

## Why It Is Hard to Find Genes Associated with Social Science Traits: Theoretical and Empirical Considerations

Christopher F. Chabris  
James J. Lee  
Daniel J. Benjamin  
Jonathan P. Beauchamp  
Edward L. Glaeser  
Gregoire Borst  
Steven Pinker  
David I. Laibson

### SUPPORTING ONLINE MATERIAL

#### *Measures*

Table S1 lists all of the phenotypes measured in this study. Any variable taking more than ten values was regarded as quantitative rather than polytomous (ordered categorical). A parenthetical N in Table S1 indicates that we were able to remove sex differences in mean and variance from a quantitative variable and then use a quantile transformation to render the resulting scores normally distributed. These transformations should increase statistical power to detect genetic associations for traits showing sex differences. Below are details on some of the behavioral phenotypes whose labels in Table S1 are not self-explanatory. (Except as noted, all economic games were played with real monetary incentives.)

**3-back.** Participants viewed a succession of words, each new word appearing every three seconds. Participants were instructed to indicate as quickly and accurately as possible whether each word matched the word seen three items previously. This task has often been employed as an indicator of working memory capacity.<sup>1</sup>

**Barratt Impulsiveness Scale (BIS).** This self-report has been found to measure three distinct factors (inattention, motor impulsiveness, and lack of planning).<sup>2</sup> We used the sum of these three factor scores as a measure of this self-report’s general factor.

**Cambridge Face Memory Test (CFMT).** Participants studied three photos of each of six target human faces and were then tested with a series of forced-choice items, each consisting of three faces, one of which was a target. This test has been shown to be a sensitive measure of prosopagnosia (a specific deficit in recognizing other people by their facial features) and also normal variability in the ability to recognize faces.<sup>3,4</sup>

**Dictator game.** Each participant was asked to imagine being randomly and anonymously paired with another participant. The participant was then asked to allocate ten dollars between the members of the pair. How much of the ten dollars each participant is willing to give away to the

other person in this task has been used as a measure of the participant’s heritable altruistic tendencies.<sup>5,6</sup> Because the distribution of allocation was almost bimodal, nearly all participants giving away either zero or five dollars, we treated this phenotype as dichotomous; all participants who gave anything at all were given the higher score.

**Discounting the future.** Participants were presented a set of choices between smaller prompt rewards and larger delayed rewards. Temporal discount rates inferred in this way, have been found to be associated with substance abuse and other outcomes.<sup>7</sup>

**General cognitive ability.** We combined the following indicators into a standardized cognitive ability composite: (1) a short form of Raven’s Advanced Progressive Matrices<sup>8</sup>, a measure of abstract reasoning ability; (2) the Arithmetic, Similarities, and Vocabulary subtests of the Multidimensional Aptitude Battery (MAB), which measure verbal ability; and (3) accuracy on a forced-choice version of the Shepard-Metzler Mental Rotation task (SMMR) a measure of spatial ability.<sup>9</sup>

**Inattentive blindness.** Participants watched a video of two teams of three players, one team wearing white shirts and the other wearing black shirts, who moved around erratically in an elevator lobby. The passes were either bounce passes or aerial passes; players would also dribble the ball, wave their arms, and make other movements. After about 45 seconds, a person wearing a gorilla costume walked through the action. The relatively high proportion of participants who report not seeing the gorilla at all is generally regarded as surprising.<sup>10</sup> The causes of individual differences in this task are unknown. This finding has achieved wide publicity, so we treated any participant who reported having seen or heard of it as a missing data point; others were classified as either noticing or missing the gorilla.

**Loss aversion.** Participants were presented with a set of choices between (1) receiving nothing or (2) a 50% chance of gaining an amount  $x$  and a 50% chance of losing an amount  $y$ . This is a standard measure of aversion to suffering financial losses.<sup>11</sup> The main loss aversion measure involved real money stakes; a separate measure was made with fictitious higher stakes.

**NEO Five-Factor Inventory.** A 60-item self-report instrument with 12 items measuring each of the following five personality factors, which constitute the most widely accepted factorization of personality: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness.

**Paired-associate recognition.** After studying a series of 25 word pairs, participants were given a recognition test in which they were given the first word in a pair and had to choose the second from among four presented alternatives. The words in the pairs were abstract and unrelated, and the distractor words were other words from the experiment, making this task difficult.<sup>12</sup>

**Religiosity.** We administered a standard scale to measure religiousness.<sup>13</sup>

**Risk aversion.** Participants were presented with a set of choices between (1) a 100% chance of receiving an amount  $x$  or (2) a 50% chance of receiving an amount  $y > x$  and a 50% chance of receiving nothing. Risk-averse choices involved turning down a larger expected value prospect

(e.g., 50% chance of receiving \$10) in favor of a smaller guaranteed amount (e.g., 100% chance of receiving \$4). This is a standard measure.<sup>11</sup>

**Shape memory.** In a study phase, participants were presented a series of irregular shapes, one at a time. In a test phase, participants then had to press one key if the shape they were viewing had already been presented in the study phase, another key if it was new.

**Social attitudes.** Items asking for attitudes toward abortion, alcohol consumption, and other social issues were taken from an existing scale.<sup>14</sup> Because the factor model postulated by the scale’s authors did not fit our data well, we analyzed each item separately.

**Spatial memory.** In a study phase, participants viewed a circular array of gray dots. Several of the dots briefly turned black, one at a time. The display continued in a test phase, where participants indicated whether each black dot had also turned black during the study phase.

**Serial Reaction Time Task (SRTT).** Participants viewed a line of four squares. During each of 384 trials a black diamond briefly appeared in one of the squares, and in response participants had to press one of four corresponding keys, using four fingers of their preferred hand. Unbeknownst to the participants, a fixed subsequence of the stimuli appeared repeatedly throughout the task, alternating with runs of stimuli chosen at random. Response time (RT) tends to decrease with each successive presentation of the repeating subsequence, although most participants do not consciously notice the repetition. The mean difference in RT between the repeating stimuli and the random stimuli was taken as a measure of implicit skill learning.

**Utilitarianism.** Participants were presented with a set of moral dilemmas in which participants rated on a 1–5 scale the appropriateness of a “utilitarian” response to the situation.<sup>15</sup> A typical item: “You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman. Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?”

**Verbal fluency.** Participants were given one minute to utter as many distinct words as possible beginning with a certain letter. Person names, places, and numbers were not counted. The letters F, A, and S were used. The counts of the uttered words beginning with these letters appeared to be equal indicators of a common factor after standardization.

**Vividness of Visual Imagery Questionnaire (VVIQ).** Participants were told to visualize certain scenes or persons and rate the vividness of distinct aspects of the mental image.<sup>16</sup>

To make the choices in the economic tasks (intertemporal choice, fairness, loss aversion, risk aversion) meaningful to the participants, we told them at the outset that their choices in these tasks might be implemented with real money. The payment policy worked as follows: At the end

of the session, each participant rolled a six-sided die. If he or she rolled a six, then a further random draw was conducted to choose one item from the behavioral-economic tasks, which was then fulfilled for that participant. For example, suppose that the participant rolled a six, and the second draw selected an item from the discounting task. If the participant expressed a preference for  $x$  dollars 30 days from now over  $y$  dollars 60 days from now, then the participant was written a check for  $x$  dollars dated 30 days from the date of the phenotyping session (or given a debit card for the same amount that would be activated on the same date). Any losses suffered in the loss aversion task came out of \$5 cash given to each participant at the beginning of the session. This \$5 was given in addition to the advertised \$50 compensation.

The MAB subtests were scored according to the instructions in the test manual.<sup>17</sup> Factor analyses of the BIS, the NEO, religiosity, utilitarianism, verbal fluency, and the VVIQ resulted in solutions with nonzero uniquenesses. For these phenotypes we estimated factor scores by Bartlett’s method, which is equivalent to maximum likelihood (ML) if the uniquenesses are normally distributed. A few participants were missing some data as a result of omits, photocopying errors, computer failures, and other administrative issues. Participants’ factor scores were treated as missing if they responded to fewer than half of a scale’s indicators. We used the OpenMx package in R to perform all factor analyses.<sup>18</sup> All coding and scoring of phenotypic measures was performed blind to participant genotypes.

Parameters describing the responses of each participant during the behavioral-economic tasks were estimated by ML, assuming choice error drawn from an extreme-value distribution. For example, an “interest rate” for discounting utility flows over time was estimated for each person and used as the phenotype for the discounting task.

### *DNA Collection, Extraction, and Genotyping*

At two points during the phenotyping session, participants provided DNA samples by washing their mouths with 10 ml of Scope mouthwash, which dislodges loose cells, and then releasing the mouthwash into a Nalgene bottle. Samples were stored either in a freezer at  $-20^{\circ}\text{C}$  or in packed dry ice until DNA extraction. Genomic DNA was extracted using a QIAamp DNA Blood Mini Kit according to the manufacturer’s recommended protocol.

Genomic DNA samples normalized to 50 ng/ $\mu\text{l}$  were genotyped at either Stanford Genome Technology Center (SGTC) or Expression Analysis (EA) in Durham, North Carolina, in four batches, using the Affymetrix Genome-Wide Human SNP Array 6.0. SNP genotypes were called using the Birdseed v2 algorithm applied to each batch individually. The median call rate before application of quality-control criteria was 99.64%. Between-batch reproducibility was assessed by genotyping both samples provided by each of two participants. Average genotype concordance between replicates was 99.7%.

Our quality-control criteria at this stage excluded all participants missing more than 7% of their genotypic data, all SNPs with minor allele frequency (MAF) less than .05, all SNPs deviating from Hardy-Weinberg equilibrium at a significance threshold of  $5 \times 10^{-8}$ , and all SNPs missing

more than 5% of their calls. We then computed the principal components of the resulting genotype matrix with the program EIGENSTRAT.<sup>19</sup> To guard against population stratification, all participants who were more than six standard-deviation units from the origin on any of the top 10 PCs were iteratively excluded (a total of 14 participants).

After application of all quality control measures, the final cleaned dataset included 401 individuals and 661,107 SNPs. Nine statistically significant principal components at a significance threshold of .05 were found. The components corresponding to the fourth and fifth largest eigenvalues weakly distinguished the two genotyping laboratories, despite the application of our quality-control steps. The first, second, third, and sixth components were significantly correlated with the geographical distance of grandparental origin from England. The seventh component tended to spread out individuals reporting non-British grandparents, whereas the eighth component tended to separate those reporting two or more British grandparents from those reporting one or none. The ninth component tended to spread out individuals reporting British grandparents, perhaps reflecting structure within Britain. To control for remaining stratification, we included all nine significant principal components as covariates in the tests for SNP-trait association.

**Table S1.** All phenotypes measured. Any phenotype measured in paper mode was administered as a traditional paper-and-pencil test. Self-report refers to questionnaire data recorded either on paper forms or a SurveyMonkey questionnaire. Phenotypes measured in computer mode were implemented as PsyScope tasks requiring participants to provide keyboard input. Physical traits were directly measured by an experimenter using either a measuring tape or a standard bathroom scale. Audio refers to sound-recorded data that was later transcribed and coded.

Phenotype	Mode	Scale
3-back accuracy	computer	quantitative (N)
3-back RT	computer	quantitative (N)
acne severity as adolescent	self-report	polytomous
acne severity as adult	self-report	polytomous
acne severity overall	self-report	polytomous
alcohol consumption frequency (last 12 months)	self-report	polytomous
alcohol drinks per drinking occasion	self-report	quantitative
alcohol total drinks in last year	self-report	quantitative (N)
allergic to animals	self-report	dichotomous
allergic to drugs	self-report	dichotomous
allergic to food	self-report	dichotomous
allergies (any)	self-report	dichotomous
anticipated remaining life expectancy	self-report	quantitative (N)
asthma as adult	self-report	dichotomous
asthma as child	self-report	dichotomous
athleticism	self-report	polytomous
attitude toward abortion on demand	self-report	polytomous
attitude toward alcohol	self-report	polytomous
attitude toward attention-drawing clothes	self-report	polytomous
attitude toward being the center of attention	self-report	polytomous
attitude toward being the leader of groups	self-report	polytomous
attitude toward big parties	self-report	polytomous
attitude toward capitalism	self-report	polytomous
attitude toward castration as sex crime punishment	self-report	polytomous
attitude toward death penalty for murder	self-report	polytomous
attitude toward doing athletic activities	self-report	polytomous
attitude toward dressing well at all times	self-report	polytomous
attitude toward education	self-report	polytomous
attitude toward exercising	self-report	polytomous
attitude toward getting along well with others	self-report	polytomous
attitude toward illegal drugs	self-report	polytomous
attitude toward legalized gambling	self-report	polytomous
attitude toward loud music	self-report	polytomous
attitude toward making racial discrimination illegal	self-report	polytomous
attitude toward open-door immigration	self-report	polytomous
attitude toward organized religion	self-report	polytomous
attitude toward playing chess	self-report	polytomous

attitude toward playing organized sports	self-report	polytomous
attitude toward public speaking	self-report	polytomous
attitude toward reading books	self-report	polytomous
attitude toward rollercoaster rides	self-report	polytomous
attitude toward smoking	self-report	polytomous
attitude toward voluntary euthanasia	self-report	polytomous
back pain	self-report	dichotomous
BIS inattention	self-report	quantitative (N)
BIS general	self-report	quantitative (N)
BIS motor	self-report	quantitative (N)
BIS nonplanning	self-report	quantitative (N)
body mass index	measured	quantitative (N)
body type (scrawny to obese)	self-report	polytomous
ca_eine mg per day	self-report	quantitative
CFMT	computer	quantitative (N)
cigarette packs per day	self-report	polytomous
cleft chin	self-report	dichotomous
co_ee cups per day	self-report	polytomous
corrective lenses needed currently	self-report	dichotomous
corrective lenses needed at any time	self-report	dichotomous
curl tongue	self-report	dichotomous
Democrat vs. Republican	self-report	polytomous
dental braces worn (ever)	self-report	dichotomous
dental braces worn or needed (ever)	self-report	dichotomous
dictator game	self-report	dichotomous
dimples	self-report	dichotomous
discounting the future	self-report	quantitative (N)
drink alcohol (ever)	self-report	dichotomous
earlobes free (vs. hanging)	self-report	dichotomous
evening person	self-report	dichotomous
exercise amount per week	self-report	polytomous
exercise intensity	self-report	polytomous
exercise regularly	self-report	dichotomous
eye color	self-report	polytomous
facial hair color	self-report	polytomous
facial hair color (red vs. not red)	self-report	dichotomous
farsighted	self-report	dichotomous
first toe longer than second toe	self-report	dichotomous
floss teeth regularly	self-report	dichotomous
freckles on face	self-report	polytomous
gambling frequency	self-report	polytomous
general cognitive ability	multiple	quantitative (N)
hair color	self-report	polytomous
hair color (red vs. not red)	self-report	dichotomous
hair curliness	self-report	polytomous
hair on middle segment of any finger	self-report	dichotomous

happiness sumscore	self-report	quantitative (N)
hay fever	self-report	dichotomous
heterosexual	self-report	dichotomous
hitchhiker's thumb	self-report	dichotomous
hours of sleep average	self-report	quantitative
hours of sleep last night	self-report	quantitative
illegal drug use	self-report	polytomous
inattentional blindness	computer	dichotomous
in-person contact with family or very close friends	self-report	dichotomous
last doctor's appointment for checkup	self-report	polytomous
liberal vs conservative	self-report	polytomous
loss aversion	self-report	quantitative
MAB Arithmetic	paper	quantitative (N)
MAB Similarities	paper	quantitative (N)
MAB Vocabulary	paper	quantitative (N)
memory problems	self-report	dichotomous
migraines at any time	self-report	dichotomous
migraine frequency	self-report	polytomous
migraine within last 12 months	self-report	dichotomous
morning person	self-report	dichotomous
multivitamin supplement	self-report	dichotomous
nearsighted	self-report	dichotomous
NEO Agreeableness	self-report	quantitative (N)
NEO Conscientiousness	self-report	quantitative (N)
NEO Extraversion	self-report	quantitative (N)
NEO Neuroticism	self-report	quantitative (N)
NEO Openness	self-report	quantitative (N)
paired-associate recognition	computer	quantitative (N)
percentage of income saved over last 3 years	self-report	quantitative
physical attractiveness	self-report	polytomous
quality of sleep	self-report	polytomous
RAPM	computer	quantitative (N)
religiosity	self-report	quantitative
right-handed	self-report	dichotomous
risk aversion	self-report	quantitative (N)
seat belt use	self-report	polytomous
shape memory accuracy	computer	quantitative (N)
shape memory response time	computer	quantitative (N)
sitting height	measured	quantitative (N)
skin color and sun exposure response	self-report	polytomous
SMMR accuracy	computer	quantitative (N)
SMMR response time	computer	quantitative (N)
smoked cigarette (ever)	self-report	dichotomous
soda cups per day	self-report	polytomous
spatial memory accuracy	computer	quantitative (N)
spatial span response time	computer	quantitative (N)

SRTT accuracy	computer	quantitative
SRTT overall RT	computer	quantitative (N)
SRTT improvement in RT	computer	quantitative (N)
standing height	measured	quantitative (N)
strength	self-report	polytomous
stress level within last 12 months	self-report	polytomous
sunscreen or protective clothing use	self-report	polytomous
tea cups per day	self-report	polytomous
time woke up this morning	self-report	quantitative (N)
tobacco use frequency (current)	self-report	polytomous
tobacco user (current)	self-report	dichotomous
tobacco user (ever)	self-report	dichotomous
unprotected sex	self-report	polytomous
utilitarianism	self-report	quantitative (N)
verbal fluency	audio	quantitative (N)
vision quality (uncorrected)	self-report	polytomous
VVIQ	self-report	quantitative (N)
weight	measured	quantitative (N)
weight (maximum)	self-report	quantitative (N)
widow’s peak	self-report	dichotomous

## References

1. Gray JR, Chabris CF, Braver TS. Neural mechanisms of general fluid intelligence. *Nature Neuroscience* 2003; 6: 316–322.
2. Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*. 1995; 51: 768–774.
3. Duchaine B, Nakayama K. The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*. 2006; 44: 576–585.
4. Wilmer JB, Germine L, Chabris CF, Chatterjee G, Williams M, et al. Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107: 5238–5241.
5. Knafo A, Israel S, Darvasi A, Bachner-Melman R, Uzefovsky F, et al. Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes, Brain and Behavior*. 2008; 7: 266–275.
6. Cesarini D, Dawes CT, Johannesson M, Lichtenstein P, Wallace B. Experimental game theory and behavior genetics. *Annals of the New York Academy of Sciences*. 2009; 1167: 66–75.
7. Chabris CF, Laibson DI, Schuldt, JP. Intertemporal choice. In: Durlauf SN, Blume LE, eds. *The new Palgrave Dictionary of Economics*. 2<sup>nd</sup> ed. London, UK: Palgrave Macmillan; 2008: 536-542.
8. Bors, DA, Stokes, TL. Raven’s advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*. 1998; 58: 382–398.
9. Shepard RN, Metzler J. Mental rotation of three-dimensional objects. *Science*. 1971; 191: 952–954.
10. Simons DJ, Chabris CF. Gorillas in our midst: Sustained inattention blindness for dynamic events. *Perception*. 1999; 28: 1059–1074.
11. Beauchamp JP, Benjamin DJ, Chabris CF, Laibson DI. How Malleable are Risk Preferences and Loss Aversion? *Harvard University Mimeo*. March 2012.
12. Wilmer JB, Chatterjee G, Chabris CF, Gerbasi ME, Germine L, Nakayama K, Duchaine, B. A brief test battery for developmental prosopagnosia. In press, *Cognitive Neuropsychology*.
13. Koenig LB, McGue M, Krueger RF, Bouchard TJ. Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of Personality*. 2005; 73: 471–488.
14. Olson JM, Vernon PA, Harris JA, Jang KL. The heritability of attitudes: A study of twins. *Journal of Personality and Social Psychology*. 2001; 80: 845–860.
15. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD. An fMRI investigation of emotional engagement in moral judgment. *Science*. 2001; 293: 2105–2108.
16. Marks DF. Visual imagery differences in the recall of pictures. *British Journal of Psychology*. 1973; 64: 17–24.
17. Jackson DN. *Multidimensional Aptitude Battery II*. 2<sup>nd</sup> ed. Port Huron, MI: Sigma Assessment Systems; 1998.
18. R Development Core Team. *R: A Language and Environment for Statistical Computing*.

Vienna, Austria: R Foundation for Statistical Computing; 2010.

19. Price AL, Patterson N, Plenge RM, Weinblatt ME, Shadick NA, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006; 38: 904–909.