An Examination of the Effects of Eye-Tracking on Behavior in Psychology Experiments

Darrell A. Worthy, Joanna N. Lahey, Samuel L. Priestley, & Marco A. Palma

Texas A&M University

(In press, *Behavior Research Methods*)

Abstract

Eye-tracking is emerging as a tool for researchers to better understand cognition and behavior. However, it is possible that experiment participants adjust their behavior when they know their eyes are being tracked. This potential change would be considered a type of Hawthorne effect, in which participants alter their behavior in response to being watched, and could potentially compromise the outcomes and conclusions of experimental studies that use eye tracking. We examined whether eye-tracking produced Hawthorne effects in six commonly used psychological scales and five behavioral tasks. The dependent measures were selected because they are widely used and cited and because they involved measures of sensitive topics, including gambling behavior, racial bias, undesirable personality characteristics, or because they require working memory or executive attention resources, which might be affected by Hawthorne effects. The only task where Hawthorne effects manifested was the mixed gambles task, in which participants accepted or rejected gambles involving a 50/50 chance of gaining or losing different monetary amounts. Participants in the eye-tracking condition accepted fewer gambles that were low in expected value, and they also took longer to respond for these low-value gambles. These results suggest that eye-tracking is not likely to produce Hawthorne effects in most common psychology laboratory tasks, except for those involving risky decisions where the probability of the outcomes from each choice are known.

Keywords: Eye-tracking, Decision-Making, Risk, Memory, Attention, Hawthorne Effects

Data and Analysis Code Available at: https://osf.io/3fdrp/

Corresponding Author:
Darrell A. Worthy
Associate Professor
Texas A&M University
4235 TAMU
College Station, TX 77845-4235
worthyda@tamu.edu
worthylab.org

Introduction

Over the past several decades, eye-tracking has become a commonly-used tool in psychological experiments to better understand the drivers behind behavior (Lejarraga, Schulte-Mecklenbeck, & Smedema, 2017; Nystrom, Andersson, Holmqvist, & Van De Weijer, 2013). It is considered safe and non-invasive, and can provide physiological data that can address research questions that cannot be answered by behavioral choice data alone. Compared to other technologies often used in psychology and neuroscience studies such as MRI, EEG, or PET, eye-tracking is less expensive, less invasive, and is associated with fewer risks (Machado & Nelson, 2011).

Although eye-tracking may be an excellent complementary tool for researchers to better understand cognition and behavior, there is a possibility that participants' knowledge that their eyes are being tracked could affect their behavior in experimental settings. This change would be considered a type of Hawthorne effect in which participants alter their behavior when they know they are being watched compared to what their behavior would have been if they did not know they were being watched (Adair, 1984). The name for this phenomenon comes from studies conducted in the 1920s and 30s by the Hawthorne plant of the Western Electric Company in Chicago, Illinois, in the United States (Wickstrom & Bendix, 2000).

The term 'Hawthorne effect' has expanded from its original meaning that referred to an effect from one knowing they are being observed to encompassing many different types of related phenomena (Adair, 1984; Merret, 2006; Chiesa & Hobbs, 2008). Some scholars have noted the importance of clarifying exactly what aspects of the research environment are putatively causing participants to be more aware that they are being observed than they would be in other experimental contexts or outside of the laboratory (Wickstrom & Bendix, 2000). Eye-tracking could produce Hawthorne effects above what would be experienced in normal laboratory settings, due to participants knowledge of their eyes being tracked. An increased Hawthorne effect in eye-tracking experiments may lead to altered behavior due to Social Desirability Bias (SBD). SDB has be observed in both field and laboratory experiments (Bateson et al. 2006; Haley & Fessler 2005; Norwood & Lusk 2011; Paulhus 2002; Sparks & Barclay 2013; Zizzo 2010). SBD may lead participants to respond inaccurately to conform to social conventions in expected ways, particularly in tasks where sensitive topics are addressed such as gambling (Schell, Godhino, & Cunningham, 2021; van der Maas, Nower, Matheson, Turner, & Mann, 2021), sexual behavior (Kelly, Soler-Hampejsek, Mensch, & Hewett, 2013), compliance with health regulations (Jensen, 2020), or other topics such as racism, ableism, lying, cheating and stealing (Charles & Dattalo, 2018).

In this paper, we examine whether eye-tracking produces Hawthorne effects by comparing behavior between groups randomly assigned to an eye-tracking and a control in several commonly used and highly cited psychological tasks and scales, listed in Table 1. To our knowledge, this is one of the first systematic examinations of whether participants' knowledge that their eye movements are being tracked affects their behavior in laboratory-based behavioral experiments. Another paper from our labs investigated whether eye-tracking produced Hawthorne effects in several economic games and found null effects across the dependent variables for most of the economics games that were administered to participants (Kee et al., 2021). However, Kee and colleagues did, unexpectedly, observe evidence of increased risk-aversion for participants in the eye-tracking condition in the Holt and Laury risk assessment task. This effect was found primarily for participants who had a large number of failed calibration

attempts within the eye-tracking condition. The effect of the eye-tracking manipulation on the Holt and Laury task was attenuated when participants who required a large number of calibration attempts were removed. Thus, there is some evidence that eye-tracking enhanced risk averse behavior, but more work is needed to determine the strength and breadth of this effect. In addition, although null results were found in incentivized economics games, psychological questionnaires are usually not incentivized and they also potentially cover more sensitive topics, and thus may have more scope for SDB than do the games in this prior study.

Table 1

Overview of Questionnaires and Experimental Tasks Used in the Present Study

Questionnaire/Task	Citation	Constructs Measured
Questionnaires		
Domain-Specific Risk-Taking (DOSPERT)	Blais & Betz (2002)	Risk-taking in financial, health/safety, recreational, ethical, and social decision-making
Big Five Inventory (BFI)	John et al. (1991)	Five main personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism
Triarchic Pyschopathy Measure (TriPM)	Patrick et al. (2009)	Boldness, meanness, and disinhibition
Center for Epidemiological Studies-Depression (CESD)	Radloff (1977)	Current depressive symptomatology
State-Trait Anxiety Inventory (STAI)	Speilberger et al. (1971)	State and trait anxiety
Sensation Seeking Scale (SSS)	Zuckerman et al. (1964)	Disinhibition, boredom, thrill seeking, and experience seeking
Experimental Tasks		
Balloon Analogue Risk (BART)	Lejuez et al. (2002)	Risk-seeking behavior
Stroop Task	Stroop (1935), Golden et al. (1978)	Goal-directed attention
Mixed Gambles Task	Tom et al. (2007)	Risk-seeking/gambling behavior
Implicit-Association Test (IAT)	Greenwald & Banaji, 1995	Implicit biases based on race
Operation Span (OSPAN)	Turner & Engle (1989)	Working-memory capacity

Note: BF_{01} indicates the support for the null hypothesis of no difference between groups.

Based on the prior results showing a potential effect of eye-tracking during decision-making tasks involving risk, as well as our hypothesis that eye-tracking affects behavior by enhancing SDB, we selected several popular questionnaires and tasks for our study based on the following criteria:

1. The task or questionnaire has been used or cited in many peer-reviewed psychology journal articles

AND

2. The task or questionnaire measures some aspect of risk-seeking or risk-averse behavior

OR

3. Behavior on the task or questionnaire is related to a sensitive topic, which could be affected by SDB.

OR

4. The task or questionnaire requires attentional or working memory resources that may be attenuated due to distraction from the eye-tracking apparatus, or alternatively enhanced from the knowledge of being observed.

In the Method section below, we describe each task or questionnaire, along with our rationale for including each measure based on one or more of the criteria listed above. We then report the results of our experiment where participants completed six questionnaires, and five experimental tasks in either an eye-tracking or control condition.

Method

Participants

We recruited 104 participants in the eye-tracking condition (57 females, 46 males, 1 Other/Prefer Not to Respond) and 110 participants in the control condition (62 females, 44 males, 4 Other/Prefer Not to Respond). As detailed below, we planned to conduct Bayesian *t*-tests between the eye-tracking and control groups on the main dependent variables from each task or scale. Because Bayesian *t*-tests are often more conservative than null-hypothesis significance tests with an alpha level of .05 (Wetzels et al., 2011), we conducted a *power analysis* for an independent samples *t*-test with an alpha level of .01, assuming a medium-sized effect (d = 0.5). This analysis suggested that for 80% power we should collect data from at least 96 participants in each condition. We targeted this sample size and ran participants through the end of the work-week.

A medium-sized effect was assumed because we had limited time and financial constraints for running the study to have sufficient power to detect small-sized effects (Fritz, Morris, & Richler, 2012). Note that, this is a limitation of our study, as small effects could still present a problem with Hawthorne effects in eye-tracking research; however, as will be shown below, there was at least moderate support for the null hypothesis for most of the scales and tasks, and most effect sizes were less than 0.2 (Cohen's d).

Materials

Triarchic Psychopathy Measure (TriPM). The Triarchic Psychopathy Measure is a 58-item psychopathy scale developed by Patrick, Fowles, & Krueger (2009). The original paper has been cited over 1,800 times. The scale includes questions for scores in three traits that are thought to underlie psychopathy: boldness, meanness, and disinhibition. Participants were asked to answer each of the 58 statements with one of four options: true, somewhat true, somewhat false, and false. The items and scoring key for this measure can be found at: patrickcnslab.psy.fsu.edu/wiki/index.php/Triarchic_Psychopathy_Measure.

This measure was chosen because many of the questions relate to sensitive topics such as drug use and other socially questionable behaviors. Participants could possibly answer many of the questions in a less revealing way if they experience SDB because of the eye-tracking apparatus.

State-Trait Anxiety Inventory (STAI). The State-Trait Anxiety inventory is a 40-item scale developed by Spielberger and colleagues (Spielberger et al., 1971). Spielberger et al. has been cited over 8,700 times. Twenty items measure state anxiety, or "how you feel right now, that is, at this moment," while the other twenty items measure trait anxiety, or how one generally would describe themselves. We were primarily interested in whether the eye-tracking condition showed greater state anxiety, due to a greater perception of being observed.

Sensation Seeking Scale (SSS). The Sensation-Seeking Scale was developed by Zuckerman and colleagues in the 1960s (Zuckerman, Kolin, Price, & Zoob, 1964). The original paper has been cited over 1,700 times. The four dimensions of the SSS are: Disinhibition, Boredom Susceptibility, Thrill and Adventure Seeking, and Experience Seeking. We included this scale because it measures aspects of risk-seeking behavior. We hypothesized that participants in the eye-tracking condition may report less risk-seeking behavior, due to SDB.

Big Five Inventory. The Big Five Inventory (BFI), developed by John, Donahue, & Kentle (1991), is a 44-item scale that measures five major personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. The original paper has been cited over 600 times. This scale was included to examine whether constructs such as agreeableness or conscientiousness were reported differently in the eye-tracking condition. For example, participants might be more likely to endorse items related to conscientiousness due to SDB.

Domain-Specific Risk-Taking (DOSPERT). The DOSPERT was created by Weber, Blais, and Betz (2002), and is used to measure risk taking in five settings: financial decision making, health/safety, recreational, ethical, and social decision making. Weber et al. (2002) has been cited over 3,800 times. This measure was included to examine whether the eye-tracking manipulation increased self-reported risk-seeking behavior.

Center for Epidemiological Studies-Depression (CESD). This scale made by Radloff (1977) is used to measure depression levels amongst the general populace. It is a self-reported scale in which individuals answer survey questions about how often they felt a certain way during a week. The original paper has been cited over 63,000 times. This measure was included to examine whether participants in the eye-tracking condition reported fewer depressive symptoms due to SDB.

Operation span (OSPAN). The OSPAN was developed by Turner and Engle (1989) to measure working memory. This paper has been cited over 3,300 times. We hypothesized that enhanced social desirability bias from the eye-tracking manipulation may either co-opt working memory resources, and serve as a distraction in this task, leading to poorer performance compared to the control group, or, alternatively that working memory may be enhanced by the manipulation, due to participants' knowledge that they are being observed.

We programmed the OSPAN using Javascript within Qualtrics, following the procedures used in the automated OSPAN (Unsworth, Heitz, Schrock, & Engle, 2005). The task consists of 75 trials where participants must solve simple math problems while remembering letters presented to them after each math problem. On each trial, participants completed a modular arithmetic math problem that involves a single-digit multiplication operation followed by a single-digit addition or subtraction operation, such as: (3 X 3) – 2 = ?." Participants had as long as they wished to respond. When they were ready to answer the math problem, they clicked their mouse on a button labeled "Answer Problem." They were then shown a number and clicked "TRUE" if the number was the correct answer to the modular arithmetic problem or "FALSE" if the number was not the correct answer. They were given a target of maintaining 85% accuracy on the modular arithmetic problems.

After each modular arithmetic problem was completed, participants would be shown a letter from the set: [F, H, J, K, L, N, P, Q, R, S, T, Y]. These letters have been used in prior work using the OSPAN because they are all pronounced with one syllable, and thus should take similar time to rehearse in working memory (Unsworth et al., 2005). After a span of 3-7 modular arithmetic-letter presentation trials, participants would see the 12 letters above printed on the

screen, in four rows of three, and asked to recall the sequence of letters they had just seen, in order. Participants completed three spans each for each span length, from 3 to 7, for a total of 75 trials.

Balloon Analogue Risk Task (BART). The BART was developed by Lejuez and colleagues (2002). This paper has been cited over 2,700 times. The BART measures the propensity to take risk via a computerized simulation of pumping up a balloon. Participants are told that each pump is worth a fixed amount of money so that more pumps lead to higher reward. However, each pump increases the chance of the balloon popping and losing all the money. In the original BART, participants would blow up the balloon, and after each pump, they would see if the balloon exploded. If it did not explode, then participants would receive additional points. Participants could stop at any point between pumps and take the points they had earned up to that moment, but if the balloon exploded, they forfeit any accumulated earnings and receive nothing for that trial.

We ran a modified version of the BART, the automated BART (Pleskac, Wallsten, Wang, & Lejuez, 2008; DeMartini et al., 2014), because given a data truncation in the original task on trials where the balloon explodes, it is not possible to determine the intended number of times participants would have pumped the balloon, had it not exploded. In our automated version, participants clicked the mouse to inflate the balloon as many times as they wished, with the balloon inflating and one point added after each click. After inflating the balloon as much as they liked, participants clicked on a button labeled "Stop. See if balloon exploded." If the balloon did not burst, then they were shown a screen labeled "The balloon did not burst! This round you earned *x* points. These points have been added to your total." where *x* was the number of points participants earned per round. If the balloon exploded, then the screen showed the balloon explode and fly away. Participants were then informed that "The balloon has exploded after pumping *x* times. You did not make any money this round."

Participants played a total of 20 rounds. They were told that the maximum number of times they could pump the balloon was 32.[1] After participants selected to stop pumping and see if the balloon exploded, the probability of the balloon exploding was the number of pumps divided by 32, the maximum number of pumps. Thus, if participants pumped the balloon 16 times, they had a 50/50 chance of the balloon exploding versus not. Sixteen is the optimal number of times to pump the balloon that maximized the expected value of 8 points (16 points times a probability of 0.5).

Following the literature, this task did not include real monetary rewards that were contingent upon performance; participants were just told to try to earn as many points as possible.

Mixed Gambles Task. The Mixed Gambles Task, developed by Tom et al. (2007), presents a gamble that has a 50/50 chance of resulting in a gain of one amount of money or a loss of another amount. The task consisted of 256 trials. Potential gain and loss sizes were adjusted independently—gains ranged from $10 to $40 in $6 increments, and losses ranged from $5 to $20 in $3 increments. The ranges were designed from findings that the normal population is on average twice as sensitive to losses as to gains (Tom et al., 2007). The possible gains and losses

[1] We used a shortened version of the BART compared to Lejeuz et al., 2002, and Pleskac et al., 2008. In their task the maximum number of pumps on each round, or trial, was 128, and the optimum number of pumps was 64. Participants played a total of 90 rounds in Lejeuz et al., 2002. The shortened version was used here to allow time for the other tasks and measures in the study.

were factorially combined so that participants completed one trial for each possible gain-loss pair.

On each trial, participants were told that they could accept or reject a 50/50 gamble of gaining or losing the gain-loss amounts for that trial. They then clicked Accept or Reject. If they clicked Reject, then they went on to the next trial. If they clicked Accept, then a random number was drawn that determined if they gained or lost points on that trial. If they gained points, they were shown a screen that said "You Gained Points!" along with the number of points they won. If the participant lost points the screen would display, in red font, "Sorry, but you lost points," along with the amount lost.

Similar to BART, participants did not gain or lose any amount of money, based on their choices, but were told to try to earn as many points as possible over the course of the task.

Stroop Task. This task was invented by John R. Stroop (1935) and updated by Golden, Freshwater, & Golden (1978). The original paper has been cited about 25,000 times. The classic Stroop tasks asks participants to name the color of the font for different color words, such "GREEN" printed in green font, or "RED" printed in yellow font. The former is an example of a congruent stimulus, while the latter is considered incongruent. Participants are asked to name the color of font the word is printed in, but there is a prepotent bias to semantically process the color words, and respond "RED" even though the word is printed in yellow font.

On each trial of our computerized version of the Stroop task, the words RED, GREEN, YELLOW, and BLUE were presented in a random order, and participants were asked to select buttons labeled "red," "blue," "green," "yellow" to indicate the font color of the words from top to bottom. Participants completed a total of 32 trials, 75% of which were congruent trials where the font color matched the meaning of the word, and 25% of trials were incongruent where the font color and word meaning mismatched. The main dependent variable of interest was the difference in response times for congruent and incongruent trials. The canonical Stroop interference finding is that response times are longer for incongruent trials. This task was included as a measure of basic goal-directed attention, to examine whether the eye-tracking apparatus led to more Stroop interference, compared to the control group.

Implicit-Association Test (IAT). The IAT measures association or stereotypes between concepts. It was created by Greenwald and Banaji (1995) and has been cited over 10,000 times. Greenwald, McGhee, & Schwartz (1998) developed an IAT to examine implicit attitudes based on race (white vs. black). The premise behind the IAT is that participants will more rapidly associate stimuli that are compatible or generally related to one another, than they will with less compatible stimuli. The black-white race IAT was included because it measures the sensitive topic of bias against black individuals. According to IAT theorists, if participants possess this type of implicit bias, then they will be slower to associate positive stimuli with faces of black individuals, compared to white individuals. We used the Black-White race IAT that was developed for use in Qualtrics, using Javascript by Carpenter and colleagues (2019). The task was created using the *iatgen* package for R (iatgen.wordpress.com).

Participants were instructed to place their left and right index fingers on the "E" and "I" keys. At the top of the screen were shown two categories; in the first block these were: African-American or Euro-American. On each trial a picture or word is shown and participants had to match it to its correct category. In Block 1 faces of black or white individuals are shown, and participants pressed E or I to place the face into the African-American or Euro-American category, respectively. If they made an incorrect classification then an "X" would appear in red

font, until they made the correct answer. In the second block the categories changed to Pleasant and Unpleasant, and participants classified words such as: Horrible, Awful, Terrible, War, Friend, Joy, Laughter, Please, into the correct category. In blocks 3 and 4, participants sorted either words or faces into two categories: 1. African-American or Pleasant or 2. Euro-American or Unpleasant, with the E and I keys, respectively. In block five participants classified faces as Euro-American or African American, but with the key associations changed, so that participants selected E for Euro-American, and I for African-American. In blocks 6-7 participants sorted words or faces into the two opposite categories from blocks 3 and 4: 1. Euro-American or Pleasant or 2. African-American or Unpleasant. The critical comparison is response times for blocks 4 and 7, with blocks 3 and 6 being practice blocks. If participants have an implicit bias favoring white individuals, then they are predicted to respond faster when Euro-American is paired with pleasant, than when it is paired with unpleasant words.

Equipment

A Tobii eye-tracker was mounted at the bottom of the monitor for each computer used in the study. Tobii X2-60 and Pro Spectrum were used for the experiment. Figure S1 shows a photograph of the computer and eye-tracker at a workstation used for running the experiment. Participants were informed that the device at the bottom of the screen was an eye-tracker, and there was a full calibration procedure for each participant in the eye-tracking condition. A staff member started the calibration for each participant and the calibration proceeded autonomously with each participant following a small dot moving along the screen to identify calibration quality at 9-points in the screen. A researcher checked that the calibration was of sufficient quality before the experiment began. Participants were aware that their eye movements were being recorded, but there was no researcher present who was monitoring participants' eye-movements in real-time.

Procedure

Lab sessions took place in the Human Behavior Lab (HBL) at Texas A&M. Participants were undergraduate students who completed the study for partial fulfillment of a course requirement. When subjects arrived at the lab, they were randomly placed at a computer station where they completed the psychology tests through Qualtrics via iMotions software. Participants also completed a demographics survey. The study was a between-subject design. Participants in the control group completed the tests without the presence of eye-tracking or web-camera. Participants in the treatment group completed the tests with the presence of eye-tracking and web-camera. Treatment conditions and the ordering of psychology tasks were randomized at the session-level.

Upon arriving at HBL, subjects were asked to read and sign one of two consent forms depending on the assignment. The treatment consent form included specific language consenting to the use of eye-tracking during the experiment. The control consent form had no language about eye-tracking procedures and equipment. Before each test, participants in the eye-tracking condition were verbally reminded about the eye-tracking equipment and it was re-calibrated. This recalibration was meant to remind subjects of the presence of the equipment in the eye-tracking condition before each test to keep the presence of the equipment salient, as would be the case if they were only doing one study at a time.

Eye-tracking calibration took approximately 15-30 seconds depending on the difficulty of the equipment to capture the pupil reflection pattern. Eye-tracking was calibrated to capture at least 80% of pupil movement during the calibration time. After five calibration attempts, the eye-

trackers were overridden. Treatment participants who were unable to be calibrated were told that the failed calibration attempts were not their fault. This was meant to ensure that the participant did not feel guilty. The number of calibration attempts were recorded for each participant and each test. If calibration had to be overridden, the override was recorded for each participant and for the specific test. Participants in the control group waited two minutes between each task to ensure a balance between the two conditions. The eye-tracking data were not analyzed, as our focus was on how the apparatus affected behavior in the tasks.

Data Analysis

For each measure, we analyzed the dependent variables that have typically been examined in prior research. For each of the questionnaires, the individual scales were scored according to the guidelines established in the studies listed in Table 1 above. For each of the experimental measures we also examined the dependent variables consistently examined in prior research. For the OSPAN task, we analyzed memory performance for each different span-length, from 3-7 letters; this is slightly different from prior research where a single score is computed across all span-lengths. For the Mixed Gambles task, we analyzed the proportion of gambles accepted, consistent with Tom et al., 2007, and we additionally examined the response times for participants from each group, which, to our knowledge, has not been examined in previous studies using this task. We consider these analyses as more exploratory because of their novelty. Our analyses of the BART, Stroop, and IAT tasks use identical dependent variables to those used in prior research (Lejeuz, et al., 2002; Golden et al., 1978; Greenwald & Banaji, 1995).

We conducted Bayesian *t*-tests between the eye-tracking and control group on the key dependent variables from each measure task and questionnaire. We used JASP software for our Bayesian analysis (jasp-stats.org; version 0.17.2.1) using the default Cauchy prior (.707). Bayes Factors can be reported in terms of evidence supporting either the null (BF_{01}) or alternate (BF_{10}) hypothesis. We mainly report Bayes Factors for null hypotheses because we predicted no differences between groups on most of the dependent variables and we were interested in quantifying the degree of support for the null hypothesis. A Bayes Factor (BF_{01}) of 3 or more is considered to indicate moderate support for the null hypothesis (Wagenmakers et al., 2018; Jeffreys, 1961), although Bayes Factors can be interpreted continuously on an odds scale. For example, a Bayes Factor (BF_{01}) of 3 suggests that the null hypothesis is three times more likely than the alternate hypothesis, given the data.[2] In addition to Bayes Factors we also report effect sizes.

We ran Bayesian general linear mixed-effects regression to analyze data from most of the experimental tasks, such as the mixed gambles, OSPAN, Stroop, and BART, using the R package called brms (Bürkner, 2017). Brms provides parameter estimates for both fixed and random effects. We examined the fixed coefficient values from models where condition (eye-tracking versus control) was used as a predictor for the main dependent variables in each task. We considered an effect to exist or be 'significant' if the 95% highest credible interval (HCI) for the predictor did not include zero (Nalborczyk et al., 2019; Byrne et al., 2020).

Our main focus in the paper was to examine differences between conditions; however, we also conducted analyses with gender included as a predictor. These are presented in the Supplemental Materials. We did not find that gender interacted with condition for any measure,

[2] Bayes Factors for the alternate and null hypotheses are inverse of each other ($BF_{10} = 1/ BF_{01}$).

except for an analysis of response times in the mixed gambles task, which we briefly discuss below.

Results

Questionnaires. We conducted Bayesian *t*-tests on the constructs from each of the six questionnaires listed in Table 2. The table lists the means for each group, along with the effect size for the difference between groups, and the Bayes Factor indicating support for the *null* hypothesis (BF_{01}) that no difference exists between groups. There was moderate support for the null hypothesis for all the constructs except for agreeableness from BFI. Participants in the control group scored about 1 point higher on agreeableness. A Bayes Factor of 1.91 suggests that the null hypothesis is slightly more supported than the alternate hypothesis, but the effect is weak (Jeffreys, 1961).

Table 2

Statistical Results for Comparisons Between Groups on Dimensions from Questionnaire Scales

Scale/Construct	Eye-Tracking M(SD)	Control M(SD)	Cohen's *d*	BF_{01}
TriPM				
Boldness	31.47 (7.57)	31.46 (7.57)	0.00	6.69
Meanness	13.60 (7.63)	12.38 (8.26)	0.15	3.72
Disinhibition	17.86 (7.46)	18.11 (7.93)	0.03	6.52
STAI				
State Anxiety	40.24 (9.98)	40.16 (12.35)	0.01	6.68
Trait Anxiety	44.39 (9.41)	44.38 (11.13	0.00	6.67
SSS				
Disinhibition	4.26 (2.60)	4.06 (2.50)	0.08	5.75
Boredom	2.69 (1.68)	2.54 (1.73)	0.09	5.47
Thrill Seeking	6.58 (2.44)	6.31 (2.56)	0.11	4.97
Experience Seeking	4.70 (1.85)	4.86 (1.94)	0.08	5.66
BFI				
Openness	34.41 (6.33)	35.14 (6.24)	0.12	4.78
Conscientiousness	31.36 (5.10)	30.77 (5.98)	0.11	5.07
Extroversion	27.54 (6.86)	26.98 (7.04)	0.08	5.68
Agreeableness	33.78 (5.26)	34.99 (5.55)	0.22	1.91
Neuroticism	24.96 (6.54)	24.82 (6.18)	0.02	6.61
DOSPERT				
Ethical	27.29 (3.33)	27.88 (3.93)	0.16	3.48
Financial	29.80 (3.56)	29.86 (3.88)	0.02	6.65
Health/Safety	26.98 (3.51)	27.37 (3.72)	0.11	4.99
Recreational	26.52 (3.85)	26.16 (3.95)	0.10	5.35
Social	29.46 (3.43)	29.04 (3.49)	0.12	4.63
CESD				
Depression	18.13 (9.48)	18.82 (11.05)	0.07	5.99

Note: BF_{01} indicates the support for the null hypothesis of no difference between groups.

OSPAN. A total of 109 and 103 participants, in the control and eye-tracking conditions, respectively, performed the OSPAN task. We first analyzed the proportion correct for each participant on the modular arithmetic problems. We removed seven participants who did not

achieve an accuracy on the math problems of at least 85%. This left 105 and 100 participants in the control and eye-tracking conditions respectively.

Next, we computed the proportion of letters correctly recalled for each span length (3-7), for participants in each condition (Figure 1). We fit a Bayesian general linear logistic regression model to predict whether participants correctly recalled the letter in the correct order and for each span. Correct recall was regressed on the interaction between condition and span length, with a random intercept for each participant. This model suggested a non-zero interaction effect, $b = 0.094$, $SE = 0.044$, *95%* HCI = [0.005, 0.179]. It appears that participants in the eye-tracking condition performed worse than participants when recalling letters from spans of 4, but slightly better than control participants when recalling letters from the other span lengths (Figure 1). We ran a reduced model on the data only for span length of 4. This model predicted correct responses from condition, with random intercepts for participants. The difference between the control and eye-tracking conditions was non-zero, $b = -0.823$, $SE = 0.413$, *95%* HCI = [-1.654, -0.028]. Participants in the eye-tracking condition were about 4% less accurate when recalling 4-letter spans compared to participants in the control condition. However, participants in the eye-tracking condition were more accurate than control condition participants for every other span length.

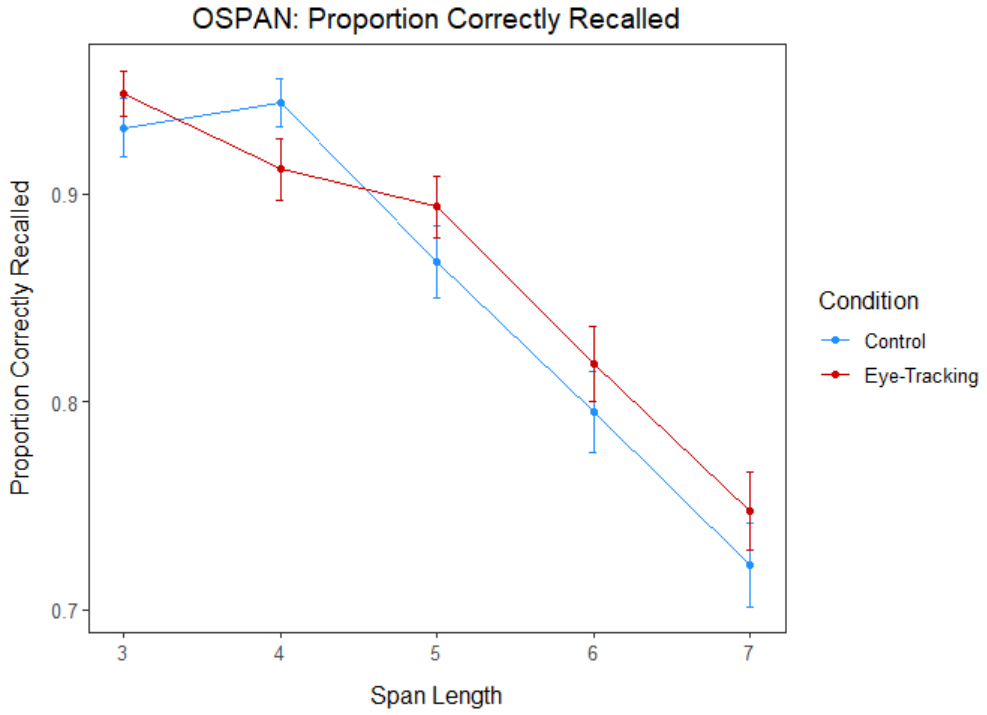


Figure 1. Proportion of letters correctly recalled for different span lengths. Error bars represented standard errors of the mean.

BART. A total of 106 and 104 participants in the control and eye-tracking conditions, respectively, completed the BART. We first examined the average pumps made by participants in each condition. A Bayesian *t*-test suggested moderate support for the null hypothesis of no difference between the eye-tracking ($M = 14.07$, $SD = 4.24$) and control ($M = 14.57$, $SD = 3.66$) conditions in average number of pumps, $d = .12$, $BF_{01} = 4.54$.

We next examined the average proportion of trials where the balloon exploded, for participants in each condition. Participants in the control condition ($M = 0.612$, $SD = 0.109$) averaged slightly more explosions than participants in the eye-tracking condition ($M = 0.575$, $SD = 0.142$.) A Bayesian t-test showed weak support for the null hypothesis, $d = .29$, $BF_{01} = 1.13$.

Mixed Gambles Task. There were a total of 108 and 103 participants in the control and eye-tracking conditions, respectively. We first computed the average proportion of trials where the gamble was accepted for each participant. One participant in the control condition accepted the gamble on all 256 trials; this participant's data set was excluded from further analyses, leaving 107 participants in the control condition. Participants in the control condition accepted the gamble on 60.2% of trials ($SD = 13.6\%$) while participants in the eye-tracking condition accepted the gamble on 58.7% of trials ($SD = 14.4\%$). A Bayesian t-test indicated moderate support for the null hypothesis of no difference in the proportion of gambles accepted, $d = 0.10$, $BF_{01} = 5.11$.

We next examined whether gambling behavior differed across trials with varying amounts that could be gained or lost, by computing the expected value (EV) of each gamble:

$$EV = 0.5 * gain - 0.5 * loss \quad (1)$$

If the amount to be gained was larger than the amount that could be lost, then the EV for a given gamble was positive; if the amount that could be lost was larger than the gain, then the EV was negative. The EV would be 0 if the amounts to be gained or lost were equal. The optimal strategy in the task is to accept every gamble that has a positive EV, and reject every gamble with a negative EV.

We ran a Bayesian general linear logistic regression model with the interaction between condition and EV predicting whether the gamble would be accepted or not. The model also included random intercepts for participants. There was a strong, non-zero interaction effect, $b = 0.050$, $SE = 0.005$, $95\%\ HCI = [0.041, 0.060]$. Participants in the eye-tracking condition accepted fewer low-EV gambles, but accepted more high-EV gambles, compared to control participants. Figure 2 shows the simple slopes predicting the probability of accepting the gamble, based on its EV.

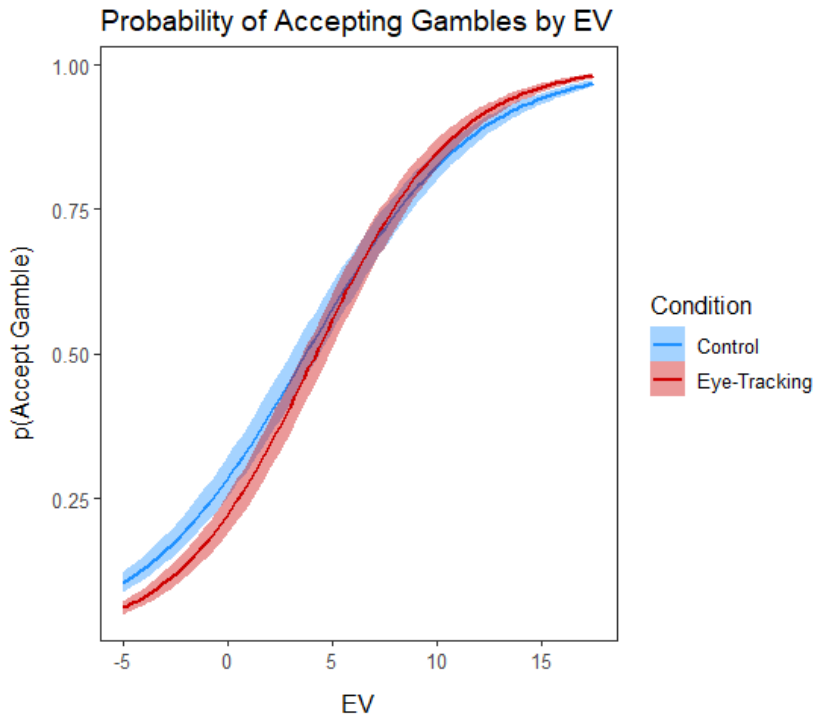


Figure 2. Simple slopes from the GLM predicting 'Accept' from the interaction between EV and condition.

We performed a similar analysis where we examined the average EV for accepted versus rejected gambles within each condition (Figure 3). We ran a Bayesian general linear logistic regression model with EV regressed on the interaction between choice type and condition, with random intercepts for participants. There was a strong, non-zero interaction effect, b = -0.665, *SE* = 0.089, *95% HCI* = [-0.840, -0.488]. On average, participants in the eye-tracking condition accepted gambles higher in EV and rejected gambles lower in EV than participants in the control condition.

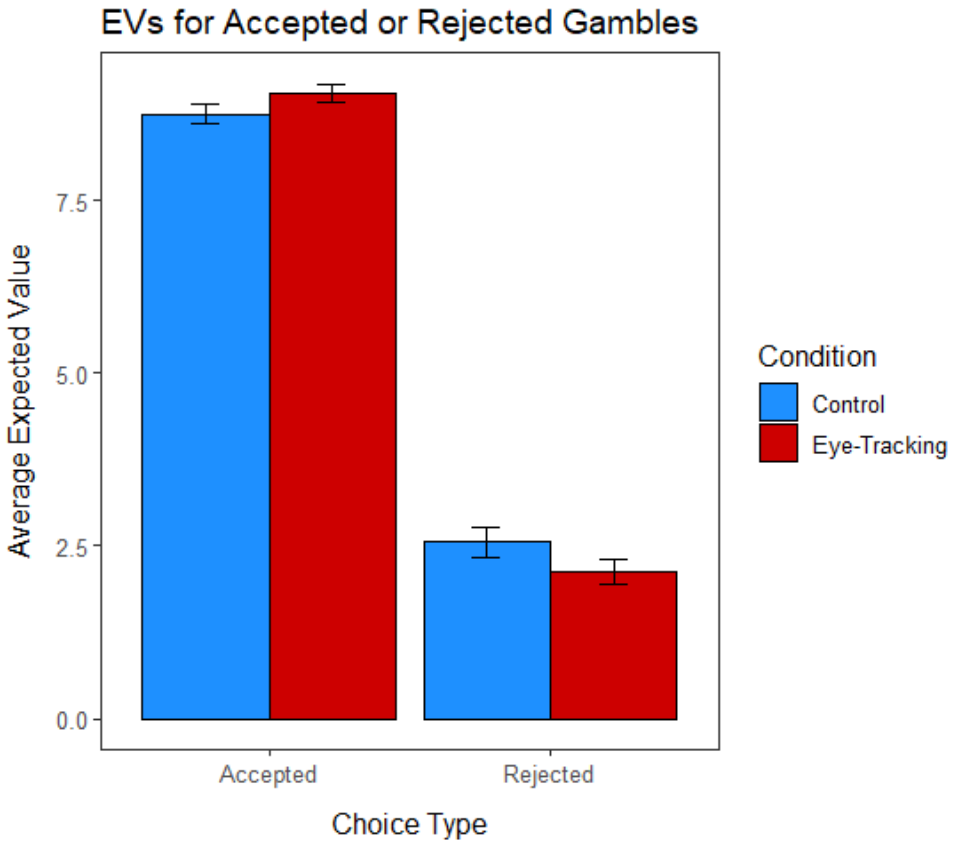


Figure 3. Average expected values (EVs) for accepted versus accepted gambles. Errors bars represent standard errors of the mean.

Next, we examined whether participants in each group differed in their response, or decision times for each gamble. We first computed the overall mean and standard deviation for response times, across all participants and trials, and removed trials with response times that were greater than two standard deviations above the mean (Ratcliff, 1993; Whelan, 2008). Out of 53,760 total trials, 2,158 were removed based on this criterion.

With the remaining trials we ran a Bayesian mixed effects model with response time regressed on the interaction between condition and EV, with random intercepts for participants. There was a strong, non-zero interaction effect, $b = 3.77$, $SE = 1.80$, *95% HCI* = [-7.19, -0.17]. Figure 4 shows the simple slopes predicting response times based on EV for each condition. Participants in the eye-tracking condition took longer to respond than participants in the control condition when the EV of the gamble was smaller. As EV increased, participants in both conditions responded quicker, with the difference in response time between conditions becoming negligible for high-EV gambles. We also ran an additive model predicting RTs from EV and condition, with random intercepts for participants, and found an effect for EV, $b = -7.33$, $SE = 0.91$, *95% HCI* = [-9.12, -5.55], where responses were faster for longer EVs, but there was no overall effect of condition, $b = 47.61$, $SE = 83.19$, *95% HCI* = [-105.99, 207.48].

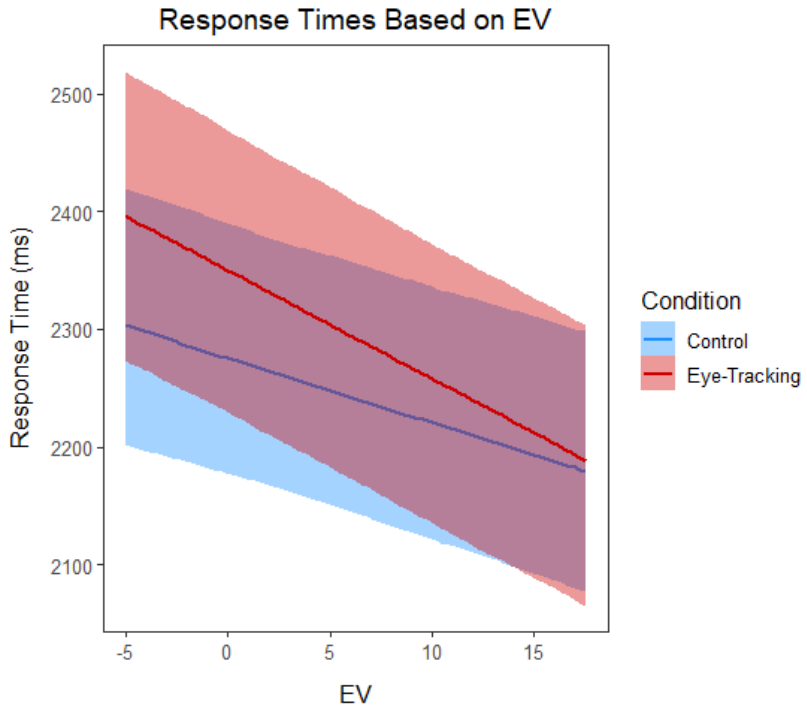


Figure 4. Regression slopes showing average response times (in ms) for gamble choices with different expected values. Participants decided to accept or reject the gamble on each trial.

It is worth noting that the results with gender included as a predictor, which are presented in the Supplemental Materials, showed that only male participants exhibited this interaction between condition and EV in predicting response times. There was no interaction between EV and condition for females. This analysis was the only one that indicated an interaction between gender and condition.

Stroop Task. The classic Stroop interference effect is that response times are longer for incongruent stimuli than for congruent stimuli. We first examined whether this effect was present and whether it differed between participants in each condition. Response times greater than two standard deviations above the mean were removed (Ratcliff, 1993; Whelan, 2008). This removed a total of 235 out of 6,752 trials. We then computed the average response times on congruent and incongruent trials for each participant and averaged these values across participants within each condition. Figure 5 plots these average response times.

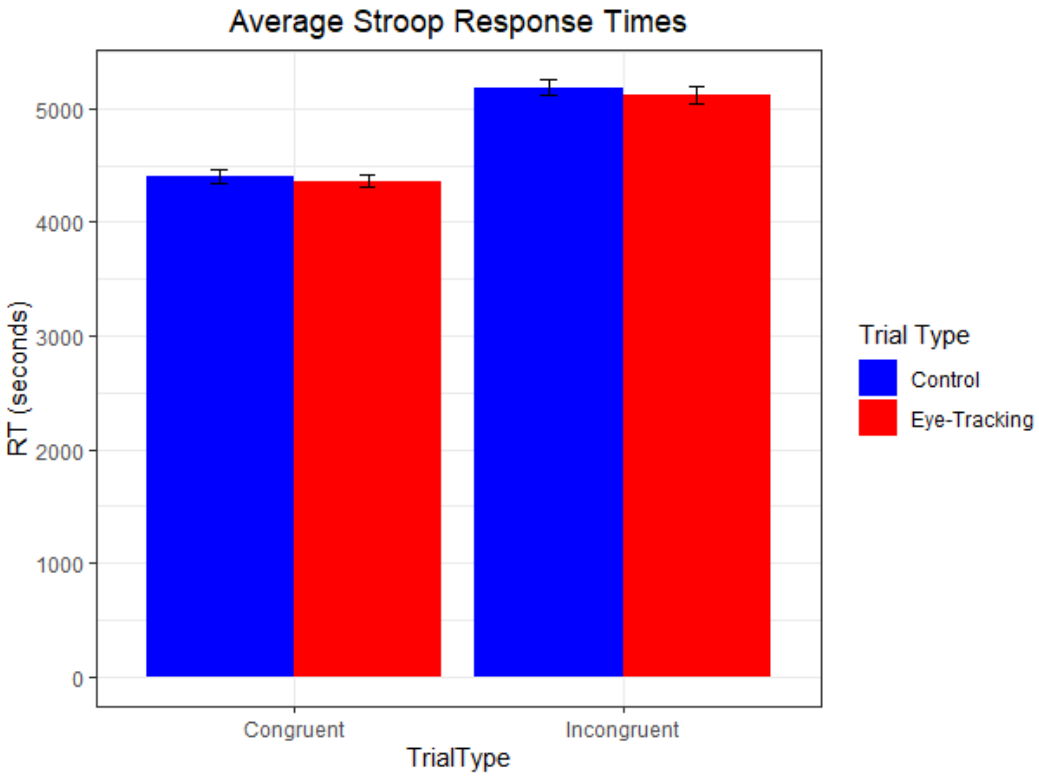


Figure 5. Average response times for congruent and incongruent trials for each condition. Error bars represent standard errors of the mean

As expected, response times were longer on incongruent than congruent trials, indicating a successful replication of the basic Stroop effect (Stroop, 1935). It also appears that the pattern or magnitude of the Stroop effect did not differ between conditions. To confirm this, we ran a Bayesian multilevel regression model where response times were predicted from the interaction between trial type and condition, with random intercepts for participants. The 95% highest credible interval for the interaction effect included zero well within its bounds, indicating no interaction between condition and trial type, b = -20.70, SE = 54.63, *95% HCI* = [-128.44, 83.92]. A similar model, but with the additive effects of trial-type and condition used as predictors for response time, indicated a large effect of trial-type, b = 758.60, SE = 26.80, *95% HCI* = [704.20, 810.47], but no effect of condition, b = -37.46, SE = 82.01, *95% HCI* = [-199.13, 122.36].

Figure 6 shows the proportion of error trials for participants in each condition for the two trial types. Although participants in the eye-tracking condition made slightly fewer errors on incongruent trials than participants in the control condition, a multilevel regression model with error as the outcome variable (coded as 1 for error trials, 0 for correct trials) and the interaction between condition and trial-type as the predictors, indicated no interaction effect, b = -0.054, SE = 0.205, *95% HCI* = [-0.468, 0.350].

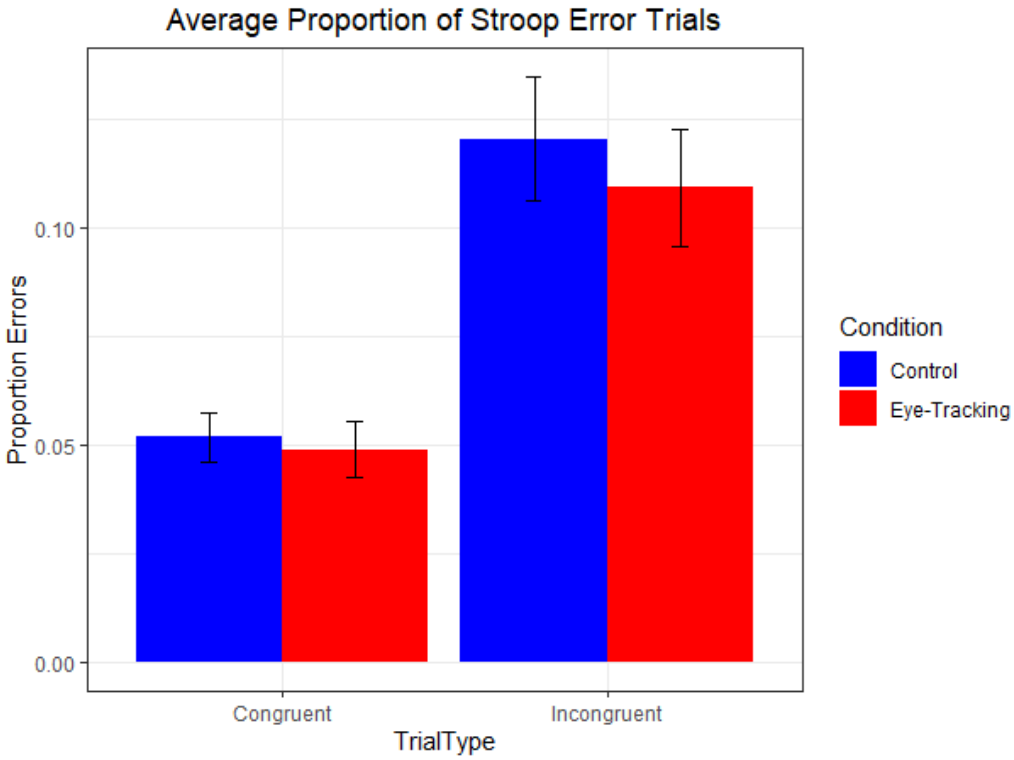


Figure 6. Average error rates for congruent and incongruent trials for each condition in the Stroop Task. Error bars represent standard errors of the mean

Implicit Attitudes Test (IAT). Code from the developers of the IATGen software package was used to analyze the Black-White race IAT data. The *D*-score data-cleaning and scoring algorithms were used that were used in prior research (Carpenter et al., 2019; Greenwald et al., 2003, Lane et al., 2007). A very small proportion of overall trials were removed for timing out (0.0002). No participants were removed for responding too fast (i.e. "button-mashing"). A total of 5.7% of trials were removed for errors, which is consistent with less than the 10% or error trials typically seen in prior work (Carpenter et al., 2019).

Figure 6 plots the average *D*-scores for participants in each condition. A Bayesian *t*-test indicated moderate support for the null hypothesis that assumes no difference between groups, $BF_{01} = 6.49$, Cohen's $d = 0.01$. A one-sample Bayesian *t*-test to test whether *D*-scores were significantly different from zero, indicated extreme support for the alternative hypothesis, $BF_{10} > 100$, Cohen's $d = 0.78$. This suggests a strong pro-White bias for participants, that did not differ between conditions.

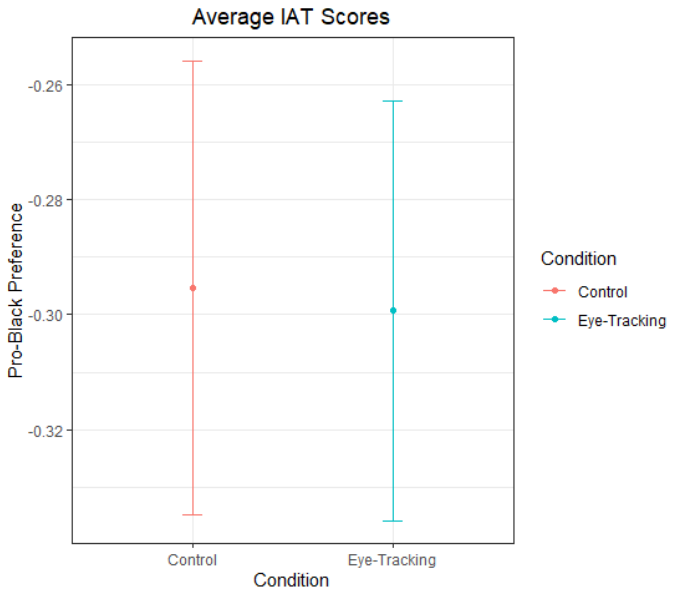


Figure 6. Average error rates for congruent and incongruent trials for each condition in the Stroop Task. Error bars represent standard errors of the mean

Discussion

We examined whether the presence of eye-tracking equipment would affect behavior in several popular psychological tasks due to SDB or distraction due to knowledge that one's eye movements were being monitored. For nearly all tasks and measures there was no effect of the manipulation, and we were able to quantify support for the null hypothesis using Bayesian statistical methods (Wagenmakers et al., 2018). However, for the mixed gambles task, we observed some strong effects of the eye-tracking manipulation. Participants in the eye-tracking group were more risk-averse than control participants, when making decisions on whether to accept gambles with small EVs. The eye-tracking manipulation led participants to perform in what could be considered a more optimal or rational manner, as they rejected more gambles that had a negative EV, than control participants, also accepted more high-EV gambles. Participants in the eye-tracking condition also took longer to respond than control participants for the low-EV gambles. This suggests that they were more deliberative and more risk averse. This difference in behavior between the two groups is consistent with the hypothesis that eye-tracking may induce an SDB, which leads participants to be more selective and restrained in their gambling behavior.

The increased risk aversion in the mixed gambles task is consistent with the results from Kee and colleagues (2021), who observed greater risk aversion for eye-tracking condition participants in the Holt and Laury risk assessment task. Thus, the presence of eye-tracking equipment may lead to more risk-averse behavior under certain conditions. Researchers should be aware of these potential effects, particularly if participants are asked to make risky or irrational decisions, such as whether to accept gambles with negative EVs. In the mixed gambles task, participants probably had somewhat clear knowledge that they would be making an irrational, and questionable decision by accepting gambles low in expected value. This task involves decision-from description, as opposed to decision from experience (Hertwig, Barron, Weber, & Erev, 2004), because participants had full knowledge of the contingencies associated

with decisions to accept or reject gambles. In tasks involving decisions from experience, there may be less of an effect of SDB because it is less clear which choices are socially desirable.

The BART is an example of a decision from experience task, because participants have to learn how the number of pumps affects the likelihood of the balloon exploding through repeated trial and error. Although the BART task involves risky decision-making, participants probably only had vague knowledge as to how many pumps were considered rational. This may have been why we did not observe any difference between conditions in the BART; there was no clearly optimal behavior, and so less pressure from SDB. Alternatively, it could be that SDB is irrelevant to the part task, because it assesses individuals' risk tolerance and rationality, which are not affected by a desire to be viewed favorably by others. DOSPERT and SSS are other measures that involve sensitivity to risky choices, but we did not observe any differences based on condition on constructs from these two scales. These are self-reported scales, and we did not find any evidence that participants in the eye-tracking condition reported different levels of potentially undesirable behavior on *any* of the scales they were given. For the self-reported scales, there was at least moderate support for the null hypothesis for every measure, except agreeableness, from the Big 5 Personality Inventory, for which there was still anecdotal support for the null. Thus, it appears that the eye-tracking apparatus had no effect on any of the self-report measures, even though several measures included questions related to sensitive topics.

The only other reliable effect was the interaction between condition and span length in the OSPAN task. Participants in the eye-tracking condition performed worse on span lengths of 4, but slightly better on all other span lengths than control participants. We did not anticipate this type of interaction, and so we urge caution in drawing conclusions from this finding. More work is needed to determine if there is a theoretical explanation for this interaction, or if it is a random effect, due to the large number of hypotheses tested throughout this paper, that is not likely to be replicated.

This study examined Hawthorne effects on several different tasks and measures that are commonly used in psychology and related fields. Nevertheless, our battery was not exhaustive, and future studies should examine whether equipment such as eye-tracking produce Hawthorne effects that might be unanticipated, or unaccounted for by researchers using the equipment to test various empirical questions. One set of tasks that could be more prone to Hawthorne effects from eye-tracking are visual-search tasks where participants must find items in different visual arrays (Wolfe, 2010; Treisman, 1982; Anderson, Laurent, & Yantis, 2011). Eye movements are more relevant to visual search tasks than the tasks and measures used in the present study, so it is possible that knowledge that one's eyes are being tracked could be a source of distraction in these types of tasks, more so than for tasks like the OSPAN or Stroop tasks. It is also important to note that our results are most externally valid for modern eye-trackers that sit at the base of a monitor and require less researcher interaction than earlier styles. Eye-trackers that require extensive calibration, chin-rests, or goggles may have different levels of salience. Thus, different eye-tracking equipment and designs could be more attention-demanding and distracting. More research is needed to determine if more invasive or noticeable eye-tracking procedures and apparatuses, such as those that study visual search, use glasses, or require more involvement with the researcher have Hawthorne effects.

Based on our results, it appears that Hawthorne effects are only likely to occur for risky decision-making tasks involving decision from description, where the potential gains and losses are clearly described before participants make each choice (Hertwig et al., 2004). The mixed gambles task we used is the only decision-from-description task used in our study. Some other

examples are delay discounting tasks or the tasks used to examine the predictions of prospect theory (Tversky & Kahneman, 1971; 1992). The mixed gambles task used in the present study involved only 50/50 gambles, but there are many other possible tasks involving different probabilities and amounts of reward that have been examined in the literature (see Erev et al., 2010). It is possible that SDB from an apparatus such as that used for eye-tracking may affect behavior in these and similar tasks as well. Overall, eye-tracking researchers should not worry much about Hawthorne effects for most behavioral tasks, but the results of this study should be considered when planning eye-tracking studies involving risky decision-making, particularly cases where participants might have a reasonable idea of what the optimal choice would be.

Declarations

Funding

The authors have no funding to declare

Conflicts of Interests/Competing Interests

The authors have no conflicts of interest to declare.

Ethics Approval

This study was approved by the Internal Review Board at Texas A&M University (IRB2021-098D).

Consent to Participate

All participants gave informed consent.

Consent for Publication

All participants were informed that the anonymized data would be published.

Availability of Data and Materials

Data and analysis materials are available on the Open Science Framework through the following link: https://osf.io/3fdrp/

Code Availability

Analysis code is available as the OSF link listed in the previous section.

Authors' Contributions

DAW designed the study, programmed the experiments, analyzed the data, and wrote the first draft of the manuscript. JNL designed the study, analyzed the data, and edited the manuscript; SLP designed the study, oversaw data collection, analyzed the data, and edited the manuscript, MAP designed the study, oversaw data collection, and edited the manuscript.

References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*(2), 334.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, *108*(25), 10367-10371.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1-28.

Byrne, K. A., Peters, C., Willis, H. C., Phan, D., Cornwall, A., & Worthy, D. A. (2020). Acute stress enhances tolerance of uncertainty during decision-making. *Cognition*, *205*, 104448.

Carpenter, T. P., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S., LaBouff, J., ... & Chakroff, A. (2019). Survey-software implicit association tests: A methodological and empirical analysis. *Behavior Research Methods*, *51*, 2194-2208.

Charles, J.L.K., & Dattalo, P.V. (2018). Minimizing social desirability bias in measuring sensitive topics: The use of forgiving language in item development.

Chiesa, M., & Hobbs, S. (2006). Making sense of social research: How useful is the Hawthorne effect? *European Journal of Social Psychology, 38*(1), 67-74.

DeMartini, K. S., Leeman, R. F., Corbin, W. R., Toll, B. A., Fucito, L. M., Lejuez, C. W., & O'Malley, S. S. (2014). A new look at risk-taking: using a translational approach to examine risk-taking behavior on the balloon analogue risk task. *Experimental and Clinical Psychopharmacology*, *22*(5), 444.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., ... & Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15-47.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2.

Golden, C., Freshwater, S. M., & Golden, Z. (1978). Stroop color and word test.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, *15*(8), 534-539.

Jeffreys, H. (1961). *Theory of Probability.* Oxford: UK, Oxford University Press.

Jensen, U. T. (2020). Is self-reported social distancing susceptible to social desirability bias? Using the crosswise model to elicit sensitive behaviors. *Journal of Behavioral Public Administration*, *3*(2).

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory. *Journal of Personality and Social Psychology*.

Kee, J., Knuth, M., Lahey, J. N., & Palma, M. A. (2021). Does eye-tracking have an effect on economic behavior?. *Plos One*, *16*(8), e0254867.

Kelly, C. A., Soler-Hampejsek, E., Mensch, B. S., & Hewett, P. C. (2013). Social desirability bias in sexual behavior reporting: Evidence from an interview mode experiment in rural Malawi. *International Perspectives on Sexual and Reproductive Health*, *39*(1), 14.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the implicit association test: IV. *Implicit Measures of Attitudes*, *59*, 102.

Larsen, M., Nyrup, J., & Petersen, M. B. (2020). Do survey estimates of the public's compliance with COVID-19 regulations suffer from social desirability bias? *Journal of Behavioral Public Administration*, *3*(2).

Lejarraga T, Schulte-Mecklenbeck M, Smedema D. (2017). The pyetribe: Simultaneous eyetracking for economic games. *Behavioral Research Methods*. 2017; 49(5): 1769–1779.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75.

Machado, C. J., & Nelson, E. E. (2011). Eye-tracking with nonhuman primates is now more accessible than ever before. *American Journal of Primatology*, *73*(6), 562-569.

Merrett, F. (2006). Reflections on the Hawthorne effect. *Educational Psychology, 26*(1), 143-146.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P. C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225-1242.

Nyström, M., Andersson, R., Holmqvist, K., & Van De Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, *45*, 272-288.

Patrick, C. J., Fowles, D. C., & Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: Developmental origins of disinhibition, boldness, and meanness. *Development and Psychopathology*, *21*(3), 913-938.

Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. W. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology*, *16*(6), 555.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Reynolds, B., & Schiffbauer, R. (2004). Measuring state changes in human delay discounting: An experiential discounting task. *Behavioural Processes*, *67*(3), 343-356.

Schell, C., Godinho, A., & Cunningham, J. A. (2021). Examining change in self-reported gambling measures over time as related to socially desirable responding bias. *Journal of Gambling Studies*, *37*, 1043-1054.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L. F., & Natalicio, D. S. (1971). The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican Journal of Psychology*, *5*(3 & 4).

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Tom, S. M., Fox, C. R., Trepel, C. & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315, 515–518.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 194.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent?. *Journal of Memory and Language*, *28*(2), 127-154.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297-323.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498-505.

van der Maas, M., Nower, L., Matheson, F. I., Turner, N. E., & Mann, R. E. (2021). Sources of bias in research on gambling among older adults: Considerations for a growing field. *Current Addiction Reports*, *8*, 208-213.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58-76.

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263-290.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291-298.

Wolfe, J. M. (2010). Visual search. *Current biology*, *20*(8), R346-R349.

Wickström, G., & Bendix, T. (2000). The" Hawthorne effect"—what did the original Hawthorne studies actually show?. *Scandinavian Journal of Work, Environment & Health*, 363-367.

Zuckerman, M., Kolin, E. A., Price, L., & Zoob, I. (1964). Development of a sensation-seeking scale. *Journal of Consulting Psychology*, *28*(6), 477.

Supplemental Materials

Equipment

Figure S1 shows a picture of a typical experimental workstation used in the experiments. The Tobii Pro X2-60 eye-tracker is outlined in red.

Figure S1. Photograph a typical computer workstation and eye-tracker used in the experiment.

Analysis of Gender Effects.

We examined main effects of self-reported gender, as well as interactions between gender and condition, to determine whether the eye-tracking manipulation affected males and females differently. Table S1 shows the means for males and females within each condition, for each construct on the scales participants completed. The two rightmost columns show the Bayes Factor for the main effect of gender and the interaction between gender and condition, from a Bayesian ANOVA conducted using JASP software. Specifically, these values are the Bayes Factor for including each effect in the model ($BF_{inclusion}$). Values greater than 3 indicate moderate support for the effect, while values less than 0.33 (1/3) indicate moderate support for the null hypothesis that no effect exists.

Table S1

Statistical Results for Comparisons Between Groups on Dimensions from Questionnaire Scales

Scale/Construct	Male Eye-Tracking	Female Eye-Tracking	Male Control	Female Control	BF Gender	BF Interaction
TriPM						
Boldness	33.22 (7.94)	30.05 (7.01)	32.75 (9.67)	30.73 (6.30)	1.62	0.09^{-}
Meanness	16.09 (8.04)	11.60 (6.70)	16.71 (8.50)	9.11 (6.58)	$>1000^{+}$	0.49
Disinhibition	18.00 (6.87)	17.75 (7.96)	19.25 (8.30)	17.02 (7.33)	0.20^{-}	0.03^{-}
STAI						
State Anxiety	38.70 (9.07)	41.55 (10.66)	37.71 (10.87)	41.19 (10.66)	0.73	0.06^{-}
Trait Anxiety	43.04 (8.35)	45.30 (10.14)	42.50 (9.58)	45.11 (11.64)	0.42	0.04^{-}
SSS						
Disinhibition	3.83 (2.44)	4.61 (2.72)	4.30 (2.38)	3.90 (2.57)	0.13^{-}	0.06^{-}
Boredom	2.85 (1.71)	2.57 (1.66)	3.39 (1.51)	1.94 (1.54)	165.18^{+}	2.46
Thrill Seeking	7.09 (2.31)	6.21 (2.50)	6.86 (2.53)	5.87 (2.50)	3.40^{+}	0.12^{-}
Experience Seeking	4.63 (1.87)	4.80 (1.83)	5.09 (2.09)	4.66 (1.81)	0.12^{-}	0.03^{-}
BFI						
Openness	35.29 (5.45)	33.88 (6.90)	34.41 (6.17)	35.31 (6.33)	0.11^{-}	0.16^{-}
Conscientiousness	30.80 (4.51)	31.70 (5.59)	29.27 (4.79)	31.82 (6.43)	1.10	0.14^{-}
Extroversion	25.82 (7.12)	29.00 (6.39)	24.21 (6.91)	29.24 (6.42)	692.71^{+}	0.20^{-}
Agreeableness	33.80 (5.48)	33.70 (5.21)	33.30 (5.90)	36.26 (5.06)	0.85	0.87
Neuroticism	22.29 (5.49)	26.82 (6.56)	21.84 (5.33)	26.60 (6.56)	$>1000^{+}$	0.12^{-}
DOSPERT						
Ethical	27.78 (3.98)	27.00 (3.31)	28.36 (3.98)	27.27 (3.82)	0.50	0.06^{-}
Financial	30.04 (3.92)	29.63 (3.19)	29.89 (4.28)	29.86 (3.63)	0.17^{-}	0.02^{-}
Health/Safety	26.00 (3.25)	27.71 (3.54)	26.66 (4.05)	27.69 (3.16)	4.19^{+}	0.13^{-}
Recreational	27.00 (3.89)	26.00 (3.79)	27.48 (3.54)	24.90 (3.75)	25.53^{+}	0.39
Social	28.53 (3.43)	30.16 (3.28)	28.41 (3.76)	29.53 (3.29)	4.46^{+}	0.16^{-}
CESD						
Depression	16.39 (8.53)	19.47 (10.12)	16.39 (9.81)	19.87 (11.03)	1.39	0.07^{-}

Note: "BF Main Effect" is the Bayes factor for the t-test with gender as the independent variable. "BF Interaction" is the Bayes Factor for the interaction effect from a Bayesian ANOVA between gender and condition. Bayes Factors greater than 3 indicate at least moderate support for the alternate hypothesis that the effect exists; values less than 0.33 indicate at least moderate support for the null hypothesis that no effect exists. $^{+}$ - indicates at least moderate support for the alternate hypothesis that the effect exists, $^{-}$ indicates at least moderate support for the null hypothesis.

OSPAN. The average proportion of letters correctly recalled are shown in Figure S1. There appears to be no main effect of gender or interactions between gender and condition. A Bayesian mixed effects regression model predicting correct responses from the interaction between condition, gender, and span length indicated that the three-way interaction effect was not different from zero, b = -0.02, SE = 0.09, *95%* HCI = [-0.19, 0.16].

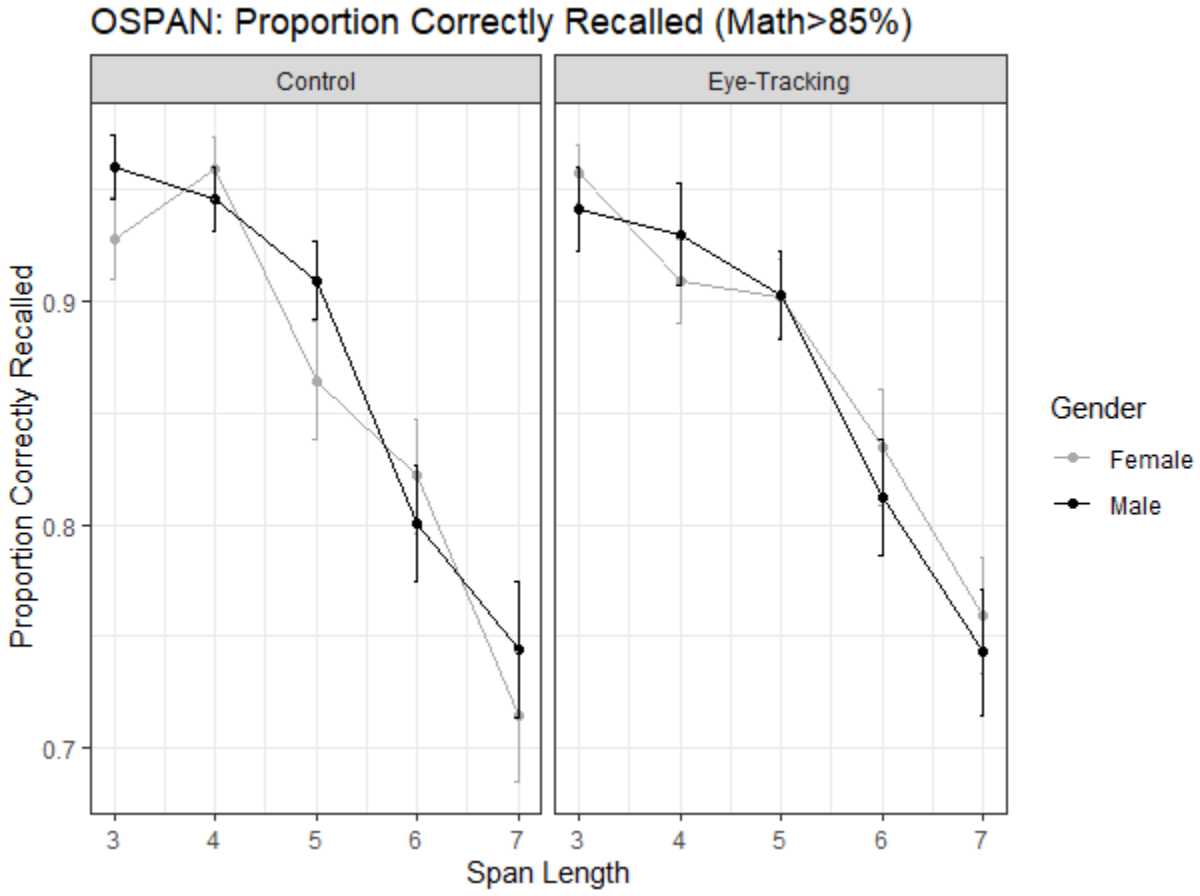


Figure S1. Proportion correctly recalled for each span length in the OSPAN task. Error bars represent standard errors of the mean.

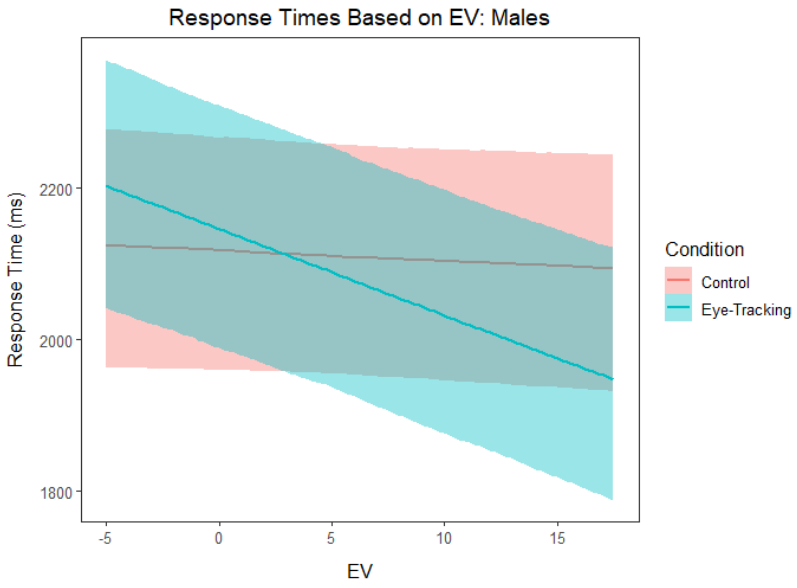
BART. A 2 X 2 Bayesian ANOVA indicated strong support for the null hypothesis that no interaction effect exists, BF_{10} = 0.09. The evidence for the null effect of gender was only anecdotal, BF_{10} = 0.76. Males made slightly more pumps on average than females with the means within each condition as follows: Male Eye-Tracking, M = 14.80, SD = 4.62; Female Eye-Tracking, M = 13.46, SD = 4.24; Male Control, M = 15.12, SD = 3.57; Female Control, M = 14.12, SD = 3.64.

Mixed Gambles Task. A 2 X 2 Bayesian ANOVA with the average proportion of accepted gambles for each participant as the dependent variable, and condition and gender as predictors showed moderate support for the null hypothesis for the interaction effect, BF_{01} = 6.88, and anecdotal support for the null hypothesis for the effect of gender, , BF_{10} = 0.15. The mean proportion of accepted gambles were: Male Eye-Tracking, *M* = 0.62, *SD* = 0.15; Female Eye-Tracking, *M* = 0.56, *SD* = 0.14; Male Control, *M* = 0.62, *SD* = 0.16; Female Control, *M* = 0.60, *SD* = 0.12.

We also ran a Bayesian mixed effects model with 'accept' choices predicted from the interaction between condition, expected value, and gender, with random intercepts for each participant. The three-way interaction effect was not different from zero, *b* = -0.01, *SE* = 0.09, *95%* HCI = [-0.19, 0.16].

Next, we examined whether the interaction between EV and condition in predicting response times differed between males a females. A Bayesian mixed effects regression model with RT regressed on the interaction between EV, condition, and gender, with random intercepts for participants indicated a non-zero interaction effect, *b* = -10.43, *SE* = 3.60, *95%* HCI = [-17.49, -3.15]. Figure S2 shows the simple slopes for each condition, for males (a) and females (b). It appears that there is an interaction between EV and condition only for males. Bayesian regression models with RT regressed on the interaction between EV and condition showed an interaction effect for males, *b* = -10.43, *SE* = 3.60, *95%* HCI = [-17.49, -3.15], but not for females, *b* = 0.46, *SE* = 2.48, *95%* HCI = [-4.44, 5.17]. Males in the eye-tracking condition were slower to make decisions for low EV trials, but much faster on the high EV trials, compared to control participants.

a.



b.

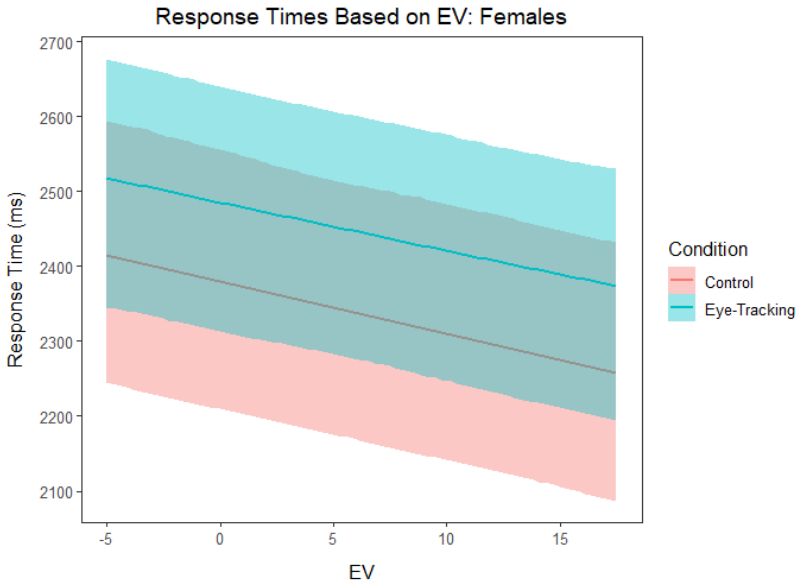


Figure S2. Simple slopes for predicting response times from expected value for a.) males, and b.) females, in each condition.

Stroop Task. Figure S2 shows the average response times for males and females in the Stroop task. A Bayesian mixed effects regession model with RT regressed on the interaction between Trial Type, Condition, and Gender, indicated no three-way interaction effect, b = -13.26, SE = 113.69, *95%* HCI = [-237.66, 215.39].

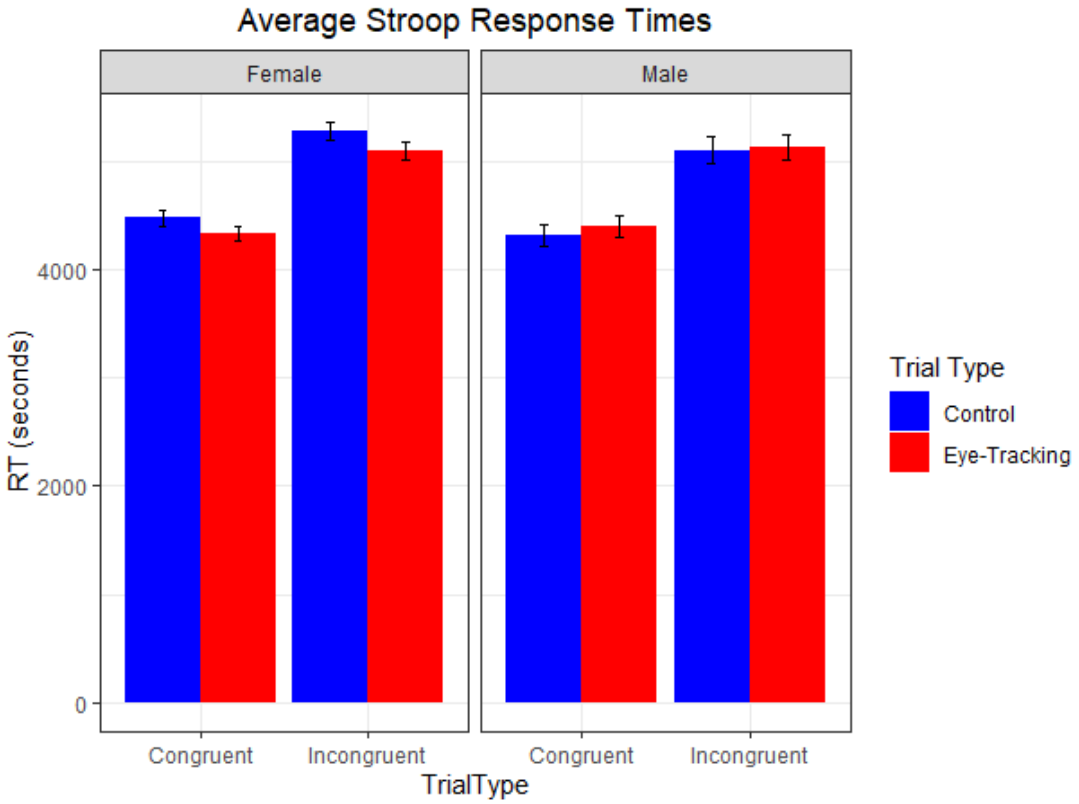


Figure S3. Average response times in the Stroop task. Error bars represent standard errors of the mean.

Figure S4 shows the average proportion of errors for males and females in different conditions and trial types. A Bayesian mixed effects regession model with RT regressed on the interaction between Trial Type, Condition, and Gender, indicated no three-way interaction effect, b = 0-.60, SE = 0.49, *95%* HCI = [-1.57, 0.34].

Figure S4. Average response times in the Stroop task. Error bars represent standard errors of the mean.

Implicit Attitudes Test (IAT). Figure S5 shows the average *D*-scores for males and females in each condition. Visual inspection of the graph suggests a possible interaction where males scores worse in the eye-tracking condition, but females scores better. However, there is substantial variance *D*-scores. A 2 (Condition) X 2 (Gender) Bayesian ANOVA with *D*-Score as the dependent variable indicated little support for the inclusion of the interaction effect, $BF_{inclusion} = 0.18$. A standard, frequentist ANOVA also suggested that the interaction effect was not significant, $F(1,194) = 2.44$, $p = 0.12$, $h^2_p = 0.01$.

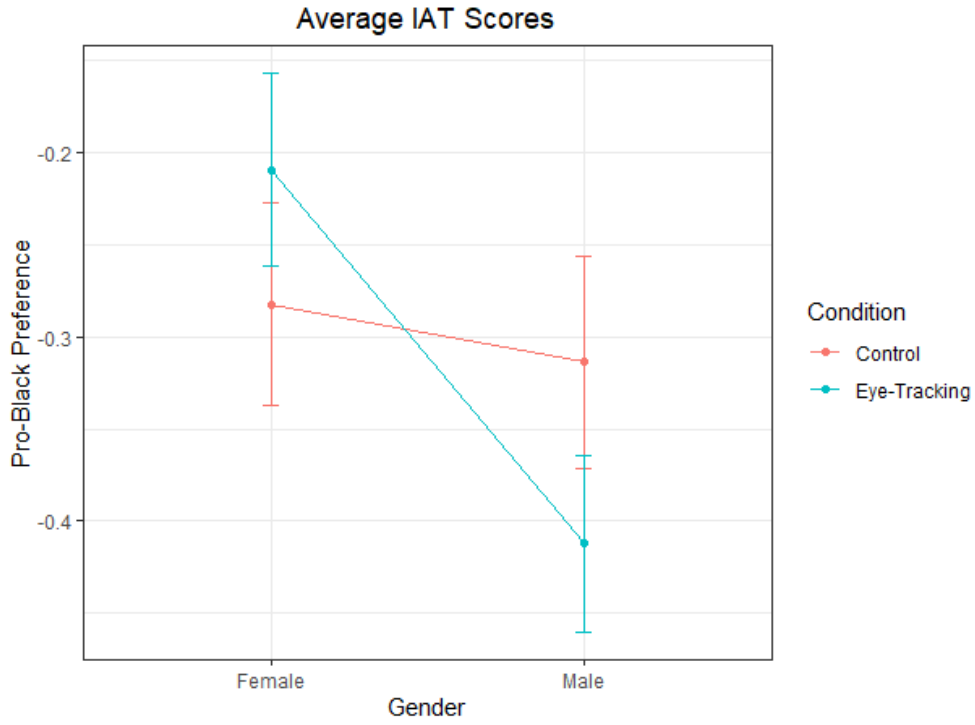


Figure S5. Average IAT scores. Error bars represent standard errors of the mean.