
<td>67 (6.63%)</td>
<td>56 (5.83%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>70 (7.00%)</td>
<td>52 (5.02%)</td>
<td>194 (9.12%)</td>
<td>17 (1.68%)</td>
<td>17 (1.77%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>49 (4.90%)</td>
<td>3 (0.29%)</td>
<td>46 (2.16%)</td>
<td>35 (3.46%)</td>
<td>35 (3.65%)</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some high school/secondary school</td>
<td>25 (2.50%)</td>
<td>8 (0.77%)</td>
<td>54 (2.54%)</td>
<td>108 (10.68%)</td>
<td>72 (7.50%)</td>
<td></td>
</tr>
<tr>
<td>High school degree/completed secondary school</td>
<td>15 (1.50%)</td>
<td>141 (13.61%)</td>
<td>513 (24.12%)</td>
<td>272 (26.9%)</td>
<td>347 (36.15%)</td>
<td></td>
</tr>
<tr>
<td>Some university</td>
<td>71 (7.10%)</td>
<td>277 (26.74%)</td>
<td>575 (27.03%)</td>
<td>127 (12.56%)</td>
<td>98 (10.21%)</td>
<td></td>
</tr>
<tr>
<td>University degree</td>
<td>308 (30.80%)</td>
<td>407 (39.29%)</td>
<td>539 (25.34%)</td>
<td>221 (21.86%)</td>
<td>228 (23.75%)</td>
<td></td>
</tr>
<tr>
<td>Some graduate/postgraduate education</td>
<td>318 (31.80%)</td>
<td>66 (6.37%)</td>
<td>129 (6.06%)</td>
<td>126 (12.46%)</td>
<td>52 (5.42%)</td>
<td></td>
</tr>
<tr>
<td>Graduate/postgraduate degree (e.g., doctoral degree)</td>
<td>238 (23.80%)</td>
<td>106 (10.23%)</td>
<td>241 (11.33%)</td>
<td>93 (9.20%)</td>
<td>88 (9.17%)</td>
<td></td>
</tr>
<tr>
<td>No response</td>
<td>25 (2.50%)</td>
<td>31 (2.99%)</td>
<td>76 (3.57%)</td>
<td>64 (6.33%)</td>
<td>75 (7.81%)</td>
<td></td>
</tr>
</tbody>
</table>
Table S14-2. Demographics for replications of the salvation prime effect.

<table>
<thead>
<tr>
<th></th>
<th>PureProfile USA</th>
<th>PureProfile Australia</th>
<th>PureProfile UK</th>
<th>Crowdsourced Labs USA</th>
<th>Crowdsourced Lab Canada</th>
<th>Crowdsourced Lab Ireland</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>514</td>
<td>298</td>
<td>312</td>
<td>563</td>
<td>91</td>
<td>80</td>
</tr>
<tr>
<td>Mean</td>
<td>47.21</td>
<td>44.53</td>
<td>47.66</td>
<td>19.54</td>
<td>21.09</td>
<td>20.23</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>16.46</td>
<td>17.19</td>
<td>15.75</td>
<td>3.15</td>
<td>2.73</td>
<td>5.18</td>
</tr>
<tr>
<td>Male</td>
<td>131 (25.49%)</td>
<td>94 (31.54%)</td>
<td>103 (33.01%)</td>
<td>220 (39.08%)</td>
<td>27 (29.67%)</td>
<td>22 (27.5%)</td>
</tr>
<tr>
<td>Female</td>
<td>378 (73.54%)</td>
<td>203 (68.12%)</td>
<td>206 (66.03%)</td>
<td>338 (60.04%)</td>
<td>63 (69.23%)</td>
<td>58 (72.5%)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (0.78%)</td>
<td>1 (0.34%)</td>
<td>2 (0.64%)</td>
<td>4 (0.71%)</td>
<td>1 (1.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>No response</td>
<td>4 (0.78%)</td>
<td>1 (0.34%)</td>
<td>2 (0.64%)</td>
<td>4 (0.71%)</td>
<td>1 (1.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Agnostic</td>
<td>51 (9.92%)</td>
<td>28 (9.4%)</td>
<td>38 (12.18%)</td>
<td>41 (7.28%)</td>
<td>7 (7.69%)</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>Atheist</td>
<td>25 (4.86%)</td>
<td>78 (26.17%)</td>
<td>69 (22.12%)</td>
<td>50 (8.88%)</td>
<td>16 (17.58%)</td>
<td>14 (17.5%)</td>
</tr>
<tr>
<td>Buddhism</td>
<td>4 (0.78%)</td>
<td>16 (5.37%)</td>
<td>0 (0%)</td>
<td>11 (1.95%)</td>
<td>3 (3.3%)</td>
<td>1 (1.25%)</td>
</tr>
<tr>
<td>Catholic</td>
<td>167 (32.49%)</td>
<td>43 (14.43%)</td>
<td>38 (12.18%)</td>
<td>207 (36.77%)</td>
<td>13 (14.29%)</td>
<td>54 (67.5%)</td>
</tr>
<tr>
<td>Islam</td>
<td>4 (0.78%)</td>
<td>4 (1.34%)</td>
<td>4 (1.28%)</td>
<td>34 (6.04%)</td>
<td>11 (12.09%)</td>
<td>1 (1.25%)</td>
</tr>
<tr>
<td>Judaism</td>
<td>17 (3.31%)</td>
<td>2 (0.67%)</td>
<td>0 (0%)</td>
<td>62 (11.01%)</td>
<td>1 (1.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Other</td>
<td>114 (22.18%)</td>
<td>86 (28.86%)</td>
<td>56 (17.95%)</td>
<td>87 (15.45%)</td>
<td>19 (20.88%)</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>Protestant</td>
<td>131 (25.49%)</td>
<td>39 (13.09%)</td>
<td>103 (33.01%)</td>
<td>57 (10.12%)</td>
<td>18 (19.78%)</td>
<td>2 (2.5%)</td>
</tr>
<tr>
<td>No response</td>
<td>1 (0.19%)</td>
<td>2 (0.67%)</td>
<td>4 (1.28%)</td>
<td>14 (2.49%)</td>
<td>3 (3.3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Less than $10,000 USD a year</td>
<td>45 (8.75%)</td>
<td>19 (6.38%)</td>
<td>37 (11.86%)</td>
<td>64 (11.37%)</td>
<td>15 (16.48%)</td>
<td>28 (35%)</td>
</tr>
<tr>
<td>USD $20,000-$40,000</td>
<td>114 (22.18%)</td>
<td>66 (22.15%)</td>
<td>112 (35.9%)</td>
<td>66 (11.72%)</td>
<td>10 (10.99%)</td>
<td>11 (13.75%)</td>
</tr>
<tr>
<td>USD $40,000-$60,000</td>
<td>91 (17.7%)</td>
<td>54 (18.12%)</td>
<td>60 (19.23%)</td>
<td>67 (11.9%)</td>
<td>20 (21.98%)</td>
<td>21 (26.25%)</td>
</tr>
<tr>
<td>USD $60,000-$80,000</td>
<td>73 (14.2%)</td>
<td>56 (18.79%)</td>
<td>27 (8.65%)</td>
<td>59 (10.48%)</td>
<td>6 (6.59%)</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>USD $80,000-$100,000</td>
<td>39 (7.59%)</td>
<td>27 (9.06%)</td>
<td>16 (5.13%)</td>
<td>68 (12.08%)</td>
<td>11 (12.09%)</td>
<td>6 (7.5%)</td>
</tr>
<tr>
<td>USD $100,000 a year or more</td>
<td>76 (14.79%)</td>
<td>39 (13.09%)</td>
<td>13 (4.17%)</td>
<td>157 (27.89%)</td>
<td>15 (16.48%)</td>
<td>2 (2.5%)</td>
</tr>
<tr>
<td>No response</td>
<td>4 (0.78%)</td>
<td>5 (1.68%)</td>
<td>6 (1.92%)</td>
<td>38 (6.75%)</td>
<td>4 (4.4%)</td>
<td>2 (2.5%)</td>
</tr>
<tr>
<td>Very Progressive/Left-wing</td>
<td>43 (8.37%)</td>
<td>12 (4.03%)</td>
<td>17 (5.45%)</td>
<td>18 (3.2%)</td>
<td>7 (7.69%)</td>
<td>3 (3.75%)</td>
</tr>
<tr>
<td>Progressive/Left-wing</td>
<td>47 (9.14%)</td>
<td>34 (11.41%)</td>
<td>42 (13.46%)</td>
<td>123 (21.85%)</td>
<td>15 (16.48%)</td>
<td>1 (1.25%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Somewhat Progressive/Left-wing</td>
<td>41 (7.98%)</td>
<td>36 (12.08%)</td>
<td>53 (16.99%)</td>
<td>65 (11.55%)</td>
<td>14 (15.38%)</td>
<td></td>
</tr>
<tr>
<td>Moderate/Centrist</td>
<td>217 (42.22%)</td>
<td>166 (55.7%)</td>
<td>141 (45.19%)</td>
<td>215 (38.19%)</td>
<td>25 (27.47%)</td>
<td></td>
</tr>
<tr>
<td>Somewhat Conservative/Right-wing</td>
<td>54 (10.51%)</td>
<td>33 (11.07%)</td>
<td>39 (12.5%)</td>
<td>66 (11.72%)</td>
<td>9 (9.89%)</td>
<td></td>
</tr>
<tr>
<td>Conservative/Right-wing</td>
<td>59 (11.48%)</td>
<td>13 (4.36%)</td>
<td>16 (5.13%)</td>
<td>35 (6.22%)</td>
<td>12 (13.19%)</td>
<td></td>
</tr>
<tr>
<td>Very Conservative/Right-wing</td>
<td>52 (10.12%)</td>
<td>4 (1.34%)</td>
<td>3 (0.96%)</td>
<td>7 (1.24%)</td>
<td>3 (3.3%)</td>
<td></td>
</tr>
<tr>
<td>No response</td>
<td>1 (0.19%)</td>
<td>(0%)</td>
<td>1 (0.32%)</td>
<td>34 (6.04%)</td>
<td>9 (9.89%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some high school/secondary school</td>
<td>17 (3.31%)</td>
<td>34 (11.41%)</td>
<td>25 (8.01%)</td>
<td>8 (1.42%)</td>
<td>3 (3.3%)</td>
<td></td>
</tr>
<tr>
<td>High school degree/completed secondary school</td>
<td>115 (22.37%)</td>
<td>85 (28.52%)</td>
<td>115 (36.86%)</td>
<td>161 (28.6%)</td>
<td>12 (13.19%)</td>
<td></td>
</tr>
<tr>
<td>Some university</td>
<td>140 (27.24%)</td>
<td>50 (16.78%)</td>
<td>32 (10.26%)</td>
<td>363 (64.48%)</td>
<td>59 (64.84%)</td>
<td></td>
</tr>
<tr>
<td>University degree</td>
<td>129 (25.1%)</td>
<td>67 (22.48%)</td>
<td>87 (27.88%)</td>
<td>24 (4.26%)</td>
<td>13 (14.29%)</td>
<td></td>
</tr>
<tr>
<td>Some graduate/postgraduate education</td>
<td>35 (6.81%)</td>
<td>40 (13.42%)</td>
<td>21 (6.73%)</td>
<td>5 (0.89%)</td>
<td>3 (3.3%)</td>
<td></td>
</tr>
<tr>
<td>Graduate/postgraduate degree (e.g., doctoral degree)</td>
<td>73 (14.2%)</td>
<td>20 (6.71%)</td>
<td>26 (8.33%)</td>
<td>2 (0.36%)</td>
<td>1 (1.1%)</td>
<td></td>
</tr>
<tr>
<td>No response</td>
<td>5 (0.97%)</td>
<td>2 (0.67%)</td>
<td>6 (1.92%)</td>
<td>(0%)</td>
<td>(0%)</td>
<td></td>
</tr>
</tbody>
</table>