# AN INVESTIGATION OF STEREOTYPE THREAT AS AN INSIGHT INTO THE REPLICATION CRISIS

**SAMANTHA ISABELLA BOOTH**
Bachelor of Science, University of Lethbridge, 2020

A thesis submitted
in partial fulfilment of the requirements for the degree of

**MASTER OF SCIENCE**

in

**PSYCHOLOGY**

Department of Psychology
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

**AN INVESTIGATION OF STEREOTYPE THREAT AS AN INSIGHT INTO THE REPLICATION CRISIS**

**SAMANTHA ISABELLA BOOTH**

Date of defence: August 15th, 2025

| | | |
|---|---|---|
| Dr. L. Barrett<br>Thesis Supervisor | Professor | Ph.D. |
| Dr. S.P. Henzi<br>Thesis Examination Committee Member | Professor | Ph.D. |
| Dr. T. Bonnell<br>Thesis Examination Committee Member<br>University of Calgary | Assistant Professor | Ph.D. |
| Dr. J. Mills<br>Chair, Thesis Examination Committee Member | Professor | Ph.D. |

## DEDICATION

A reader quick, keen, and leery

Did wonder, ponder, and query

When results clean and tight

Fit predications just right

If the data preceded the theory

-- Kerr, N. L., (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196-217.

## ABSTRACT

This thesis explores the replication crisis, using stereotype threat as a demonstrative example. First, I conducted a quality control analysis using papers on stereotype threat's effect on women's math performance, to determine if methods and reporting had improved in the time period between 2001 and 2023. I did not find a significant improvement for the majority of the variables I tested, with the exception of sample size, average group size, and open data practices, which improved over the tested time period. Then, I designed and ran an experiment, testing the effect of implicit and explicit stereotype threat on women's performance on a novel click accuracy task. Using this data, I first conducted a pervasiveness analysis, to determine the prevalence and magnitude of stereotype threat's effect on participants' scores. Within my sample, stereotype threat did not seem to cause a significant reduction in task performance compared to the control condition. Then, using Bayesian modelling, I conducted an exploratory analysis to test several ways that my results could be manipulated. First, I explored the results using four different outcome variables and found that the strength and confidence of my findings varied depending on this choice. Then, I tested the effect using two different pairs of experimenters, to determine if individual differences have an impact on the results. I found that individual experimenter variation can have a significant impact on the strength and direction of results, which may be misattributed to experimenter sex in some studies on stereotype threat. This thesis contributes to a body of work that aims to explore the causes of stereotype threat and suggests several methods for improving the quality of psychological research.

## ETHICS STATEMENT

Work described in Chapter 3 of this thesis received research ethics approval from the University of Alberta Research Ethics Board, Project Name “STEREOTYPE THREAT AND THE REPLICATION CRISIS: COMPARING THE RESULTS OF STUDY MANIPULATIONS”, Pro00139620, October 7th, 2024.

## ACKNOWLEDGEMENTS

First and foremost, I would like to like to thank my supervisor, Louise Barrett, for inspiring my topic, reading every word of this thesis many times over, and providing invaluable and ingenious insight every step of the way. This thesis would have sucked without you.

I would also like to thank the other members of my committee, Peter Henzi and Tyler Bonnell. Peter, you gave me the idea for my canonical model (which turned into my pervasiveness analysis), and you also reminded me about the importance of relaxing and not taking work too seriously. Tyler, thank you for meeting with me so many times to help me improve my models – statistical analysis is hard, but you make the process feel much less daunting.

Thank you to Anne Jones, who was the first person I ever worked with in the lab. You have taught me so much, and I cherish the time we spent together at the playground, cursing the sun, the bugs, and observational data collection. Thank you for letting me stay in your guest room once a week when I had to come to Lethbridge to TA, I will always miss our tradition of hot yoga, A&W, and a sleepover. Finally, thank you for Zoom Pomodoros and three non-negotiables per day, you were instrumental in providing the motivation needed to finish this thesis.

Thank you to Anna Quinn and Madison Clarke, for being the people I could always turn to when I needed to commiserate about the horrors, or if I just needed a nice, long yap session. I am also very grateful that you both defended before me – both of your theses have been open on my computer for the past six months, both as reference and inspiration.

Thank you to everyone else that I've had the pleasure of working with and becoming friends with in the Banzi lab, for inspiring me to be a better scientist, and a more interesting, well-rounded person. I consider myself very lucky to have had a cohort of intelligent, resourceful people to surround myself with. Thank you to Anna Quinn, Madison Clarke, Rachel Warken, Dylan LaValley, Joe Dyck, Tyler Coulouras, and Drayton Pratt for helping me collect data for my experiment, and persevering through the technical issues. Thank you again to Tyler Coulouras for redacting 50 papers for me and for not missing a single year or citation. I couldn't have asked for a better research assistant. Thank you to Holly Kalyn Bogard for providing me with many useful references, for helping me conduct a factor analysis, and for being the only other person in the lab studying the behavior of scientists. To everyone else in the lab that I did not thank by name, just know that I appreciate you, and your support meant the world to me.

Thank you to my husband, Calvin, for supporting me through two years and eight months of this Masters degree. I literally could not have found the strength to do this without you. Thank you for always listening to my ideas and letting me show you all of my graphs, even if they were boring and made no sense to you.

To my Dad and Grandparents, for still thinking, after all this time, that this degree had anything to do with clinical psychology. Thank you for always checking up on my progress and telling me you were proud of me.

Finally, thank you to my Mom, for raising me to believe that I can do anything I set my mind to. I wouldn't be the person I am today if it weren't for your unwavering love and support.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| QRP | Questionable Research Practice |
| NHST | Null-Hypothesis Significance Testing |
| HARK | Hypothesize After Results are Known |
| ST | Stereotype Threat |
| SIAS | Social Identities and Attitudes Scale |
| GI | Gender Identification |
| GSC | Gender Stigma Consciousness |
| DI | Domain Identification |
| DSC | Domain Self-Concept |
| NA | Negative Affect |
| GRE | Graduate Record Examination |
| CI | Credible Interval |
| ESS | Effective Sample Size |

# CHAPTER 1: GENERAL INTRODUCTION

Most scientists would agree that replication, the process of independently verifying the results of a study, is one of the most fundamental pillars upholding the integrity of science. Simons (2014) contends that the only way to have confidence in the reliability of a scientific finding is the independent replication of that finding by a laboratory other than the one that conducted the original research. In this way, we can start to separate out studies containing measurement or statistical errors, or those whose results came about through questionable research practices (QRPs), from studies that offer reliable evidence for how the world works. It is not necessarily true that an effect will be found by all researchers who test it, even if the thing they are testing has real-world significance; random chance, small effect sizes, and unrecognized confounds can all play a role. Generally speaking, however, we should prefer to see a scientific finding confirmed independently before placing too much confidence in it.

Replication has always been a crucial step in conducting trustworthy, reliable science. But when did we begin talking about a crisis of replication in psychology? Researchers have been writing about problems that occur within psychology for many years, and specifically, much work has been done to flag those issues that result in non-replicability of psychological results. At the heart of many of these issues is the misunderstanding and misuse of the *p*-value, a term popularized by Fisher (1925). *P*-values are often used as strict decision criteria for whether a finding is 'significant' or not; *p*-values above 0.05 are considered non-significant, while findings with *p*-values below 0.05 are called significant - this standard is known as null hypothesis significance testing (NHST). A *p*-value of 0.05 means that there is a 5% chance of achieving a result at least

as extreme as the observed findings if the null-hypothesis was true; that is, if your suggested hypothesis was not true, there is a 5% chance that you would achieve the results that you did or results even more extreme. Fisher (1925) emphasized that using a *p*-value of 0.05 in a test of significance was a convenient metric to understand the probability of achieving your results by chance alone but should not be used to make decisions about whether a result should be considered worthy of publication.

Sterling (1959) raised concerns that relying blindly on NHST as a criterion for publication would create a body of work where statistically significant results were over-represented, which is an issue since significance may be achieved by chance alone, and non-significant results would remain unknown to the scientific community, since they are unlikely to be published. That is, under these conditions, the published psychological literature would become biased and untrustworthy since it would reveal only half of the story. Bakan (1966) also discusses issues that arise from NHST; first, not only are journal editors much less likely to accept non-significant results, but researchers also choose not submit such results for publishing in the first place. Instead, they selectively put forth only their significant results, and discard data that does not serve to bring their *p*-values below the 0.05 threshold of significance. This is a problem, Bakan (1966) goes on to say, because it canonizes Type 1 errors, or false positive results, which are notoriously difficult to remove from the literature due to the exceptionally low rate of replication studies. Similarly, Armitage at al. (1969) warned about the dangers of conducting repeated significance tests on data as it continues to be collected, as this practice drastically increases the chances of a Type 1 error; *p*-values can fluctuate drastically,

especially with small sample sizes, so collecting more data could reveal that the effect is not significant after all.

Nearly forty years later, Ioannidis (2005) discussed the possibility that Type 1 errors may make up the bulk of published research. He simulated the likelihood of a study correctly identifying a true positive result based on typical sample sizes and effect sizes, the number of tested relationships in a given field of research, the flexibility of study designs and statistical methods, and sources of bias within a scientific field. Disturbingly, Ioannidis (2005) concluded that the probability of a research finding being true can be astronomically small under certain circumstances. So, even prior to psychology's recent troubles, researchers have been concerned about the way that we conduct research and have predicted many of the issues that contribute to the current crisis of replicability.

Despite such long-standing concerns, there are those who claim that reference to a 'crisis' in psychology is an exaggeration, and psychologists who refer to it as such are victims of the base rate fallacy (Bird, 2021) — the tendency to ignore the underlying base rate of a phenomenon when presented with case-specific information. For example, when presented with statistics about the prevalence of irreproducible studies, people will be more likely to attribute irreproducibility to the replication crisis rather than considering the naturally existing frequency, or the base rate, of false hypotheses in the field. If there is already a high chance that any given hypothesis will be false within psychology, then it is to be expected that there will be a high rate of false positive results, based on chance alone. I have a few issues with this argument, however. First, even if the high rate of false positives in the literature can be attributed to a high base rate of false hypotheses alone, that does not mean that these false positives should go uncorrected. Makel, et al. (2012)

found that only 1.07% of all studies published in 100 high-impact-factor psychology journals since the year 1900 were replication studies, so clearly the issue is more complex than a simple base rate fallacy. Additionally, if psychologists are truly so bad at generating hypotheses that are likely to be true, then it is possible that we are basing our research on weak theories, which can be partially attributed to the fact that psychologists are rewarded for publishing exciting, innovative research rather than building incrementally on an established body of research. This, to me, indicates that the reason for the high base rate of false hypotheses is not simple randomness, but a human-driven problem. Bird (2021) ultimately suggests that there is no replication crisis within psychology because low replication is to be expected based on how we generate our hypotheses. Of course, scientists will frequently get things wrong, as there are many more ways to be wrong than there are to be right, but if we base our new hypotheses on theories that have held up to previous replication attempts, then we may yet create a body of literature that does not have quite so high a percentage of false positive results. Regardless of the source of false positive results in the literature, we should still strive to improve our standards by increasing the rate of replication studies, which we know are uncommon (Makel et al., 2012), and by reducing the rates of QRPs and fraud, which we know happen frequently (John et al., 2012).

Some groups have attempted to measure the consequences of allowing Type 1 errors to accumulate in the published literature. Following the increased attention that irreproducible studies began to receive in 2011, a group of scientists created the Reproducibility Project, a large-scale effort to conduct replications of studies that had been published in three prominent psychology journals (Open Science Collaboration,

2012). Ultimately, 100 replications were conducted by 270 researchers, yet only 36% of replications found significant results, although, as the authors point out, a replication of a result does not necessarily mean the original finding is correct, nor does a failure to replicate mean that the original finding was a false positive (Open Science Collaboration, 2015). That being said, a 36% replication rate certainly does not inspire confidence in our current methods and standards in psychology.

Other fields of science have been experiencing replication crises of their own. For example, a pharmaceutical research team at Bayer Healthcare analyzed data from 67 projects (70% of which were from the field of oncology) and found inconsistencies in nearly two thirds of results (Prinz et al., 2011), and researchers at the biopharmaceutical company Amgen identified and repeated 53 extremely influential preclinical oncology papers but were only able to replicate 11% of them (Begley & Ellis, 2012). Within the field of economics, Camerer et al. (2016) attempted replications of 18 studies published in high-impact economics journals, and despite taking immense care to avoid bias and keep statistical power high, they were only able to replicate 11 of them. Clearly, psychologists are not alone in facing a crisis of confidence in published results, although the field of psychology has been at the forefront of most public discussions of this issue. Positively, this can be partially attributed to the diligence with which psychologists have worked to focus their efforts on systematically investigating and solving the crisis within their field. However, psychology may also be especially susceptible to the causes of non-replicability due to the inherent difficulty of measuring complex human emotions, thoughts, personalities, perceptions, and other phenomena that are not directly observable.

Despite the unique challenges that psychologists face while testing their theories, it may be a simple matter of intuition to determine which theories seem most likely to stand up to replication, and which theories do not. Camerer et al. (2016), in addition to the replications that they carried out on 18 influential economics papers, also measured the beliefs of economists about the likelihood of these findings being confirmed, as another replicability indicator. By participating in a prediction market, economists familiar with common economics methodologies were able to place bets on their perceived likelihood that each of the 18 studies would successfully replicate. They found that prediction market beliefs were positively correlated with actual replication success, demonstrating that if a result seems too good to be true, then it probably is.

One factor that these researchers may be using to determine the likelihood of replication is whether a finding is based on a solid theoretical foundation. Many authors have written about a 'theory crisis' in psychology, the idea that psychologists are not very good at coming up with theories that stand up to rigorous testing, due to their vague and abstract nature (Eronen & Bringmann, 2021). There are several reasons why developing solid psychological theories is so difficult, which Eronen and Bringmann (2021) outline. First, in psychology there are few stable, immutable phenomena that can inform our theories, and so oftentimes the data is insufficient to provide strong proof that a theory is true. Second, psychologists are often more inclined to come up with something new and exciting to study than to undertake the boring task of validating existing constructs. Finally, determining the causality of psychological relationships can be extremely difficult, given the impossibility of manipulating psychological variables - you cannot directly manipulate a person's thoughts, you can only manipulate the information that

they receive and infer that they interpret this information in a certain way. It is also difficult to measure outcomes - you must rely on self-report or behavioral data, as many psychological phenomena, such as thoughts, cannot be directly measured. Taking together this theory crisis in psychology, as well as the idea that researchers can intuitively determine which papers are likely to replicate, it seems that psychologists may need to take a step back and participate in more grounded, albeit more boring, research in order to preserve the integrity of their findings, rather than building upon theories that seem to be crumbling around us.

### 1.1 Standards of Evidence

As I have already noted, psychologists have been writing about these issues for decades, but in 2011 the narrative around irreproducible results shifted, in part due to the publication of a paper that seemed to defy the laws of physics. Bem (2011), using completely standard methods and statistics, claimed to have found evidence that people had the capacity to see into the future. Bem (2011) designed several experiments to test whether people could predict the future, including an experiment where a group of undergraduates predicted whether curtain A or curtain B was covering an image. The position of the hidden image was not randomly generated until after the participant had made their guess, and thus the test was said to measure precognition. In this experiment, participants were apparently able to identify the curtain behind which an erotic image would be found at above chance levels (53.1%). In total, Bem (2011) reports on nine experiments on precognition, eight of which report a statistically significant result. Many criticisms of this work have been published, including Wagenmakers et al. (2011), who point out that Bem (2011) presents exploratory analyses as confirmatory ones, and that

these results disappear when re-analyzed within a Bayesian framework. Criticisms of Bem's methods aside, the publication of these experiments suggested that the standards of experimentation, statistics, and scientific reporting in psychology were lacking. If the editors and reviewers were unable to reject a paper that defies our understanding of physical laws simply because it met all of the accepted standards of psychological research, then clearly those standards should be interrogated.

This is what Simmons et al., (2011) aimed to accomplish. Their paper served to demonstrate that the requirements for reporting psychological research meant that it would be extremely easy to cover up unethical research conduct, while also showing how researcher degrees of freedom, or decisions made by researchers during the many stages of conducting a study, can inadvertently lead to false positive results. Researchers need to make many decisions at every level when designing and conducting a psychological study. Which hypotheses are they testing? What are the predictions? Who will make up their participant pool? How many participants? How will they collect their data? What statistical tests will they conduct? How will they interpret and report their results? And so on. Each time a decision is made, there are many other paths that go unexplored. Ultimately, these decisions can be ambiguous, but researcher decisions are not arbitrary; the choices made are often those that lead towards statistical significance, as Simmons et al., (2011) demonstrated. First, using simulation, Simmons et al., (2011) found that, by combining the effects of four common degrees of freedom, it was possible for a researcher to have over a 60% chance of achieving a false positive result. Then, by taking advantage of these researcher degrees of freedom in a real study, they were able to find evidence for a necessarily false result: they demonstrated statistically that listening to

“When I’m Sixty-Four" by the Beatles resulted in participants becoming a year and a half younger when compared to participants in a control condition. To produce this result, they engaged in several forms of data manipulation, including asking participants a total of twelve questions but only reporting the results of two of these, and analyzing their results for statistical significance every ten participants and stopping data collection when they achieved a *p*-value of less than 0.05. Simmons et al., (2011) was a concrete demonstration of how researchers could easily, and often inadvertently, steer their results towards significance. That is, many of the forms of data manipulation they engaged in, and the manner in which they reported their results, were deemed acceptable by the standards of many journals, and need not reflect any attempt to deceive or falsify results with intent. In other words, the standards of reporting on a study facilitated (and also obscured) the way in which certain details need not be reported, in ways that could change the interpretation of the results completely.

### 1.2 Perverse Incentives

The current crisis of replicability is a nuanced, multifaceted issue consisting of many complicated and interrelated factors, but none more predominantly than publication bias. Publication bias refers to the tendency for statistically significant, novel findings to be published at a disproportionately higher rate than studies that support the null hypothesis, offer negative results, or are perceived as lacking novelty, such as replication studies (Fergusen & Heene, 2012). Fanelli (2010) found that papers published in psychological journals were more likely to report a positive result than papers published in any other discipline of science. Is this because psychologists are better than other scientists at generating and testing correct hypotheses? A more likely explanation, as

noted above, is that positive results are being disproportionately favored for publication over null results. Martin and Clarke (2017) reviewed the publishing policies of 1151 psychological journals and found that only 3% of journals actively encouraged replications, while 63% took a more neutral stance, neither encouraging nor discouraging replication studies. More concerning, they also found that 33% of journals implicitly discouraged replication studies by calling for "original research", and 1% of journals explicitly did not accept replication studies for publication – this is a third of journals whose policies disincline researchers from conducting important replications of previous work. Martin and Clarke's (2017) findings were not constrained by specific branches of psychology, nor were they influenced by the impact factor of the journal, a metric that evaluates the relative importance of a journal within its field by quantifying how often articles within the journal are cited (Sharma at al., 2014).

It is not difficult to see how journal policies of this nature will lead to distortions of the scientific record. If journals do not encourage researchers to submit replication studies for publication, then researchers will be much less likely to conduct them; why spend time and resources conducting a study that has an inherently low likelihood of publication? In addition to their reluctance to publish replication studies, journals are also not likely to publish null findings, so researchers who obtain a null result from a study are incentivized to either file it away with other studies that are conducted but never reported (known as "file-drawering" a study: Rosenthal, (1979)), or they can massage their null or negative finding into a positive one in order to increase chances of publication. Both reactions to publication bias contribute to a body of literature that cannot be falsified; if negative results are always tucked away or manipulated into significance, then Type 1

errors in the published literature do not face opposition and are thus canonized and unfalsifiable (Nissen et al., 2016).

The missing piece that keeps this feedback loop going is known as the "publish-or-perish" culture, where researchers must choose between publishing prolifically, or "perishing" into obscurity by missing out on the promotions and grants afforded by achieving a high number of publications (Remus, 1978). Lilienfeld (2017), refers to this dynamic as "the grant culture", where researchers are rewarded with grant funding proportional to the number of publications they author, which in turn leads to more opportunities for promotions, salary increases, laboratory space, graduate student positions, and other invaluable resources. Many scientists respond to this publish-or-perish dynamic by working toward increasing publications via ethical routes, such as allocating more time and resources towards projects that are more promising, delegating resources and responsibilities appropriately throughout their laboratories, and collaborating with other researchers in their field. However, as Lilienfeld (2017) points out, incentives that reward publications also reward relying on questionable research practices (QRP) to achieve those publications more reliably and with less effort, for both researchers who are inclined toward unethical practices, as well as those researchers who simply do not know any better.

### 1.3 Questionable Research Practices

The term questionable research practices (QRPs) can be used to describe actions taken by researchers that, while not considered outright fraud, take advantage of ethical grey areas regarding acceptable research methods. One study conducted by John et al. (2012) suggests that not only are QRPs extremely common, but they may actually be the

scientific norm. After surveying over 2,000 psychologists, John et al., (2012) concludes that some of the most common QRPs included failing to report all of a study's dependent measures, reporting an unexpected finding as an expected one (also known as HARKing), deciding whether or not to continue data collection depending on whether current results were significant, and failing to report all of a study's conditions. The survey also revealed that even a practice as indefensible as falsifying data was self-reported, albeit rarely. Given the incentives that researchers face to publish their results, it comes as no surprise that these QRPs are used as a means for researchers to avoid null results in their studies.

### 1.3.1 HARKing

The scientific method dictates that when designing a study, all hypotheses and predictions must be finalized before any data are collected to test them. Hypothesizing After Results are Known, also known as HARKing, is the practice of generating a hypothesis that fits your results, and then presenting said hypothesis as if it had been generated prior to data collection (Kerr, 1998). The danger of HARKing, according to Kerr (1998), is that it can solidify Type 1 errors in the published literature. If the results of a study have been obtained through statistical errors, random chance, or even through unethical actions, then presenting them as if they were predicted by previous research lends credibility to the false finding. Further, if researchers conduct their studies and do not find support for their original hypothesis, this is still informative; it tells the reader that their specific methodology did not work to find evidence for their specific predictions. When future researchers design their own study, they will know to explore different methodological paths if they want to investigate the question further. Unfortunately, in our publish-or-perish culture, HARKing can be rewarded, and some

may view it as an easy way to save their study from the file drawer if the results did not support their original hypothesis.

### 1.3.2 *p*-hacking

Another QRP with strong incentives is *p*-hacking, which is the practice of collecting, selecting, or analyzing data in such a way that non-significant results become significant (Head at al., 2015). Researchers may analyze data during data collection and decide to stop collecting data as soon as significance is achieved, they may collect responses to many variables and only report those that have a significant impact on their results, or they could exclude or include outliers or subgroups based on whether they bring the results below $p = 0.05$ (Head et al., 2015). There are many ways for researchers to *p*-hack their results, but the outcome is the same; a non-significant result has been artificially transformed into a significant one.

### 1.3.3 Fraud

The QRPs discussed so far certainly constitute unethical behavior, but often the individuals engaging in them are not fully cognizant of the consequences that HARKing or *p*-hacking have on the integrity of the scientific record. However, the same cannot be said for those individuals who knowingly fabricate data or commit fraud. One metric that helps us estimate the prevalence of fraudulent data is the Retraction Watch Database (2018), which lists over 50,000 retracted papers, as of August 17th, 2024. It is important to note that retraction does not necessarily mean that fraud has taken place, and in fact, retraction is an important part of the system of checks and balances that encourages science to be self-correcting. In fact, Lu et al. (2013) found that 21.9% of retractions from

the Web of Science Database were self-reported by the authors, suggesting that there are some honest scientists who use retraction to correct their own mistakes. This honesty seems to pay off; analysis of citation losses on prior work after a retraction takes place revealed that authors who self-retract do not suffer the same reduction in citations that authors who experience non-self-reported retractions face (Lu et al., 2013). It is vital that we have a system in place for removing papers from the scientific literature, to address honest errors that make their way into the published literature. However, fraud still makes up a sizable portion of retractions. One review of retractions in biomedical and life science journals revealed that 43.4% of retractions listed the reason as fraud (Fang et al., 2012). In the same year, Grieneisen and Zhang (2012) investigated the reasons behind the full or partial retraction of 3,631 articles from 42 online databases and found that, of those papers that listed a reason for retraction, 20% contained fraudulent or fabricated data, or other research misconduct. Other reasons for retraction included several forms of publishing misconduct, which made up a further 46% of retractions, and questionable data or interpretations which represented another 25% of retractions (Grieneisen & Zhang, 2012). Bik et al. (2016) estimated the prevalence of a different kind of fraud: inappropriate image duplication in biomedical research, which points to the authors' attempts to somehow skew interpretations of their findings, much like duplicating data from a participant whose results were favorable to the hypothesis. They found that, of 20,621 papers from 40 journals dating from 1995 to 2014, 3.8% contained duplicated images, and more than half of these bore the marks of deliberate manipulation.

Based on investigations by concerned scientists, it seems that fraud is alarmingly common within the published literature. But what rates of fraud are scientists self-

reporting? Fanelli (2009) attempted to quantify an answer to this question by combining the results of 18 surveys, all of which asked researchers about research misconduct, both perpetrated and observed. 1.06% of researchers admitted to fabricating and/or falsifying data themselves, and 12.6% admitted to observing the fabrication and/or falsification of data by other researchers. Of course, self-report data are likely to underestimate the true prevalence of this problem, so we can assume the real numbers are likely higher. John et al. (2012) also endeavored to determine rates of data falsification among scientists. They asked participants whether they themselves had falsified data, and what their estimates were for the general prevalence of this behavior, believing the true prevalence to be somewhere in the middle. Based on their findings, they estimated that 9% of researchers may be guilty of falsifying data – and according to the authors, this could be a conservative estimate, due to the negative stigma surrounding data falsification. Clearly fraud and other unethical research practices have tarnished the published literature, but what can be done to rectify the situation in which we find ourselves?

## 1.4 Solutions

### 1.4.1 Unbiased Publication

One seemingly simple solution that would address the problem at the root is for journal editors to disincentivize the causes of non-replicability by committing to certain standards for the quality of articles that they publish. Nosek et al. (2015) provides a set of guidelines for journals that they believe can help promote transparency, openness, and reproducibility in the published literature going forward. These guidelines include standards for the open availability of a paper's data, code, and other materials, requiring the preregistration of study design and analysis plans, as well as using Registered Reports

for replication studies, such that publication can be guaranteed for authors, regardless of their findings. With these standards in place, researchers would have a more difficult time manipulating aspects of their study to encourage publication and would be further disincentivized from this behavior by having guarantees that they would be able to publish their work regardless of the finding, as long as their methods were appropriate.

### 1.4.2 Open Data

The "Mertonian norms" are a set of guidelines meant to encourage scientific integrity and development, and include the norm of communality, which states that science would not be possible without collaboration between scientists, and thus the findings and products of science should be communal; individual scientists should not claim ownership over their results or processes (Merton, 1925). The concept of open science seems intuitive. Why should scientists have to reinvent the wheel every time they wish to learn something new about the world? Doesn't it make more sense to iterate on the findings of other scientists? And what better way to do that than to make freely available all of the materials that were used in your scientific research? However common sense this idea may seem to us today, the concept of widely and openly disseminating scientific knowledge was not popularized until the late sixteenth century, prior to which pervaded a culture of secrecy and individualism (David, 2008). The birth of the scientific journal in 1665 marked the beginning of our formal system of scientific knowledge distribution, providing new incentives for early scientific researchers to communicate their findings to the public, such as the avoidance of wasted effort duplicating existing research, and maximizing society's growth of reliable information (David, 2008).

The practice of openly sharing your data and code for all to download and peruse also serves as an excellent deterrent for those scientists who may be inclined to engage in unethical research practices and also helps us detect those scientists who are bold enough to do it anyway. This is just what Simonsohn (2013) did; noticing inconsistencies in the summary statistics of two papers, and after simulations proved that the means and standard deviations were necessarily incongruous, he requested the raw data from the authors of both papers. Both authors complied, and Simonsohn was able to analyze the data and confirm his suspicions. Simonsohn (2013) took several precautions to ensure that these inconsistent findings were irrefutably the result of data manipulation rather than errors on the part of the researcher, including replicating his findings across several studies, and contacting authors privately to give them an opportunity to address his concerns. After these precautions were taken and Simonsohn's suspicions were not dissuaded, he discreetly conveyed his suspicions to the appropriate investigating authorities, and the researchers behind these fraudulent papers were forced to resign. A third researcher was also solicited by Simonsohn to provide the raw data used in their research, but the author claimed to have lost them, thus escaping the same retribution that the other authors faced. It seems that if all of the data was required to be publicly posted in the first place, then Simonsohn wouldn't have had to go through as much trouble to acquire the files from the authors, and the third author wouldn't have been able to hide behind the excuse of negligence to avoid punishment. If journals required authors to submit their data in order to publish their work, it is true that potential fraudsters might simply become better at falsifying data to avoid detection. But for those who are not committing deliberate falsification, open data practices may deter sloppy science and some of the less blatant forms of QRPs. To be sure, it is not realistic to expect that every

dataset be re-analyzed before publication to ensure there are no errors or inconsistencies, but it would be much easier to investigate these potential issues if the data were already publicly available by default.

### 1.4.3 Simmons et al. Suggestions

In addition to demonstrating the dangers of undisclosed flexibility during the design, analysis, and reporting of a study, Simmons et al., (2011) also provided a set of guidelines for authors and reviewers that would help solve the problem with minimal effort exerted from either side. For authors, Simmons et al., (2011) suggested making the decision about when to stop data collection prior to collecting data; collecting at least 20 observations per study condition; listing all variables and experimental conditions included in a study; reporting what the results would be including and excluding any observations they ultimately eliminated; and, if the study included a covariate, also reporting the results without it. For reviewers, they suggested enforcing all of the aforementioned author requirements; being more tolerant of results that do not show significant effects; requiring authors to demonstrate how their results would change if they analyzed their data differently; and in extreme cases, reviewers should require authors to complete an exact replication of their study if their justifications for decisions are not convincing. Using these requirements, Simmons et al., (2011) demonstrated just how difficult it becomes to get away with many forms of scientific misconduct, such as *p*-hacking, although dedicated fraudsters would no doubt be able to bypass many of these guidelines with a carefully crafted fake dataset.

### 1.4.4 Registered Reports

Nosek and Lakens (2014) edited a special issue of the journal *Social Psychology* which follows a typical formula for registered reports: researchers submit proposals to the journal and, if accepted, the proposals and all study materials are registered online prior to data collection. Researchers then have the assurance that if they proceed with data collection as outlined in their proposal, without unjustified deviations, their results will be published in the special issue. Registered reports serve multiple purposes: first, they encourage researchers to pursue replication studies, as they would have the promise of publication before they go through the trouble of collecting and analyzing data, not to mention writing it up into a publishable paper. The second purpose served by registered reports is to discourage bad actors from participating in the age-old tradition of HARKing, since there would be publicly accessible information detailing the hypotheses that the researcher intended to test. In fact, registered reports prevent most of the ill effects of undisclosed flexibility that Simmons et al., (2011) discuss. By determining beforehand their expected sample size, researchers would have to justify deviations from this number, which would make it more difficult to analyze their results during data collection and simply decide to stop once they happen to reach statistical significance. Researchers would also be unable to collect data on many variables and only report the effects of the few that achieved statistical significance. A registered report that also includes a plan for data analysis can prevent the authors from presenting an exploratory analysis, where the researcher freely explores the data, as a confirmatory one – a more rigorous process whereby the researcher begins with an a priori hypothesis and follows a strict protocol to adequately test that hypothesis (Kimmelman et al., 2014). Clearly there

are many benefits to registered reports, both for the researcher and for the integrity of published research.

### 1.4.5 Abandoning Statistical Significance and The Value of Bayesian Analysis

Many of the solutions previously discussed have been designed to solve issues of HARKing. However, one contributor to the replication crisis that many of these solutions overlook is *p*-hacking. The issue of *p*-hacking - finagling your results to fall below $p = 0.05$ - serves to highlight the arbitrary nature of *p*-values, and many scientists are in favor of abolishing *p*-values altogether. One group of scientists argue that an over-reliance on *p*-values and statistical significance can often cause us to erroneously dismiss many findings as showing no effect, pointing out that a statistically non-significant result does not "prove" the truth of the null hypothesis (Amrhein et al., 2019). Amrhein et al, (2019), along with more than 800 other scientists argue that there are certainly some use cases for *p*-values, but in general, scientists should cease using them as a strict yes or no answer to the question of whether a result supports a hypothesis. Rather than blindly deferring to the *p*-value to determine the worth of their findings, Amrhein et al. (2019) suggests that researchers should evaluate other aspects of their work, such as prior evidence, study design, and confidence in the quality of their data to decide what their findings mean.

The other main alternative to the *p*-value that many scientists advocate for is the adoption of Bayesian statistical analysis, as opposed to frequentist statistics that rely on *p*-values. Bayesian statistics can offer several advantages: within a Bayesian framework, scientists can use prior evidence for a phenomenon to update their statistical models, interpretations of Bayesian results more closely reflect most people's intuitive understanding of statistics, and Bayesian models are much less susceptible to small

sample sizes (McElreath, 2020). Despite having many virtues, the Bayesian approach certainly is not a panacea to all problems of reproducibility. Simmons et al., (2011) point out that adopting a Bayesian statistical approach can actually increase researcher degrees of freedom, as it introduces even more judgements that fall on the researcher, such as choosing a prior distribution to use in your model—although inappropriate use of priors in Bayesian analysis would be easily detected. Clearly, there are no one size fits all solutions to our current crisis of confidence and solving it will take a concerted effort on all of our parts, requiring a massive shift in the scientific norms to which we are all beholden.

### 1.5 Stereotype Threat

Like many popular psychological phenomena, stereotype threat has been scrutinized for irreproducible findings. Stereotype threat is the phenomenon whereby the fear of conforming to negative stereotypes about one's group may hinder task performance, first described in 1995 by Steele and Aronson. Stereotype threat was first investigated within the context of race; Steele and Aronson (1995) asked whether Black undergraduate students underperformed on a difficult test compared to their white student counterparts if the test was described as being diagnostic of intellectual ability, a metric for which negative stereotypes exist about Black individuals. They found that, after adjusting for SAT scores, Black participants performed worse than white participants in the diagnostic condition, but not in the control condition. Steele and Aronson (1995) found no effect when testing whether gender played a role in inhibiting task performance under threat. However, Spencer et al. (1999) later tested the effect of describing a mathematics test as producing gender differences on the performance of equally qualified

men and women and found women's performance to be greatly reduced in the stereotype threat condition. Since then, stereotype threat effects have been tested on many other stereotyped groups within society, such as the elderly (Lamont, et al., 2015), low socioeconomic status individuals (Harrison et al., 2006), athletes (Dee, 2014), and people facing weight stigma (Guardabassi & Tomasetto, 2018), to list just a few examples.

### 1.5.1 Causes

Despite being studied extensively over the last 30 years, the underlying mechanisms of stereotype threat are still poorly understood. Two competing theories have dominated the literature. Schmader and Johns (2003) posit that stereotype threat reduces task performance by placing a burden on cognitive resources, thereby reducing the working memory available to succeed at a task. Alternatively, Harkins (2006), explains that the possibility of evaluation on a task causes those being assessed to more frequently select a prepotent response, which can inhibit performance on complex tasks, but may actually facilitate performance on simpler ones. This theory is often called the "mere effort" account, or the "motivational account" of stereotype threat. Many studies have since attempted to parse out which of these mechanisms is responsible for performance decrements under stereotype threat, and compelling evidence has been published to support both the working memory theory (Schmader & Johns, 2003; Johns et al., 2008; Rydell et al., 2009) and the mere effort account of stereotype threat (Harkins, 2006; Jamieson & Harkins, 2007; Seitchik & Harkins, 2015). The mere effort account certainly seems to be more popular in the literature, but it seems as though the most likely explanation for stereotype threat is a combination of these two theories.

### 1.5.2 Problems

Stereotype threat is one of the most widely researched topics within social science and has been cited as a significant contributing factor in the underperformance of stereotyped groups in domains such as mathematics examinations (Spencer et al., 1999), mental rotation tasks (Moe & Pazzaglia, 2006), and motor skill learning (Mousavi et al., 2021). Some researchers, however, have called for a critical re-examination of the stereotype threat literature. Stereotype threat has recently come under scrutiny for its inconsistent results, showing a seemingly strong effect in certain studies (Spencer at al., 1999), but being notably absent from the results of others (Pennington et al., 2019; Flore et al., 2018).

Sackett et al. (2004) caution against taking stereotype threat as the *only* explanation for performance differences between groups. They noted that the vast majority of academic and non-academic references to the original stereotype threat research conducted by Steele and Aronson (1995) have misinterpreted the results of the study. These publications incorrectly assert that the performance differences between Black and white individuals completely disappear in the non-stereotype threat condition. In fact, subgroup differences still exist in the absence of stereotype threat, but these differences are exacerbated in the presence stereotype threat condition, which is why Steele and Aronson (1995) adjusted the results by participant's SAT scores. At a baseline, there were differences between Black and white participants in terms of math performance, so the suggestion that addressing stereotype threat will solve all performance differences for marginalized groups is a misinterpretation of the source material and also overstates the real-world effects of stereotype threat.

Existing research on stereotype threat has also been criticized for containing fundamental methodological issues, which contribute to the exaggeration of the effects of stereotype threat on minority groups' performance. Stoet and Geary (2012) critically examined the effects of stereotype threat on women's mathematics performance, by first discussing the original three experiments published on the topic by Spencer et al., (1999), followed by a meta-analysis of all replication attempts. First, in the original paper that tested the effect of stereotype threat on women in mathematics, Stoet and Geary (2012) identified several methodological shortcomings. First, the paper failed to successfully discriminate between the stereotype threat explanation and other explanations for performance differences. Second, in one of the experiments, more than half the data were discarded. Finally, they excluded some participants for not making a reasonable effort on the test based on the time taken to complete it but retained other participants who scored a zero or lower, which could also indicate a lack of effort. Next, Stoet and Geary (2012) combed through the literature, and after excluding studies that fundamentally deviated from Spencer at al., (1999), they were left with 10 true replication attempts, of which only three successfully replicated the original results. This seems to indicate that stereotype threat could potentially play a role in suppressing women's mathematics performance under some circumstances, but it certainly doesn't inspire the confidence to say that stereotype threat is the main contributor to the performance gap between men and women, or for any measurable performance deficits for other marginalized groups.

One final problem that affects the stereotype threat literature is publication bias. Flore and Wicherts (2015) conducted a meta-analysis to estimate the effects of stereotype threat on young girls, in an effort to determine whether the detrimental effects of

stereotype threat only emerge in adulthood. In addition to this, they conducted tests to determine whether their subset of studies was suffering from publication bias, which had been suggested by previous stereotype threat publications (e.g., Ganley et al., 2013; Stoet & Geary, 2012). Using a common method to assess publication bias, they produced a funnel plot of effect sizes, which demonstrated asymmetry. This suggests that papers with effect sizes outside of certain thresholds had either been file-drawered or somehow manipulated into significance. They also determined that the published literature contains more significant effects than one would predict based on the cumulative power of the included samples. Overall, most of the publication bias tests that Flore and Wicherts (2015) conducted on their subsample of the literature indicated that the effect size of stereotype threat is likely to be exaggerated.

### 1.6 Thesis Outline

My thesis aims to address some of the causes of non-replicability within psychology. For my first data chapter, I wanted to explore whether study methodology and reporting has improved since 2011, when the replication crisis became a mainstream topic of conversation. Many potential solutions to the replication crisis have been suggested, but only a few have been successfully implemented. I was interested in comparing the methods and results of papers published within the last few decades, to determine if scientists have become better at reporting their methodologies in the interest of transparency and replicability, and to find out if there has been an improvement in actual research practices. With this in mind, I collected papers published between 1999 and 2024 on the effects of stereotype threat on the mathematics performance of women and girls. Looking only at the methods and results sections of the papers, I rated the

transparency of reporting exhibited by each paper, as well as the research methodologies to determine whether there have been any improvements over time. Then, by visualizing my data, I looked specifically at two time periods, namely pre-and-post 2011, to get a sense of methodological standards before and after the replication crisis entered mainstream discussion.

For my second data chapter, I conducted an experiment that overtly investigated the effects of implicit and explicit stereotype threat on undergraduate's performance on a simple click-accuracy task. However, inspired by Simmons et al., (2011), I conducted an exploratory analysis on the data to examine the outcomes of multiple branching paths from decisions made throughout the process of the study. We know from Simmons et al., (2011) that scientists conducting a study tend to make decisions that lead them towards statistical significance, so the goal of my analysis was to demonstrate how the results may be interpreted differently if alternate choices had theoretically been made throughout the process.

# CHAPTER 2: QUALITY CONTROL ANALYSIS OF THE STEREOTYPE THREAT LITERATURE

## 2.1 Introduction

In the years since the replication crisis became a topic of concern, efforts have been made to improve standards of scientific research and reporting, as well as encourage greater academic integrity among scientists. Nosek and Bar-Anan (2012), for example, put forward the idea of open, continuous online peer review, which would serve to bolster current review processes, rather than replace them. With this system in place, articles would still make their way through a formal editor-based review process, but there would also be an informal forum linked to each article where readers could post comments, questions, and critiques, so as to encourage discussion, illuminate errors, or provide directions for future research.

Building on this concept, Chambers (2013) introduced the idea of the registered report, where researchers submit an outline of their proposed research, including their hypotheses, proposed data collection plan, and their plan for data analysis, and the journal accepts and reviews the studies that they find the most promising. Once the study is complete, the authors submit their finalized manuscript as well as their raw data, and if these are appropriately aligned with their initial report, the study is published regardless of the results, which helps prevent publication bias. It is important to make a distinction between preregistration and the registered report. If preregistration is the general process of detailing research plans prior to conducting said research, then the registered report takes it a step further, making preregistration a required step in the process of publication. Nosek et al. (2018) discussed the state of preregistration, pointing out that although there

had been a slow and steady increase in preregistration education, incentives, and services, widespread adoption had not yet been achieved. Preregistration constitutes just one of the ways that psychologists have attempted to rectify the issue of non-replicability. Other ways that the quality of published literature could be improved on include increasing sample sizes and effect sizes to reduce uncertainty in results, encouraging more detailed methodological reporting to make replications easier to conduct, and eliminating known confounds within the control group.

In order to determine if research practices have improved over time, with special interest taken in the years since 2011, I conducted a quality control analysis on a subset of the stereotype threat literature. Specifically, I collected papers reporting on studies that tested the effects of stereotype threat on women's mathematics performance. I chose several metrics by which to measure improvements to methodological reporting, as well as improvements to research methods, which I will outline here.

### 2.1.1 Methodological Reporting Improvements

One of the questions I asked of my set of papers was if methodological reporting was improving over time. Thoroughly and exhaustively reporting all methodologies used throughout the course of the study is imperative for replication. The methods section acts as a set of instructions for the next author(s) to follow in order to achieve experimental replication. If these instructions are unclear or incomplete, then replication becomes difficult if not impossible. Emailing the original authors of the paper for clarification offers its own barriers, as the authors often do not respond, or, if the original paper was published some time ago, they may not remember the exact methodology that they

followed. With this in mind, I chose several metrics to estimate if standards of reporting had improved over time within my sample.

#### 2.1.1.1 Reporting Stereotype Threat Activation Prompt

In order to properly replicate a study on stereotype threat, one would certainly need to know how the original authors worded their stereotype threat activation prompt to reduce any confounds that may arise from the nuances of word choice and threat framing. Failure to disclose the exact statement shown to participants during a study leaves a lot of room for uncertainty when other researchers attempt to design their own studies on the effects of stereotype threat on women's mathematics performance. This raises the key question of whether more recent papers were more likely to disclose their exact stereotype threat activation statement, in the interest of greater transparency.

#### 2.1.1.2 Reporting and Controlling for Experimenter Gender

This quality control analysis includes only studies where stereotype threat is activated explicitly, by the participant hearing or reading a statement that states outright that some groups perform more poorly on the given task. However, some studies have found that stereotype threat can be activated implicitly, whereby the participant's gender is made salient and threatened with a more subtle cue. For example, Inzlicht and Ben-Zeev (2000) investigated whether the gender composition of a group could act as a stereotype threat activation cue. They tested this by recruiting women to complete a mathematics test either in the presence of two other female participants, or in the presence of two male confederates. Inzlicht and Ben-Zeev (2000) found that the women who completed the mathematics test in the gender minority condition performed significantly

worse than those in the same-sex group composition. If the mere presence of males can induce a stereotype threat activation effect, then it seems as though studies on this topic might attempt to control for this confound. This could be done by using multiple male and female experimenters and testing whether experimenter gender has an effect on the results, or possibly by using only female experimenters so as not to unintentionally activate stereotype threat within the control condition. At the very least, authors should report the gender of their experimenters in the interest of transparency, so that readers can make an informed comparison between the results of stereotype threat studies.

#### 2.1.1.3 Reporting Mathematics Test

When designing a study to test the effects of stereotype threat on women's mathematics performance, how does the researcher decide which mathematical questions to use? Many researchers in my sample used questions from, or questions similar to, the Graduate Record Examination (GRE) (e.g. Brodish & Devine, 2009); a test often used during the admission process for graduate school in the United States and Canada (Educational Testing Service, 2025). Some studies in my sample used modular arithmetic questions (e.g. Beilock et al., 2007), and many other studies did not report the source of their mathematics questions at all (e.g. Cadinu et al., 2003). But does it matter what type of mathematics questions are used in a study on stereotype threat and mathematics? O'Brien and Crandall (2003) tested whether the difficulty of the mathematics test moderated the effect of stereotype threat on women's mathematics performance, and found that women under stereotype threat, when compared to women in the control condition, performed worse on a difficult mathematics test; but surprisingly, they performed *better* on an easy mathematics test. This suggests that if the mathematics exam

is not sufficiently difficult, it may not be able to elucidate an effect of stereotype threat, and in fact, if the test is too easy, women under threat may actually experience a boost to performance. If the difficulty of the mathematics questions used in stereotype threat research does indeed make a difference to the results, it seems as though authors should include the full details of the mathematics exam that participants were given. In this way, researchers hoping to replicate the results of another paper can be sure that differences in their results are not due to an unforeseen difference in the difficulty of the mathematics exam administered to participants.

#### 2.1.1.4 Effect Size Reporting

Within a frequentist framework, *p*-values are commonly used to report whether the difference between groups is statistically significant by calculating the probability of attaining the observed results if there was no true difference between study groups. An effect can be statistically significant, but have minimal real-world consequences, as in, we may find a reliable, repeatable effect for something, but if the effect size is very small, then this effect may not actually be noticeable in the real world. It is very important to report the effect size of your findings so that readers may judge the real-world implications of a result. I wanted to investigate my sample to determine how often authors leave this crucial statistic out, and if effect size reporting has any relationship with year of publication.

#### 2.1.1.5 Open Data

The practice of researchers openly disseminating their data is one that can have many advantages in terms of transparency, replication, and collaboration. The open access

movement, which began with the promotion of widespread access to scientific publications, has evolved, and today's proponents of the movement argue for the sharing of research data wherever possible (Mauthner & Parry, 2013). As a follow-up to an earlier survey on the topic, Tenopir et al. (2015) investigated common perceptions of the open data movement, and found that in general, researchers were becoming more open to the idea of sharing their data, and adoption of the practice was on the rise. The ability to access data collected by other scientists encourages progress within science; it allows new investigators to explore old data with new questions, or new methods, compare or combine data from different studies on the same topic, and scrutinize data for evidence of errors or QRPs (Mauthner & Parry, 2013). It may also make the process of replicating a study simpler; the ability to peruse the original study's data may illuminate how the original researchers collected, coded, and defined their data variables. I wanted to investigate my sample to determine if I could find the same trend that Tenopir et al. (2015) found with regard to an increase in open data sharing.

#### 2.1.1.6 Preregistration

As previously discussed, preregistration is an effective way to promote trustworthy science, but adoption of this practice has not been particularly widespread (Nosek et al., 2018). I wanted to know if preregistration was common within my sample, and if more recently conducted research was more likely to be preregistered.

### 2.1.2 Methodological Improvements

I have investigated whether authors are becoming more transparent in their methodological reporting, which is an important step in solving the problem of

irreproducibility in psychology. Now I will explore whether papers in my sample have improved their actual methodology over time. Often, due to researcher degrees of freedom and the seemingly infinite number of choices that can be made while designing a study, it is difficult to know which forking path will make for a more rigorous methodology. However, there are some measures that clearly improve the reliability of a study, such as collecting data on a larger sample size, or blinding the researcher to the participant's condition. Improving the rigor with which we conduct studies allows us to have more confidence that our scientific findings are trustworthy, which not only places us on sturdier ground for comparison between studies but also improves public trust in science. With this in mind, I chose several metrics by which the rigor of study design could be measured and tested.

#### 2.1.2.1 Sample Size

One of the simplest ways to measure the quality of a study is to look at the sample size. In general, the larger the sample size, the more likely the study is to detect an effect if it is there, and also to correctly identify if the effect is not present; thus, a sufficiently large sample size protects against both false positives and false negatives. If a researcher wanted to improve their methodology in the easiest possible way, I predict that they would simply collect data from a greater number of participants. I investigated my sample of papers to determine if researchers publishing papers more recently used a larger subject pool compared to earlier papers on the same topic.

#### 2.1.2.2 Study Group Size

Related to sample size, I wanted to determine if the average number of participants per study group has also increased. If more recent studies have increased their average sample size, but also increased the number of different conditions tested, then this would not necessarily constitute an improvement on methodological rigor. For example, a study with 100 participants, and 2 study groups would have a total of 50 participants per study group, which would likely be sufficient to detect an effect. However, a study with 100 participants and 8 study groups would only have an average of 12.5 participants per group, and would therefore be much more susceptible to the pitfalls of random chance.

#### 2.1.2.3 Researcher Blindness

Within clinical research, one of the most crucial aspects of study design is that the researchers be blind to the condition of the participant they are testing, as this can mitigate a large source of bias within a study (Monaghan et al., 2021). By blinding the researcher to a participant's group assignment, it prevents the researcher from inadvertently treating that participant differently and also reduces the possibility of a researcher interpreting ambiguous data in a biased manner. Since this principle is applied almost universally to clinical research, it is difficult to argue against its merits within psychological research, wherever possible. I therefore investigated my sample of papers to determine if blinding the researcher to the participants' condition was more common in more recent research.

#### 2.1.2.4 Effect Size

I have argued that reporting the effect size of a study is imperative to contextualize the results in terms of real-world consequence. Next, I want to know if the magnitude of effect sizes are changing in more recent years. I would not expect effect sizes to be increasing, as this would suggest that stereotype threat was harming women's math performance more profoundly in more recent years, or that new research was less rigorous and therefore producing increasingly inflated effect sizes. If, however, effect sizes are decreasing, that would suggest to me that early stereotype threat research was more likely to find large effect sizes due to their small sample sizes and less rigorous study designs; in other words, these large effect sizes are a reflection of poor methodology, not true stereotype threat.

### 2.1.3 Does the Paper Find an Effect of Stereotype Threat?

When initially choosing a topic that might serve as a tool to investigate the replication crisis, I was interested in stereotype threat due to the contentious nature of the subject, with some systematic reviews concluding that stereotype threat has strong real-world implications (e.g Pennington et al., 2016) and other studies discovering that the effects of stereotype threat might not be as strong as earlier papers tended to indicate (e.g. Flore et al., 2018). I hypothesized that, with the adoption of better research practices, perhaps more recently published research would find an effect of stereotype less often, suggesting that stereotype threat was not the robust, real-world phenomena that earlier research had made it out to be.

### 2.1.4 Other Problems in the Literature

#### 2.1.4.1 Missing Participants

While compiling my sample of stereotype threat papers, I noticed some other trends that may help explain some of the discrepancies in the strength and direction of stereotype threat findings within the literature. First, I noticed that several papers seemed to report some of their results with a different sample size than they started with, even after accounting for any participants the researchers explicitly excluded. Thus, I wanted to investigate if the phenomenon of missing participants was common in my sample, and if older research was more likely to suffer from this issue, whether the participants were deliberately excluded from analyses to skew the results towards significance, or if they were carelessly dropped due to researcher negligence.

#### 2.1.4.2 "Gender Differences" vs. "Males Perform Better"

Another issue I encountered in my sample is the different ways that researchers worded their stereotype threat activation statements. Researchers tend to assume that, in the participant's mind, "gender differences" equals "men perform better", so it is quite common for researchers to simply inform participants that the task they were about to complete had demonstrated some nebulous gender difference, rather than explicitly state which gender is supposed to have performed worse. Clearly stating which gender exhibits poorer task performance removes ambiguity and creates a more reliable comparison between findings, since you can be certain that all participants received the same information unbiased by their interpretation of the stereotype threat statement. This discrepancy in methodology may partially explain why some papers seem to find such a

strong effect of stereotype threat, and others find no measureable difference between conditions.

#### 2.1.4.3 Inappropriate Control Group

A final issue that I noted within my sample of stereotype threat papers was a tendancy toward inappropriate or confounded control groups. There were several factors that I deemed inappropriate. First, as we have already established, the mere presence of males can implicitly activate stereotype threat within women, so any control group in which males were present, either in the form of the researcher, other participants, or confederates, was coded as inappropriate. Next, any study with a control statement that mentioned gender was also deemed inappropriate. Danaher and Crandall (2008), reanalyzing data from a paper by Stricker and Ward (2004), concluded that presenting demographic questions prior to a mathematics test reduced women's performance on the task. That is, by reminding the participant of their gender, it seems that this may render the control statement ineffective. It is important to note that some studies, such as the original Stricker and Ward (2004) paper, and Inglis and O'Hagan (2022) concluded that there was no effect of presenting demographics questions before the test versus after, but the possibility that this may factor in to priming stereotype threat seems like a good reason to keep the control statement neutral.

## 2.2 Data and Methods

### 2.2.1 Paper Selection

In order for a paper to initially be considered for inclusion, it needed to report on a study in which a group of female individuals completed a mathematics test, with at least one of the study groups being exposed to stereotype threat before testing. Studies could initially include female individuals of any age, ranging from elementary school to adulthood. The studies could also include a male sample, as long as it also included a female sample.

I used two databases listed on the University of Lethbridge Library's Psychology Resources page: MEDLINE [via OVID] and Web of Science Core Collection. These two sources were chosen due to their ease of use, and for the differing subjects covered by each database in the hopes of capturing a wider range of articles. My last download from MEDLINE [via OVID] was May 22, 2024, and my last download from Web of Science Core Collection was August 12, 2024. For both database searches, I used the search term "stereotype threat" and selected all English language journal articles. No other restrictions or filters were used.

In order to select articles to include in my review, I imported the articles from both databases into Rayyan (Ouzzani et al., 2016), a free website that allows for easy screening of articles for inclusion in systematic reviews. I then went through each article featuring the words "math", "mathematics", or "mathematical" and tagged every article that met my initial inclusion criteria. This gave me a list of 194 articles, which I imported into Zotero (Corporation for Digital Scholarship, 2023), a reference manager software. Using Zotero's function to find available PDFs, I was able to automatically add the PDF

to over half of the citations. The rest I added manually, using both the University of Lethbridge Library Summons function, and Google Scholar. However, I was unable to access the PDF of every paper, and after deleting the citations to which I did not have full access, I had a list of 145 articles.

Since my aim was to determine if study design and methodology has improved alongside the heightened awareness of issues of non-replicability, I attempted to mitigate any bias that may have arisen from knowledge of when a paper was published. To that end, I recruited an undergraduate research assistant to reduce each paper to only the methods and results sections, as well as to redact all citations or other information that may have indicated when the paper was published. This was accomplished using Adobe Acrobat (Adobe Inc., 2024). Each paper was assigned a number in sequence, so that the redacted papers could be linked back to the original citation. One file was corrupted and was unable to be redacted with the Adobe Acrobat software, so it was excluded, leaving me with a total of 144 papers.

### 2.2.2 Data Collection

For the purposes of comparison, I ultimately decided to include only studies that were conducted in a laboratory setting, with undergraduate participants, where explicit stereotype threat was a testing condition. This meant excluding any studies that took place online or in classrooms, with a subject pool recruited from elementary, middle, and high schools, as well as studies that used only subtle or implicit threat cues. After these exclusions, a total of 50 papers remained. See Appendix A.1 for a full bibliography of included papers.

After these exclusions were complete, I read the methods and results sections of the remaining 50 anonymized papers, and gathered information on a total of 38 variables, including sample size, effect size, and whether or not the paper found an effect of stereotype threat. For a full list of variables and their definitions, see Appendix A.2. Once these data were collected from all 50 papers, identifying information was added back in, such as the article title, author, year, and journal. Then, with the articles' complete information, I was able to collect data on a further four variables: was the paper a replication study, was the study preregistered, did the study disclose the full mathematics test that participants completed, and did the paper post the raw data online?

### 2.2.3 Data Analysis

I constructed 15 models in a Bayesian framework, using the 'brms' package (Bürkner, 2017), in R (v4.3.2; R Core Team, 2023). Each was a simple logistic regression. Model 1 through Model 13 use year as the predictor variable, with the following response variables: (1) whether the study reported its stereotype threat activation statement (yes/no), (2) whether the study reported its experimenter gender(s) (yes/no), (3) whether the study accounted for experimenter gender(s) (yes/no), (4) whether the study reported its mathematics test in full (yes/no), (5) whether the study reported its effect size (yes/no), (6) whether the study had openly available data (yes/no), (7) whether the study was preregistered (yes/no), (8) sample size, (9) average group size, (10) whether the experimenter was blind to participants' condition (yes/no), (11) effect size, (12) whether the study found a simple effect of stereotype threat (yes/no), and whether the study found any effect of stereotype threat (yes/no). Model 14 and Model 15 both use whether the study found a simple effect of stereotype threat (yes/no) as the response variable, but used

the following predictor variables: (14) whether the stereotype threat activation statement explicitly stated that men perform better (yes/no), and (15) whether the control group was appropriate (yes/no). A Bernoulli distribution was specified for most models, with the exception of Model 8, which used a negative binomial distribution as sample size is an integer variable, Model 9, which used a gamma distribution with a log link as group size represents an average value, and Model 11, which used a cumulative link distribution with a logit link since effect size is an ordered categorical variable. All models were run with 4 chains, and 2000 iterations were set for most models, with the exception of Model 4 and Model 7, which benefitted from 4000 iterations in order to reduce divergent transitions. The R code for this analysis is available here: https://github.com/samantha-lloyd/Analysis/blob/main/Quality-Control-Analysis.

To prepare my data for analysis, I scaled and mean-centered my year variable. All models were run with weakly informative priors (normal (0, 1)). Model convergence was confirmed using $\hat{R}$s ($\hat{R}$s=1.00), and model performance was assessed using the 'pp_check' function in 'bayesplot' (Gabry & Mahr, 2017). Credible intervals were set to 95% for ease of interpretation. Figures were generated using the 'ggplot2' (Wickham, 2009) and 'ggridges' (Wilke, 2018) packages. Conditional and marginal $R^2$ values were calculated using the 'bayes_R2' function (Bürkner, 2017).

### 2.2.4 Descriptive Statistics

#### 2.2.3.1 When Were the Papers Published?

The oldest paper in my sample was published in 2001, and the most recent was published in 2023 (Figure 2.1). Both the median and the mean year for my sample was 2010.

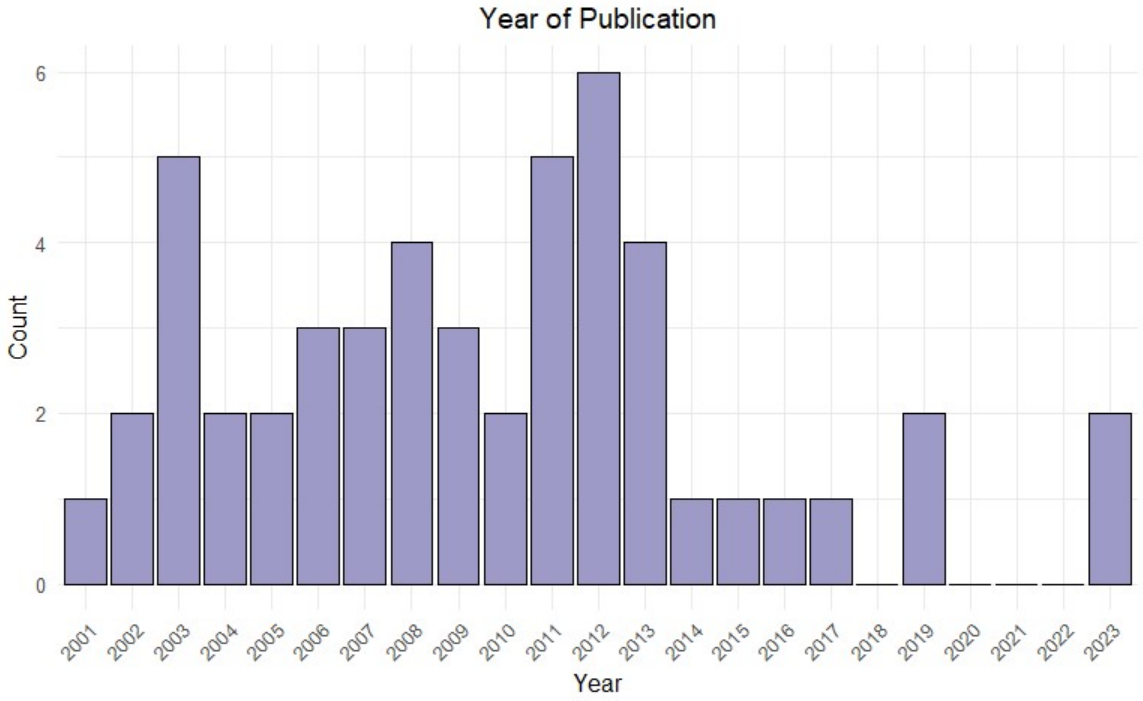



**Figure 2.1** Year of publication for papers used in my quality control analysis.

### 2.2.3.2 Where Were the Data Collected?

Data were primarily collected in the United States, but several other countries are represented, such as Italy and the United Kingdom (Figure 2.2). In total, these papers represent data from 10 countries.

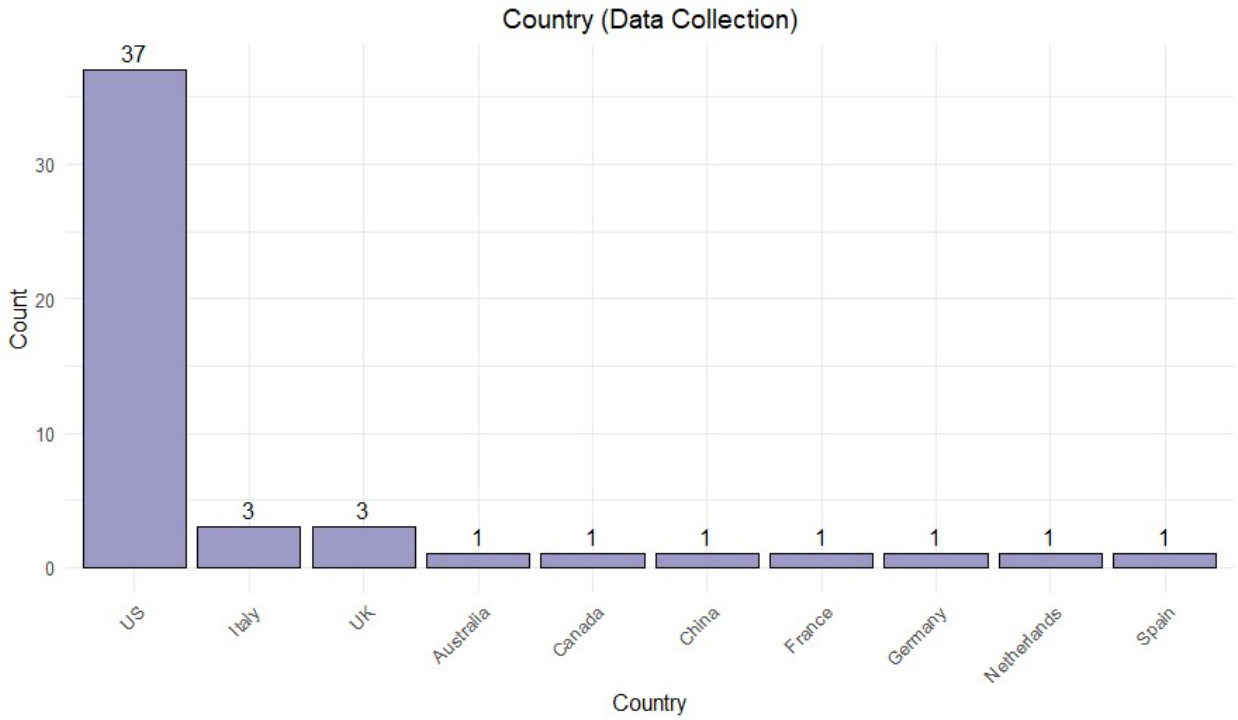



**Figure 2.2** Country in which data collection took place for the papers in my quality control analysis.

### 2.2.3.3 In Which Journals Were These Papers Published?

In total, my 50 papers represent publications from 28 journals, many of which are in the field of social psychology. See Appendix A.3 for the full list of journals.

## 2.3 Results

### 2.3.1 Is Methodological Reporting Improving?

#### 2.3.1.1 Does the Study Explicitly Report a Stereotype Threat Activation Prompt?

Only 14 (28%) papers reported their exact stereotype threat statement wording, with the remaining papers only offering a general statement regarding the threat that participants were exposed to during the study. Figure 2.3 shows whether these 14 papers were published in the period between 2001-2011, or between 2012-2023. I found a very weak effect of year on the likelihood that an author will report their stereotype threat activation statement explicitly, in that more recent papers appear slightly more likely to explicitly report the statement they used, but this effect is quite uncertain ($\beta = 0.25$, 95% CI [-0.33, 0.84], Figure 2.4). It therefore seems possible that papers in my sample are trending towards greater methodological transparency, but not in a meaningful way.

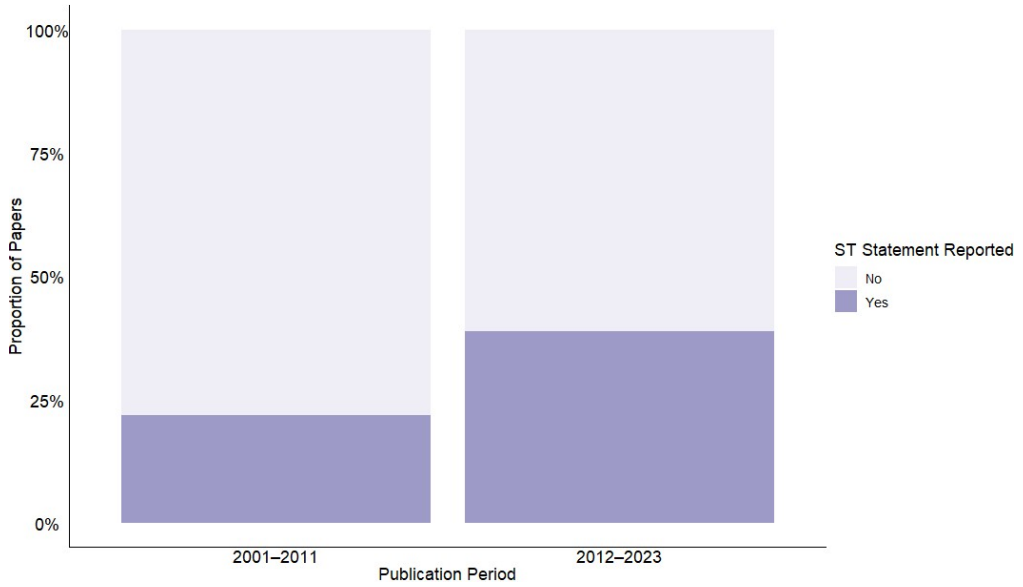



**Figure 2.3** Proportion of papers in my sample explicitly reporting stereotype activation statement, categorized into pre-2011 and post-2011 time periods.

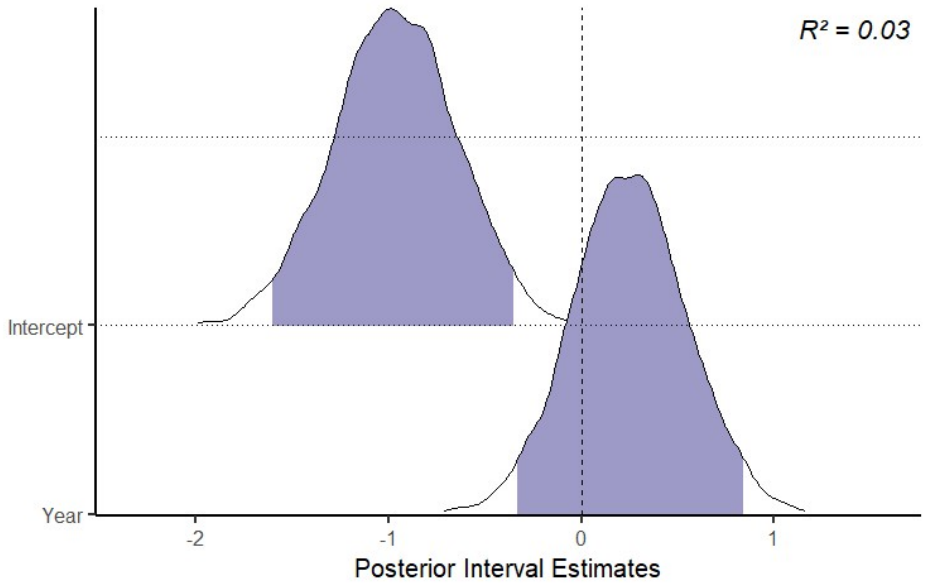



**Figure 2.4** Model 1. Posterior density estimates for the effect of year on the likelihood of a paper reporting their stereotype threat activation statement explicitly. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.1 for model code, summary table, pp_check, and conditional effects.

#### 2.3.1.2 Does the Study Report the Gender of the Experimenter?

23 papers (46%) reported the gender of their experimenters, and a mere three papers (6%) attempted to control for experimenter gender in their results. Figure 2.5 and Figure 2.7 respectively show whether these papers were published in the period between 2001-2011, or between 2012-2023. Again, we can consider if more recent papers are more likely to report the gender of the researcher or take it into consideration when analyzing their data. Here, there was a very weak negative effect of year: more recently published papers were *less* likely to report the gender of their researcher(s) ($\beta$ = -0.36, 95% CI [-0.97, 0.22], Figure 2.6). Once again, however, this effect is not particularly meaningful, as the 95% credible interval (CIs) crosses zero.

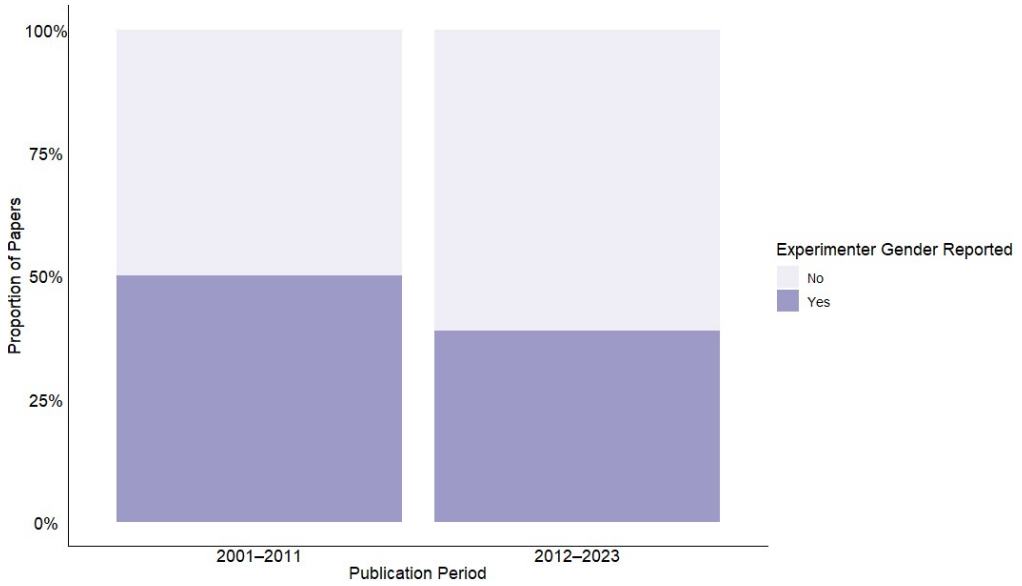



**Figure 2.5** Proportion of papers in my sample reporting experimenter gender, categorized into pre-2011 and post-2011 time periods.

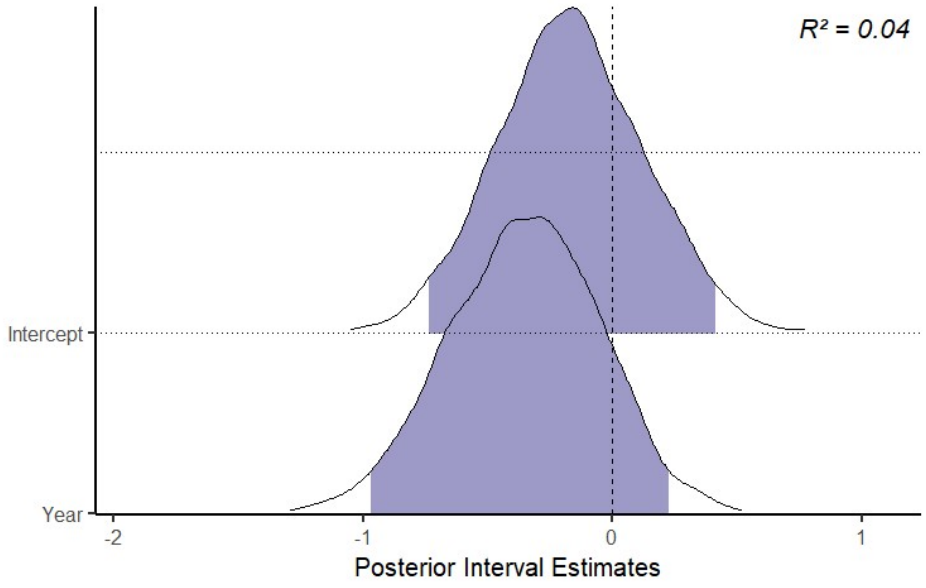



**Figure 2.6** Model 2. Posterior density estimates for the effect of year on the likelihood of a paper reporting researcher gender. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.2 for model code, summary table, pp_check, and conditional effects.

As to whether more recently published papers are more likely to attempt to control for the effect of researcher gender, I found a very weak positive effect for year: more

recent papers were more likely to attempt to control for gender in their analysis, but again, this effect is uncertain (β = 0.40, 95% CI [-0.56, 1.31], Figure 2.8).

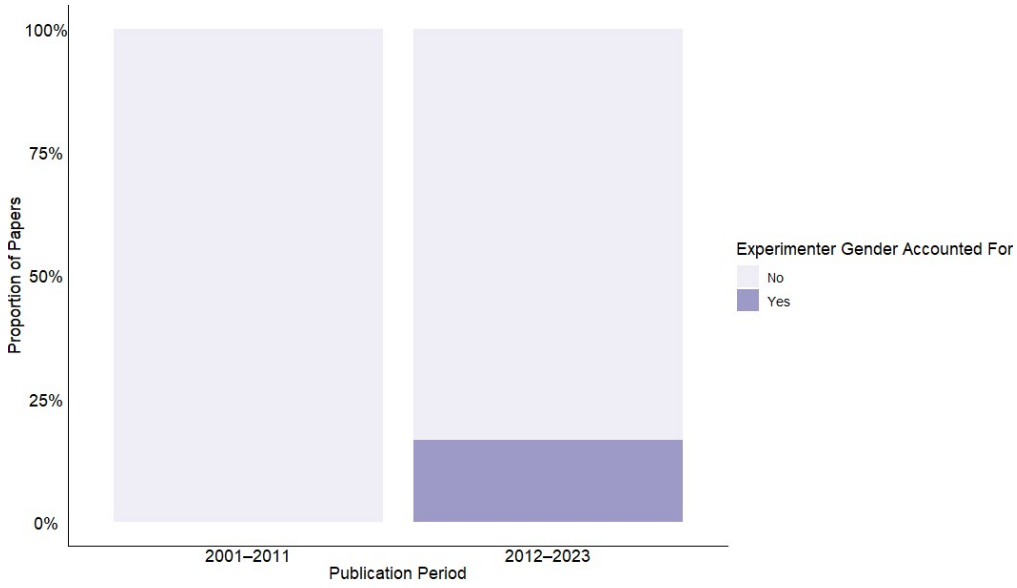



**Figure 2.7** Proportion of papers in my sample accounting for experimenter gender, categorized into pre-2011 and post-2011 time periods.

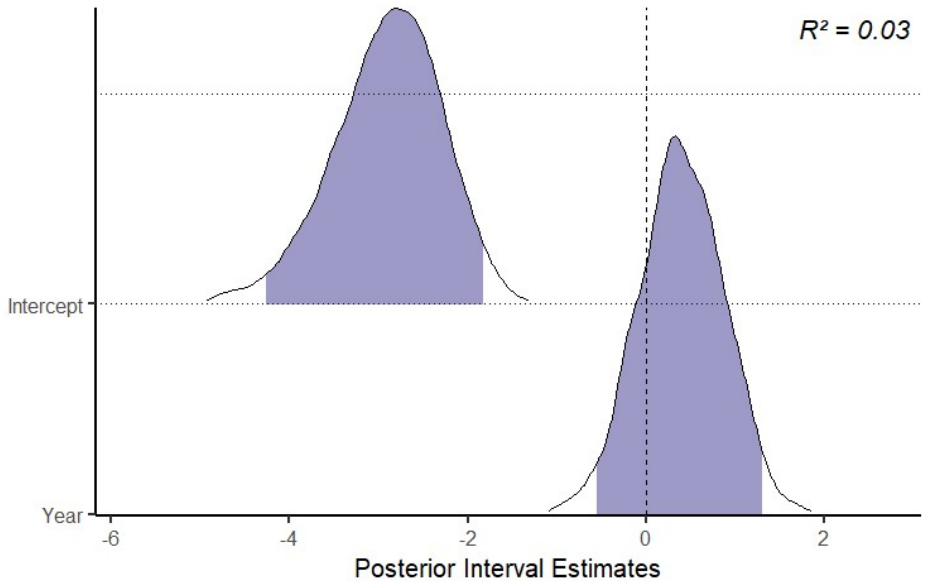



**Figure 2.8** Model 3. Posterior density estimates for the effect of year on the likelihood of a paper controlling for researcher gender. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.3 for model code, summary table, pp_check, and conditional effects.

#### 2.3.1.3 Does the Study Report the Details of the Mathematics Test?

Only 1 paper in my sample (2%) openly shared the exact mathematics test that their participants completed. Most other papers either made no mention of, or gave only a single example of, their mathematics questions. Figure 2.9 shows whether this paper was published in the period between 2001-2011, or between 2012-2023. Examining the effect of year on the probability of reporting the mathematics question showed a seemingly positive effect, in that more recent papers were more likely to report the details of the test used ($\beta = 1.05$, 95% CI [-0.18, 2.36], Figure 2.10). This is consistent with my sample, as the only paper within my sample that included its mathematics exam was published recently, in 2023. However, given that only a single datapoint exists to inform this result, this finding is not robust.

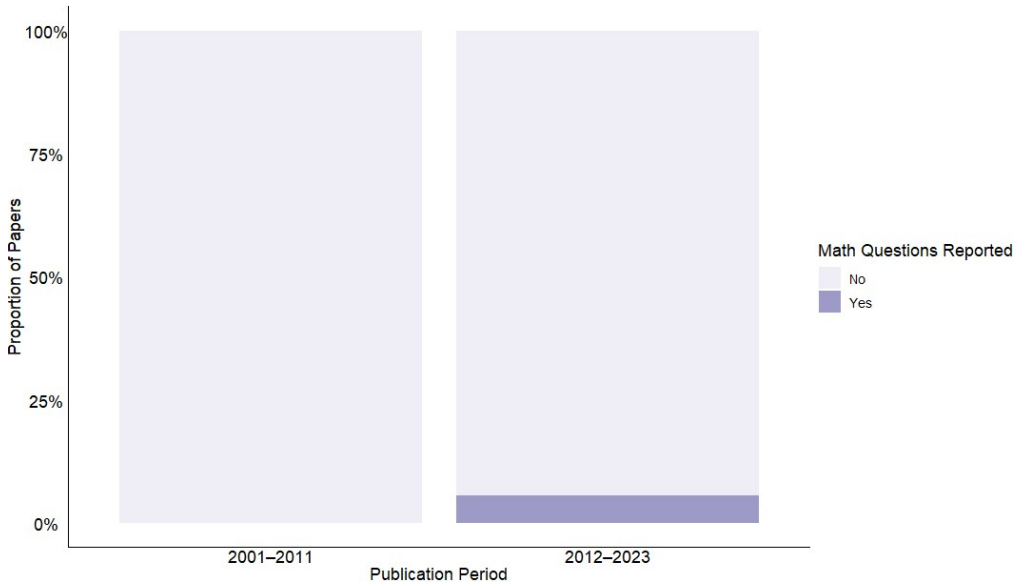



**Figure 2.9** Proportion of papers in my sample reporting the full mathematics exam used, categorized into pre-2011 and post-2011 time periods.

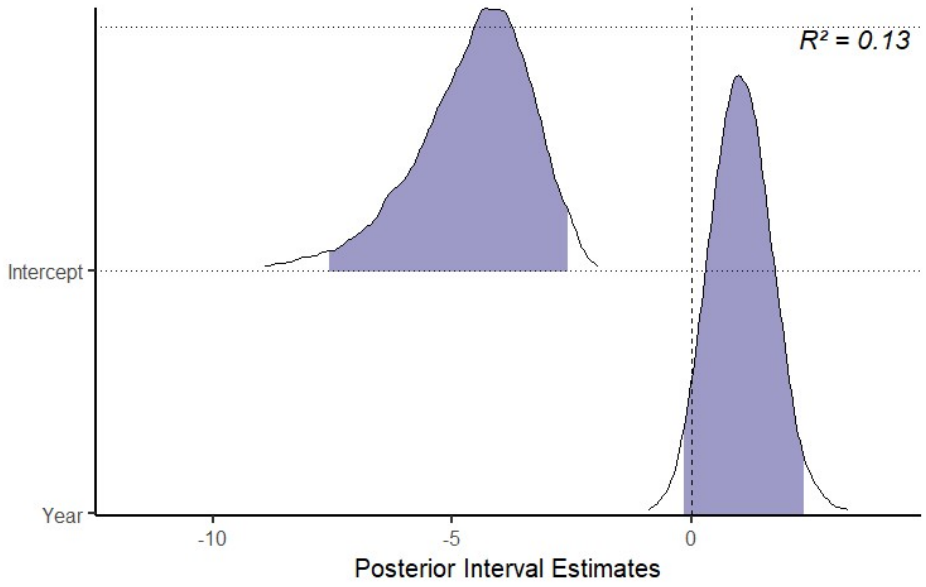



**Figure 2.10** Model 4. Posterior density estimates for the effect of year on the likelihood of a paper reporting its mathematics test. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.4 for model code, summary table, pp_check, and conditional effects.

#### 2.3.1.4 Does the Study Report Effect Size?

36 papers in my sample reported their effect size by some metric (72%), with Cohen's *d* being the most common (13 papers, or 26%), followed by partial eta squared ($\eta_p^2$) (10 papers, or 20%). Figure 2.11 shows whether these 36 papers were published in the period between 2001-2011, or between 2012-2023. More recent papers are slightly more likely to report the effect size of their stereotype threat finding, but once again, the 95% CIs cross over zero so this effect is uncertain (β = 0.30, 95% CI [-0.30, 0.92], Figure 2.12).

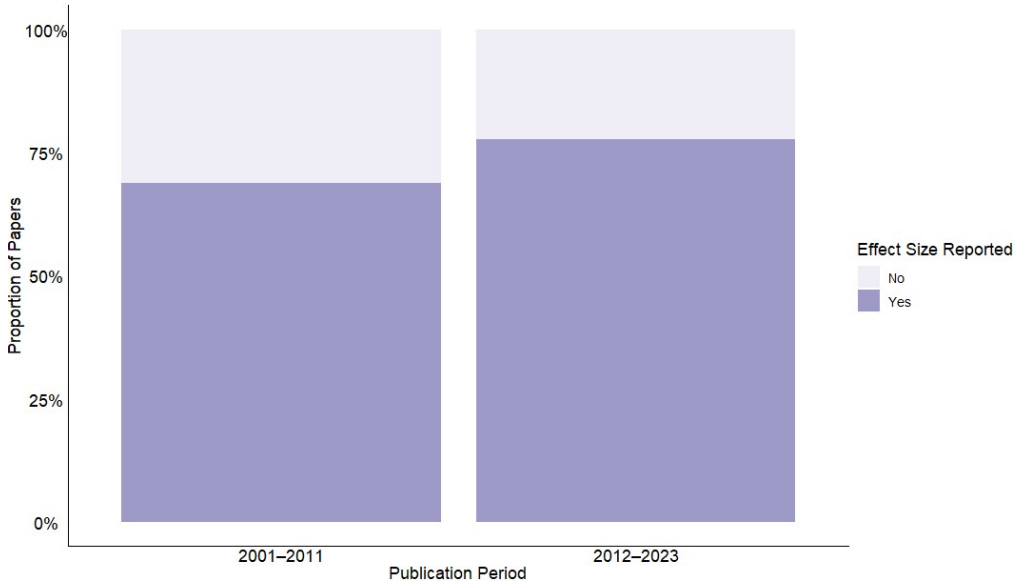



**Figure 2.11** Proportion of papers in my sample reporting effect size, categorized into pre-2011 and post-2011 time periods.

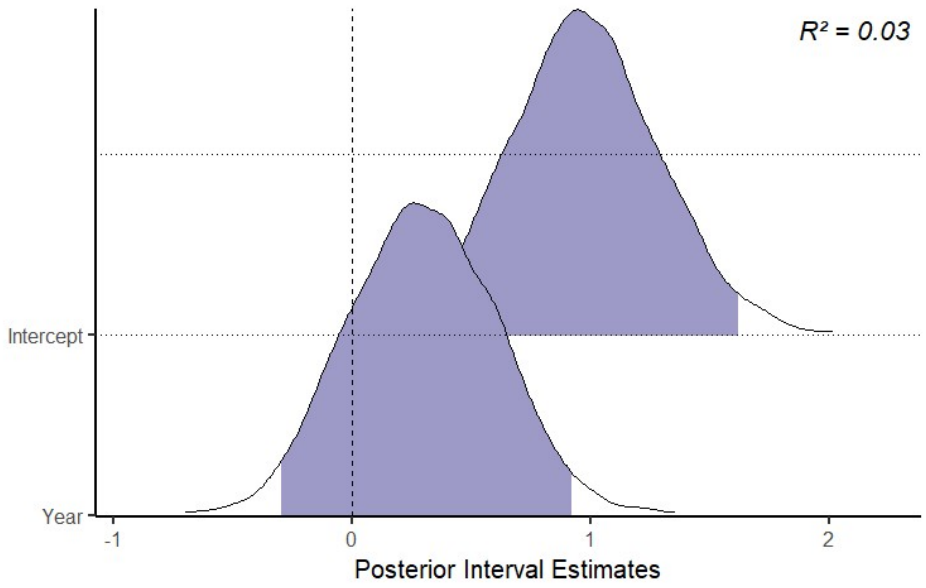



**Figure 2.12** Model 5. Posterior density estimates for the effect of year on the likelihood of a paper reporting the effect size of its stereotype threat finding. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.5 for model code, summary table, pp_check, and conditional effects.

#### 2.3.1.5 Do Studies Make Their Data Openly Available?

Within my sample, only two papers (4%) uploaded their data files to an online repository, and two additional papers (4%) indicated that data were available upon request. As emailing authors to request data has its own problems (for example, the authors could simply not respond, or the data may eventually be lost, corrupted, or deleted), I ran a model coding as "1" only those papers whose data were currently available online, and coding all other papers as "0". Figure 2.13 shows whether these two papers were published in the period between 2001-2011, or between 2012-2023. My results indicate that there is a positive effect of year on the likelihood that a paper will have openly available data; that is, more recently published papers are more likely to share their data online, likely attributable to more widespread use of the internet. ($\beta$ = 1.29, 95% CI [0.15, 2.47], Figure 2.14). The posterior density estimate does not cross 0, so we can be fairly confident in this result.

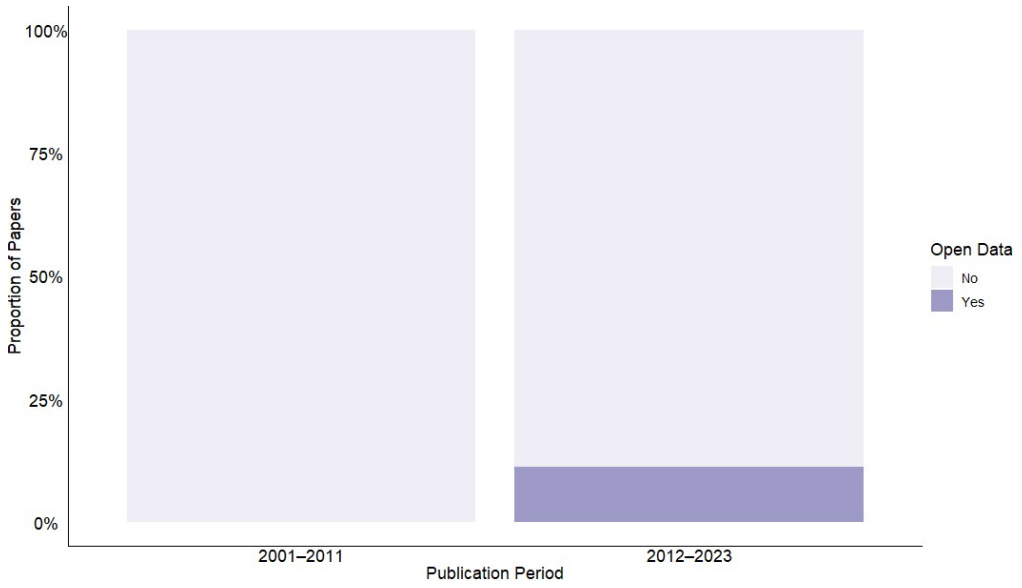



**Figure 2.13** Proportion of papers in my sample with openly available data, categorized into pre-2011 and post-2011 time periods.

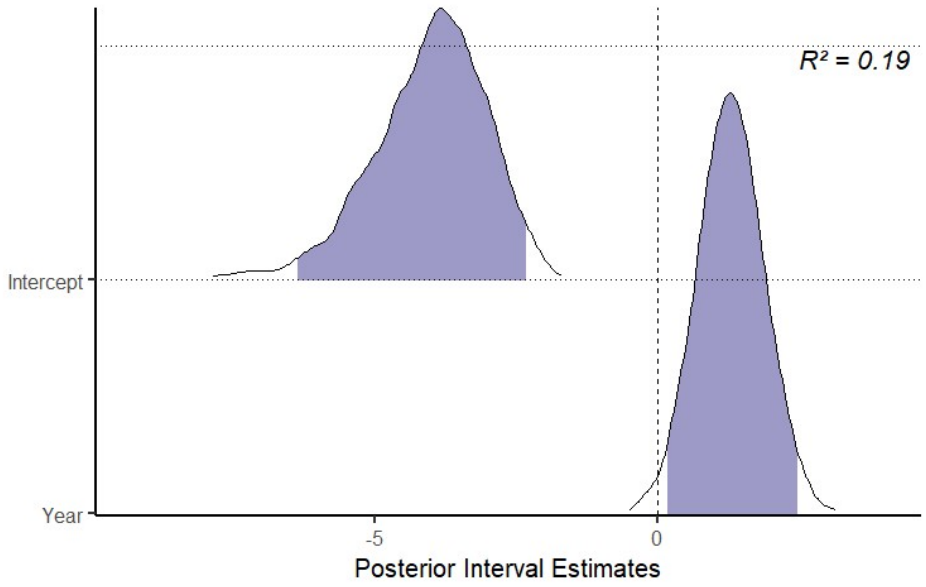



**Figure 2.14** Model 6. Posterior density estimates for the effect of year on the likelihood of a paper having openly available data. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.6 for model code, summary table, pp_check, and conditional effects.

#### 2.3.1.6 Do Authors Preregister Their Studies?

Within my sample of 50 papers, only 1 paper (2%) in my sample was preregistered. Figure 2.15 shows whether this paper was published in the period between 2001-2011, or between 2012-2023. My model indicates that more recent publications are indeed more likely to preregister their methods, which is consistent with my data as the single preregistered paper was published in 2023 ($\beta = 1.04$, 95% CI [-0.20, 2.33], Figure 2.16). As with most other models, the 95% CIs cross over zero so this effect is technically uncertain.

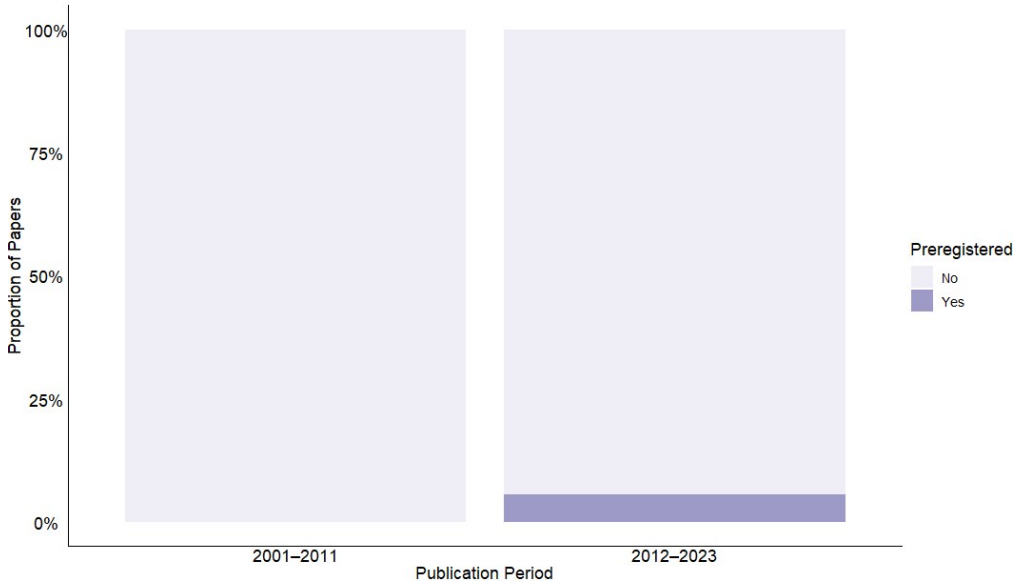



**Figure 2.15** Proportion of preregistered papers in my sample, categorized into pre-2011 and post-2011 time periods.

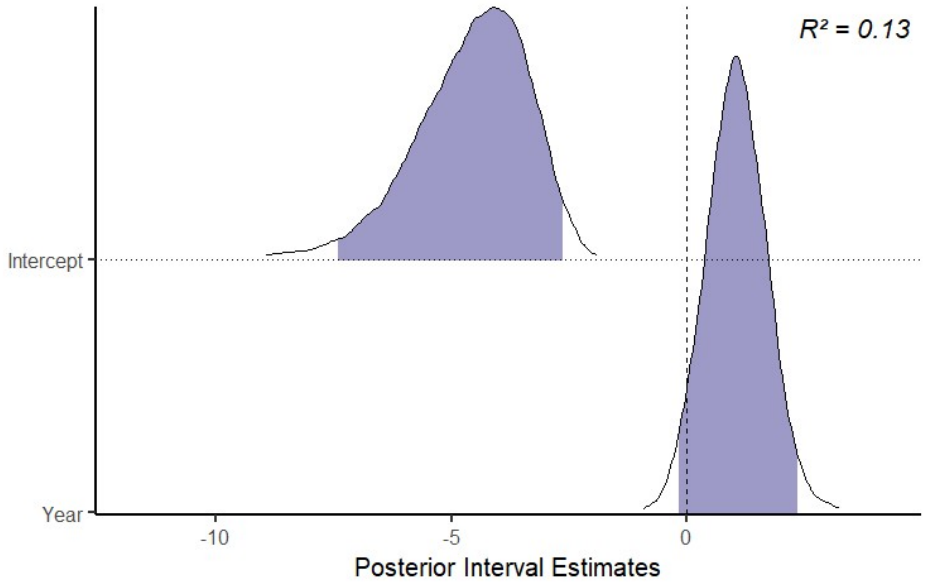



**Figure 2.16** Model 7. Posterior density estimates for the effect of year on the likelihood of a paper being preregistered. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.7 for model code, summary table, pp_check, and conditional effects.

### 2.3.2 Is Methodology Improving?

#### 2.3.2.1 Are Sample Sizes Getting Larger?

Sample sizes of the studies used in my analysis can be found in Figure 2.17. Based on my small sample of the stereotype threat literature, it appears that more recent publications are indeed more likely to have a larger sample size (β = 0.33, 95% CI [0.20, 0.46], Figure 2.18). It appears that authors of more recent publications have made an effort to improve the design of their studies and collect additional data in order to increase the robustness of their findings.

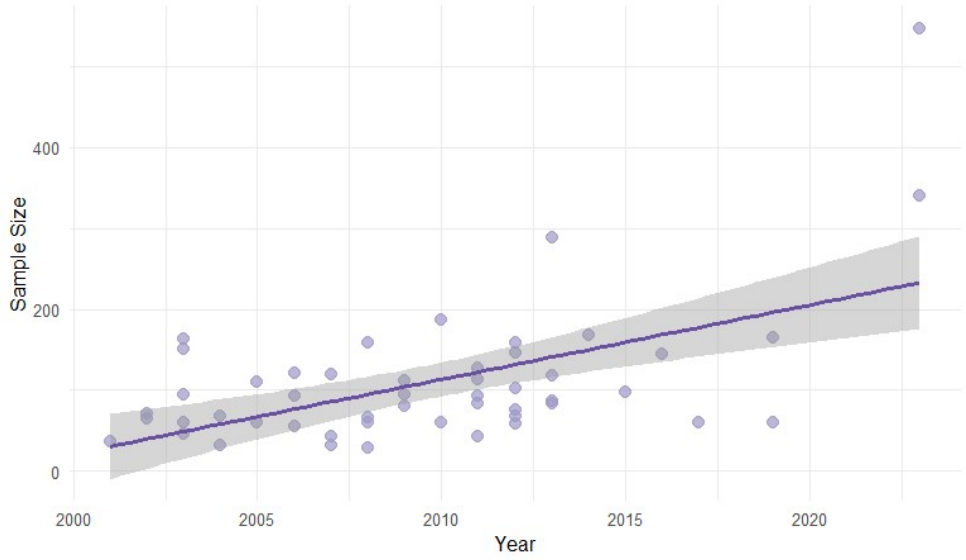



**Figure 2.17** Sample sizes of papers in my sample, from 2001-2023.

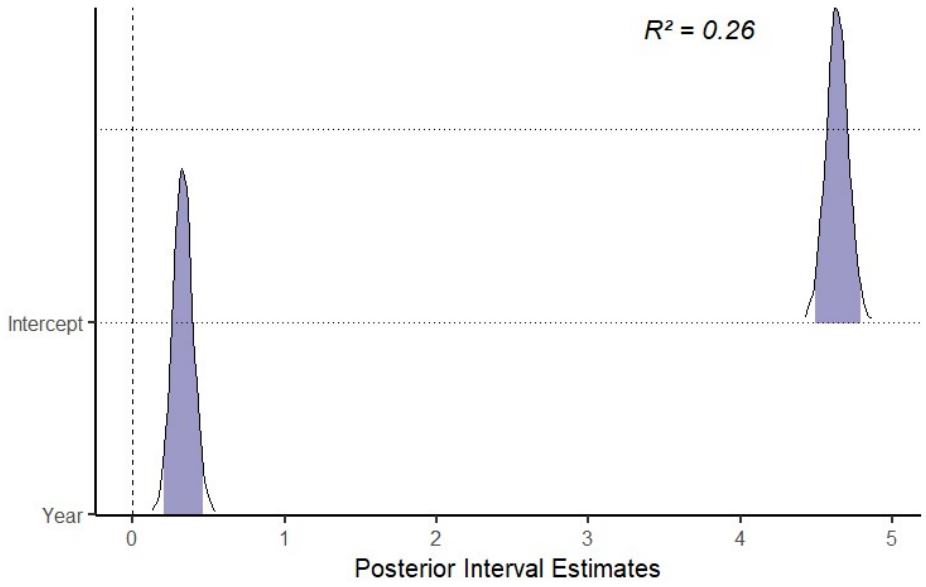



**Figure 2.18** Model 8. Posterior density estimates for the effect of year on a study's sample size. Purple fill is truncated to indicate the 95% credible intervals. model code, summary table, pp_check, and conditional effects.

### 2.3.2.2 Are Study Groups Getting Larger?

Average group sizes of the studies used in my analysis can be found in Figure 2.19. The average number of participants per test group increased through time along with sample size, meaning that it appears as though more recent studies took greater care to improve study design ($\beta = 0.32$, 95% CI [0.19, 0.45], Figure 2.20).

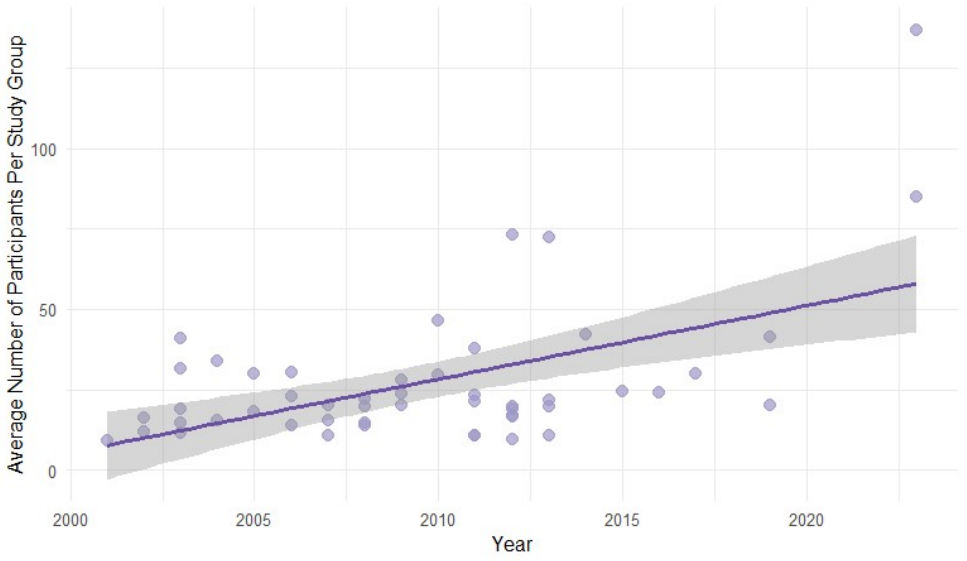



**Figure 2.19** Average number of participants per study group for papers in my sample, from 2001-2023.

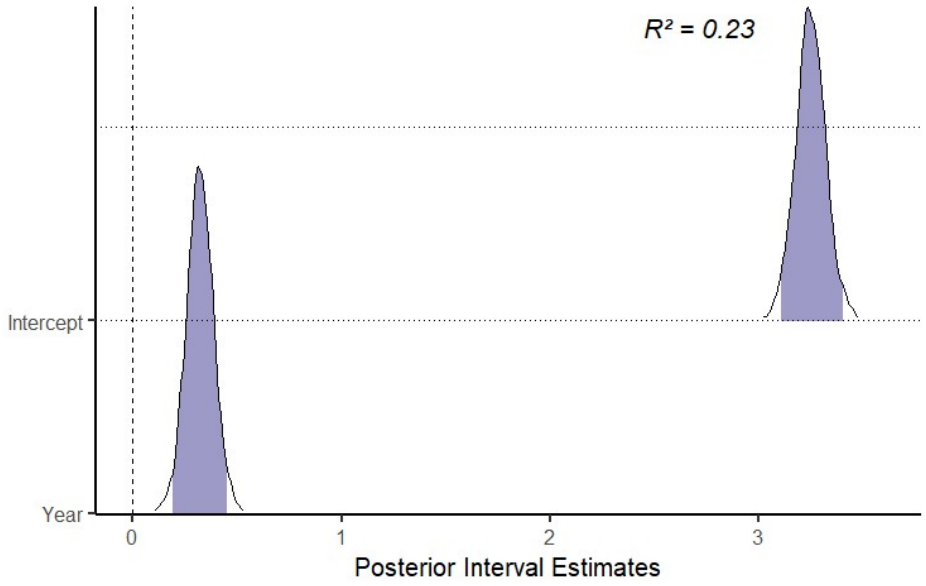



**Figure 2.20** Model 9. Posterior density estimates for the effect of year on a study's average number of participants per group. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.9 for model code, summary table, pp_check, and conditional effects.

#### 2.3.2.3 Are Researchers More Likely to Be Blind to Condition?

Within my sample of papers, only 21 (42%) reported whether the researcher was blind to the intervention, and of those 21 papers, only 10 explicitly blinded the researcher to the condition of the participant they were testing. Figure 2.21 shows whether these papers were published in the period between 2001-2011, or between 2012-2023. Testing only those 21 papers where researcher blindness was reported, I asked if researchers were more likely to be blind to condition in more recently published papers, and I found that they were ($\beta = 0.60$, 95% CI [-0.35, 1.65], Figure 2.22). However, as with many findings within this chapter, this result is uncertain, as the 95% CIs cross zero. This suggests that newer publications may be more interested in reducing sources of bias, but not in a particularly meaningful way.

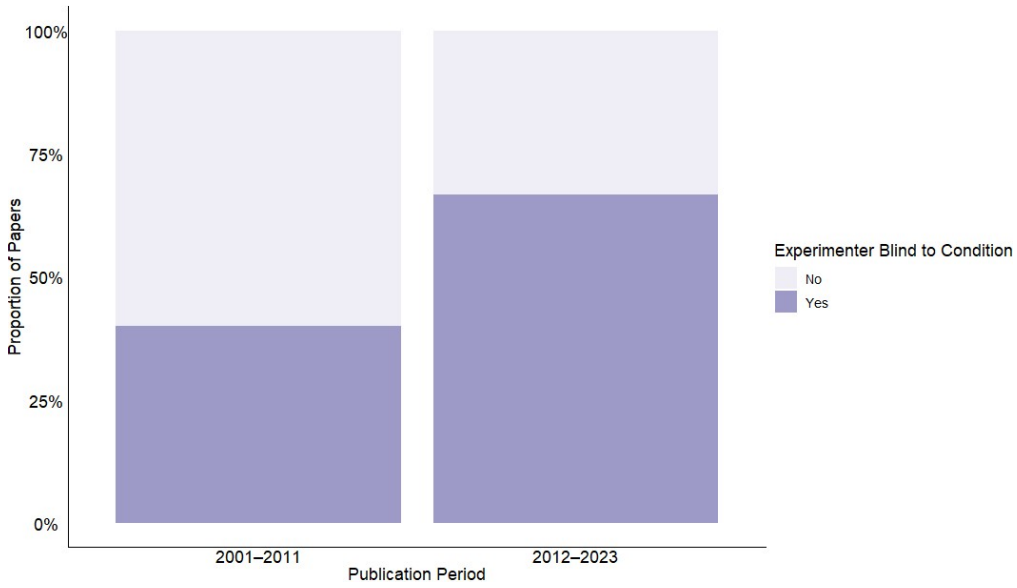



**Figure 2.21** Proportion of papers in my sample with the experimenter(s) blinded to participant condition, categorized into pre-2011 and post-2011 time periods.

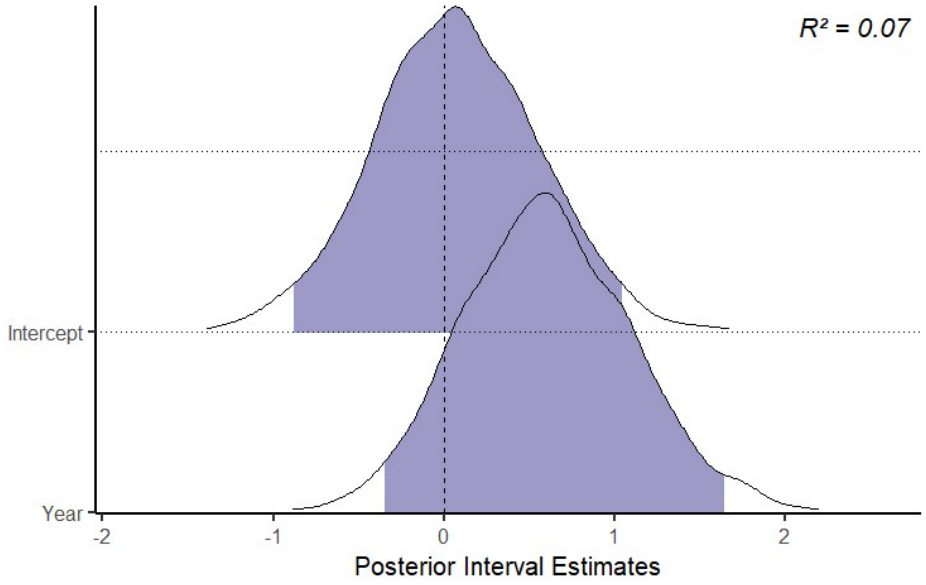



**Figure 2.22** Model 10. Posterior density estimates for the effect of year on the likelihood of researcher being blind to intervention. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.10 for model code, summary table, pp_check, and conditional effects.

#### 2.3.2.4 Are Effect Sizes Getting Larger?

You may recall that of the 50 studies included in my sample, 36 papers reported their effect size. Using typical standards for each type of effect size[1], I categorized each into small, medium, or large effect sizes. Within these 36 papers, 13 reported a small effect, 15 reported a medium effect, and 8 reported a large effect. Figure 2.23 shows the proportion of small, medium, and large papers published in the period between 2001-2011, or between 2012-2023. Referring to Figure 2.24, we can see that the log-odds of a

[1] Cohen's d (small = 0.2, medium = 0.5, large = 0.8)
Eta squared ($\eta^2$) (small = 0.01, medium = 0.06, large = 0.14)
Partial eta squared ($\eta^2p$) (small = 0.01, medium = 0.06, large = 0.14)
Pearson's r (small = 0.1, medium = 0.3, large = 0.5)
R-squared ($R^2$) (small = 0.01, medium = 0.09, large = 0.25)

paper reporting a small effect size versus a medium or large effect is negative, indicating that in general, within the 36 papers that reported their effect size, it is more likely that they will report a medium or large effect size, as opposed to a small one. The log-odds of a study reporting a small or medium effect size versus a large effect size, however, is positively skewed, demonstrating that papers are more likely to report a small or medium effect size, rather than a large one. Taken together, medium effect sizes are the most common, which is consistent with my data. As for the main question, this model demonstrates that more recently published articles are less likely to report larger effect sizes, which indicates that more rigorous methodologies may paint a less exaggerated picture of how stereotype threat affects women's mathematics performance, but again, this effect is not certain ($\beta$ = -0.42, 95% CI [-1.08, 0.26], Figure 2.24).

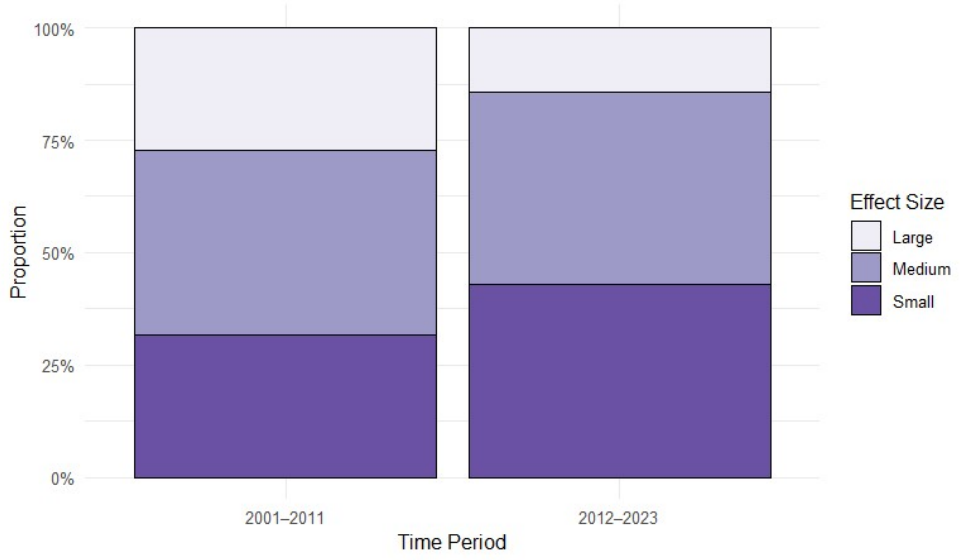



**Figure 2.23** Proportion of papers in my sample with small, medium, and large effect sizes, categorized into pre-2011 and post-2011 time periods.

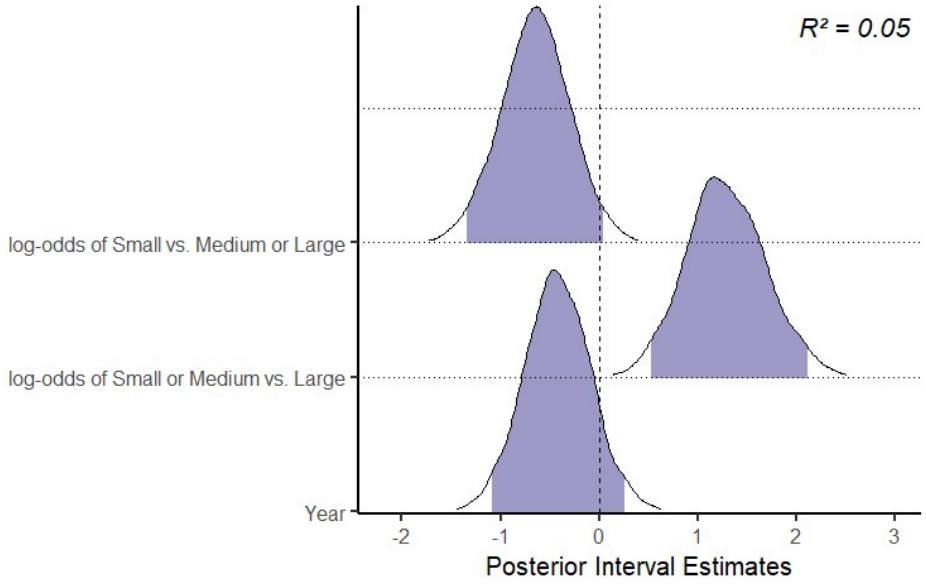



**Figure 2.24** Model 11. Posterior density estimates for the effect of year on the likelihood of larger effect sizes. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.11 for model code, summary table, pp_check, and conditional effects.

### 2.3.3 Are More Recent Papers More or Less Likely to Find an Effect of Stereotype Threat?

Within my sample of 50 papers, 26 (52%) found an outright effect of stereotype threat on women's mathematics performance. Figure 2.25 shows whether these papers were published in the period between 2001-2011, or between 2012-2023. Of the remaining 24 (48%) papers that did not report a basic effect for stereotype threat, a further 15 (30%) reported that stereotype threat had an interaction effect with another variable to reduce women's mathematics performance. That means that only 9 papers (18%) in my sample found no effect of stereotype threat whatsoever. I analyzed these data in two ways: first, I looked at the effect of year on the likelihood of a study finding a straightforward effect of stereotype threat. My results revealed an extremely uncertain but possibly positive effect, meaning that more recent papers may be slightly more likely to report a

straightforward effect of stereotype threat, but this effect is nearly centred on zero, so the likely answer to this question, at least within my sample, is that old and new papers alike are equally as likely to report an effect of stereotype threat on women's mathematics performance (β = 0.12, 95% CI [-0.44, 0.71], Figure 2.26).

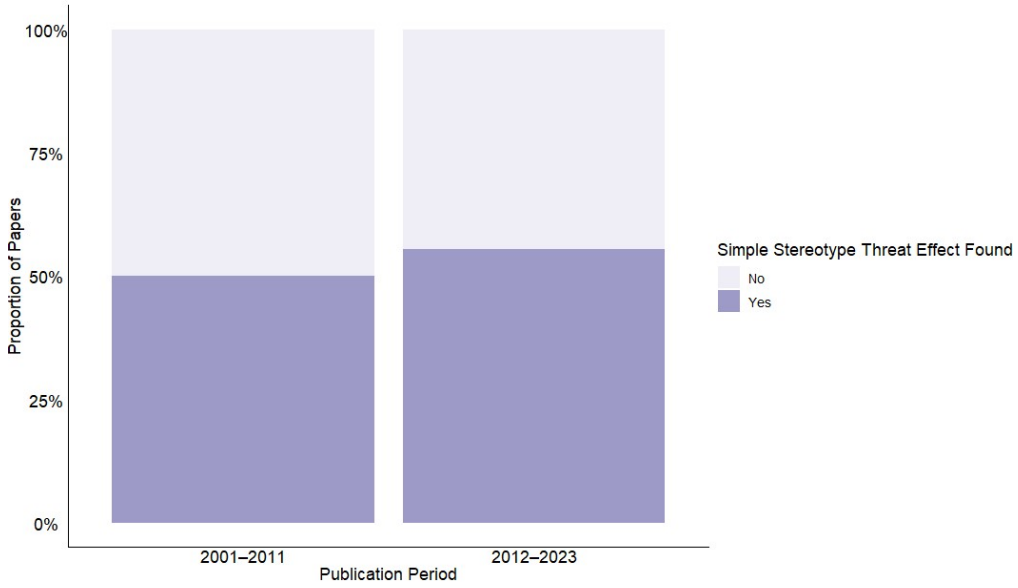



**Figure 2.25** Proportion of papers in my sample reporting a simple stereotype threat effect, categorized into pre-2011 and post-2011 time periods.

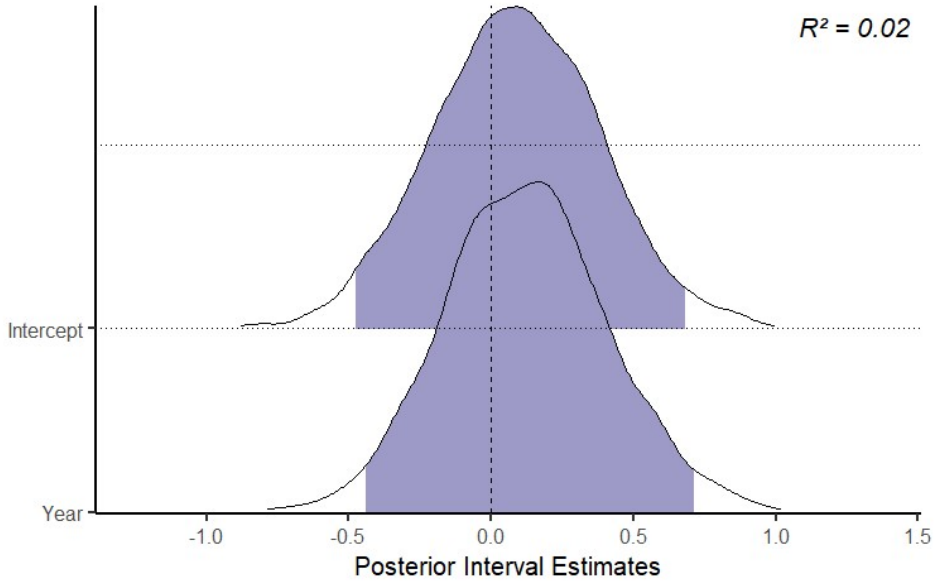



**Figure 2.26** Model 12. Posterior density estimates for the effect of year on the likelihood of reporting a simple effect of stereotype threat. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.12 for model code, summary table, pp_check, and conditional effects.

The second way that I analyzed these data is by considering whether more recent papers are more likely to report *either* a simple stereotype threat effect, or an effect of stereotype threat in interaction with another variable. Other variables include the participant's mathematics self-concept, the participant's testosterone levels, and the type of mathematics question (solve vs. comparison). Figure 2.27 shows whether these papers were published in the period between 2001-2011, or between 2012-2023. When analyzed in this manner, it appears that more recent papers are less likely to report any effect of stereotype threat, whether simple or as an interaction ($\beta$ = -0.44, 95% CI [-1.12, 0.17], Figure 2.28). This effect, like most reported in this chapter, is uncertain, but points towards a trend of more recent papers being less likely to report a finding that supports stereotype threat.

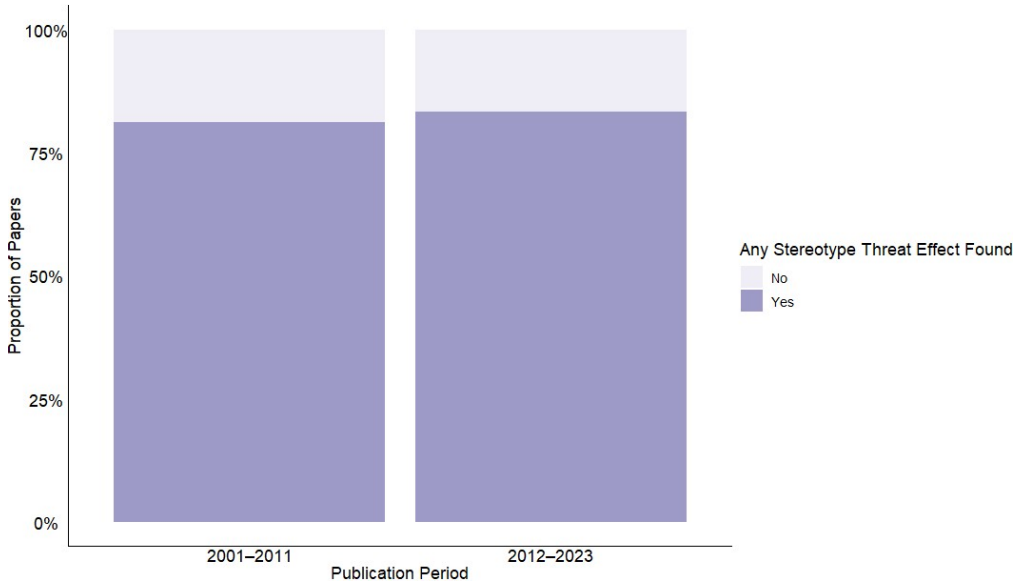



**Figure 2.27** Proportion of papers in my sample reporting any stereotype threat effect, categorized into pre-2011 and post-2011 time periods.

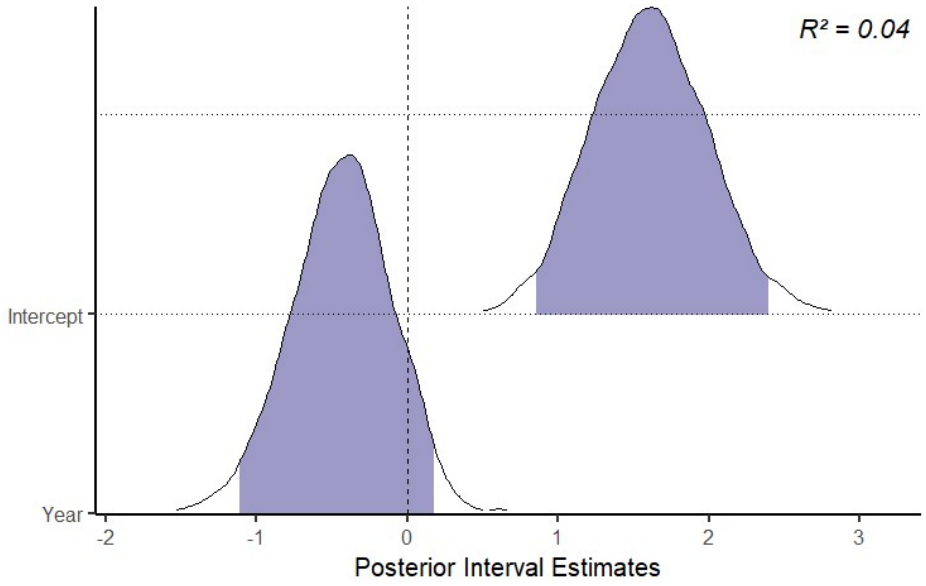



**Figure 2.28** Model 13. Posterior density estimates for the effect of year on the likelihood of reporting any effect of stereotype threat (simple or as an interaction with another variable). Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.13 for model code, summary table, pp_check, and conditional effects.

### 2.3.4 Other Problems Within the Stereotype Threat Literature

#### 2.3.4.1 Missing Participants

5 papers in my sample (10%) reported their result for the effect of stereotype threat on women's mathematics performance with some number of participants seemingly unaccounted for. For 3 papers in my sample (6%), the incomplete reporting of the results made it impossible to tell if all participants were included in the analysis. Finally, for 1 paper (2%) in my sample, it appears that the authors analyzed their results with *more* participants than they initially recruited. This likely isn't a major factor in discrepancies within the stereotype threat literature, but it points to a problem of carelessness that may be contributing to problems of irreproducibility within this research area, and possibly within psychology as a whole.

#### 2.3.4.2 "Gender Differences" vs. "Males Perform Better"

Within my sample, 24 papers (48%) simply informed participants that the mathematics test they were about to take has demonstrated gender differences, leaving it up to the participants' interpretation which gender is supposed to perform better. 23 papers (46%) in the sample explicitly stated that males perform better on the given task. Finally, three papers (6%) in my sample reported such little information about their threat statement that it was impossible to determine whether or not they specified that males performed better on the task. I found that if a study's stereotype threat statement explicitly indicated that males perform better on the task at hand, the study was also potentially more likely to find a simple effect of stereotype threat ($\beta = 0.59$, 95% CI [-0.40, 1.63], Figure 2.29).

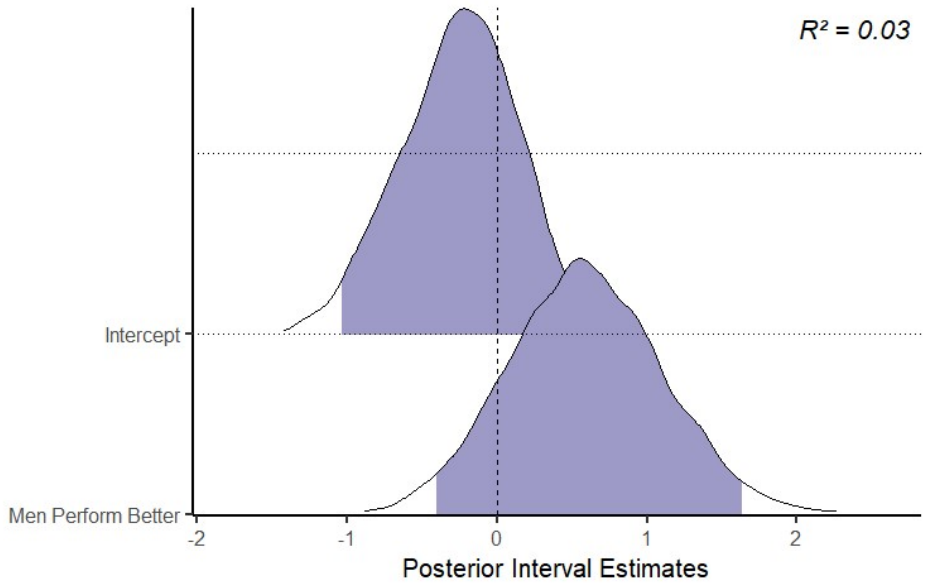



**Figure 2.29** Model 14. Posterior density estimates for the effect of stereotype threat statement wording ("Men Perform Better" vs. "Gender Differences") on the likelihood of finding a simple stereotype threat effect. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.14 for model code, summary table, pp_check, and conditional effects.

#### 2.3.4.3 Inappropriate Control Condition

A mere 6 papers (12%) had a control group that I deemed appropriate to parse out an effect of stereotype threat, 33 (66%) had a control group that I deemed inappropriate, and a further 11 papers (22%) did not provide enough information to determine if the control group was appropriate or not. Using a subset of the data which included only those studies which could be categorized into appropriate or inappropriate control groups, I found that studies with control groups that I deemed appropriate were potentially more likely to find a positive effect of stereotype threat on women's math performance, although the effect is quite uncertain ($\beta = 0.25$, 95% CI [-0.89, 1.40], Figure 2.30).

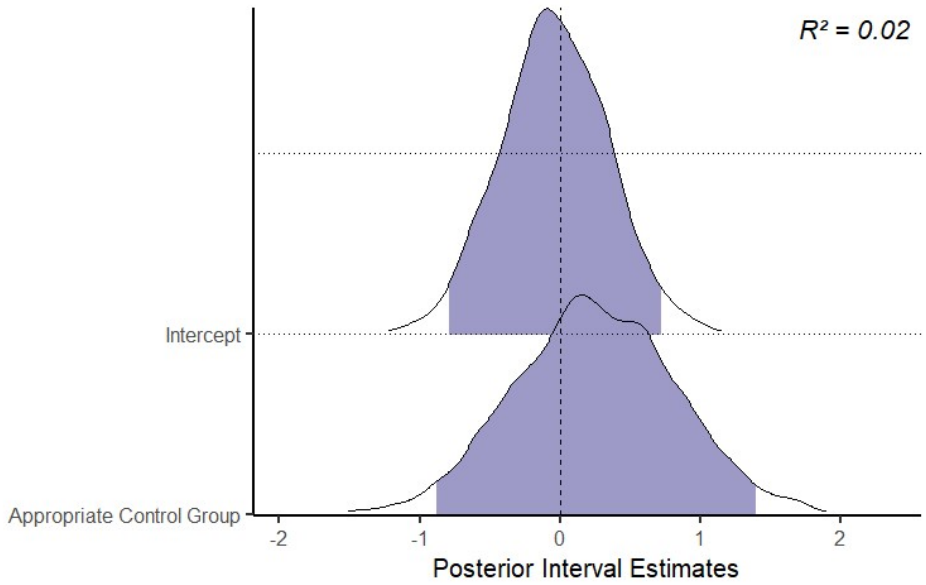



**Figure 2.30** Model 15. Posterior density estimates for the effect of appropriate vs. inappropriate control group on the likelihood of finding a simple stereotype threat effect. Purple fill is truncated to indicate the 95% credible intervals. See Appendix A.4.15 for model code, summary table, pp_check, and conditional effects.

### 2.3.5 Evidence of Replication

Within my sample, I had hoped to find papers that were explicitly replicating prior work on the relationship between stereotype threat and women's math performance. I checked for evidence of replication in the methods and results sections of the papers, as well as by searching for variations of the word "replication" within the full text of each article, and I was only able to find papers that had partially replicated the methods of previous papers on the topic. A partial replication could include stating that the paper was conceptually but not experimentally replicating earlier work, using the same stereotype threat manipulation as a previous paper, or employing the same math test used by other researchers. A total of 14 papers in my sample were partially replicating some methodological aspect of another paper, but I found no papers that described a full replication of previous work on the subject.

### 2.3.6 Results Summary

**Table 2.1** Results Summary Table for Models with Year

| Model | Parameter | Estimate | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|
| 1 | ST Statement Reported | 0.25 | -0.33 | 0.84 |
| 2 | Experimenter Gender Reported | -0.36 | -0.97 | 0.22 |
| 3 | Accounts for Experimenter Gender | 0.40 | -0.56 | 1.31 |
| 4 | Mathematics Test Reported | 1.05 | -0.18 | 2.36 |
| 5 | Effect Size Reported | 0.30 | -0.30 | 0.92 |
| 6* | Open Data | 1.29 | 0.15 | 2.47 |
| 7 | Preregistered | 1.04 | -0.20 | 2.33 |
| 8* | Sample Size | 0.33 | 0.20 | 0.46 |
| 9* | Group Size | 0.32 | 0.19 | 0.45 |
| 10 | Experimenters Blinded | 0.60 | -0.35 | 1.65 |
| 11 | Effect Size | -0.42 | -1.11 | 0.24 |
| 12 | Simple ST Effect | 0.12 | -0.44 | 0.71 |
| 13 | Any ST Effect | -0.44 | -1.12 | 0.17 |

*Note* All models have $\widehat{R}$ = 1.00 and ESS (Effective Sample Size) >1000. * indicates models in which the lower and upper 95% CI (Credible Intervals) do not cross zero.

**Table 2.2** Results Summary Table for Models with Other Predictor Variables

| Model | Parameter | Predictor Variable | Estimate | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|---|
| 14 | ST Effect Found | "Men Perform Better" in ST Statement | 0.59 | -0.40 | 1.63 |
| 15 | ST Effect Found | Appropriate Control Group | -0.36 | -0.97 | 0.22 |

*Note* All models have $\widehat{R}$ = 1.00 and ESS (Effective Sample Size) >1000.

## 2.4 Discussion

Taking into account the factors I have described throughout this chapter, it seems that many of them demonstrate weakly positive effects, but with a high degree of uncertainty around them. This hints at the possibility that studies within psychology, or at the very least, studies within my small subsample of the stereotype threat literature, are slowly moving towards changes that may improve the trustworthiness and replicability of

the published psychological literature. This includes making changes in terms of greater transparency of study methods and results, open dissemination of study data, and preregistration of studies. It also includes improvements to actual methodology, such as increasing sample size and properly blinding researchers to participants' assigned study condition. Taking into account all of my findings, the only definitively positive results that I was able to detect was that sample sizes, and by extension, study groups are increasing in size over time. This makes logical sense, because the easiest way to improve statistical rigor in a study is to simply collect more data.

Other findings that demonstrated improvement over time included the adoption of open data and preregistration, and the likelihood of including in the supplementary material the mathematics test that was used in the study, but these results were based on only one or two papers having done so, which makes these results particularly susceptible to the effects of random chance. Every other finding that I described within this chapter was more or less uncertain, indicating that while small changes might be underway, there is still a long way to go before we can definitively say that we have improved the quality of the published psychological literature.

# CHAPTER 3: HOW RESEARCHER DEGREES OF FREEDOM AFFECT STUDY OUTCOMES

As previously discussed, my research was inspired in part by an influential paper in the replication crisis literature: "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" by Simmons, Nelson, and Simonsohn (2011). In this paper, the authors demonstrated, using simulations as well as real-world experimentation, the ease with which researchers can use ambiguity in research design and analysis to steer their results towards statistical significance. In this paper, Simmons et al. (2011) reported on an experiment in which they showed that listening to "When I'm Sixty-Four" by The Beatles made people younger. To achieve this obviously impossible "finding", they took advantage of flexibility in data collection and analysis to increase the likelihood of finding a false positive effect and then used selective reporting of their methods and results to present the findings in the most convincing way. Specifically, they removed 14 participants from the analysis without reporting the original sample size; they reported the results based on analysis of only two songs when three songs were included in the original study design; they did not decide in advance when to terminate data collection and instead, analyzed the data for every 10 participants as they took part, stopping data collection as soon as they achieved the results that they wanted; they asked participants twelve different questions, all related to age or age-based stereotypes, yet they reported only on two variables that gave them the results they sought; and finally, they did not report results without the covariate of father's age, thus preventing readers from discerning its contribution to the significance of the result. All of these manipulations adhered to the accepted analytical standards of the time.

Taking inspiration from Simmons et al. (2011), I designed a study to explore the effects of researcher degrees of freedom on the outcomes of statistical analyses, by varying the choices made about study design and data analysis. I chose to investigate this through the lens of stereotype threat, since this is a topic that is well established in the literature—the first paper on stereotype threat was published 30 years ago (Steele & Aronson, 1995)—and also due to the contentious nature of the topic: recent studies have called into question the validity of the published literature due to improper statistical methods and publication bias (Flore et al., 2018). My aim is to shed light on problems that have potentially occurred within studies of stereotype threat, and by extension, psychology as a whole.

## 3.1 Study Design

To design a study that could serve as a demonstration for the ignorant and unscrupulous ways that scientists can take advantage of researcher degrees of freedom, I had to consider the potential manipulations that would have the greatest effects on the results. I also tried to avoid common pitfalls that have been implicated as possibly contributary factors to the disparate nature of stereotype threat study results. With these in mind, I designed a study to test women's performance on a video-game-like click-accuracy task under 3 conditions: explicit stereotype threat, implicit stereotype threat, and a control condition with no stereotype threat.

### 3.1.1 Task

Since Steele and Aronson published the first study of stereotype threat in 1995, hundreds of studies have been conducted on the topic. Schimmack (2022) reports a total of 256 articles with the phrase "stereotype threat" in the title or abstract in a database of

121 psychology journals. In my own reading, I came across several articles which reported that over 300 laboratory studies had been published on the effects of stereotype threat, citing Walton and Spencer (2009) as their source. However, when I read Walton and Spencer (2009) I was unable to locate this exact quote. The precise number of studies that have been published on the phenomena of stereotype threat is therefore unclear, but it is apparent that it is in the order of hundreds. Mathematics seems to be the most commonly studied topic within stereotype threat – my search for studies that focused on how stereotype threat affects math performance for women and girls yielded 144 papers, many of which reported on multiple experiments. Spatial abilities, such as mental rotation, are another commonly studied topic in terms of the effects of stereotype threat on task performance (e.g. Campbell & Collaer, 2009). When designing my study, I wanted to avoid tasks relating explicitly to math and spatial abilities, given the strong association between these types of tasks and stereotype threat; I did not want to inadvertently prime stereotype threat in my control condition simply as a result of the chosen task. A lesser studied topic within the stereotype threat literature is motor performance. In fact, when searching for papers, I was only able to find 33 papers regarding stereotype threat and motor skills. In these papers, commonly chosen tasks included golf putting (e.g. Stone et al., 1999; Beilock et al., 2006; Stone & McWhinnie, 2008), soccer dribbling (e.g. Chalabaev et al., 2008; Hermann and Vollmeyer, 2016; Nahidi et al., 2023), and even simple motor tasks such as muscle contraction (e.g. Chalabaev et al., 2013; Laurin et al., 2022). Combining concepts that have shown susceptibility to stereotype threat in the past, such as muscle contraction (e.g. Chalabaev et al., 2013) and reaction time (e.g. Jamieson & Harkins, 2007), I chose to investigate stereotype threat by using a task that measured the speed and accuracy of clicking on

targets displayed on a computer screen, in the digital aim training game KovaaK's (KovaaK's Games, 2018). Kovaak's was chosen due to its customizability, and also because gameplay data are automatically exported to .csv. With this task, I hoped to capture any effects of stereotype threat that may inhibit muscle contractions and reaction time when moving the mouse cursor to the chosen target, as well as any effects on cognitive processing that may occur when perceiving and choosing a target to click on. See Appendix B.1 for a visualization of the KovaaK's aim training game.

### 3.1.2 Task Difficulty

Another consideration in my study design was the difficulty of the task. As discussed in Chapter 2, previous studies have found that insufficiently difficult tasks may not be susceptible to the detrimental effects of stereotype threat and may, in fact, promote improved performance under threat conditions (Spencer et al., 2009; O'Brien & Crandall, 2003). For this reason, I included two versions of my chosen click-accuracy task that varied in task difficulty. This was achieved by reducing the size of the targets for the more difficult version of the test. Again, refer to Appendix B.1 for a visualization of the Easy and Difficult versions of the task.

### 3.1.3 Researcher Sex

Another factor with an established effect on the outcome of stereotype threat research on women is the presence or absence of males, whether in the form of a male researcher, or male participants being tested simultaneously with female subjects (Inzlicht and Ben-Zeev, 2000). For this reason, I wanted to vary the sex of the researcher. In my examination of a subset of the stereotype threat literature, I found that many studies employ only a single researcher without regard for their sex, or they may sometimes

employ a single representative of each sex, but rarely do they report having multiple researchers of each sex. Even if a study attempts to control for the effects of researcher sex on task performance by using both sexes, a single representative of each does not seem sufficient to parse out the confound that arises when a single exemplar is used in a study. Using only a single male and a single female experimenter does not account for individual differences that may affect how a participant performs on the given task, and also constitutes an example of pseudoreplication, since the effect of the individual researcher and the effect of researcher sex cannot be statistically separated. For that reason, I used multiple male and multiple female researchers, in an attempt to average out individual differences and interrogate actual performance differences brought about by the sex of the researcher.

### 3.1.4 Social Identities and Attitudes Scale (SIAS)

In addition to task difficulty and researcher sex, several other factors have been shown to moderate how stereotype threat affects task performance, including gender identification (Schmader, 2002), gender stigma consciousness (Pinel, 1999), domain identification (Smith & White, 2001), domain self-concept (Keller, 2007), and negative affect (Cadinu et al., 2005). These factors and how they relate to stereotype threat have all been studied individually, but prior to 2011, their effects had not been studied collectively. To address this deficit in the literature, Picho and Brown (2011) created and validated the Social Identities and Attitudes Scale (SIAS), containing these key stereotype threat dimensions, as well as questions pertaining to ethnic group identification. The original SIAS (Picho & Brown, 2011) focused on mathematics, but Picho and Brown (2011) encouraged readers to adapt the SIAS as needed to capture their own domain of study. As

such, I used a modified version of the SIAS, removing questions related to ethnicity, as it was not my area of focus, and adjusting questions that mentioned mathematics to refer instead to "competitive tasks." See Appendix B.2 for the modified SIAS.

## 3.2 Methods

### 3.2.1 Participants

130 students enrolled in Psychology courses at the University of Lethbridge during the Fall semester of 2024 were recruited for this study. Registration was advertised to non-male participants to reduce the number of study groups required. However, there were no restrictions in place to prevent students from enrolling in the study regardless of sex or gender, and several students were excluded based on their answers to the demographics (more on participant exclusion below). See Appendix B.3 for ethnicity and age distributions for my sample. Participants were randomly assigned to one of eight conditions, as outlined in Tables 3.1 and 3.2.

**Table 3.1** Description of Study Groups Before Participant Exclusion

| Researcher Sex | Task Difficulty | *n* (subtotal) | Stereotype Threat | *n* |
|---|---|---|---|---|
| Female | Easy | 32 | Yes | 8 |
| | | | No | 24 |
| | Difficult | 32 | Yes | 20 |
| | | | No | 12 |
| Male | Easy | 32 | Yes | 15 |
| | | | No | 17 |
| | Difficult | 34 | Yes | 21 |
| | | | No | 13 |

*Note. n* is based on the original sample size of 130. '***n* (subtotal)**' describes the number of participants in each group before accounting for the stereotype threat condition. Group sizes were kept approximately equal, but the randomization of stereotype threat statement introduced group imbalance.

**Table 3.2** Description of Study Groups After Participant Exclusion

| Researcher Sex | Task Difficulty | *n* (subtotal) | Stereotype Threat | *n* |
|---|---|---|---|---|
| Female | Easy | 13 | Yes | 4 |
| | | | No | 9 |
| | Difficult | 18 | Yes | 12 |
| | | | No | 6 |
| Male | Easy | 13 | Yes | 7 |
| | | | No | 6 |
| | Difficult | 19 | Yes | 12 |
| | | | No | 7 |

*Note.* *n* is based on the final sample size of 63. '***n* (subtotal)**' describes the number of participants in each group before accounting for the stereotype threat condition. Group sizes were kept approximately equal, but the randomization of stereotype threat statement introduced group imbalance.

### 3.2.2 Recruitment

Participants were recruited from SONA, the University of Lethbridge's online study recruitment system. Recruitment materials employed deception to prevent participants from knowing the true nature of the study; potential participants read that they would be completing a study investigating "hand-eye coordination, by testing speed and accuracy in a short computer task". Debriefing took place at the very end of the study. See Appendices B.4 and B.5 for recruitment materials and debriefing documentation.

### 3.2.3 Condition Randomization and Counterbalancing

Counterbalancing of researcher sex was based on the availability of the individual researchers. Eight researchers (four male, four female) were placed on a weekly schedule, with approximately equal timeslots, although availability of researchers often changed from week to week, so researchers sometimes ran participants outside of their regularly scheduled timeslots. Task difficulty was counter-balanced by alternating the difficulty between easy and difficult. If a particular category was under-represented, often due to

no-shows unevenly affecting one particular task difficulty, researchers would only conduct one task difficulty until the numbers evened out. Through this method, study group sizes were kept approximately equal throughout the study, before knowing which participants were assigned to each stereotype threat condition. Randomization for stereotype threat condition was achieved through the administration of a Qualtrics survey. This automatically randomized between a control statement and a statement designed to illicit stereotype threat and was shown to participants before they completed the main task, but after they completed the pre-test. See Appendix B.6 for the control and stereotype threat statements. Researchers were instructed not to read the statement shown to participants, but they did not leave the room while the statement was displayed, so it is technically possible that some participants may have had a researcher who was aware of their study group. Participation in this study was rewarded with a 1% bonus course credit allocated to an eligible course of the participant's choosing.

### 3.2.4 Procedure

Participants were tested individually. Upon arrival at the laboratory, participants signed an informed consent form. The researcher then explained the task, following a script to ensure each participant heard the same information (see Appendix B.7 for the researcher script). Once the participants had any clarifying questions addressed, they would begin. First, they would play through a thirty-second-long version of the task three times in a row, which allowed participants to familiarize themselves with the procedure, in addition to giving me a baseline measurement of their skill at this task. These three pre-test trials were completed using the easy task. Next, participants were instructed to read a short statement on the computer. The researcher would navigate to a Qualtrics survey and

instruct the participant to click the "Next" button when they were ready to view the statement. The researcher was instructed to look away from the screen to avoid learning which study group to which the participant had been randomly assigned. After participants had read the statement, they entered their SONA ID to acknowledge that they had read and understood the statement. This also enabled me to keep track of which statement they had been shown. Participants then completed a ninety-second-long version of the task, either easy or difficult, depending on their study group. After testing was completed, participants would complete a survey that asked questions pertaining to their prior experience with similar tasks, such as using a computer mouse and playing video games (See Appendix B.8), a demographics questionnaire (See Appendix B.9), and the modified version of the Social Identities and Attitudes Scale (SIAS).

### 3.2.5 Participant Exclusion

A total of 67 participants were excluded from my study, bringing my initial sample size of 130 down to 63 (Table 3.3). Note that the total number of participants listed sums to 72, since some participants were excluded from the study for multiple reasons.

**Table 3.3** Reasons for Participant Exclusion

| Reason For Exclusion | Number of Participants Excluded |
|---|---|
| Participants took an earlier version of the study. | 27 |
| Participants attended meetings where I explained the true purpose of the study. | 2 |
| Participants indicated on the demographics survey that their sex or gender was male. | 2 |
| Participants did not read the stereotype threat or control statement. | 2 |
| Participants were underage. | 2 |
| Participants experienced a Kovaak's glitch during their pre-test trial. | 22 |
| Participants experienced a Kovaak's glitch during their main trial. | 15 |

### 3.2.6 Data Analysis

I constructed six models in a Bayesian framework, using the 'brms' package (Bürkner, 2017), in R (v4.3.2; R Core Team, 2023). Each was a simple logistic regression. Each model used participant's test group (consisting of researcher sex, task difficulty, and stereotype threat statement) as the predictor variable, with the following response variables: (16) Accuracy score, (17) Performance score, (18) Performance Change score, (19) Hits.i|trials(Clicks.i) score. Model 20 and Model 21 are identical to Model 19, but the analysis was performed on subsets of the data, with Model 20 representing data only from Sam and Tyler, and Model 21 representing data only from Anna and Drayton. A student distribution was specified for Model 16 through Model 18, since Accuracy, Performance, and Performance Change are all continuous and positively skewed, and a binomial distribution was specified for Models 19 through Model 21 since Hits.i|trials(Clicks.i) represents success and failure counts out of a known number of

trials. For Model 16 through Model 21, four chains and 8000 iterations were specified. The R code for this analysis is available here: https://github.com/samantha-lloyd/Analysis/blob/main/Manipulation-Analysis.

To prepare my data for analysis, I scaled and mean-centered all continuous variables. All models were run with weakly informative priors (normal (0, 1)). Model convergence was confirmed using $\hat{R}$s ($\hat{R}$s=1.00), and model performance was assessed using the 'pp_check' function in 'bayesplot' (Gabry & Mahr, 2017). Credible intervals were set to 95% for ease of interpretation. Figures were generated using the 'ggplot2' (Wickham, 2009) and 'ggridges' (Wilke, 2018) packages. Conditional and marginal $R^2$ values were calculated using the 'bayes_R2' function (Bürkner, 2017).

### 3.3 Results

I wish to introduce this results section with a justification for its format. In many ways, my approach did not follow the typical formula for a psychological study, and the results section reflects those idiosyncrasies. First, I wanted to present my findings in in a straightforward descriptive fashion and then take advantage of researcher degrees of freedom to illustrate how the same study might have yielded different results depending on choices made throughout the process. Speelman and McGann (2020) have introduced the idea of the pervasiveness analysis, an approach that aims to address the ergodic fallacy in psychology: the use of group-based aggregate statistics to draw conclusions about the individuals within a sample. A pervasiveness analysis is quite simple on its face: we need only to know the proportion of individuals in a sample that show the given effect. This method of presenting results sidesteps the potential pitfalls of researcher degrees of freedom, as the experimenter needs only to make a few choices before the

findings are presented, namely, defining the desired effect of the study, and deciding on the threshold for that effect being present in a group. Speelman and McGann (2020) suggest a threshold of 80% as offering convincing evidence that the majority of people show an effect. This pervasiveness analysis provides an uncomplicated way of comparing participants' performance across different study groups to determine how researcher sex, task difficulty, and stereotype threat affect task performance on the actual individuals in my sample. Following this, I will use a series of Bayesian models to demonstrate the effects of different study design and analysis choices.

### 3.4.2 Pervasiveness Analysis

To begin my pervasiveness analysis, I calculated the difference in performance between pre-test and actual test for each participant. Participants displaying a performance change of 10% or more in either direction were said to have increased or decreased their performance on the task, whereas differences within the 10% window on either side of 0 were said to show no meaningful difference in performance. I then calculated the proportion of individuals demonstrating a performance decrease, a performance increase, and no change in performance for each of the 8 study groups, outlined in Table 3.4. First, there are no groups for which the majority (80% or more) of participants demonstrated a decrease in performance between the pre-test and the test. There are several groups, however, in which the majority of participants experienced an *increase* in performance on the actual test: 100% of participants in the Male/Easy/Control group showed an increase in performance ($M$ = +64%), while 86% of participants in the Male/Easy/Stereotype Threat statement improved their task performance ($M$ = +44%).

**Table 3.4** Proportion of individuals showing a performance decrease, a performance increase, and no change in performance by study group.

| Researcher Sex | Female | | | | Male | | | |
|---|---|---|---|---|---|---|---|---|
| Task Difficulty | Easy | | Difficult | | Easy | | Difficult | |
| Stereotype Threat | Yes | No | Yes | No | Yes | No | Yes | No |
| **Proportion of individuals showing a performance reduction** | **25%** | **0%** | **33%** | **33%** | **0%** | **0%** | **42%** | **71%** |
| Proportion of individuals showing a performance increase | 50% | 67% | 25% | 33% | *86% | *100% | 50% | 14% |
| Proportion of individuals showing no difference in performance | 25% | 33% | 42% | 33% | 14% | 0% | 8% | 14% |
| Mean performance difference (%) | +21% | +26% | -1% | +6% | +44% | +64% | +4% | -12% |

*Note.* The proportion of individuals showing a performance reduction is in **bold**, as this is the target effect. * indicates effects that the majority of participants in a given group exhibited.

While it is true that there are no groups for which the majority of participants demonstrated a drop in performance, there is an odd effect for those participants in the Male/Difficult conditions: 71% of participants who read the control statement performed worse on the task ($M$ = -12%), whereas only 42% of those participants reading the stereotype threat statement performed worse ($M$ = +4%). This runs counter to the typical stereotype threat hypothesis: one would expect that reading the stereotype threat

statement would cause a higher proportion of participants to perform worse on the task, not the other way around. If we do not split the results by the sex of the experimenter, this effect nearly disappears, demonstrating that conclusions drawn from the analysis of pervasiveness can be susceptible to manipulations (Table 3.5).

**Table 3.5** Proportion of individuals showing a performance decrease, a performance increase, and no change in performance by study group. This graph represents the results not split by experimenter sex.

| Task Difficulty | Easy | | Difficult | |
|---|---|---|---|---|
| Stereotype Threat | Yes | No | Yes | No |
| **Proportion of individuals showing a performance reduction** | **7%** | **0%** | **37%** | **48%** |
| Proportion of individuals showing a performance increase | *80% | *90% | 37% | 38% |
| Proportion of individuals showing no difference in performance | 13% | 10% | 27% | 14% |
| Mean performance difference (%) | +42% | +45% | +1% | +4% |

*Note.* The proportion of individuals showing a performance reduction is in bold, as this is the target effect I was most interested in. * indicates effects that the majority of participants in a given group exhibited.

### 3.4.3 Pervasiveness Analysis Discussion

There are several interesting points to discuss with regard to the pervasiveness analysis. First, stereotype threat statements were predicted to cause a performance reduction compared to the control statement. However, the only group for which this was true was the Female/Easy condition, where 25% of participants who read the stereotype threat statement exhibited a performance reduction ($M$ = +21%), compared to 0% of participants in the control statement condition ($M$ = +26%). This is an interesting finding,

as the female experimenter theoretically represents the less threatening condition, as does the easy task. It is important to note that this result is based on an *n* of 4, which means that only a single participant exhibited this reduction in performance - not exactly a robust effect. I can say with confidence, therefore, that within my sample, the expected stereotype threat effect was not present.

Indeed, one of the groups displayed behavior counter to the expected stereotype threat effect: for the participants in the Male/Difficult group, individuals that read the stereotype threat statement were actually less likely to exhibit a performance reduction than those that were given the control statement (42% vs. 71%). This effect is interesting because one would expect that, in the most theoretically threatening condition, (male experimenter, difficult task), the stereotype threat statement would have a more profoundly negative effect on participants' performance. Other stereotype threat literature supports a dual process model of stereotype threat, which proposes that explicit and implicit stereotype threat cues operate through independent mechanisms, but both have detrimental effects on performance (Stone & McWhinnie, 2008). My findings, however, are supported by Nguyen and Ryan's (2008) meta-analysis, in which they found that overall, subtle stereotype threat cues yielded the largest effect sizes when compared to moderate and blatant cues. Perhaps when the stereotype threat is too overt, participants do not internalize its effects, and instead attempt to rebel against them, resulting in increased performance instead.

For participants who took the easy version of the task, the presence of a male experimenter resulted in a greater proportion of participants increasing their task performance when compared to a female experimenter for both the control condition

(67% increased performance with a female experimenter (*M* = +26%) compared to 100% with a male experimenter (*M* = +64%)), and the stereotype threat condition (50% increased performance with a female experimenter (*M* = +21%) compared to 86% with a male experimenter (*M* = +44%)). This suggests one of two things: either a male experimenter was perceived as less threatening than a female experimenter while completing an easy task, or the presence of a male experimenter was motivational to participants, causing them to try harder on the task in an attempt to defy negative stereotypes.

### 3.4.4 Varying Researcher Degrees of Freedom

#### 3.4.4.1 Varying the Outcome Variable

To investigate this, I created four models, changing only the outcome variable and the model family when necessary, but keeping all other model parameters constant. The four outcome variables I chose were: accuracy (which is simply hits divided by clicks); performance (accuracy multiplied by the inverse of the time taken per successful hit as high accuracy and low time are better); performance change (task performance subtracted from pre-test performance); and Hits.i|trials(Clicks.i) (proportion of successful hits given the number of clicks, moderated by the number of clicks attempted). For each model, the reference group (intercept) is the Female/Easy/Control condition, as this should theoretically be the least threatening condition.

From the accuracy model (Figure 3.1), we can conclude that there was a meaningful negative effect on participant's accuracy in both the Male/Difficult/Control condition (β = -0.08, 95% CI [-0.16, -0.00]), and the Male/Difficult/Stereotype Threat

condition (β = -0.07, 95% CI [-0.14, -0.00]). That is, completing a difficult task in the presence or a male experimenter decreased task performance, regardless of the nature of the statement. No other meaningful effects were detected. Study group explains about 25% of the variance in participants' accuracy.

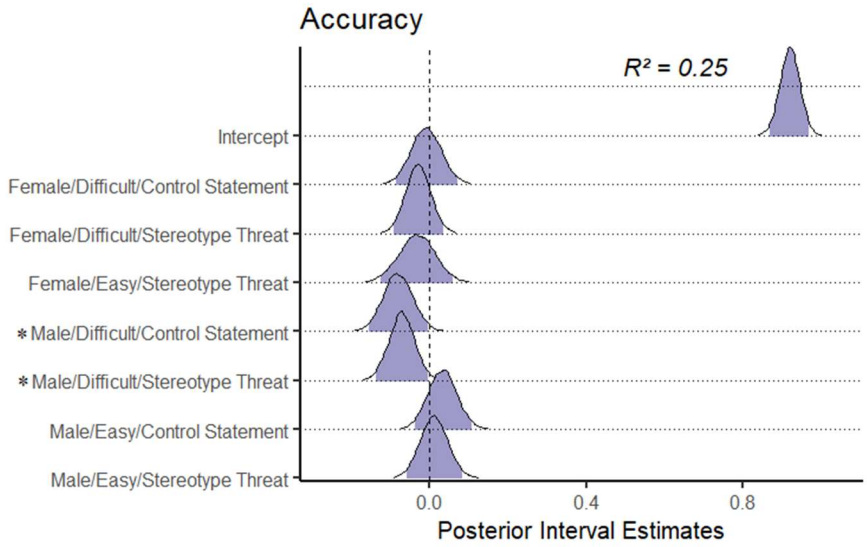



**Figure 3.1** Posterior density estimates for the effect of condition on participants' accuracy. Purple fill is truncated to indicate the 95% credible intervals. See Appendix B.10.1 for model code, summary table, and pp_check. * indicates effects whose CIs do not cross zero.

The performance model (Figure 3.2) showed a meaningful negative effect of task difficulty. Specifically, an effect was detected in all four groups that took the difficult version of the task: Female/Difficult/Control (β = -0.16, 95% CI [-0.30, -0.01]), Female/Difficult/Stereotype Threat (β = -0.24, 95% CI [-0.36, -0.12]), Male/Difficult/Control (β = -0.25, 95% CI [-0.39, -0.11]), and Male/Difficult/Stereotype Threat (β = -0.26, 95% CI [-0.38, -0.13]). We can also say that the decrease in performance is slightly larger when participants' read the stereotype threat statement compared to the control statement, and that this difference is more exaggerated with a

female experimenter. No other meaningful effects were detected. Study group explains about 51% of the variance in participants' performance.

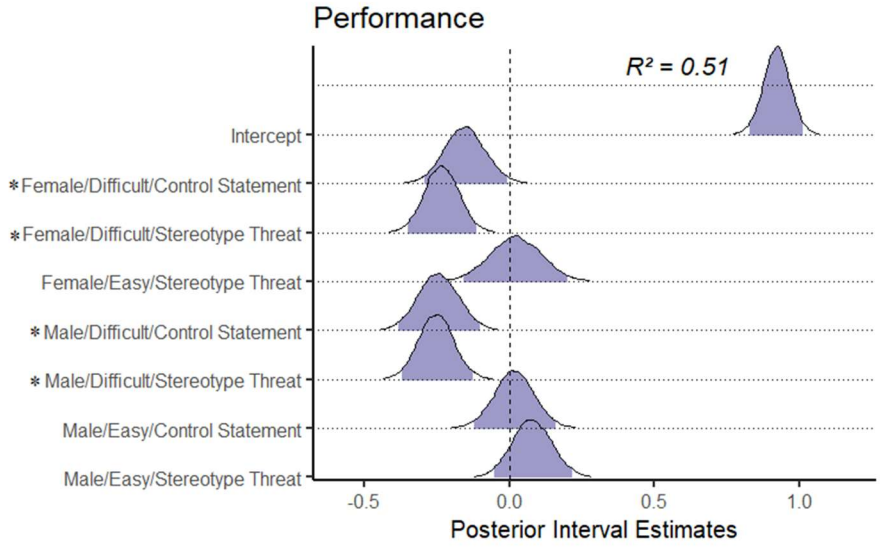



**Figure 3.2** Posterior density estimates for the effect of condition on participants' performance. Purple fill is truncated to indicate the 95% credible intervals. See Appendix B.10.2 for model code, summary table, and pp_check. * indicates effects whose CIs do not cross zero.

Moving on to the performance change model (Figure 3.3), only two groups displayed a meaningful change in performance: Female/Difficult/Stereotype Threat condition (β = -0.19, 95% CI [-0.37, -0.02]), and the Male/Difficult/Control condition (β = -0.29, 95% CI [-0.49, -0.08]). In the difficult condition, reading the stereotype threat statement was more threatening in the presence of a female experimenter, but reading the control statement was more threatening in the presence of a male experimenter. No other meaningful effects were detected. Study group explains about 38% of the variance in participants' change in performance.

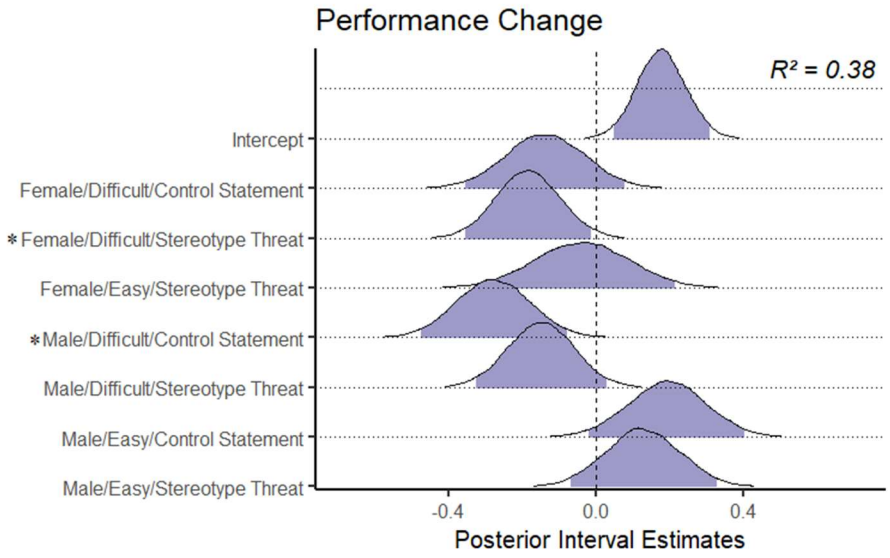



**Figure 3.3** Posterior density estimates for the effect of condition on participants' performance change, calculated by subtracting task performance from pre-test performance. Purple fill is truncated to indicate the 95% credible intervals. See Appendix B.10.3 for model code, summary table, and pp_check. * indicates effects whose CIs do not cross zero.

Lastly, we can look at the Hits.i|trials(Clicks.i) model (Figure 3.4). Here, five conditions were shown to have an meaningful effect on participants performance: Female/Difficult/Stereotype Threat (β = -0.34, 95% CI [-0.65, -0.04]), Female/Easy/Stereotype Threat (β = -0.45, 95% CI [-0.81, -0.08]), Male/Difficult/Control (β = -0.84, 95% CI [-1.15, -0.53]), and Male/Difficult/Stereotype Threat (β = -0.83, 95% CI [-1.12, -0.55]). There was also a positive meaningful effect for the Male/Easy/Control condition (β = 0.67, 95% CI [0.21, 1.14]). As the Female/Difficult/Stereotype Threat and Female/Easy/Stereotype Threat groups both experienced a reduction in score compared to Female/Easy/Control, this suggests that, when a female experimenter is present, reading the stereotype threat statement reduces performance no matter the difficulty of the task. Next, since the Male/Difficult/Control and Male/Difficult/Stereotype Threat had an

approximately equal reduction, we can conclude that the difficult task in the presence of a male researcher is equally threatening, regardless of the statement read. Finally, since the Male/Easy/Control statement experienced an improvement to their score compared to the Female/Easy/Control condition, this means that a male experimenter is either less threatening, or more motivating than a female experimenter, when the task is easy and there is no explicit stereotype threat. No other meaningful effects were detected. Study group explains about 84% of the variance in participants' score.

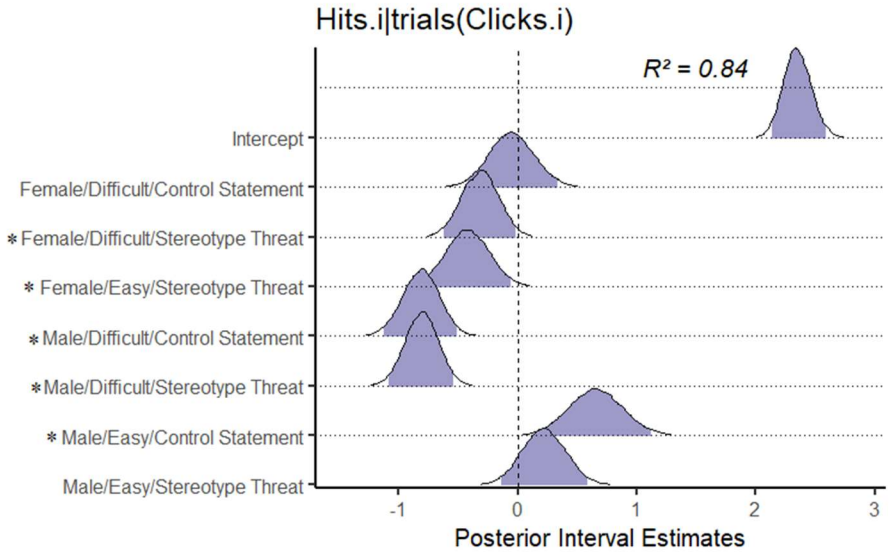



**Figure 3.4** Posterior density estimates for the effect of condition on participants' proportion of successful hits out of their total number of clicks, moderated by total number of clicks. Purple fill is truncated to indicate the 95% credible intervals. See Appendix B.10.4 for model code, summary table, and pp_check. * indicates effects whose CIs do not cross zero.

### 3.4.4.2 Varying the Individual Researchers Included in the Analysis

In my study, I used four male and four female researchers in order to control for pseudoreplication. Here, I investigate what the results might have looked like if I had only used a single experimenter of each sex. I used Hits.i|trials(Clicks.i) as the outcome

variable, based on the model above. so that results presented in this section can be compared to the original model with all researchers included.

For the Female/Difficult/Control group, the results for the Sam(F)-Tyler(M) pairing (Figure 3.5) show a decrease in participants' score (β = -0.88, 95% CI [-1.38, -0.38]), whereas the results from the Anna(F)-Drayton(M) pairing (Figure 3.6) show an increase in participants' score (β = 0.97, 95% CI [0.18, 1.83]). The Male/Difficult/Stereotype Threat condition reveals similar effects, with Sam-Tyler yielding a negative effect (β = -2.07, 95% CI [-2.52, -1.60]) and Anna-Drayton displaying a positive effect (β = 0.64, 95% CI [0.12, 1.16]). While there were no other meaningful effects, the same trend is present for the other groups, with Sam-Tyler leading to generally reduced performance, and Anna-Drayton leading to generally increased performance.

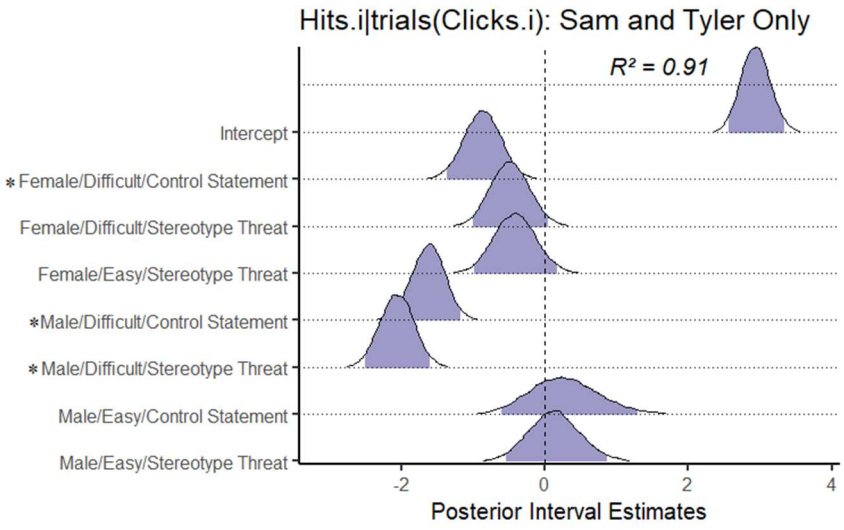



**Figure 3.5** Posterior density estimates for the effect of condition on participants' proportion of successful hits out of their total number of clicks, moderated by total number of clicks. This model represents data from only one female researcher (Sam) and one male researcher (Tyler). Purple fill is truncated to indicate the 95% credible intervals. See Appendix B.10.5 for model code, summary table, and pp_check. * indicates effects whose CIs do not cross zero.

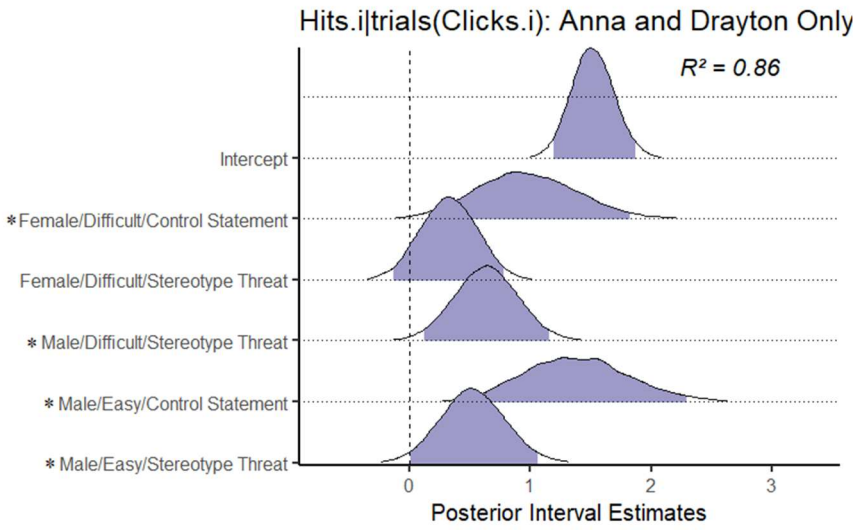



**Figure 3.6** Posterior density estimates for the effect of condition on participants' proportion of successful hits out of their total number of clicks, moderated by total number of clicks. This model represents data from only one female researcher (Anna) and one male researcher (Drayton). Purple fill is truncated to indicate the 95% credible intervals. See Appendix B.10.6 for model code, summary table, and pp_check. * indicates effects whose CIs do not cross zero.

#### 3.4.4.3 SIAS: Continuous vs. Median Split

As part of my experiment, I administered a modified SIAS questionnaire to participants, which asked questions related to their gender identification, gender stigma consciousness, domain identification, domain self-concept, and their negative affect. This questionnaire has been found in the past to indicate susceptibility to stereotype threat, in that higher scores are correlated with a heightened sensitivity to stereotype threat cues. Here I present two ways of examining the relationships between SIAS and explicit stereotype threat as it relates to task performance.

First, we can look at a simple line graph plotting SIAS score against performance, for both the control group and for the explicit stereotype threat group (Figure 3.7). For participants in the control condition, there seems to be a weak positive relationship

between SIAS score and task performance. However, for participants who read the stereotype threat statement, this relationship between SIAS and performance score disappears. This may indicate that participants with high SIAS scores indeed are more susceptible to stereotype threat; high SIAS correlates with lower scores for participants in the ST condition than for those in the control group.

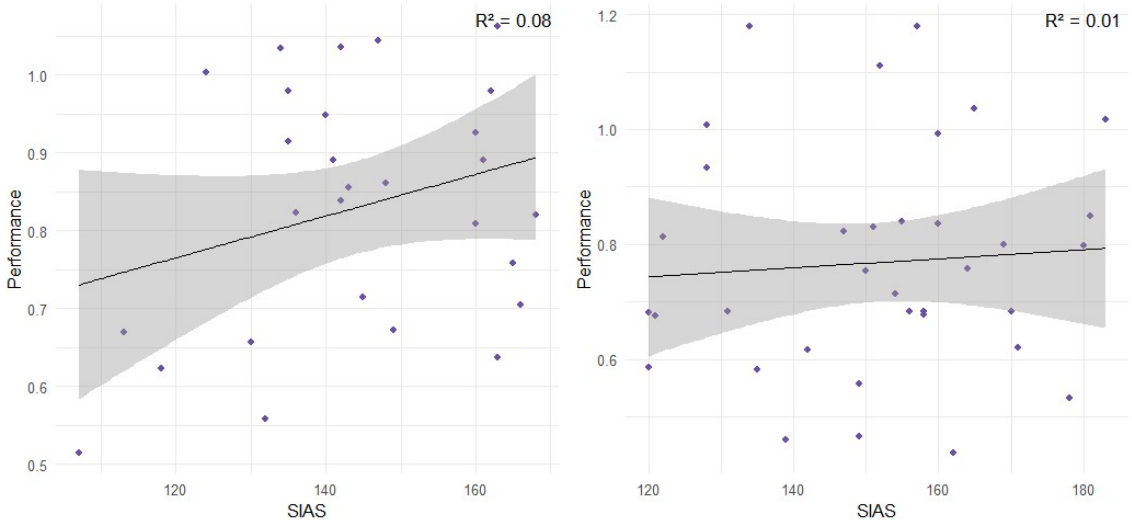



**Figure 3.7** Performance score as a function of SIAS score. Graphs represent participants who read the control statement (left) and the stereotype threat statement (right).

Next, we can explore what happens if we adopt an alternative analytical approach. Here, I split participant responses into low and high scores based on the median SIAS score and then assessed the performance scores for each group (see Figure 3.8). This shows that, for participants in the High SIAS group, explicit stereotype threat reduced performance compared to the control condition. The same pattern emerges in the low SIAS group, but the performance decrease in the stereotype threat condition looks to be more pronounced. This is contradictory to the first set of results; when we perform a

median split on the data, it appears as though low SIAS scores, rather than high ones, cause participants' performance to be more vulnerable to stereotype threat.

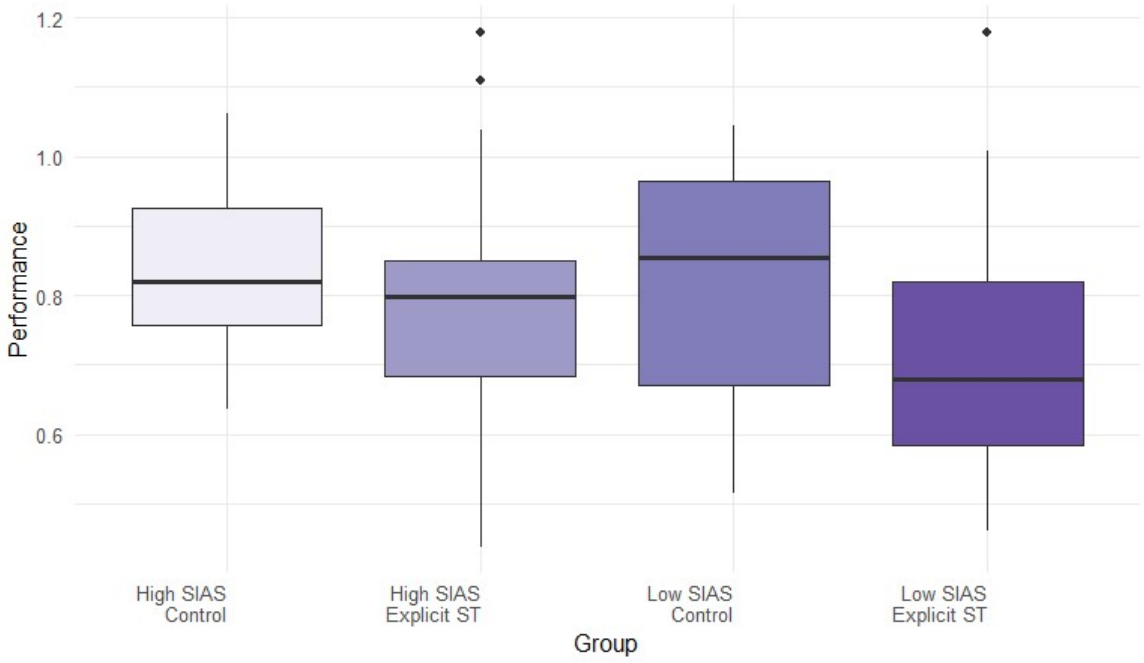



**Figure 3.8** Performance score as a function of low and high SIAS scores, split by stereotype threat condition.

### 3.4.5 Implications of Varying Researcher Degrees of Freedom

Theoretically, there are an endless number of ways I could have analyzed my data to demonstrate the effects of different choices. I selected only a few examples to demonstrate how outcomes could vary, so these results are intended to be illustrative, rather than definitive. That is, while it is reasonable to suggest that a compelling argument could be made for the use of each of the four outcome variables that I constructed, the same could also be said for other variables that I did not consider.

I began with accuracy, as this seemed to be the simplest way to determine a participant's score on the task. However, I think this is the least defensible of the four outcome variables, since it does not take into account the time taken to click on each target. This is also my worst performing model, with an $R^2$ value of 0.25. Next, we have performance, which is the metric that I came up with to determine a participant's score, which accounts for their accuracy while clicking on targets, and also the time it took to click on each target. This seems to help explain the data better, which is reflected by the higher $R^2$ value of 0.51, but this variable caused some issues with model convergence and might not be the best way to quantify a participant's task score. Then I attempted to use the same variable that I used in my pervasiveness analysis, which is the change in performance between the pre-test and the actual test. Looking at the $R^2$ value of 0.38, this did not perform as well in the model as performance on its own, which tells me that the change in performance from the pre-test to the test does not explain the data as well as simply looking at the performance score for the test itself. Last, I used a variable called Hits.i|trials(Clicks.i), which looks at the successful number of hits out of the total number of clicks, (which is simply accuracy), but it also controls for the number of total clicks. This prevents the issue of a participant getting 100% accuracy by simply clicking on one or two targets and doing nothing for the rest of the trial and achieves something similar to my original performance metric. Of all my models, this outcome variable performed the best, with an $R^2$ value of 0.84. Not only do each of these outcome variables result in vastly different $R^2$ values, even with everything else in the model kept constant, they also lead to different and sometimes contrasting interpretations of the data.

Across all four models, the only result that we can be consistently confident about is that participants in the Male/Difficult/Control group decreased their score compared to the Female/Easy/Control group for all metrics of task performance. All other groups vary drastically between the different outcome variables, sometimes displaying a clear effect in one direction, and other times showing no effect at all (i.e., sitting squarely across 0). The point of this demonstration is that, for my study, there is no single right answer to which outcome variable should be chosen. Arguments can be made for or against each of those I chose, so how does the experimenter decide what is the best, most robust option? Even within the stereotype threat and math literature, some papers simply analyze the participants' raw score on the math test, while others deduct points for wrong answers to discourage guessing, and other studies also include the time taken per question as part of the analysis. Thus, even the most basic decision, such as what to measure and how it should be measured, can be ambiguous and can lead researchers to draw different conclusions from their analyses.

However, all of this assumes that the researcher simply chooses the outcome variable that they deem to be the most appropriate, oblivious to the effects of this choice on the results of their study. We must also consider the researcher who uses the ambiguity surrounding the choice of outcome variable to their advantage; that is, by testing all four variables and selecting the one that is most favorable to their hypothesis. As with many choices made throughout a study, there is truly no correct answer for which outcome variable should be selected, and each of the four variables I chose yielded different results. This ambiguity will always exist for scientists, both honest and fraudulent, and there are no catch-all solutions to this problem. Perhaps if preregistration was the norm,

researchers would be forced to choose their outcome variable prior to data collection and statistical analysis, and there should be a requirement that their selection be justified based on theory and not based on the effect of their choice on the results.

I also analyzed how individual identity could influence the results by varying the identity of the experimenters. This is not a watertight demonstration; due to my low sample size, removing the data from all but two researchers leaves me with a vanishingly small number of participants in some of my groups. For example, there were no participants in the Female/Easy/Stereotype Threat condition for the Anna-Drayton subset. If I had only used two researchers in total for my actual study, then each group would have ended up with many more data points. I still consider this to be an interesting comparison to make, but we must be cautious about placing too much stock in these findings.

When pairing Anna and Drayton, I did so because Anna had a comparatively low average participant performance compared to the other female researchers, and Drayton had a relatively high average participant performance. I chose the data from Tyler and Sam since they displayed the opposite pattern. By choosing these two pairings, I wanted to investigate whether I could find opposing results, which is what I accomplished. Data from only Anna and Drayton skews all results towards the positive side of zero, and data from only Sam and Tyler skews the results in a negative direction. Not only does the initial selection of researchers have the potential to affect the outcome of a study, but a dishonest researcher has the opportunity to remove researchers from their analysis whose results do not support the hypothesis, and current standards of reporting would allow for this to occur without alerting even the most thorough reader.

Lastly, I looked at the relationship between SIAS score and task performance for both the control and stereotype threat condition. The manipulation I explored with this variable was to contrast the results with SIAS as a continuous variable with the results I might see if instead I performed a median split on SIAS. Leaving SIAS as a continuous variable, it seems that the predicted relationship between high SIAS scores and stereotype threat susceptibility is present in the data; participants with higher SIAS scores who read the stereotype threat statement experienced a reduced performance compared to higher SIAS scores in the control condition. This may be indicative of an increased susceptibility to stereotype threat as SIAS score rises. If we instead split SIAS into a categorical variable, then the opposite trend seems to appear, with low SIAS individuals in the stereotype threat group experiencing a greater decrease to their performance when compared to high SIAS individuals reading the same statement. One factor that may cloud interpretations of this relationship between SIAS and task performance under stereotype threat is that the SIAS questionnaire was administered after the stereotype threat statement was read and the task was completed. It is possible that reading the stereotype threat activation statement may have an effect on a participants' answers to the SIAS questionnaire, meaning that those who read the stereotype threat statement were more likely to have higher SIAS scores, and thus SIAS score can vary based on the statement a participant reads. In this case we may not be saying the SIAS scores indicate an inherent immutable trait within a person that determines their susceptibility to stereotype threat, but we are instead saying that activating stereotype threat leads participants to attain higher SIAS scores.

## 3.5 Discussion

Comparing the pervasiveness analysis findings to the results of my Bayesian models, an interesting trend arises. The pervasiveness analysis did not reveal the predicted effect of stereotype threat on women's task performance; participants who were given the stereotype threat treatment did not perform any worse than those who read the control statement, with the exception of the Female/Easy condition—although this latter finding is based only a single participant exhibiting this performance decrease, so the finding is not robust. In contrast, the Bayesian models lend themselves to other possible interpretations. For my model using performance as the outcome variable, the only factor causing low performance was the difficulty of the task. However, if we examine the results of the model using accuracy as the outcome variable, we see that the effect of task difficulty was only present when a male researcher administered the task. The performance change model shows that, with a difficult task, the stereotype threat statement reduces performance more than the control statement when a female researcher is present, but the control statement reduces performance more than the stereotype threat statement when a male researcher is present. Finally, looking at the Hits.i|trials(Clicks.i) model, we see that stereotype threat effects are present, but only when a female researcher administers a difficult task, since the Female/Difficult/Stereotype Threat group performed worse than the Female/Difficult/Control group, or when a male researcher administers an easy task, since the Male/Easy/Stereotype Threat group performed worse than the Male/Easy/Control statement. The point here is that since there is no single right answer to which outcome variable a researcher can choose, it is a bit alarming that these different

outcomes, and hence interpretations, can be pulled out of the models, depending on the variables chosen.

I suggest that an analysis of pervasiveness should be performed as the first step in any analysis, to determine if the individuals in the sample exhibit the predicted effect. If the individuals in the sample do not show consistent behavioural differences, then there is no reason to conduct further group-level analyses. My pervasiveness analysis showed that there were no consistent reductions to performance as a result of stereotype threat, but by aggregating the data and using inferential statistics, I was able to produce results that could be used to argue for some ST effects. This is a common problem in psychology; Speelman and McGann (2025) call this the Generalization Delusion, or the inappropriate application of group-level findings to individuals within that population. If the individuals within a sample are not demonstrating the effect, determined by an analysis of pervasiveness, then how can we use that same data to attempt to find evidence of this effect at the group level?

I set out to design a study to test how manipulating study design and analysis could potentially yield different (and possibly contradictory) results. Throughout the process of a typical psychological study, it is likely that a researcher would use the information at their disposal to choose one direction to explore, and let the other potential choices fall by the wayside. Sometimes the researcher might conduct a preliminary exploration of each potential choice, and then decide which one to pursue, based on whether the decision brings them closer to attaining results that align with their hypothesis. There is nothing inherently wrong with the fact that researchers have these degrees of freedom when designing and analyzing their studies, and making these choices

is an unavoidable part of doing science. There would be no point in conducting research if we already knew the right answer. The goal here is not to demonstrate that researcher degrees of freedom are somehow harmful, rather I wanted to draw attention to the fact that seemingly small decisions can produce very different results, so researchers should ensure that they have a logical defense for their choices and aren't simply choosing the path that is most likely to give them the result they are hoping for.

# CHAPTER 4: GENERAL DISCUSSION

In this thesis I aimed to gain insight into some of the ways that normative research practices within psychology can inadvertently contribute to a body of literature that cannot be replicated, and how this has, in part, led to the replicability crisis that psychologists find themselves in today. In Chapter 1, I provided an in-depth overview of the replication crisis, outlining some of the biggest contributors to non-replicability, and discussed a few of the solutions that have been proposed to encourage transparency and rigor within psychological research. Instead of looking at psychology as a whole to examine some of these key causes of non-replicability, I chose to narrow my focus to the stereotype threat (ST) literature, which is of interest due to some controversial and opposed findings within the field. I therefore conducted an analysis of the ST literature to determine if efforts to improve replicability have been successful, and I also explored experimentally how researcher's choices can affect study outcomes. Here, I will summarize my findings, discuss the limitations of my research, and suggest future directions for study.

## 4.1 Evidence of Improvement in the Stereotype Threat Literature

In Chapter 2, I took a close look at studies which tested the effect of stereotype threat on women's math performance. I hypothesized that several events taking place in 2011 would catalyze an improvement to research methodology and reporting that would reveal itself when examining papers published pre- and post-2011. Unfortunately, I was not able to effectively investigate my papers in this manner (see below), but I was able to look at general trends over time. For the majority of variables, year did not have a strong effect in either direction, meaning that more recently published papers were not

significantly better or worse than early papers on the topic in terms of rigor and transparency. Variables that did not exhibit definitive changes within my subsample of the literature from 2001-2023 were as follows: whether the authors explicitly reported the stereotype threat statement used in the study; whether the authors reported the gender of the experimenters who collected data for the study; whether the authors reported controlling for experimenter gender in their analysis; whether the authors reported the mathematics test that they administered to participants in their study; whether the results of the study included the effect size; whether the study was preregistered; whether the experimenters were blind to the participants' condition while collecting data; the size of the effect; and whether or not the study found any effect of stereotype threat on women's math performance. Variables that did appear to improve over time were: average sample size and average study group size were both larger in more recently published papers, and open data was more common. Other problems that I found during my exploration of the literature included some studies dropping participants from the results with no explanation, and several factors that may compromise the control groups of some stereotype threat studies, including the presence of male individuals, and the wording of the control statements. I looked for replication studies within my subsample and was only able to identify partial replications of previous research; no true replication studies were found. This suggests to me that psychologists, at least those who study stereotype threat, have not made any drastic changes to the way that they conduct or report on their research in the past three decades.

## 4.2 An Exploration of Researcher Degrees of Freedom

In Chapter 3, I reported on my own experiment, where I looked at the effects of implicit and explicit stereotype threat on women's performance on a novel click accuracy task. Using these data, I conducted an exploratory analysis to demonstrate the different outcomes that are possible by following multiple branching paths of decision trees through to the end. I was able to show how different effects become more or less certain based on specific choices, which can change our confidence in results, and in some cases, can completely shift interpretations. First, I explored my results by varying the outcome variable used in a simple Bayesian model. The only finding that was consistent between my four outcome variables (accuracy, performance, performance change, and Hits.i|trials(Clicks.i)) was that participants' score was always negatively affected in the Male/Difficult/Control group compared to the Female/Easy/Control group. All other effects vary in their certainty between the four models, landing firmly on one side of zero in some models, and including both negative and positive values in others. The four outcome variables I chose to explore are all more or less viable options, and one could make an argument for or against any one of them. The point here is that there is often a certain degree of ambiguity in the choices that a researcher must make, so it is important to have a strong justification for those choices.

Next, I explored the results that may have been achieved if I had chosen only one male and one female researcher to collect data for my study. I explored this by varying the subset of data used in my Hits.i|trials(Clicks.i) model, selecting two male/female pairs from my larger dataset. By doing this, I was able to generate two model outputs that were diametrically opposed, with one pair of researchers displaying mostly positive scores

compared to the Female/Easy/Control group, and the other pair showing mostly negative results. The point of this analysis was to show that individual variation may play a larger role than variation based on sex, so if a researcher wants to make an honest attempt to control for researcher sex, they must employ more than a single individual from each group.

## 4.1 Limitations

In this thesis, I have discussed the dangers of researcher degrees of freedom ad nauseum, but I will reiterate the point once more. When designing a study, a near infinite number of branching paths lie before a researcher. While some choices may have an obvious correct answer, many decisions remain arbitrary, and it is up to the researcher to use the resources available to them to decide which path to explore. Ultimately, these decisions can be influenced by many factors, not all of which contribute to making the study more robust or reproduceable. Researchers may make decisions to reduce the amount of time spent collecting data, such as choosing to drop certain variables that are difficult or time-consuming to gather data on, or they may use methodologies that sacrifice precision in favor of convenience. They may pursue cheaper methods to fit within a limited budget, when a more expensive option may yield more accurate results. Researchers can also be influenced by the commonly used methodologies of other scientists in their field or laboratory, even though other methods may be more appropriate. Access to individuals well-versed in statistics can also be a factor, as some researchers may be able to drastically improve their statistical rigor by consulting a statistically talented lab-mate. However, even those individuals who have access to a team of helpful geniuses, abundant funding, and all the patience in the world will still

face decisions to which there is no single solution, and no objectively correct answer. The process of conducting research for this thesis revealed to me the extent to which science can be messy, inexact, and objectively unobjective. I encountered a host of limitations while conducting both my quality control analysis and my stereotype threat study, which I will outline below.

### 4.1.1 Literature Review: Limitations

When choosing a topic on which to conduct my quality control analysis, I wanted to choose one with an abundance of published research. Perusing the literature on studies of the effect of stereotype threat on the math performance of women and girls initially yielded 194 papers, which seemed a respectable sample size. However, after encountering issues with access to many of these articles, as well as narrowing my inclusion criteria in order to be better able to compare my chosen articles, I ended up with a sample size of only 50 papers, which was fewer than I had hoped for. I believe that I would have better been able to understand the trends I was interested in with a larger sample size.

Another limitation of my quality control analysis is that I could not effectively separate my papers into pre- and post-2011. Clearly, all papers published in 2010 or earlier would fall into the pre-2011 category. However, if my goal was to determine if the events of 2011 had an effect on the ways that researchers design, conduct, and report on their research, I would need to know exactly when the study was designed, when all of the data was collected, and when the manuscript was written and finalized. This is next to impossible without preregistration, and since only a single paper in my sample was preregistered, I was unable to determine the precise year that the papers dated 2011 and later were designed, conducted, and written up. Since it can take a period of many

months, or even possibly years for a paper to complete the process from conception to publication, I did not have enough certainty to sort my papers into the two categories necessary to determine if 2011 had a particular effect on researcher behavior.

### 4.1.2 Experimental Investigation of ST: Limitations

When designing a laboratory study with multiple researchers, human participants, and various software programs involved, something is sure to go awry. One of the biggest issues that compromised the integrity of my data was a problem that I had with the KovaaK's game software both glitching and crashing. A total of 37 participants experienced an issue during either their pre-test or the main trial of the task. The most common issue was the game seemingly pausing itself, forcing the researcher to have to take control of the computer to resume the trial. Worse, a few participants experienced a full KovaaK's crash. Despite several troubleshooting attempts, the cause of these glitches was never discovered.

A further 27 participants needed to be excluded from the final dataset due to taking an earlier version of the study. Two main differences existed between the two versions; first, I had originally used a 5-point Likert scale rather than the appropriate 7-point Likert scale for my SIAS questionnaire, and second, my original control statement (which I modelled after other papers in the stereotype threat literature) was potentially priming stereotype threat, in that it made mention of gender, and I decided to change it to more neutral wording. Considering the two major data issues mentioned in this section, plus a few minor ones, data from a total of 67 participants was excluded from the study, over half of my original participants.

When randomizing and counterbalancing the three variables that contribute to the participants' overall study condition, I was able to adjust researcher sex and task difficulty adaptively to compensate for imbalances and keep each variable approximately equal throughout the study. However, the version of the stereotype threat statement that the participants were exposed to was randomized by a Qualtrics survey, which was not interfered with throughout the course of data collection. As a result, some of the study groups have an exceedingly small number of participants, and others have a disproportionately large number of participants. For example, the Female/Easy/Stereotype Threat group had only eight participants, whereas Female/Easy/Control had 24 participants, both based on the original sample size of 130.

Another limitation that may hinder my ability to compare the results of my study to other experiments on the topic is the fact that I only collected data from female participants. Collecting data from male undergraduates as well as female would have allowed me to complete a more in-depth analysis to help me fully understand the effects of stereotype threat on women's math performance. A performance decrease with one particular researcher, for example, may currently be attributable to stereotype threat, but with the addition of male participants (who should not be experiencing stereotype threat in the same way), the effect may have proven to be something else entirely.

Another issue that arose with my study was choosing a task performance metric on which to compare participant's scores. Simply using click accuracy was insufficient, since it is possible to get 100% accuracy by successfully clicking on a single target and then doing nothing else for the remainder of the trial. Since the participants were instructed to click on targets as fast as they could while also being as accurate as possible,

I wanted to find some metric of performance that captured both of those measures. Taking the participants' accuracy score and multiplying it by the inverse of their average time to click on a target (since better performance would involve a high accuracy score, but a low time score), I calculated my own measure of performance. However, this resulted in some issues with model convergence, and other performance measures needed to be explored. This issue would have been avoided had I used a more traditional task from the stereotype threat literature, such as mathematics score.

My study procedure had the experimenter greet participants at the laboratory door before showing them to an office where the study was to take place. Since all eight experimenters for my study were either Undergraduate or Graduate students at the University of Lethbridge, sometimes the experimenter was greeting a participant that they knew personally. If other experimenters were available at the lab, they could test the participant instead, to avoid any confounds that a pre-existing participant/experimenter relationship could introduce. This in itself has the potential to cause issues, since the participant is still encountering someone they know moments before participating in a study, which might have an effect on their anxiety. If there were no other experimenters available to test the participant, then this problem may become even more pronounced.

### 4.1.3 Conclusions and Future Directions

When I began the work for this thesis, I wanted to gain a deeper understanding of how the replication crisis came to be, and how psychology researchers have both contributed to, and attempted to solve, issues of non-replicability. I did this by looking at trends in rigor and transparency over time within a subtopic of psychology to summarize common issues, and to quantify trends in how psychologists have attempted to fix these

issues. I found very weak and uncertain effects, which did not provide compelling evidence to indicate that any significant changes have been successfully implemented. I also ran my own experiment, to get more hands-on experience with how these issues may arise, and how difficult it may be to anticipate and address these problems. I found that it was exceedingly difficult to anticipate and avoid the problems that would cause my findings to be unreliable, and it was not overly challenging to manipulate the strength and interpretation of my model outcomes. This does not paint an optimistic picture about the future of psychology. However, Speelman and McGann (2025), in a new book on this topic, suggest several routes for improvement, many of which address the points made throughout this thesis.

First, they acknowledge that the incentives which serve to reward bad behavior in psychological research are ever-present, but nevertheless they are optimistic about psychology's future. Speelman and McGann (2025) posit that open and honest discussion is now a standard practice within psychology, leading to improved methodologies, including better ways of framing research questions, and answering them. They believe that in order to solve the replication crisis, this openness should be encouraged across scientific disciplines, in order to cultivate critical thinking, and to encourage scientists to explore scientific practices that may be more appropriate to the scale at which they are working. This will give scientists a more diverse toolbox with which to conduct their research, and this inter-disciplinary approach can help psychologists better understand the questions they are asking and the methodologies required to actually answer them.

Speelman and McGann (2025) also provide insight into the ways that psychologists can work to strengthen psychological theory so that future research can be

built upon more stable ground. They point out that a rigorous ecosystem of psychology can be built upon a large number of small-scale studies, which can help us better be able to make generalizations; evidence of a phenomenon derived from many small-scale studies can be just as convincing, if not more, of the generalizability of a phenomenon than one large-scale study on the topic. This lends itself well to the analysis of pervasiveness, which addresses the crux of the issue: how many individuals actually show the effect we are looking for, and to what extent. Our current way of doing science, which involves running models and talking about group averages, is not particularly conducive to understanding real-world individual behavior. The analysis of pervasiveness can help psychologists make a clearer connection between the questions they are asking, and the data they will collect to directly answer those questions in terms of real-world relevance.

Speelman and McGann (2025) suggest that we can start by refocusing our energy to psychological behaviors that may seem obvious, and to resist the temptation to study surprising and novel behaviors that may bring glory to the researcher, but may not contribute to an understanding of robust, generalizable behaviors. This suggests, to me, that psychology should pause its pursuit of new and interesting psychological behaviors and instead focus on strengthening our baseline understanding of how individuals actually experience the world and behave within it. Speelman and McGann (2025) are particularly critical of the unjustified aggregation of data and the use of inferential statistics, as they believe that these methods create a fundamental divide between our psychological studies and our understanding of actual human behavior. If there is no clear reason for analyzing aggregated data with inferential statistics, Speelman and McGann argue, then these analyses should not be done. Finally, they warn psychologists against assuming that

human behaviors can be explained by some underlying mechanism that awaits discovery, as this can lead psychologists to make sweeping generalizations that are ultimately unsupported by the current scientific evidence.

I believe that within this thesis, I have demonstrated that efforts to improve the standards of psychological research have been slow, and it is easy to fall victim to the pitfalls that reduce our confidence in psychological findings. However, I believe that psychologists are slowly but steadily working their way towards a more robust body of work, upon which we can continue to build our psychological theories with greater confidence. We psychologists need to make a more concerted effort to understand the questions we want to ask, and design studies that can directly and effectively answer those questions, and we need to hold steadfast against incentives that reward us for laziness, complacency, and bad research practices.

## REFERENCES


Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature (London)*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A. General*, *132*(2), 235–244. https://doi.org/10.2307/2343787

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature (London)*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Beilock, S. L., Jellison, W. A., Rydell, R. J., McConnell, A. R., & Carr, T. H. (2006). On the causal mechanisms of stereotype threat: can skills that don't rely heavily on working memory still be threatened? *Personality and Social Psychology Bulletin*, *32*(8), 1059–1071. https://doi.org/10.1177/0146167206288489

Beilock, S., Rydell, R., & McConnell, A. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal Of Experimental Psychology-General*, *136*(2), 256–276. https://doi.org/10.1037/0096-3445.136.2.256

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bik, E. M., Casadevall, A., & Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, *7*(3). https://doi.org/10.1128/mbio.00809-16

Bird, A. (2021). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, *72*(4), 965–993. https://doi.org/10.1093/bjps/axy051

Brodish, A., & Devine, P. (2009). The role of performance-avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal Of Experimental Social Psychology*, *45*(1), 180–185. https://doi.org/10.1016/j.jesp.2008.08.005

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1). doi:10.18637/jss.v080.i01

Cadinu. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science*. https://doi.org/10.1111/j.0956-7976.2005.01577.x

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology*, *33*(2), 267–285. https://doi.org/10.1002/ejsp.145

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science (American Association for the Advancement of Science)*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Campbell, S. M., & Collaer, M. L. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychology of Women Quarterly*, *33*(4), 437–444. https://doi.org/10.1111/j.1471-6402.2009.01521.x

Chalabaev, A., Brisswalter, J., Radel, R., Coombes, S. A., Easthope, C., & Clément-Guillotin, C. (2013). Can stereotype threat affect motor performance in the absence of explicit monitoring processes? Evidence using a strength task. *Journal of Sport & Exercise Psychology*, *35*(2), 211–215. https://doi.org/10.1123/jsep.35.2.211

Chalabaev, A., Sarrazinr, P., Stone, J., & Cury, F. (2008). Do achievement goals mediate stereotype threat?: An investigation on females' soccer performance. *Journal of Sport & Exercise Psychology*, *30*(2), 143–158. https://doi.org/10.1123/jsep.30.2.143

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology*, *38*(6), 1639–1655. https://doi.org/10.1111/j.1559-1816.2008.00362.x

David, P. A. (2008). The historical origins of "open science": an essay on patronage, reputation and common agency contracting in the scientific revolution. *Capitalism and Society*, *3*(2). https://doi.org/10.2202/1932-0213.1040

Dee, T. S. (2014). Stereotype threat and the student athlete. *Economic Inquiry*, *52*(1), 173–182. https://doi.org/10.1111/ecin.12006

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, *4*(5), e5738–e5738. https://doi.org/10.1371/journal.pone.0005738

Fanelli, D. (2010). "Positive" results increase down the Hierarchy of the Sciences. *PloS One*, *5*(4), e10068–e10068. https://doi.org/10.1371/journal.pone.0010068

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences - PNAS*, *109*(42), 17028–17033. https://doi.org/10.1073/pnas.1212247109

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fisher, R. (1925). *Statistical Methods for Research Workers* (11th ed. rev.). Oliver and Boyd.

Flore, P. C., Mulder, J., & Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(2), 140–174. https://doi.org/10.1080/23743603.2018.1559647

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Gabry, J., & Mahr, T. (2017). bayesplot: Plotting for Bayesian models. *R package version 1.10.0, 1*(0).

Ganley, C., Mingle, L., Ryan, A., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology*, *49*(10), 1886–1897. https://doi.org/10.1037/a0031412

Grieneisen, M. L., & Zhang, M. (2012). A comprehensive survey of retracted articles from the scholarly literature. *PloS One*, *7*(10), e44118–e44118. https://doi.org/10.1371/journal.pone.0044118

Guardabassi, V., & Tomasetto, C. (2018). Does weight stigma reduce working memory? Evidence of stereotype threat susceptibility in adults with obesity. *International Journal of Obesity*, *42*(8), 1500–1507. https://doi.org/10.1038/s41366-018-0121-2

Harkins, S. G. (2006). Mere effort as the mediator of the evaluation-performance relationship. *Journal of Personality and Social Psychology*, *91*(3), 436–455. https://doi.org/10.1037/0022-3514.91.3.436

Harrison, L. A., Stevens, C. M., Monty, A. N., & Coakley, C. A. (2006). The consequences of stereotype threat on the academic performance of White and non-White lower income

college students. *Social Psychology of Education*, *9*(3), 341–357. https://doi.org/10.1007/s11218-005-5456-6

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106–e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hermann, J. M., & Vollmeyer, R. (2016). "Girls should cook, rather than kick!" – Female soccer players under stereotype threat. *Psychology of Sport and Exercise*, *26*, 94–101. https://doi.org/10.1016/j.psychsport.2016.06.010

Inglis, M., & O'Hagan, S. (2022). Stereotype threat, gender and mathematics attainment: A conceptual replication of Stricker & Ward. *PLOS ONE*, *17*(5). https://doi.org/10.1371/journal.pone.0267699

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*(5), 365–371. https://doi.org/10.1111/1467-9280.00272

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jamieson, J. P., & Harkins, S. G. (2007). Mere effort and stereotype threat performance effects. *Journal of Personality and Social Psychology*, *93*(4), 544–564. https://doi.org/10.1037/0022-3514.93.4.544

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology. General*, *137*(4), 691–705. https://doi.org/10.1037/a0013834

Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology*, *77*(2), 323–338. https://doi.org/10.1348/000709906X113662

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kimmelman, J., Mogil, J. S., & Dirnagl, U. (2014). Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biology*, *12*(5), e1001863–e1001863. https://doi.org/10.1371/journal.pbio.1001863

KovaaK's [Video game]. (2018). The Meta Game Inc.

Lamont, R. A., Swift, H. J., & Abrams, D. (2015). A review and meta-analysis of age-based stereotype threat: Negative stereotypes, not facts, do the damage. *Psychology and Aging*, *30*(1), 180–193. https://doi.org/10.1037/a0038586

Laurin, R., Renard-Moulard, M., & Cometti, C. (2022). Stereotype threat effect on a simple motor task: An investigation of the visuo-spatial working memory. *Research Quarterly for Exercise and Sport*, *93*(2), 423–428. https://doi.org/10.1080/02701367.2020.1826391

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*(4), 660–664. https://doi.org/10.1177/1745691616687745

Lu, S. F., Jin, G. Z., Uzzi, B., & Jones, B. (2013). The retraction penalty: Evidence from the web of science. *Scientific Reports*, *3*(1), 3146–3146. https://doi.org/10.1038/srep03146

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*, 523–523. https://doi.org/10.3389/fpsyg.2017.00523

Mauthner, N. S., & Parry, O. (2013). Open access digital data sharing: Principles, policies and practices. *Social Epistemology*, *27*(1), 47–67. https://doi.org/10.1080/02691728.2012.760663

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in r and stan* (2nd ed.). CRC Press. https://doi.org/10.1201/9780429029608

Merton, R. K. (1925). *The normative structure of science* (12th ed. rev.). Oliver and Boyd.

Moe, A., & Pazzaglia, F. (2006). Following the instructions! Effects of gender beliefs in mental rotation. *Learning and Individual Differences*, *16*(4), 369. https://doi.org/10.1016/j.lindif.2007.01.002

Monaghan, T. F., Agudelo, C. W., Rahman, S. N., Wein, A. J., Lazar, J. M., Everaert, K., & Dmochowski, R. R. (2021). Blinding in clinical trials: Seeing the big picture. *Medicina (Kaunas, Lithuania)*, *57*(7), 647. https://doi.org/10.3390/medicina57070647

Mousavi, S. M., Gray, L., Beik, S., & Deshayes, M. (2021). "You kick like a girl!" The effects of gender stereotypes on motor skill learning in young adolescents. *Journal of Sport & Exercise Psychology*, *43*(6), 450–458. https://doi.org/10.1123/jsep.2020-0255

Nahidi, N., Saemi, E., Doustan, M., Aronson, J., & Laurin, R. (2023). The effect of gender stereotype threat and conceptions of ability on motor learning and working memory.

*Journal of Motor Learning and Development*, *11*(2), 338. https://doi.org/10.1123/jmld.2022-0047

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *eLife*, *5*(Journal Article). https://doi.org/10.7554/elife.21451

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science (American Association for the Advancement of Science)*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*(3), 217–243. https://doi.org/10.1080/1047840X.2012.692215

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences - PNAS*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology (Göttingen, Germany)*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

O'Brien, L., & Crandall, C. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality And Social Psychology Bulletin*, *29*(6), 782–789. https://doi.org/10.1177/0146167203029006010

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science (American Association for the Advancement of Science)*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan - A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210–210. https://doi.org/10.1186/s13643-016-0384-4

Pennington, C. R., Litchfield, D., McLatchie, N., & Heim, D. (2019). Stereotype threat may not impact women's inhibitory control or mathematical performance: Providing support for the null hypothesis. *European Journal of Social Psychology*. https://doi.org/10.1002/ejsp.2540

Pennington, C. R., Heim, D., Levy, A. R., & Larkin, D. T. (2016). Twenty years of stereotype threat research: a review of psychological mediators. *PloS One*, *11*(1), e0146487–e0146487. https://doi.org/10.1371/journal.pone.0146487

Picho, K., & Brown, S. W. (2011). Can stereotype threat be measured? A validation of the Social Identities and Attitudes Scale (SIAS). *Journal of Advanced Academics*, *22*(3), 374–411. https://doi.org/10.1177/1932202X1102200302

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology*, *76*(1), 114–128. https://doi.org/10.1037/0022-3514.76.1.114

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews. Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Remus, W. (1978). Strategies for a publish or perish world. *IEEE Transactions on Professional Communication*, *PC-21*(4), 141–143. https://doi.org/10.1109/TPC.1978.6594213

R-Core-Team. (2023). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Rydell. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal Of Personality and Social Psychology*. https://doi.org/10.1037/a0014846

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American-white differences on cognitive tests. *The American Psychologist*, *59*(1), 7–13. https://doi.org/10.1037/0003-066X.59.1.7

Schimmack, U. (2022, February 15). Publication bias in the stereotype threat literature. *Replicability-Index*. https://replicationindex.com/2022/02/15/rr22-stereotype-threat/

Schmader. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal Of Experimental Social Psychology*. https://doi.org/10.1006/jesp.2001.1500

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal Of Personality and Social Psychology*, *85*(3), 440–452. https://doi.org/10.1037/0022-3514.85.3.440

Seitchik, A. E., & Harkins, S. G. (2015). Stereotype threat, mental arithmetic, and the mere effort account. *Journal of Experimental Social Psychology*, *61*, 19–30. https://doi.org/10.1016/j.jesp.2015.06.006

Sharma, M., Sarin, A., Gupta, P., Sachdeva, S., & Desai, A. (2014). Journal impact factor: Its use, significance and limitations. *World Journal of Nuclear Medicine*, *13*(2), 146–146. https://doi.org/10.4103/1450-1147.139151

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–1888. https://doi.org/10.1177/0956797613480366

Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement*, *61*(6), 1040–1057. https://doi.org/10.1177/00131640121971635

Speelman, C. P., & McGann, M. (2020). Statements about the pervasiveness of behavior require data about the pervasiveness of behavior. *Frontiers in Psychology*, *11*, 1–16. https://doi.org/10.3389/fpsyg.2020.594675

Speelman, C. P., & McGann, M. (2025). The Great Psychology Delusion: Missteps, Pitfalls, and How to Make a More Successful Psychological Science [Unpublished manuscript]

Spencer, S., Steele, C., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal Of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*(1), 93–102. https://doi.org/10.1037/a0026617

Stone, J., Lynch, C. I., Sjomeling, M., Darley, J. M., & Insko, C. A. (1999). Stereotype threat effects on Black and white athletic performance. *Journal of Personality and Social Psychology*, *77*(6), 1213–1227. https://doi.org/10.1037/0022-3514.77.6.1213

Stone, J., & McWhinnie, C. (2008). Evidence that blatant versus subtle stereotype threat cues impact performance through dual processes. *Journal of Experimental Social Psychology*, *44*(2), 445–452. https://doi.org/10.1016/j.jesp.2007.02.006

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. Journal of Applied Social Psychology, 34(4), 665–693. https://doi.org/10.1111/j.1559-1816.2004.tb02564.x

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS One*, *10*(8), 1–24. https://doi.org/10.1371/journal.pone.0134826

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*(9), 1132–1139. https://doi.org/10.1111/j.1467-9280.2009.02417.x

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wilke, C. O. (2018). ggridges: ridgeline plots in 'ggplot2'. *R package version 0.5.6, 1,* 483.

# APPENDICES

## Appendix A: Chapter 2

**A.1** Papers Included in My Quality Control Analysis


Beilock, S., Rydell, R., & McConnell, A. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology-General*, *136*(2), 256–276. https://doi.org/10.1037/0096-3445.136.2.256

Bonnot, V., & Croizet, J. (2011). Stereotype threat and stereotype endorsement: Their joint influence on women's math performance. *Revue Internationale De Psychologie Sociale-International Review of Social Psychology*, *24*(2), 105-120. https://shs.cairn.info/journal-revue-internationale-de-psychologie-sociale-2011-2-page-105?lang=en.

Brodish, A., & Devine, P. (2009). The role of performance-avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal of Experimental Social Psychology*, *45*(1), 180–185. https://doi.org/10.1016/j.jesp.2008.08.005

Brown, R., & Pinel, E. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology*, *39*(6), 626–633. https://doi.org/10.1016/S0022-1031(03)00039-8

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology*, *33*(2), 267–285. https://doi.org/10.1002/ejsp.145

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science*, *16*(7), 572–578. https://doi.org/10.1111/j.0956-7976.2005.01577.x

Chalabaev, A., Major, B., Sarrazin, P., & Cury, F. (2012). When avoiding failure improves performance: Stereotype threat and the impact of performance goals. *Motivation and Emotion*, *36*(2), 130–142. https://doi.org/10.1007/s11031-011-9241-x

Dar-Nimrod, I., & Heine, S. J. (2006). Exposure to scientific theories affects women's math performance. *Science*, *314*(5798), 435. https://doi.org/10.1126/science.1131100

Elizaga, R., & Markman, K. (2008). Peers and performance: How in-group and out-group comparisons moderate stereotype threat effects. *Current Psychology*, *27*(4), 290–300. https://doi.org/10.1007/s12144-008-9041-y

Fogliati, V., & Bussey, K. (2013). Stereotype threat reduces motivation to improve: Effects of stereotype threat and feedback on women's intentions to improve mathematical ability. *Psychology of Women Quarterly*, *37*(3), 310–324. https://doi.org/10.1177/0361684313480045

Ford, T., Ferguson, M., Brooks, J., & Hagadone, K. (2004). Coping sense of humor reduces effects of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin*, *30*(5), 643–653. https://doi.org/10.1177/0146167203262851

Gerstenberg, F., Imhoff, R., & Schmitt, M. (2012). 'Women are bad at math, but I'M not, am I?' Fragile mathematical self-concept predicts vulnerability to a stereotype threat effect on mathematical performance. *European Journal of Personality*, *26*(6), 588–599. https://doi.org/10.1002/per.1836

Hutter, R., Davies, L., Sedikides, C., & Conner, M. (2019). Women's stereotype threat-based performance motivation and prepotent inhibitory ability. *British Journal of Social Psychology*, *58*(3), 691–713. https://doi.org/10.1111/bjso.12298

Jamieson, J., & Harkins, S. (2012). Distinguishing between the effects of stereotype priming and stereotype threat on math performance. *Group Processes & Intergroup Relations*, *15*(3), 291–304. https://doi.org/10.1177/1368430211417833

John-Henderson, N., Rheinschmidt, M., & Mendoza-Denton, R. (2015). Cytokine responses and math performance: The role of stereotype threat and anxiety reappraisals. *Journal of Experimental Social Psychology*, *56*, 203–206. https://doi.org/10.1016/j.jesp.2014.10.002

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science*, *16*(3), 175–179. https://doi.org/10.1111/j.0956-7976.2005.00799.x

Jones, P. R. (2011). Reducing the impact of stereotype threat on women's math performance: Are two strategies better than one? *Electronic Journal of Research in Educational Psychology*, *9*(2), 587–616. https://doi.org/10.25115/ejrep.v9i24.1458

Jordano, M., & Touron, D. (2017). Priming performance-related concerns induces task-related mind-wandering. *Consciousness And Cognition*, *55*, 126–135. https://doi.org/10.1016/j.concog.2017.08.002

Josephs, R., Newman, M., Brown, R., & Beer, J. (2003). Status, testosterone, and human intellectual performance: Stereotype threat as status concern. *Psychological Science*, *14*(2), 158–163. https://doi.org/10.1111/1467-9280.t01-1-01435

Krendl, A., Richeson, J., Kelley, W., & Heatherton, T. (2008). The negative consequences of threat: A functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math. *Psychological Science*, *19*(2), 168–175. https://doi.org/10.1111/j.1467-9280.2008.02063.x

Leitner, J., Jones, J., & Hehman, E. (2013). Succeeding in the face of stereotype threat: The adaptive role of engagement regulation. *Personality And Social Psychology Bulletin*, *39*(1), 17–27. https://doi.org/10.1177/0146167212463083

Lesko, A., & Corpus, J. (2006). Discounting the difficult: How high math-identified women respond to stereotype threat. *Sex Roles*, *54*(1), 113–125. https://doi.org/10.1007/s11199-005-8873-2

Mangels, J., Good, C., Whiteman, R., Maniscalco, B., & Dweck, C. (2012). Emotion blocks the path to learning under stereotype threat. *Social Cognitive and Affective Neuroscience*, *7*(2), 230–241. https://doi.org/10.1093/scan/nsq100

McGlone, M. S., & Aronson, J. (2007). Forewarning and forearming stereotype-threatened students. *Communication Education*, *56*(2), 119–133. https://doi.org/10.1080/03634520601158681

O'Brien, L., & Crandall, C. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, *29*(6), 782–789. https://doi.org/10.1177/0146167203029006010

Osborne, J. (2007). Linking stereotype threat and anxiety. *Educational Psychology*, *27*(1), 135–154. https://doi.org/10.1080/01443410601069929

Penner, A. M., & Willer, R. (2011). Stigma and glucose levels: Testing ego depletion and arousal explanations of stereotype threat effects. *Current Research in Social Psychology*, *16*(3), 1-15.

Pennington, C., & Heim, D. (2016). Creating a critical mass eliminates the effects of stereotype threat on women's mathematical performance. *British Journal of Educational Psychology*, *86*(3), 353–368. https://doi.org/10.1111/bjep.12110

Pennington, C., Litchfield, D., McLatchie, N., & Heim, D. (2019). Stereotype threat may not impact women's inhibitory control or mathematical performance: Providing support for the null hypothesis. *European Journal of Social Psychology*, *49*(4), 717–734. https://doi.org/10.1002/ejsp.2540

Perry, S., & Skitka, L. (2009). Making lemonade? Defensive coping style moderates the effect of stereotype threat on women's math test performance. *Journal of Research in Personality*, *43*(5), 918–920. https://doi.org/10.1016/j.jrp.2009.05.013

Quinn, D., & Spencer, S. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues*, *57*(1), 55–71. https://doi.org/10.1111/0022-4537.00201

Rivardo, M. G., Rhodes, M. E., Camaione, T. C., & Legg, J. M. (2011). Stereotype threat leads to reduction in number of math problems in women attempt. *North American Journal of Psychology*, *13*(1), 5–16.

Rosenthal, H., & Crisp, R. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Personality and Social Psychology Bulletin*, *32*(4), 501–511. https://doi.org/10.1177/0146167205281009

Rydell, R., & Boucher, K. (2010). Capitalizing on multiple social identities to prevent stereotype threat: The moderating role of self-esteem. *Personality and Social Psychology Bulletin*, *36*(2), 239–250. https://doi.org/10.1177/0146167209355062

Rydell, R., Loo, K., & Boucher, K. (2014). Stereotype threat and executive functions: Which functions mediate different threat-related outcomes? *Personality and Social Psychology Bulletin*, *40*(3), 377–390. https://doi.org/10.1177/0146167213513475

Rydell, R., McConnell, A., & Beilock, S. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology*, *96*(5), 949–966. https://doi.org/10.1037/a0014846

Rydell, R., Rydell, M., & Boucher, K. (2010). The effect of negative performance stereotypes on learning. *Journal of Personality and Social Psychology*, *99*(6), 883–896. https://doi.org/10.1037/a0021139

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology*, *38*(2), 194–201. https://doi.org/10.1006/jesp.2001.1500

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal Of Personality and Social Psychology*, *85*(3), 440–452. https://doi.org/10.1037/0022-3514.85.3.440

Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles*, *50*(11), 835–850. https://doi.org/10.1023/B:SERS.0000029101.74557.a0

Sebastián-Tirado, A., Félix-Esbri, S., Forn, C., & Sanchis-Segura, C. (2023). Are gender-science stereotypes barriers for women in science, technology, engineering, and mathematics? Exploring when, how, and to whom in an experimentally controlled setting. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1219012

Smith, J., & White, P. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles*, *47*(3), 179–191. https://doi.org/10.1023/A:1021051223441

Stahl, T., Van Laar, C., & Ellemers, N. (2012). The role of prevention focus under stereotype threat: Initial cognitive mobilization is followed by depletion. *Journal of Personality and Social Psychology*, *102*(6), 1239–1251. https://doi.org/10.1037/a0027678

Steinberg, J., Okun, M., & Aiken, L. (2012). Calculus GPA and math identification as moderators of stereotype threat in highly persistent women. *Basic and Applied Social Psychology*, *34*(6), 534–543. https://doi.org/10.1080/01973533.2012.727319

Taylor, C., Lord, C., McIntyre, R., & Paulson, R. (2011). The Hillary Clinton effect: When the same role model inspires or fails to inspire improved performance under stereotype threat. *Group Processes & Intergroup Relations*, *14*(4), 447–459. https://doi.org/10.1177/1368430210382680

Thoman, D., White, P., Yamawaki, N., & Koishi, H. (2008). Variations of gender-math stereotype content affect women's vulnerability to stereotype threat. *Sex Roles*, *58*(9), 702–712. https://doi.org/10.1007/s11199-008-9390-x

Tomasetto, C., & Appoloni, S. (2013). A lesson not to be learned? Understanding stereotype threat does not protect women from stereotype threat. *Social Psychology of Education*, *16*(2), 199–213. https://doi.org/10.1007/s11218-012-9210-6

Van Loo, K., & Rydell, R. (2013). On the experience of feeling powerful: Perceived power moderates the effect of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin*, *39*(3), 387–400. https://doi.org/10.1177/0146167212475320

Wout, D., Danso, H., Jackson, J., & Spencer, S. (2008). The many faces of stereotype threat: Group- and self-threat. *Journal of Experimental Social Psychology*, *44*(3), 792–799. https://doi.org/10.1016/j.jesp.2007.07.005

Wu, S., & Cai, X. (2023). Adding up peer beliefs: Experimental and field evidence on the effect of peer influence on math performance. *Psychological Science*. https://doi.org/10.1177/09567976231180881

**A.2** Quality Control Analysis Variable Descriptions

1. Number of Studies Reported in the Paper
    - Some papers reported on multiple studies, so I recorded how many studies were included in each publication.
2. Study Number Used
    - I recorded data from the first study reported in the paper that met my selection criteria (i.e. undergraduate women taking a mathematics test with explicit stereotype threat).
3. Country
    - The country that the research took place in.
4. Lab Experiment
    - Whether the data were collected in a laboratory (as opposed to classroom or online studies).
5. Mathematics Test
    - Whether the participants took a mathematics test.
6. Undergraduates
    - Whether the participants were recruited from an Undergraduate participant pool.
7. Number of Initial Participants
    - The number of participants that were initially recruited for the study.
8. Number of Female Participants
    - The number of female participants initially recruited for the study.
9. Number of Male Participants
    - The number of male participants initially recruited for the study. This number could be zero if no males were included in the study.
10. Participants Explicitly Thrown Out
    - The number of participants excluded from the results for any reason, explicitly stated by the authors at any point in the paper.
11. Final Sample Size (Calculated)
    - Number of Initial Participants minus Participants Explicitly Thrown Out
12. Final Sample Size (Actual)
    - Number of participants reported in the results of the paper, particularly for the result of the stereotype manipulation on the participant's mathematics score.
    - Could be derived from a table reporting *n* for each study condition. Alternatively, could be roughly estimated through the reporting of the results statistics, by using information such as *F(1, 100)*, or *t(100)* combined with the number of study groups. E.g. A study has 2 study conditions (stereotype threat and control). If the results state *F(1, 100)*, then the sample size should be approximately 102.
13. All Participants Accounted For?
    - Were there any large discrepancies between the calculated sample size and the actual sample size? Discrepancies of 1 were not coded as participants missing, since deriving the sample size from the reported results was sometimes not perfectly accurate.

14. Number of Study Groups
    - The number of conditions that participants were divided into.
15. Average Number of Participants Per Study Group
    - Final Sample Size (Calculated) divided by the Number of Study Groups. I used the calculated sample size rather than the actual sample size reported in the results since several papers did not indicate sample size in their results.
16. Randomized Groups
    - Whether the study mentioned that participants were randomly assigned to condition.
17. Researcher Blind to Condition
    - Whether the researcher was aware of the condition that a given participant was assigned to.
18. Stereotype Prime Method
    - Whether the participant received the stereotype threat instructions via reading (either on paper or on a computer), audio or video recording, or directly from the researcher.
19. Explicit Stereotype Threat
    - Whether there was an explicit activation of stereotype threat (as opposed to implicit).
20. Stereotype Threat Statement Reported
    - Whether they reported the stereotype threat statement explicitly (i.e. did they include the exact wording that they used for each condition), or in general (i.e. they stated the general wording of the stereotype threat statement but not an exact quote).
21. "Men Perform Better"
    - Whether the stereotype threat activation statement indicated that males performed better than females on the given task, as opposed to simply stating that there are undefined gender differences.
22. Experimenter Gender
    - Male (M), Female (F), Male or Female (M or F), or Unreported (U)
23. Mixed Gender Groups
    - Whether participants were tested in mixed gender groups.
24. Males Present
    - Whether any males were present during testing, either in the form of the experimenter, other participants, or confederates.
25. Demographics Questions Before Test (Control Condition)
    - Whether female participants in the control condition were asked any questions prior to testing that may have primed gender, such as being asked to indicate their gender on the front page, or reading statements related to gender.
26. Implicit Stereotype Threat (Control Condition)
    - Whether the participants in the control condition either read a demographics statement prior to testing or took the test in the presence of males.
27. Control Group
    - Whether the study included a control condition.
28. Control Statement Mentioned Gender?

- Whether the control statement said something like “males and females perform equally on this task,” or made any other mention of gender.

29. Appropriate Control Condition?
    - Whether the participants in the control condition were exposed to implicit stereotype threat or read a statement mentioning gender.
30. Reason for Inappropriate Control Group
    - If the control group was inappropriate, what was the reason?
31. Effect Size Reported
    - Whether an effect size was reported for the finding regarding the effect of stereotype threat on women’s mathematics performance.
32. Effect Size
    - If it was reported, what was the effect size?
33. Effect Size Qualifier
    - Whether the effect size was small, medium, or large based on typical conventions for the type of effect size reported (i.e. Cohen’s d).
34. Power Calculation Reported
    - Whether the authors calculated the sample size required to detect an effect.
35. Account for Implicit Stereotype Threat
    - Whether the authors accounted for implicit stereotype threat in their results. This could be done by including the gender of the researcher in their analysis to test for an effect.
36. Raw Stereotype Threat Found
    - Whether they found a simple effect of stereotype threat on women’s mathematics performance.
37. Stereotype Threat Found Only with Interaction
    - If they did not find pure stereotype threat, did they find an effect of stereotype threat with any other variable as an interaction?
38. Interaction
    - If stereotype threat was only found as an interaction with another variable, what was that variable?
39. Replication Study
    - Whether the paper was a replication of another study.
40. Preregistered
    - Whether the study was preregistered before data was collected.
41. Raw Data
    - Whether the authors linked to a website that hosts the raw data.
42. Mathematics Questions Reported
    - Whether the paper reported the exact mathematics test that the participants took.

**A.3** Quality Control Analysis Journals

| Journal | n |
|---|---|
| PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN | 7 |
| JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY | 5 |
| PSYCHOLOGICAL SCIENCE | 5 |
| JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY | 4 |
| SEX ROLES | 4 |
| EUROPEAN JOURNAL OF SOCIAL PSYCHOLOGY | 2 |
| GROUP PROCESSES & INTERGROUP RELATIONS | 2 |
| BASIC AND APPLIED SOCIAL PSYCHOLOGY | 1 |
| BRITISH JOURNAL OF EDUCATIONAL PSYCHOLOGY | 1 |
| BRITISH JOURNAL OF SOCIAL PSYCHOLOGY | 1 |
| COMMUNICATION EDUCATION | 1 |
| CONSCIOUSNESS AND COGNITION | 1 |
| CURRENT PSYCHOLOGY | 1 |
| CURRENT RESEARCH IN SOCIAL PSYCHOLOGY | 1 |
| EDUCATIONAL PSYCHOLOGY | 1 |
| ELECTRONIC JOURNAL OF RESEARCH IN EDUCATIONAL PSYCHOLOGY | 1 |
| EUROPEAN JOURNAL OF PERSONALITY | 1 |
| FRONTIERS IN PSYCHOLOGY | 1 |
| INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY | 1 |
| JOURNAL OF EXPERIMENTAL PSYCHOLOGY-GENERAL | 1 |
| JOURNAL OF RESEARCH IN PERSONALITY | 1 |
| JOURNAL OF SOCIAL ISSUES | 1 |
| MOTIVATION AND EMOTION | 1 |
| NORTH AMERICAN JOURNAL OF PSYCHOLOGY | 1 |
| PSYCHOLOGY OF WOMEN QUARTERLY | 1 |
| SCIENCE | 1 |
| SOCIAL COGNITIVE AND AFFECTIVE NEUROSCIENCE | 1 |
| SOCIAL PSYCHOLOGY OF EDUCATION | 1 |

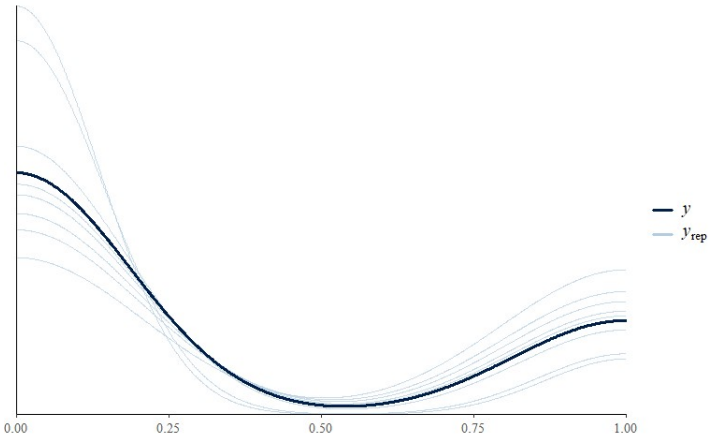
**A.4** Quality Control Analysis Code, Summary Tables, Posterior Predictive Checks, and Conditional Effects

### A.4.1 Model 1: Stereotype Threat Statement Reported ~ Year

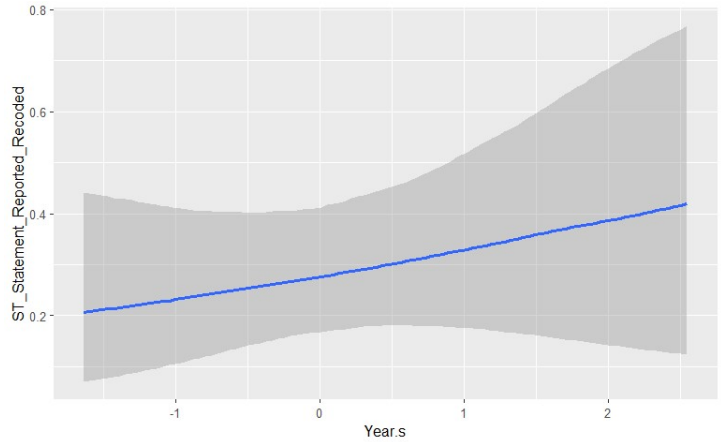
```
Model1 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = ST_Statement_Reported_Recoded ~ Year.s,
     iter = 2000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -0.97 | 0.32 | -1.61 | -0.36 | 1.00 | 2029 | 1860 |
| Year.s | 0.25 | 0.30 | -0.33 | 0.84 | 1.00 | 2194 | 1811 |

Model 1 Population-level effects.




Model 1 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



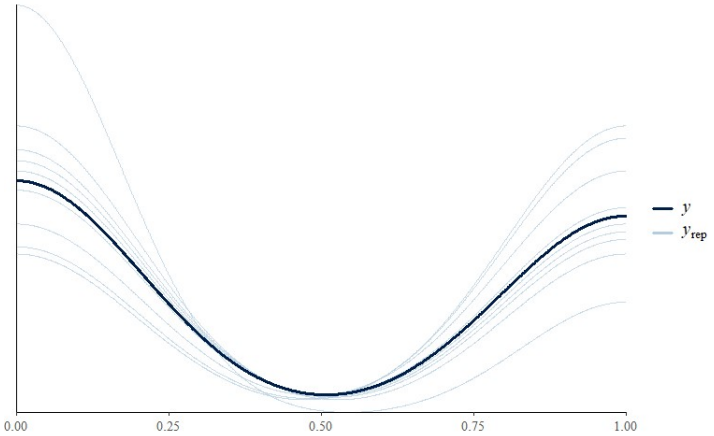

Model 1 Conditional Effects.

### A.4.2 Model 2: Reports Experimenter Gender ~ Year

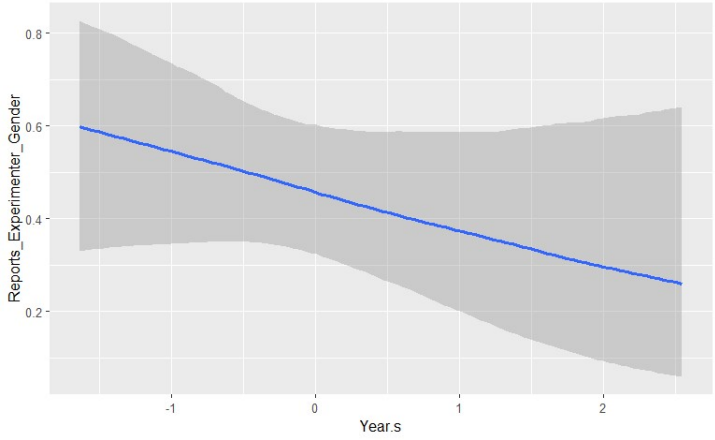
```
Model2 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Reports_Experimenter_Gender ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -0.17 | 0.29 | -0.74 | 0.41 | 1.00 | 1651 | 1369 |
| Year.s | -0.36 | 0.31 | -0.97 | 0.22 | 1.00 | 1906 | 1685 |

Model 2 Population-level effects.




Model 2 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



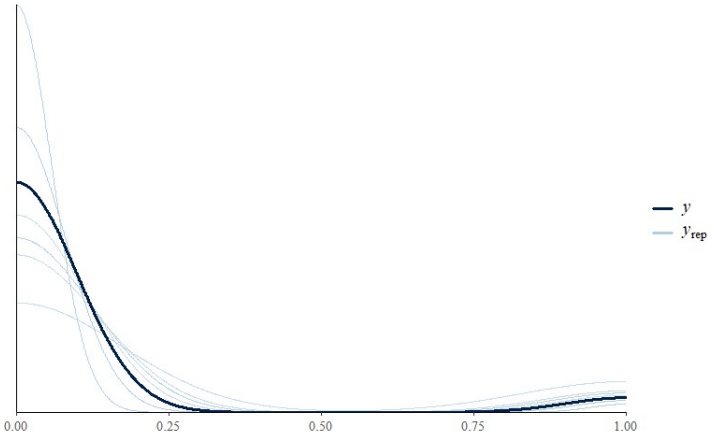

Model 2 Conditional Effects.

### A.4.3 Model 3: Accounts for Experimenter Gender ~ Year

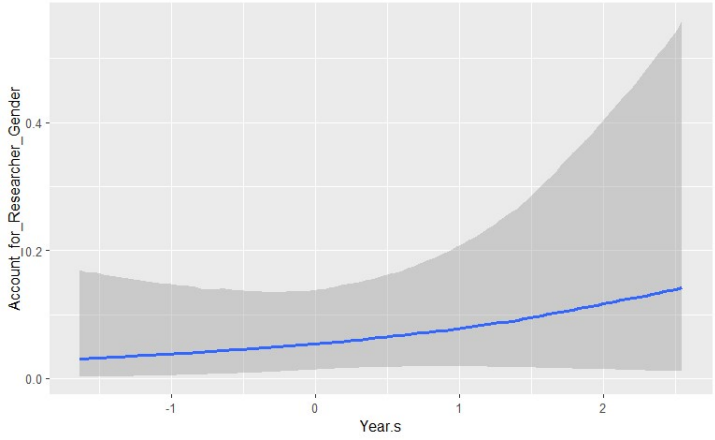
```
Model3 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Account_for_Researcher_Gender ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -2.91 | 0.62 | -4.26 | -1.83 | 1.00 | 1775 | 1956 |
| Year.s | 0.40 | 0.48 | -0.56 | 1.31 | 1.00 | 1698 | 1718 |

Model 3 Population-level effects.




Model 3 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



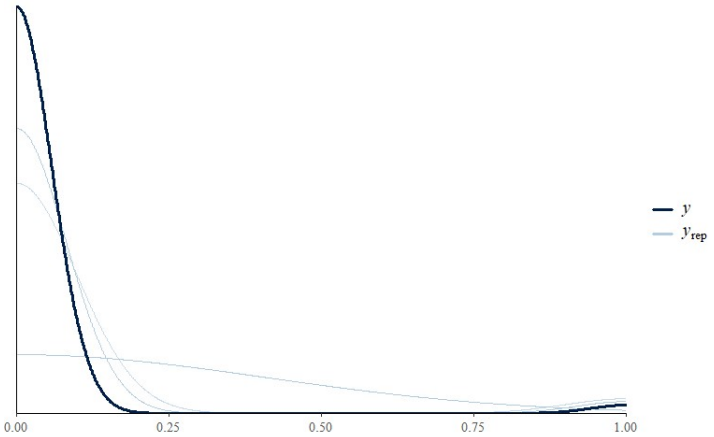

Model 3 Conditional Effects.

### A.4.4 Model 4: Math Questions Reported ~ Year

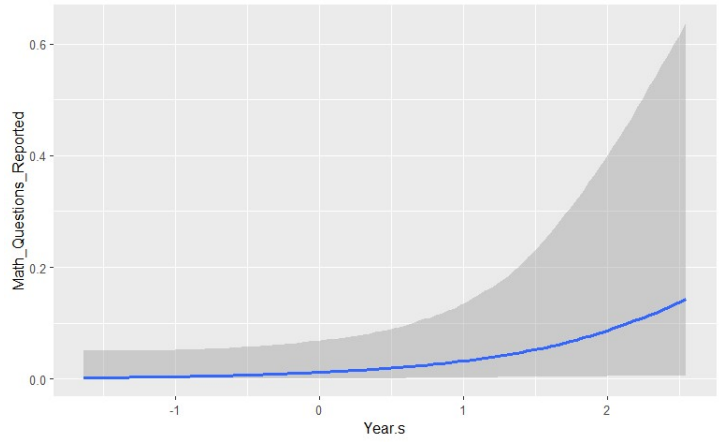
```
Model4 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Math_Questions_Reported ~ Year.s,
     iter = 4000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -4.58 | 1.24 | -7.56 | -2.60 | 1.00 | 2331 | 2147 |
| Year.s | 1.05 | 0.65 | -0.18 | 2.36 | 1.00 | 2467 | 2602 |

Model 4 Population-level effects.




Model 4 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



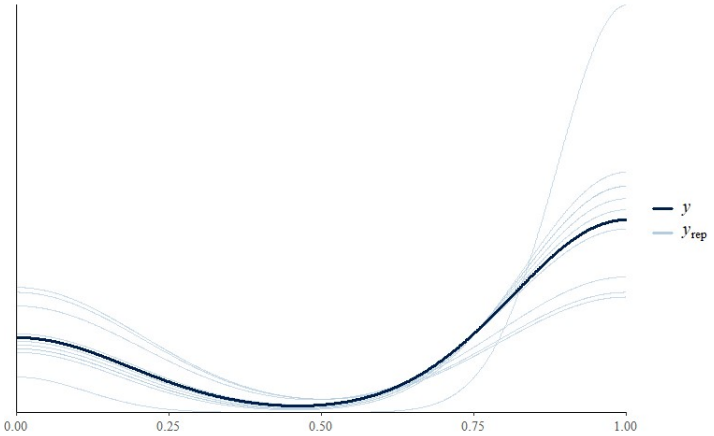

Model 4 Conditional Effects.

### A.4.5 Model 5: Effect Size Reported ~ Year

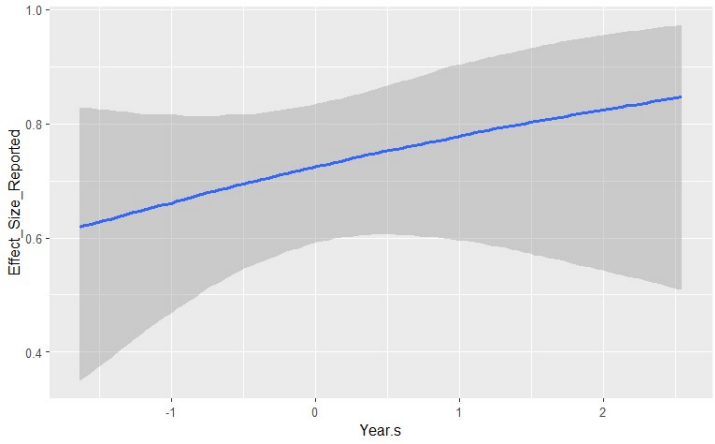
```
Model5 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Effect_Size_Reported ~ Year.s,
     iter = 2000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 0.97 | 0.32 | 0.37 | 1.61 | 1.00 | 2043 | 1803 |
| Year.s | 0.30 | 0.32 | -0.30 | 0.92 | 1.00 | 2169 | 2105 |

Model 5 Population-level effects.




Model 5 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



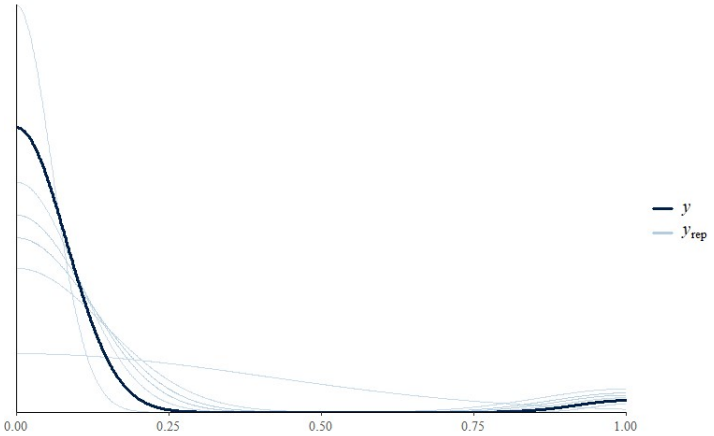

Model 5 Conditional Effects.

### A.4.6 Model 6: Open Data ~ Year

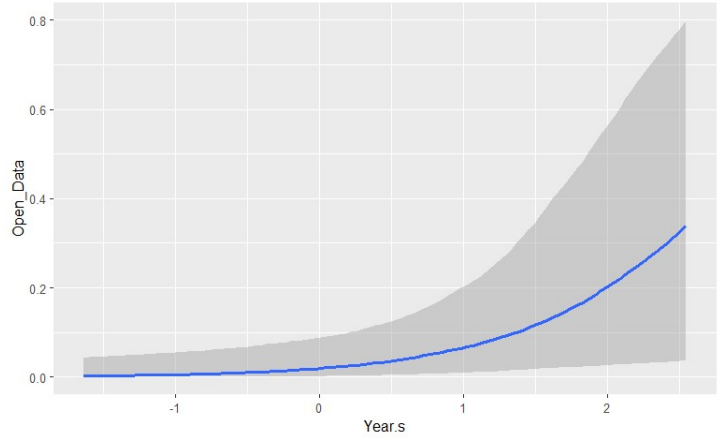
```
Model6 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Open_Data ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -4.04 | 1.03 | -6.38 | -2.34 | 1.00 | 1170 | 1100 |
| Year.s | 1.29 | 0.59 | 0.15 | 2.47 | 1.00 | 1045 | 1286 |

Model 6 Population-level effects.




Model 6 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



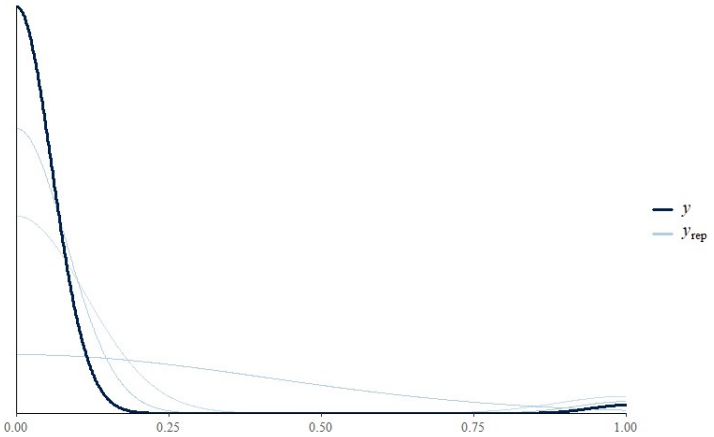

Model 6 Conditional Effects.

### A.4.7 Model 7: Preregistered ~ Year

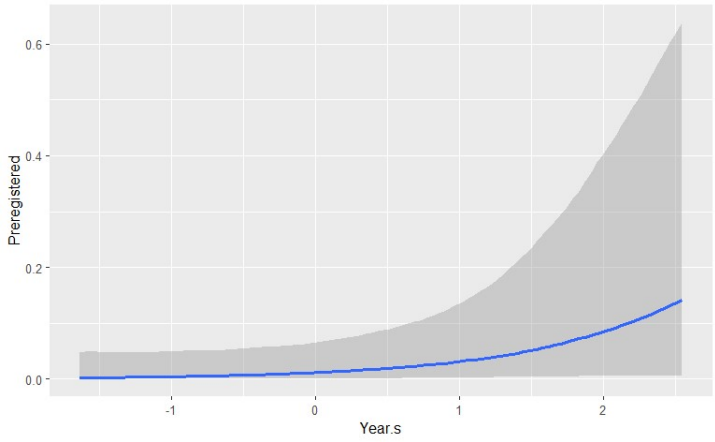
```
Model7 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Preregistered ~ Year.s,
     iter = 4000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -4.59 | 1.22 | -7.41 | -2.66 | 1.00 | 2508 | 2981 |
| Year.s | 1.04 | 0.65 | -0.20 | 2.33 | 1.00 | 2529 | 3215 |

Model 7 Population-level effects.




Model 7 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



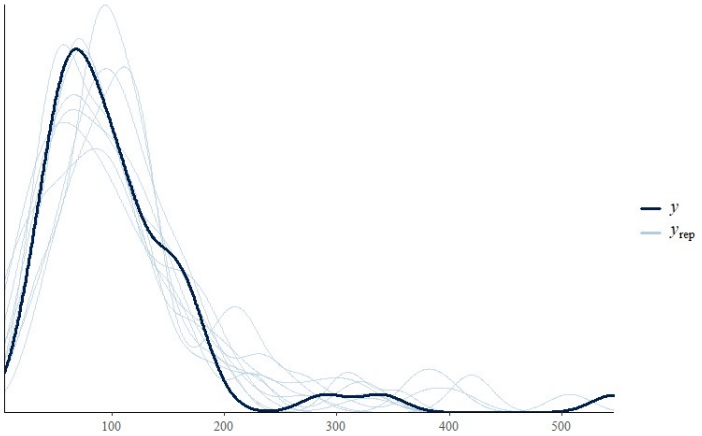

Model 7 Conditional Effects.

### A.4.8 Model 8: Sample Size ~ Year

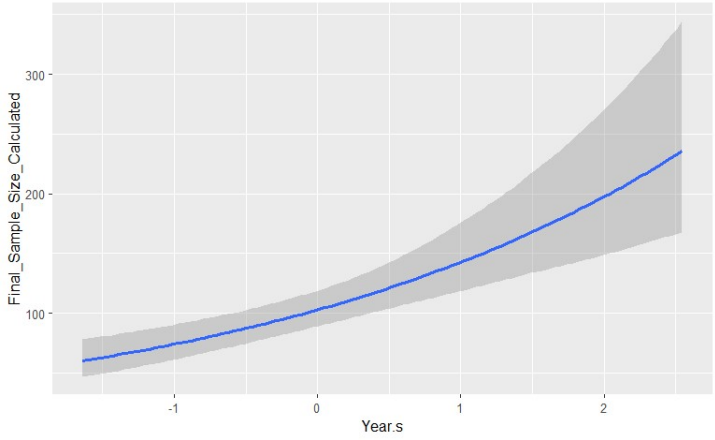
```
Model8 <-
 brm(data = QCAdf,
     family = negbinomial(),
     formula = Final_Sample_Size_Calculated ~ Year.s,
     iter = 2000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 4.63 | 0.07 | 4.49 | 4.78 | 1.00 | 2794 | 2135 |
| Year.s | 0.33 | 0.07 | 0.20 | 0.46 | 1.00 | 2750 | 2062 |

Model 8 Population-level effects.




Model 8 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



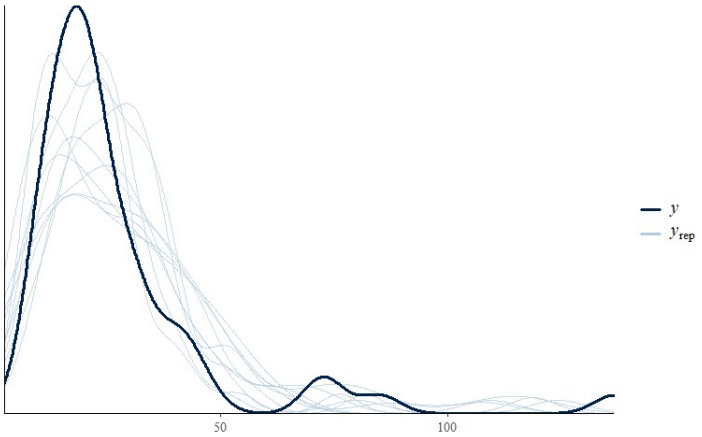

Model 8 Conditional Effects.

### A.4.9 Model 9: Study Group Size ~ Year

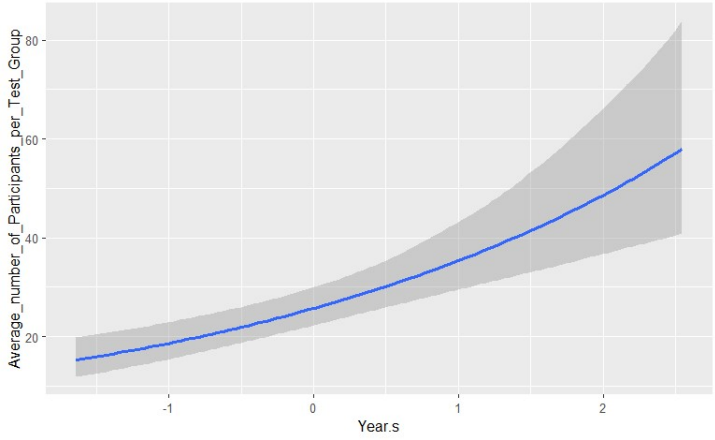
```
Model9 <-
 brm(data = QCAdf,
     family = Gamma(link = "log"),
     formula = Average_number_of_Participants_per_Test_Group ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 3.25 | 0.07 | 3.10 | 3.40 | 1.00 | 2536 | 2081 |
| Year.s | 0.32 | 0.07 | 0.19 | 0.45 | 1.00 | 2825 | 2301 |

Model 9 Population-level effects.




Model 9 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



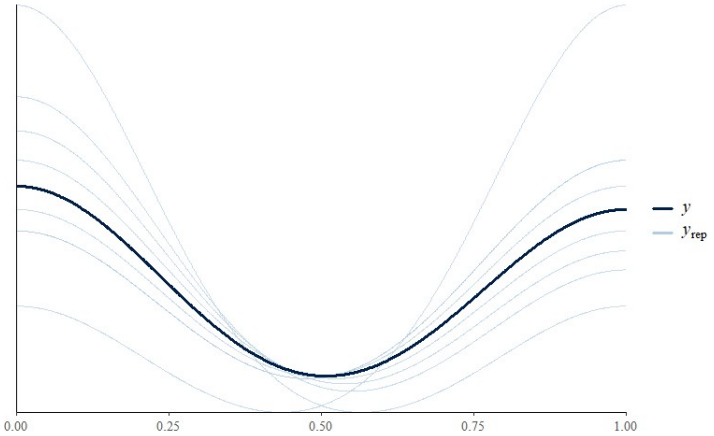

Model 9 Conditional Effects

### A.4.10 Model 10: Experimenter Blindness ~ Year

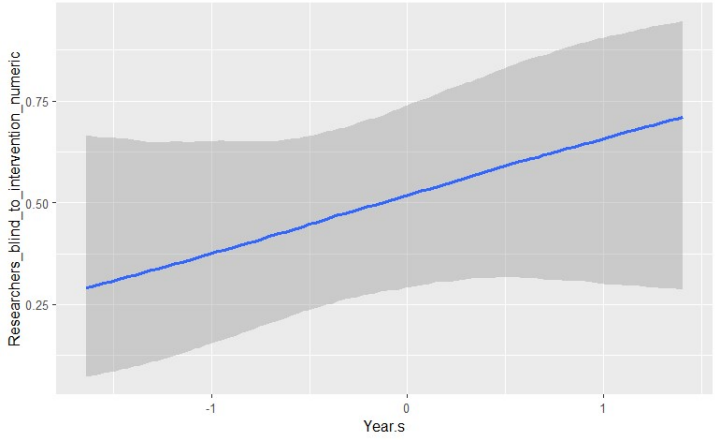
```
Model10 <-
 brm(data = QCAdf.blind,
     family = bernoulli(),
     formula = Researchers_blind_to_intervention_numeric ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 0.08 | 0.49 | -0.88 | 1.04 | 1.00 | 1871 | 1714 |
| Year.s | 0.60 | 0.50 | -0.35 | 1.65 | 1.00 | 2023 | 1950 |

Model 10 Population-level effects.




Model 10 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



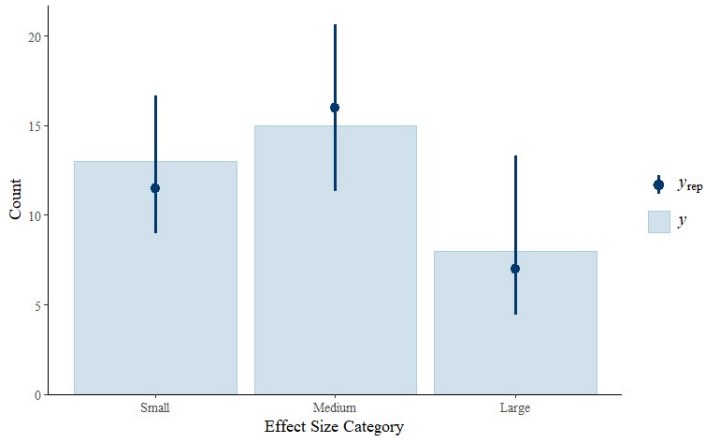

Model 10 Conditional Effects.

### A.4.11 Model 11: Effect Size ~ Year

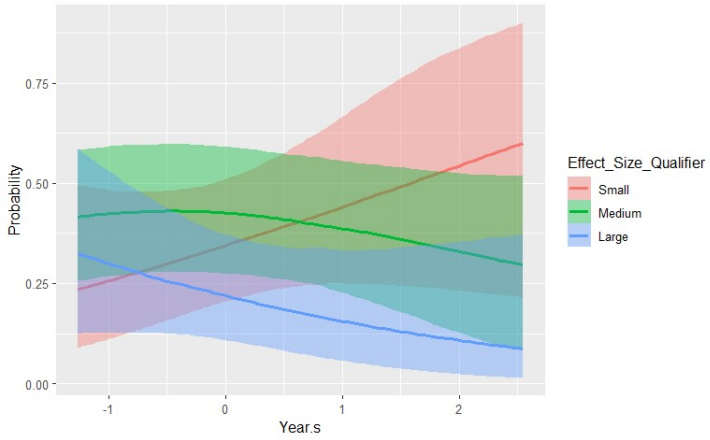
```
Model11 <-
 brm(data = QCAdf.effectsize,
     family = cumulative(link = "logit"),
     formula = Effect_Size_Qualifier ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept [1] | -0.65 | 0.35 | -1.37 | 0.04 | 1.00 | 2292 | 1868 |
| Intercept [2] | 1.30 | 0.41 | 0.54 | 2.13 | 1.00 | 3742 | 2908 |
| Year.s | -0.42 | 0.34 | -1.11 | 0.24 | 1.00 | 2697 | 2399 |

Model 11 Population-level effects. Intercept [1] = log-odds of Small vs. Medium or Large. Intercept [2] = log-odds of Small or Medium vs. Large.




Model 11 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



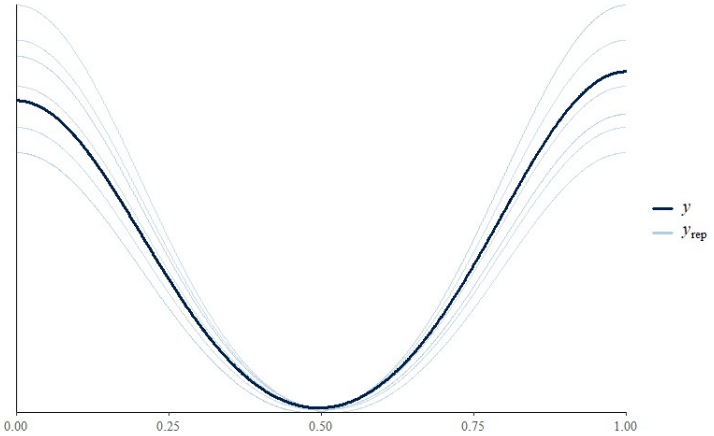

Model 11 Conditional Effects.

### A.4.12 Model 12: Simple Stereotype Threat Effect Found ~ Year

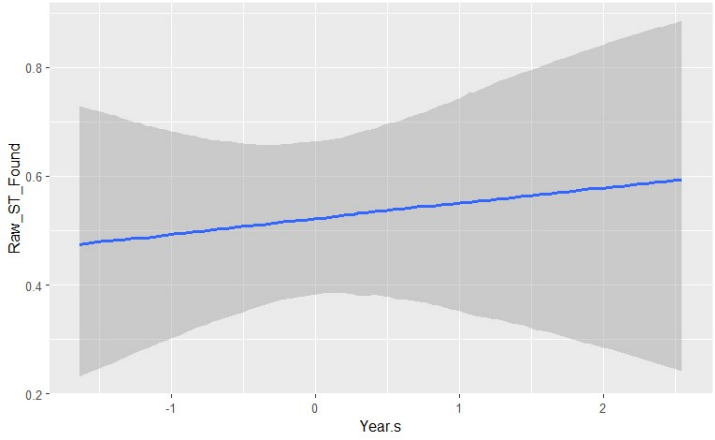
```
Model12 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Raw_ST_Found ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 0.09 | 0.30 | -0.48 | 0.68 | 1.00 | 1707 | 1779 |
| Year.s | 0.12 | 0.29 | -0.44 | 0.71 | 1.00 | 1839 | 1684 |

Model 12 Population-level effects.




Model 12 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



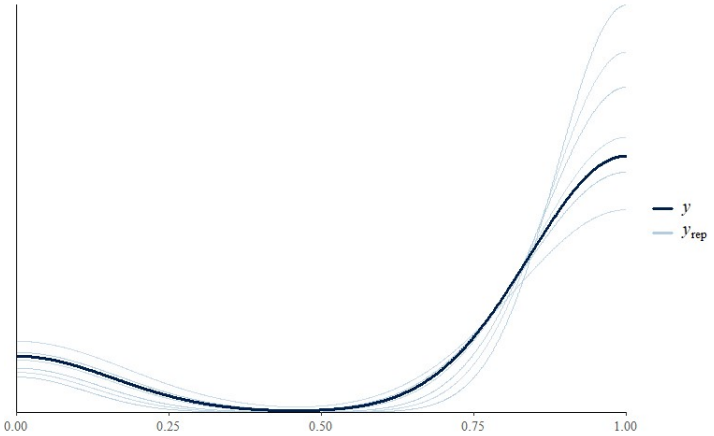

Model 12 Conditional Effects

### A.4.13 Model 13: Any Stereotype Threat Effect Found ~ Year

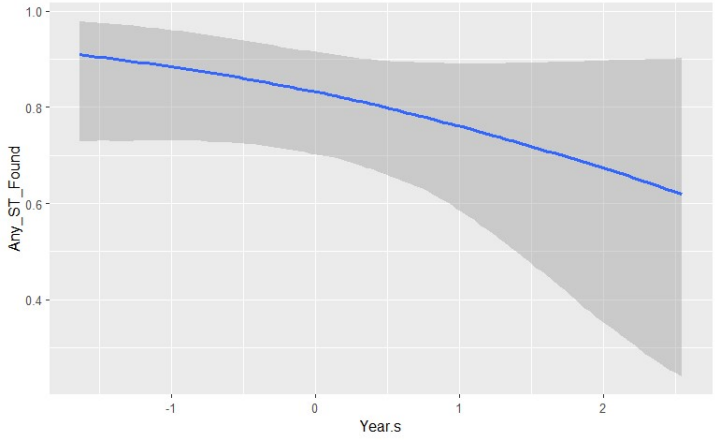
```
Model13 <-
 brm(data = QCAdf,
     family = bernoulli(),
     formula = Any_ST_Found ~ Year.s,
     iter = 2000, chains = 4, cores=4,  prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 1.61 | 0.39 | 0.86 | 2.39 | 1.00 | 1889 | 1894 |
| Year.s | -0.44 | 0.34 | -1.12 | 0.17 | 1.00 | 1801 | 1708 |

Model 13 Population-level effects.




Model 13 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



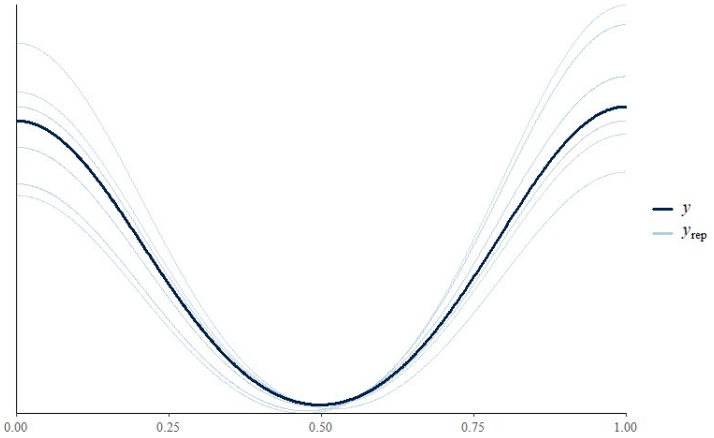

Model 13 Conditional Effects.

### A.4.14 Model 14: Simple Stereotype Threat Effect Found ~ Stereotype Threat Activation Statement States "Men Perform Better"

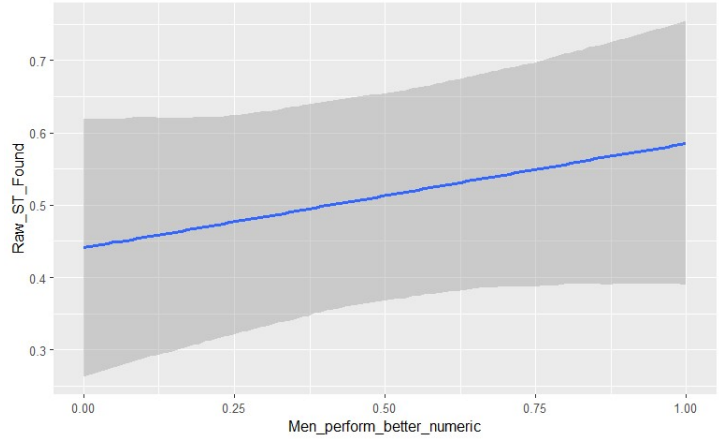
```
Model14 <-
 brm(data = QCAdf2,
     family = bernoulli(),
     formula = Raw_ST_Found ~ Men_perform_better_numeric,
     iter = 2000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ ESS | Tail_ ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -0.25 | 0.39 | -1.03 | 0.48 | 1.00 | 1911 | 1954 |
| Men_perform_better _numeric | 0.59 | 0.52 | -0.40 | 1.63 | 1.00 | 1764 | 1563 |

Model 14 Population-level effects.




Model 14 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.



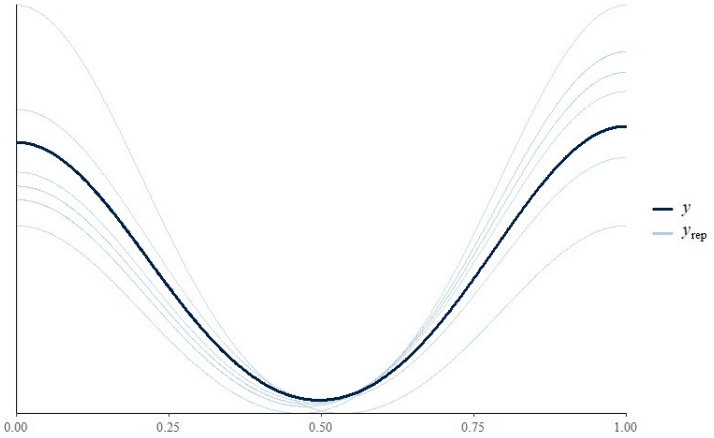

Model 14 Conditional Effects.

### A.4.15 Model 15: Simple Stereotype Threat Effect Found ~ Appropriate Control Group

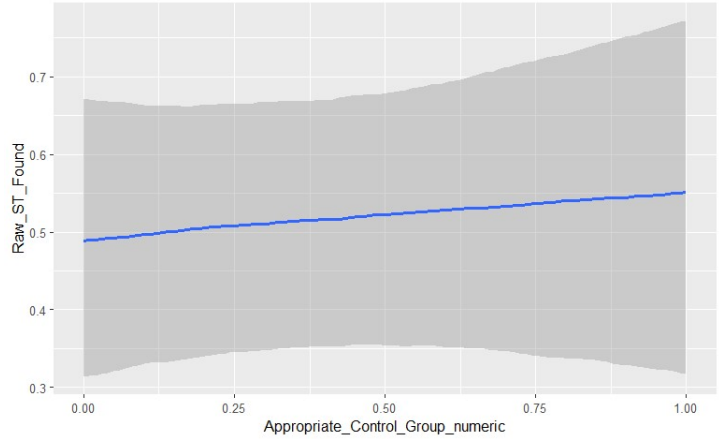
```
Model15 <-
 brm(data = QCAdf3,
     family = bernoulli(),
     formula = Raw_ST_Found ~ Appropriate_Control_Group_numeric,
     iter = 2000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ ESS | Tail_ ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -0.04 | 0.39 | -0.79 | 0.72 | 1.00 | 2119 | 1886 |
| Appropriate_Control _Group_numeric | 0.25 | 0.58 | -0.89 | 1.40 | 1.00 | 2005 | 2064 |

Model 15 Population-level effects.




Model 15 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.

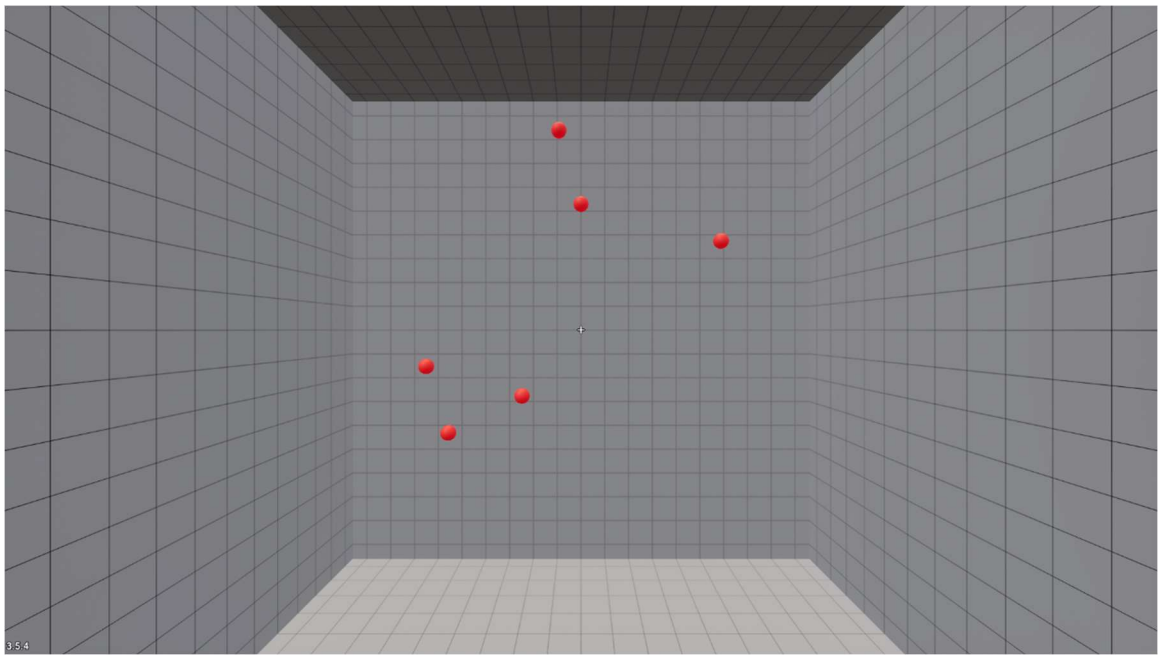



Model 15 Conditional Effects.

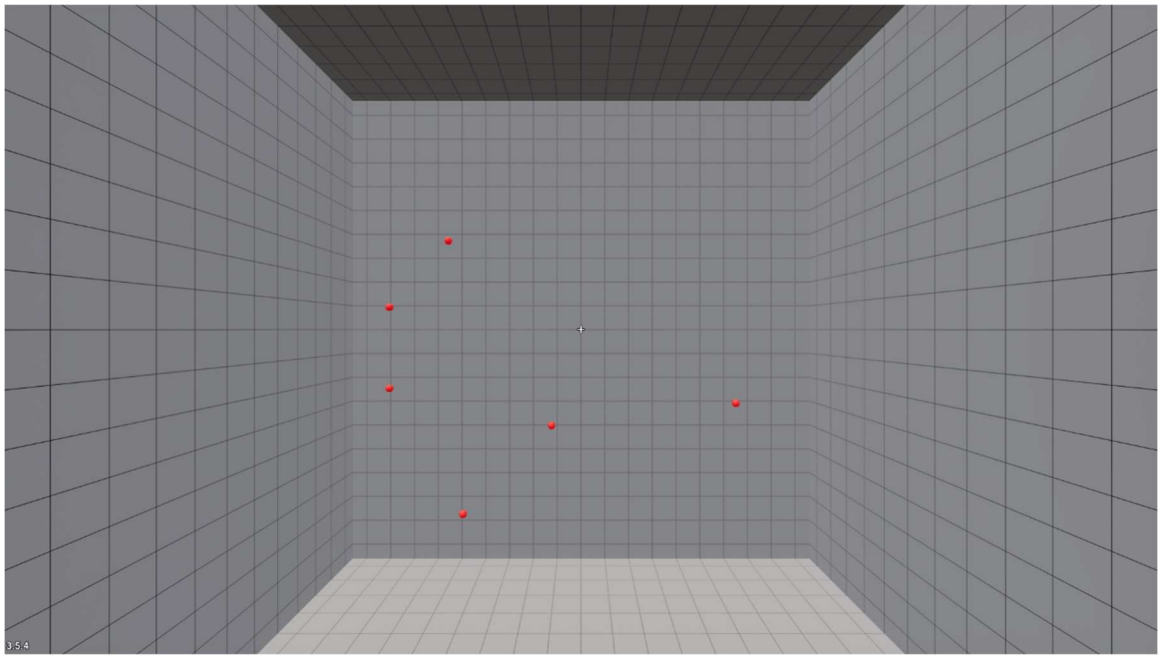
## Appendix B: Chapter 3

**B.1** In-game screenshots of the KovaaK's aim training game.

### B.1.A Easy condition.



### B.1.B Difficult condition.



**B.2** Modified SIAS.

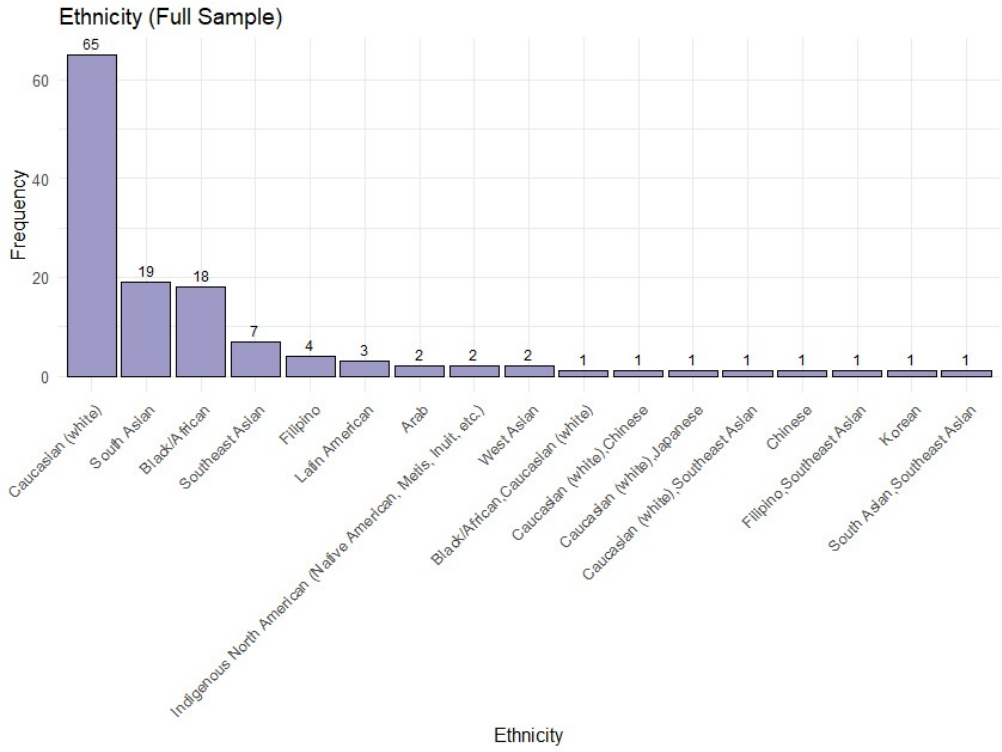
Each question was answered on a 7-point Likert scale, with the following options: Strongly Disagree, Disagree, Slightly Disagree, Neutral, Slightly Agree, Agree, Strongly Agree.

Question categories are as follows: Gender Identification (GI), Gender Stigma Consciousness (GSC), Domain Identification (DI), Domain Self-Concept (DSC), and Negative Affect (NA)

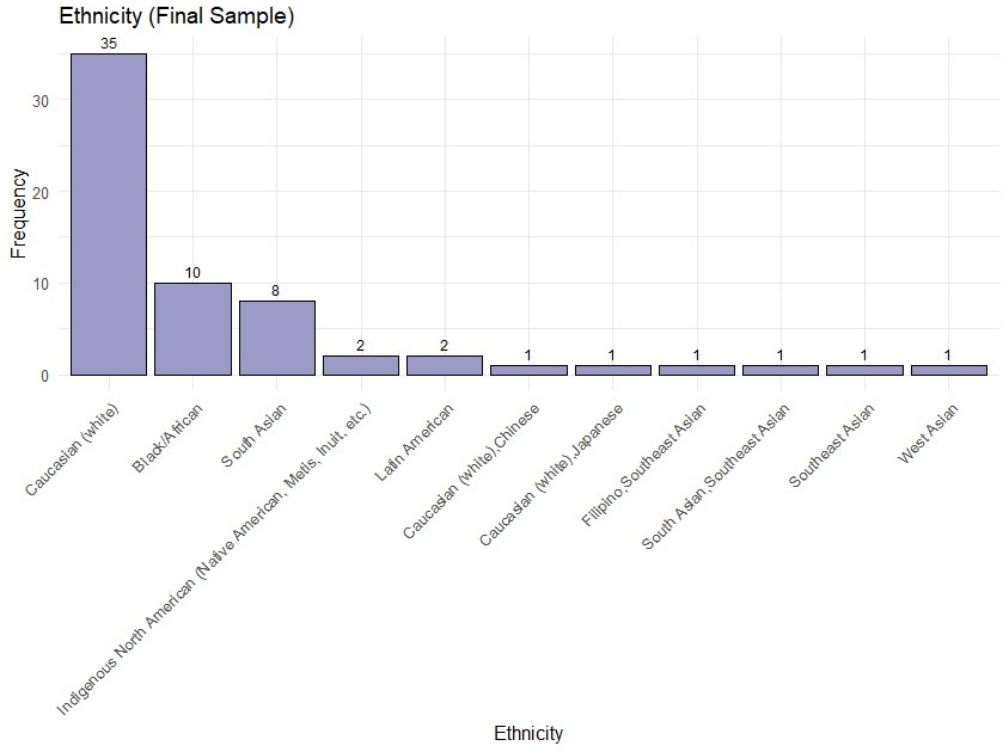
1. My gender influences how I feel about myself. (GI)
2. Being good at competitive tasks is important to me. (DI)
3. My gender contributes to my self-confidence. (GI)
4. My gender influences how teachers interpret my behavior. (GSC)
5. I have always done well at competitive tasks. (DSC)
6. I am good at competitive tasks. (DSC)
7. My gender is central to defining who I am. (GI)
8. Being good at competitive tasks will be useful to me in my future. (DI)
9. Most people judge me on the basis of my gender. (GSC)
10. I learn competitive tasks quickly. (DSC)
11. My identity is strongly tied to my gender. (GI)
12. My gender affects how people treat me. (GSC)
13. I am unable to do well in competitive tasks. (DSC)
14. Stereotypes about my gender bother me. (GSC)
15. I have strong skills at competitive tasks. (DSC)
16. My gender affects how people act towards me. (GSC)
17. My abilities at competitive tasks are important to my success. (DI)
18. Most people have unexpressed sexist thoughts. (GSC)
19. I can easily master difficult competitive tasks. (DSC)
20. Doing well in competitive tasks matters to me. (DI)
21. Members of the opposite sex interpret my behavior based on my gender. (GSC)
22. I value competitive tasks. (DI)
23. I am capable of excelling in at competitive tasks. (DI)
24. Doing well at competitive tasks is critical to my future success. (DI)
25. I experience doubt about my abilities at competitive tasks. (NA)
26. I sometimes experience feelings of frustration. (NA)
27. I sometimes feel like am letting myself down. (NA)
28. I sometimes start to lose confidence in my abilities. (NA)
29. I sometimes feel like a failure. (NA)
30. I sometimes feel hopeless. (NA)
31. I sometimes feel like giving up. (NA)

**B.3** Study sample demographics.

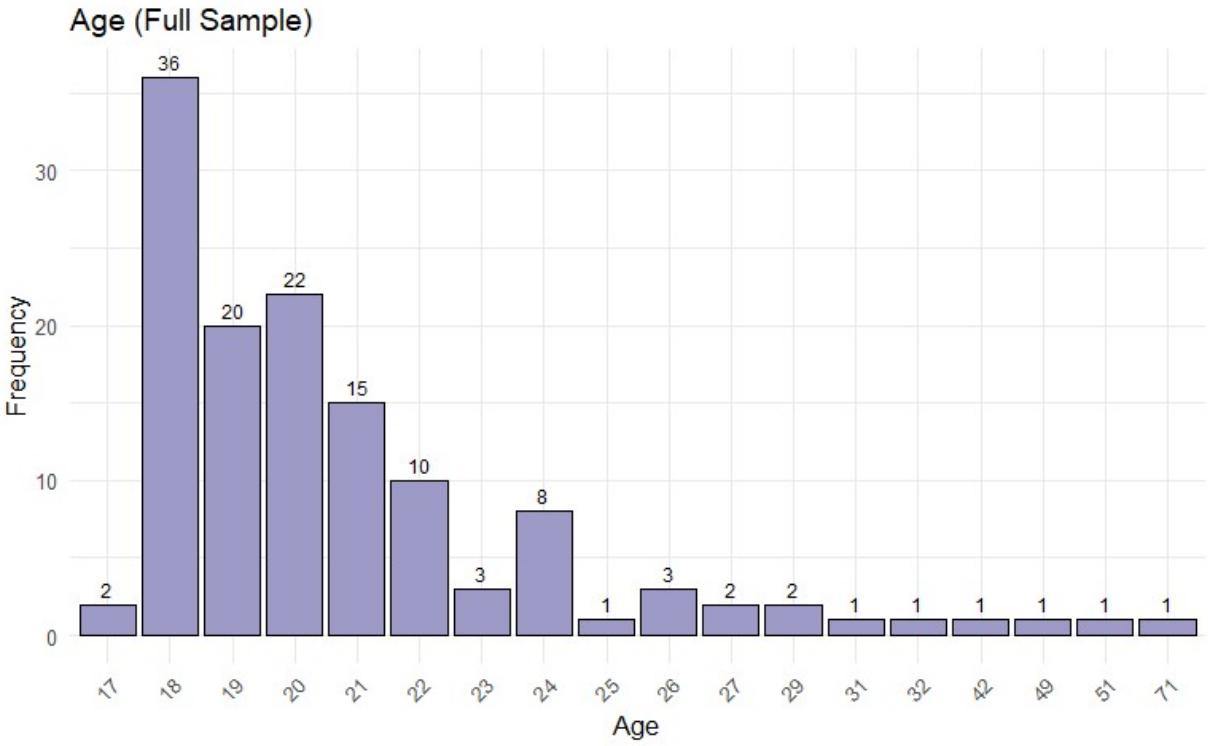
## B.3.A Full sample ethnicity distribution



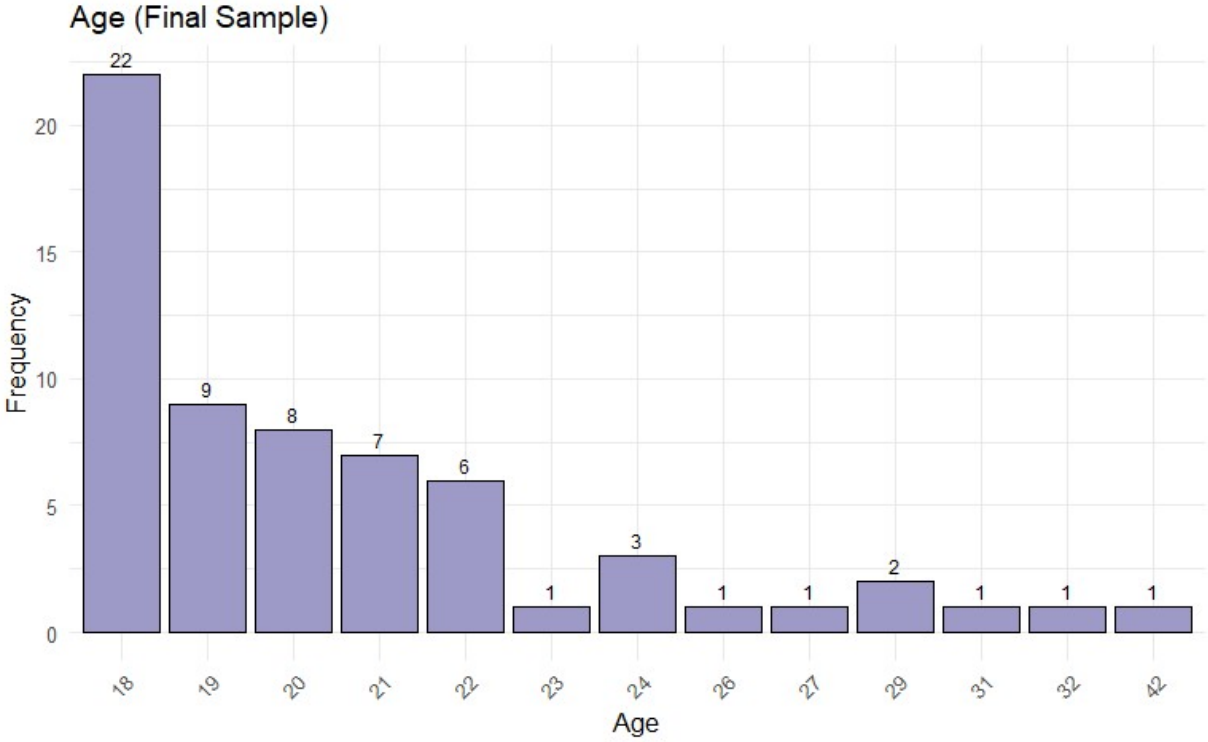

## B.3.B Final sample ethnicity distribution



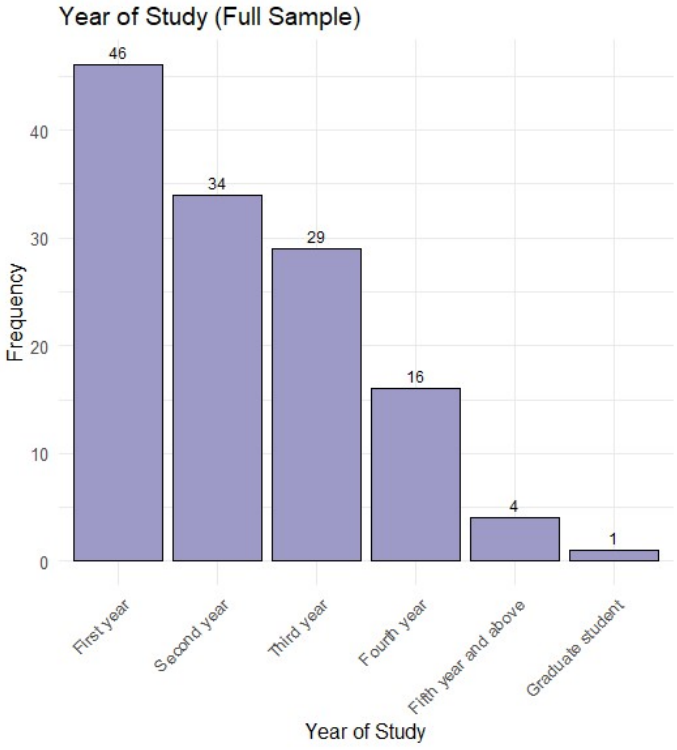
### B.3.C Full sample age distribution



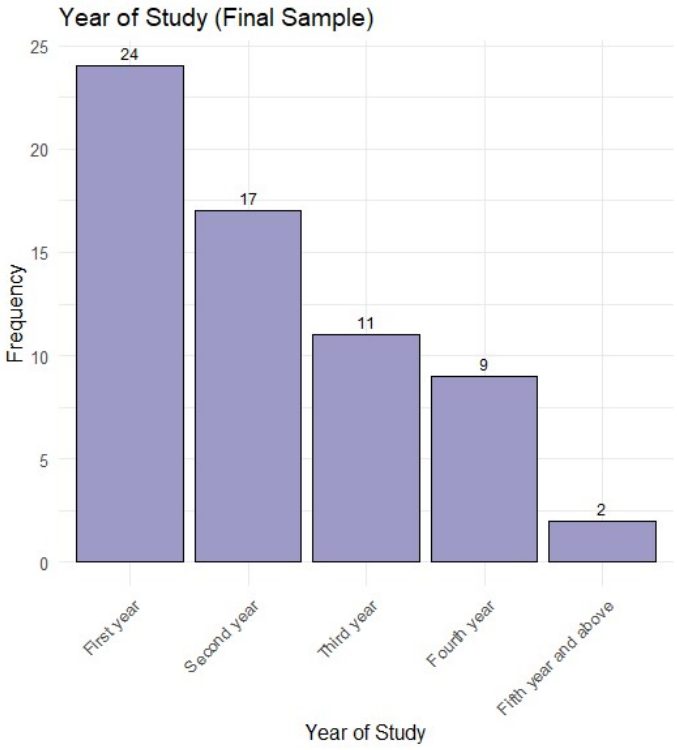

### B.3.D Final sample age distribution



## B.3.E Full sample year of study distribution




## B.3.F Final sample year of study distribution



## **B.4** Study Recruitment Materials

| | |
|---|---|
| Study Name | A Simple Test of Hand-Eye Coordination and Cognitive Awareness |
| Study Type | This study is an in-person lab study. To participate: sign up, then come to the assigned room at the time you register for. |
| Credits | 1 credit |
| Duration | 30 minutes |
| Abstract | The purpose of this study is to examine hand-eye coordination, by testing speed and accuracy in a short computer task. |
| Description | You are invited to participate in a study examining hand-eye coordination and awareness.<br><br>As a participant in this study, you will be asked to come into the lab to take part in a short computer task designed to test click speed and accuracy.<br><br>If you consent to participate, the researcher will begin with three short (30-second) click accuracy tests. These tests are meant to provide a baseline of abilities related to this task, as well as provide an opportunity to familiarize yourself with the task. You will then be given an opportunity to complete the task once more in a true test of task performance. After this task, you will fill out a short questionnaire.<br><br>If you wish to participate in this study, please sign up, and then come to the assigned room at the time you register for.<br><br>Please note, if you are in more than one Psychology course that offers credits, you must login & assign this study to the course you want your credits to go to. There will be no transferring of credits. Any questions regarding SONA, please contact psychology@uleth.ca |
| Eligibility Requirements | English-speaking, non-male students at the University of Lethbridge. 18 years or older. |
| Researcher | Sam Booth |
| Principal Investigator | Louise Barrett |

**B.5** Study Debriefing Document

**Real Study Title: Stereotype Threat and the Replication Crisis: Comparing the Results of Study Manipulations**

Thank you for your participation in this research study. For this study, it was important that we withheld some information from you and provided you with incorrect information about some aspects of the study. You can now be fully debriefed about the purpose of your participation in this research, and you will now have a chance to ask any questions you may have.

**Due to the nature of this study, and since this study will also be running in future semesters, it is very important that you do not divulge any information about the true nature of this study to any individuals who may be in the current or future participant pool (i.e. other students at the University of Lethbridge).**

The purpose of my research was to demonstrate how poor study design can lead to false-positive results in studies, which in turn leads to irreproducible studies being published in the literature. In other words, I wanted to design a study where I manipulated specific aspects to show the different outcomes of these manipulations. I chose to investigate this through the lens of stereotype threat, which is the phenomenon whereby the fear of conforming to negative stereotypes about one's group may hinder task performance (Steele & Aronson, 1995). Stereotype threat has recently come under scrutiny for its inconsistent results, showing a seemingly strong effect in certain studies (Spencer, Steele, & Quinn, 1999), but being notably absent from the results of others (Pennington et al., 2019). With this in mind, I designed a study to manipulate the gender of the researcher and the difficulty of the given task to test the outcomes of these manipulations. I also measured a baseline skill at the task, as well as collected information about an individual's susceptibility to stereotype threat, to see if other factors could better explain my results.

Since I was measuring the effects of stereotype threat, I could not alert participants to the fact that this was what I was studying, as this would result in both the experimental group and the control group being primed for stereotype threat, invalidating my results.

**Did you know anything about the true nature of this study before coming into the lab today?**

- No
- Yes:
  - If you answered yes, what did you know?

    _____
    _____
    _____
    _

**Consent to Submit Data**

Due to the anonymous nature of this study, you will only be able to withdraw your data up to two weeks after the semester ends (i.e. two weeks after April 30th, 2024). If you wish to withdraw your data after today, but before two weeks after April 30th, 2024, you may email me at booths@uleth.ca with your SONA ID number and your data will be withdrawn.

If you do not wish to have your data used in this research, you may alert the researcher now and your responses will not be included. You will still receive course credit for your participation in this study, whether you choose to withdraw your data or not.

If you wish to have your data included in our results, please check YES and include your SONA ID below.

Do you consent to having your data included in our results?

- YES
- NO

SONA ID _____

The plan for this study has been reviewed for its adherence to ethical guidelines and approved by Research Ethics Board 2 at the University of Alberta (Pro00139620). For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at reoffice@ualberta.ca.

**B.6** Study Statements

### B.6.A Control Statement

"The task you are about to complete is a test of hand-eye coordination. There has been some controversy about whether individuals exhibit variation in performance on this task. The task in which you are about to participate has demonstrated differences between individuals. By entering your SONA ID below, you agree that you have read and understood this statement. Click the next button when you are finished."

### B.6.B Stereotype Threat Statement

"The task you are about to complete is a test of hand-eye coordination. There has been some controversy about whether there are gender differences on this task. Previous research has demonstrated that women perform worse on the task in which you are about to participate. By entering your SONA ID below, you agree that you have read and understood this statement. Click the next button when you are finished."

**B.7** Study Researcher Script

**When the participant arrives:**

1. Greet them and instruct them to sit in the chair in front of the monitor.
2. Give them the Informed Consent form and ask them to fill out their SONA ID on the top of the form, which they can find in their email. (We could also just get their SONA ID from the Participant Schedule, but I want to ensure that we're giving credit to the correct person, so if they write down their SONA ID, we can make sure they match.)
3. Instruct them to read and complete the Informed Consent form.
    - If they raise any concerns about the deception aspect of the study, reassure them that it is standard within psychology. They will get all the information about the study's true nature at the end, where they will have a chance to withdraw their data.
4. Ensure the participant has the mouse pointer on their monitor before they start the task.
5. Instruct the participant:
    a) "Now you will complete a competitive task that measures your speed and accuracy when clicking.
    b) First, to familiarize yourself with the task, you will complete a 30-second-long version of this task, three times in a row.
    c) You will be scored on the number of targets you click on, *and* how fast you click on them, so try to be as fast as you can, while also being as accurate as possible.
    d) In between each of these 30-second-long tests, you will have to click the green "Next" button on the bottom right of the post-trial screen to move on to the next task.
    e) Once you have completed three short trials, you will then read a short statement, which I will show you on the computer.
    f) Once you have acknowledged that you have read and understood the statement, you will complete a 90-second-long version of the task you just did. Again, remember that speed and accuracy are being measured.
    g) When you are ready, you can begin by clicking the green play button in the bottom right corner. Once you click the play button, you will have 5 seconds, and then the task will begin. This 5 second timer will appear at the start of each of the trials.
    h) Do you have any questions? You may begin when you are ready."
6. Allow the participant to do three 30-second long click accuracy tests.
7. Then, you will navigate to Google Chrome for them (you can use the mouse or the laptop track pad.) There they will read a statement. Make sure they input their SONA ID so I can keep track of which statement they saw, and make sure they click next to submit it.
8. Once they have done that, you can navigate back to KovaaK's and click on whatever group they are in (check the Participant Schedule if you aren't sure.) Do the same thing as before and click on the name of the playlist (i.e., "Group A").

- "Now you will complete a 90-second-long version of the same task you completed before. When you're ready, you can click the green play button in the bottom right corner of the screen."

9. Once testing is complete:
    - Navigate back to Google Chrome and open the other desktop shortcut that you opened, called "Questionnaire". Instruct them to complete all questions (there are 3 pages and once they submit, it will say "We thank you for your time spent taking this survey. Your response has been recorded." Make sure they get to this screen, so their response is submitted)
    - Give the participant the Debriefing Document and make sure they fill it out.

**B.8** Study Computer Usage Questionnaire

1. SONA ID

_____

2. Participant #

_____

3. How many hours a week do you spend using a computer (laptop, desktop, tablet, etc.) (NOT INCLUDING smartphone use)?
O zero to ten
O eleven to twenty
O twenty-one to thirty
O thirty-one to forty
O forty-one to fifty
O more than fifty

4. When using a computer, what do you primarily use to navigate? If you use multiple inputs equally, check all that apply.
▢ A mouse
▢ A trackpad
▢ A touchscreen
▢ A stylus
▢ Other: _____

5. Have you ever played a video game?
O Yes
O No

6. Do you play video games recreationally (for fun)?
O Yes
O No

7. If you play video games recreationally, how many hours a week do you spend playing video games?
O I don't play video games recreationally
O zero
O one to five
O six to ten
O eleven to fifteen
O sixteen to twenty
O more than twenty

8. If you play video games recreationally, do you typically play on:
O I don't play video games recreationally

O PC
O Console
O Both

9. If you play video games recreationally on PC, do you typically use a controller, or a mouse and keyboard?
O I don't play video games recreationally on PC
O Controller
O Mouse and keyboard
O I use both equally

10. If you play video games recreationally, what genre of games do you typically play? (Check all that apply)
▢ I don't play video games recreationally
▢ First-person shooters
▢ Platformers
▢ Sports games
▢ MMOs
▢ Fighting games
▢ Puzzle games
▢ Racing games
▢ RPGs
▢ Other (please specify) _____

11. Have you ever played a first-person shooter game?
O Yes
O No

12. Do you play first-person shooter games recreationally?
O Yes
O No

13. If you play video games recreationally, list the 3 games you play most frequently:
O I don't play video games recreationally
O The video games that I play most frequently are:
_____

**B.9** Study Demographics Questionnaire

1. What is your current age?

_____

2. What is your current year of study?
O First year
O Second year
O Third year
O Fourth year
O Fifth year and above
O Graduate student

3. Which ethnicity do you most identify with? Please choose all that apply.
▢ Arab
▢ Black/African
▢ Caucasian (white)
▢ Chinese
▢ Filipino
▢ Indigenous North American (Native American, Metis, Inuit, etc.)
▢ Japanese
▢ Korean
▢ Latin American (Latino/Hispanic)
▢ Pacific Islander
▢ South Asian
▢ Southeast Asian
▢ West Asian

4. What gender do you most identify with?
O Male
O Female
O Non-binary/Two-Spirit/Agender/Other

5. What was your biological sex at birth?
O Male
O Female

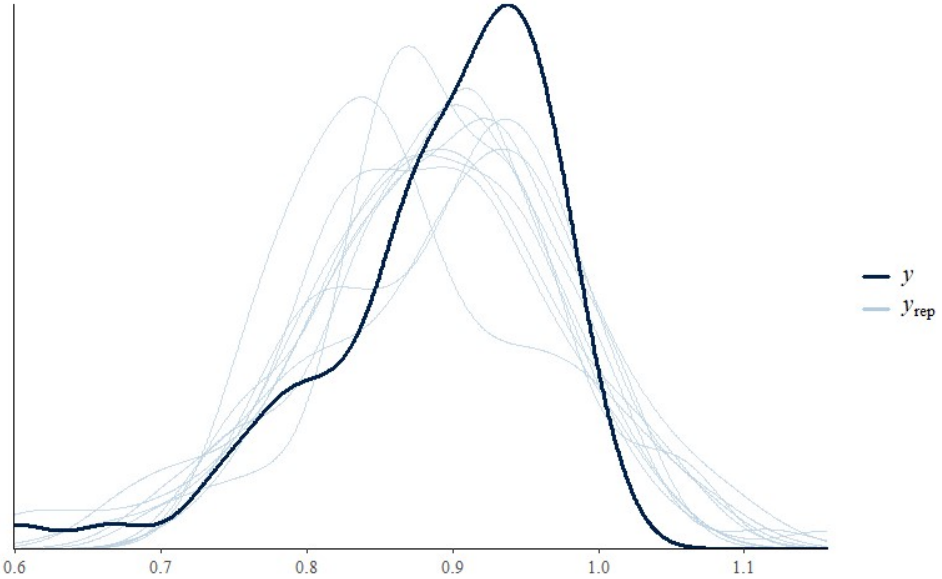
**B.10** Model Code, Summary Tables, and Posterior Predictive Checks, for Chapter 3 Bayesian Models

### B.10.1 Model 16: Accuracy Model

```
Model16 <-
brm(data = df.no.glitch.main.trials.only,
    family = student,
    formula = Accuracy ~ Test_Group.f,
    iter = 8000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 0.92 | 0.03 | 0.87 | 0.97 | 1.00 | 4475 | 6463 |
| F/D/C | -0.01 | 0.04 | -0.08 | 0.07 | 1.00 | 6529 | 9127 |
| F/D/ST | -0.03 | 0.03 | -0.09 | 0.04 | 1.00 | 6052 | 8756 |
| F/E/ST | -0.03 | 0.05 | -0.12 | 0.06 | 1.00 | 7697 | 10233 |
| M/D/C* | -0.08 | 0.04 | -0.16 | -0.01 | 1.00 | 6959 | 9237 |
| M/D/ST* | -0.07 | 0.03 | -0.13 | -0.01 | 1.00 | 6028 | 9280 |
| M/E/C | 0.03 | 0.04 | -0.04 | 0.11 | 1.00 | 6407 | 9010 |
| M/E/ST | 0.01 | 0.04 | -0.06 | 0.08 | 1.00 | 6460 | 9498 |

Model 16 Population-level effects. F = Female, M=Male, D = Difficult, E = Easy, C = Control, ST = Stereotype Threat. Stars (*) indicate effects for which the lower and upper 95% CI (Credible Intervals) do not cross zero.



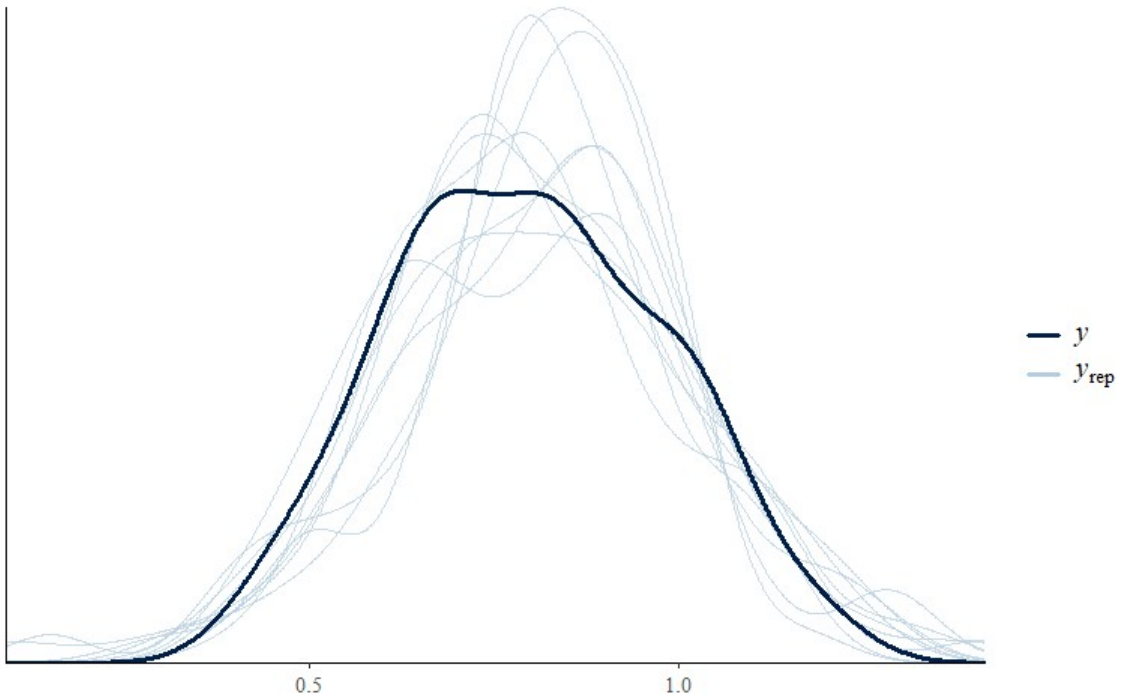

Model 16 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.

### B.10.2 Model 17: Performance Model

```
Model17 <-
brm(data = df.no.glitch.main.trials.only,
    family = student,
    formula = Performance ~ Test_Group.f,
    iter = 8000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 0.92 | 0.05 | 0.83 | 1.01 | 1.00 | 4481 | 6138 |
| F/D/C* | -0.15 | 0.07 | -0.29 | -0.01 | 1.00 | 6466 | 9135 |
| F/D/ST* | -0.23 | 0.06 | -0.35 | -0.11 | 1.00 | 5709 | 8300 |
| F/E/ST | 0.02 | 0.09 | -0.16 | 0.20 | 1.00 | 7633 | 9938 |
| M/D/C* | -0.25 | 0.07 | -0.38 | -0.10 | 1.00 | 6388 | 8389 |
| M/D/ST* | -0.25 | 0.06 | -0.37 | -0.13 | 1.00 | 6049 | 8626 |
| M/E/C | 0.02 | 0.07 | -0.13 | 0.16 | 1.00 | 6502 | 9079 |
| M/E/ST | 0.08 | 0.07 | -0.06 | 0.22 | 1.00 | 6830 | 8756 |

Model 17 Population-level effects. F = Female, M=Male, D = Difficult, E = Easy, C = Control, ST = Stereotype Threat. Stars (*) indicate effects for which the lower and upper 95% CI (Credible Intervals) do not cross zero.



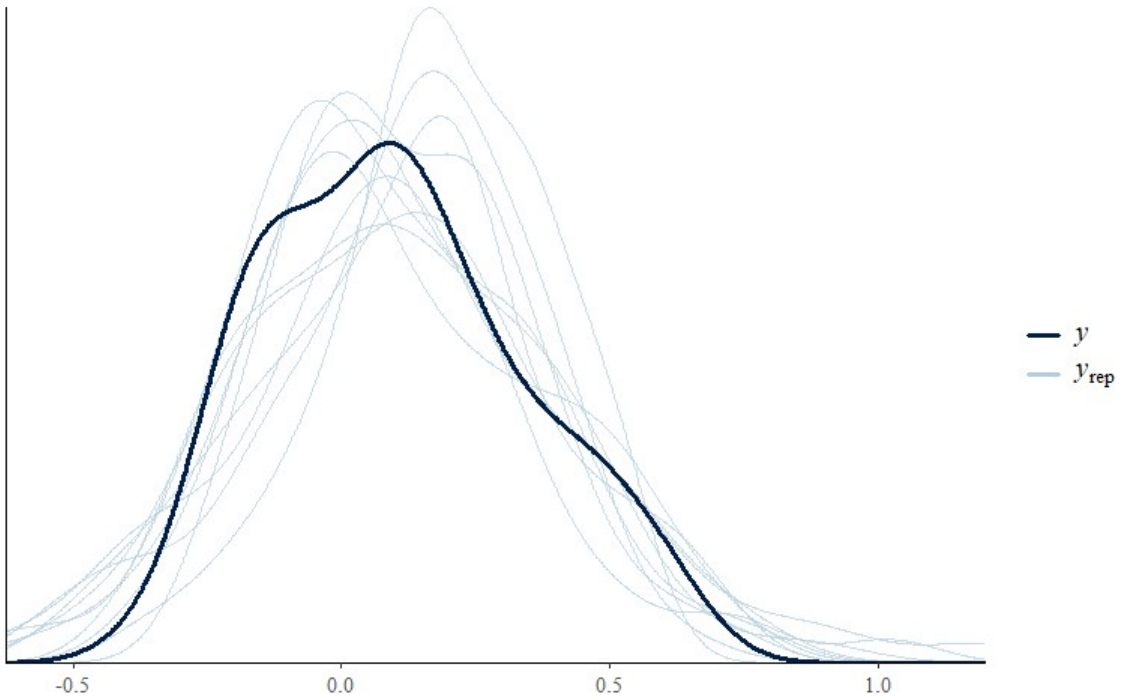

Model 17 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.

### B.10.3 Model 18: Performance Change Model

```
Model18 <-
brm(data = df.no.glitch.main.trials.only,
    family = student,
    formula = Performance_Change ~ Test_Group.f,
    iter = 8000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 0.18 | 0.07 | 0.05 | 0.32 | 1.00 | 4773 | 7221 |
| F/D/C | -0.15 | 0.11 | -0.37 | 0.08 | 1.00 | 7716 | 9068 |
| F/D/ST* | -0.19 | 0.09 | -0.37 | -0.02 | 1.00 | 6205 | 9179 |
| F/E/ST | -0.04 | 0.13 | -0.30 | 0.22 | 1.00 | 9217 | 10533 |
| M/D/C* | -0.29 | 0.10 | -0.49 | -0.08 | 1.00 | 7206 | 9894 |
| M/D/ST | -0.16 | 0.09 | -0.34 | 0.03 | 1.00 | 6725 | 9391 |
| M/E/C | 0.19 | 0.11 | -0.03 | 0.40 | 1.00 | 7307 | 9786 |
| M/E/ST | 0.12 | 0.11 | -0.09 | 0.33 | 1.00 | 6704 | 9078 |

Model 18 Population-level effects. F = Female, M=Male, D = Difficult, E = Easy, C = Control, ST = Stereotype Threat. Stars (*) indicate effects for which the lower and upper 95% CI (Credible Intervals) do not cross zero.



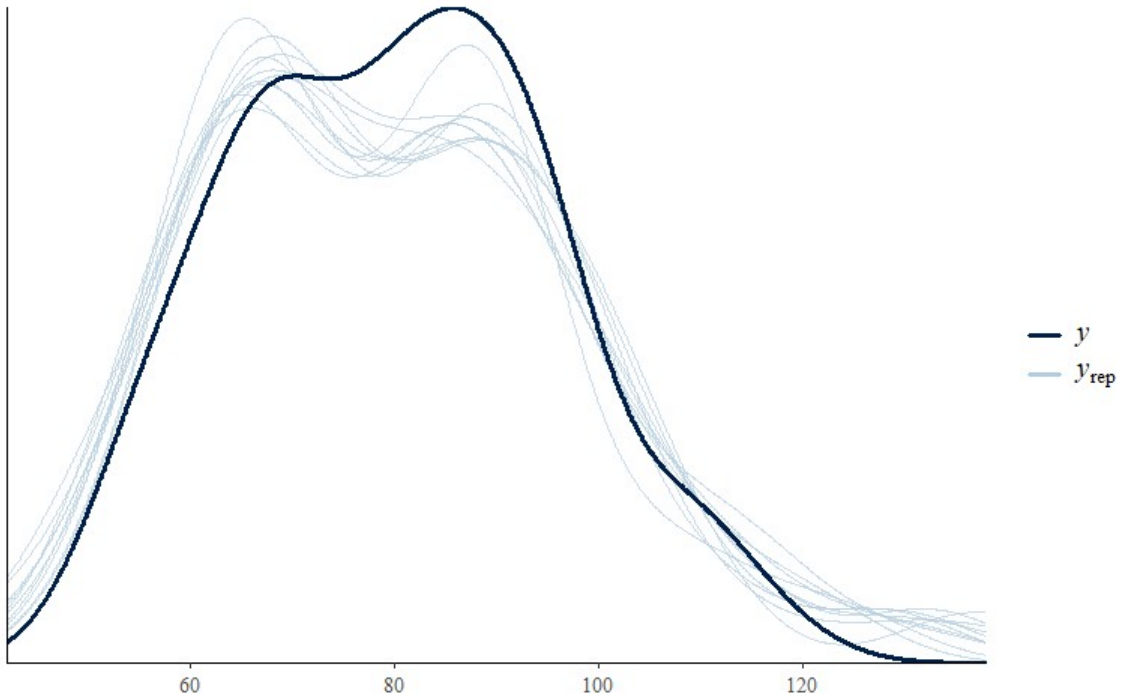

Model 18 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.

### B.10.4 Model 19: Hits.i|trials(Clicks.i) Model

```
Model19 <-
brm(data = df.no.glitch.main.trials.only,
    family = binomial,
    formula = Hits.i|trials(Clicks.i) ~ Test_Group.f,
    iter = 8000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 2.37 | 0.12 | 2.14 | 2.61 | 1.00 | 3938 | 6150 |
| F/D/C | -0.07 | 0.20 | -0.45 | 0.32 | 1.00 | 6177 | 8713 |
| F/D/ST* | -0.34 | 0.16 | -0.65 | -0.04 | 1.00 | 4998 | 7557 |
| F/E/ST* | -0.45 | 0.19 | -0.81 | -0.08 | 1.00 | 6293 | 8848 |
| M/D/C* | -0.84 | 0.16 | -1.15 | -0.53 | 1.00 | 5190 | 8007 |
| M/D/ST* | -0.83 | 0.14 | -1.12 | -0.55 | 1.00 | 4789 | 7938 |
| M/E/C* | 0.67 | 0.24 | 0.21 | 1.14 | 1.00 | 6636 | 9054 |
| M/E/ST | 0.20 | 0.19 | -0.16 | 0.57 | 1.00 | 5752 | 9237 |

Model 19 Population-level effects. F = Female, M=Male, D = Difficult, E = Easy, C = Control, ST = Stereotype Threat. Stars (*) indicate effects for which the lower and upper 95% CI (Credible Intervals) do not cross zero.



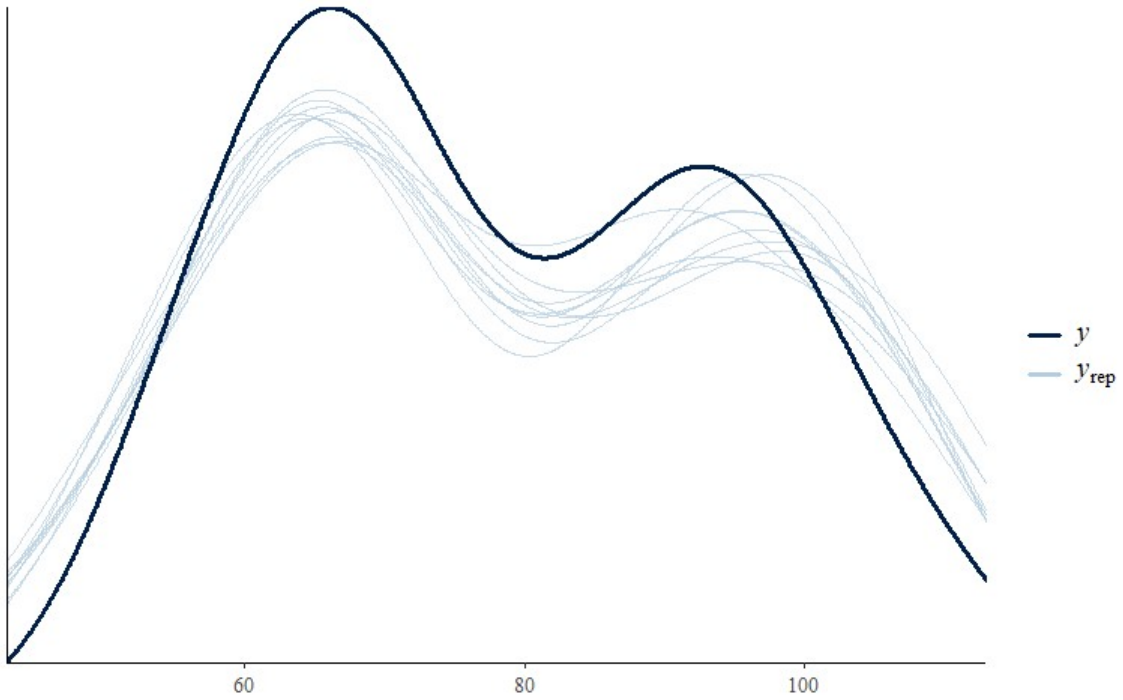

Model 19 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.

### B.10.5 Model 20: Hits.i|trials(Clicks.i) Model: Sam and Tyler Only

```
Model20 <-
brm(data = df.no.glitch.main.trials.only.Sam.and.Tyler,
    family = binomial,
    formula = Hits.i|trials(Clicks.i) ~ Test_Group.f,
    iter = 8000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 2.94 | 0.20 | 2.56 | 3.34 | 1.00 | 3924 | 6172 |
| F/D/C* | -0.88 | 0.26 | -1.39 | -0.36 | 1.00 | 5026 | 8436 |
| F/D/ST | -0.49 | 0.27 | -1.03 | 0.04 | 1.00 | 5509 | 8239 |
| F/E/ST | -0.42 | 0.29 | -0.99 | 0.16 | 1.00 | 6127 | 9345 |
| M/D/C* | -1.63 | 0.23 | -2.09 | -1.18 | 1.00 | 4644 | 7391 |
| M/D/ST* | -2.06 | 0.24 | -2.53 | -1.61 | 1.00 | 4659 | 7552 |
| M/E/C | 0.27 | 0.48 | -0.64 | 1.27 | 1.00 | 9486 | 10497 |
| M/E/ST | 0.13 | 0.36 | -0.55 | 0.87 | 1.00 | 7016 | 8575 |

Model 20 Population-level effects. F = Female, M=Male, D = Difficult, E = Easy, C = Control, ST = Stereotype Threat. Stars (*) indicate effects for which the lower and upper 95% CI (Credible Intervals) do not cross zero.



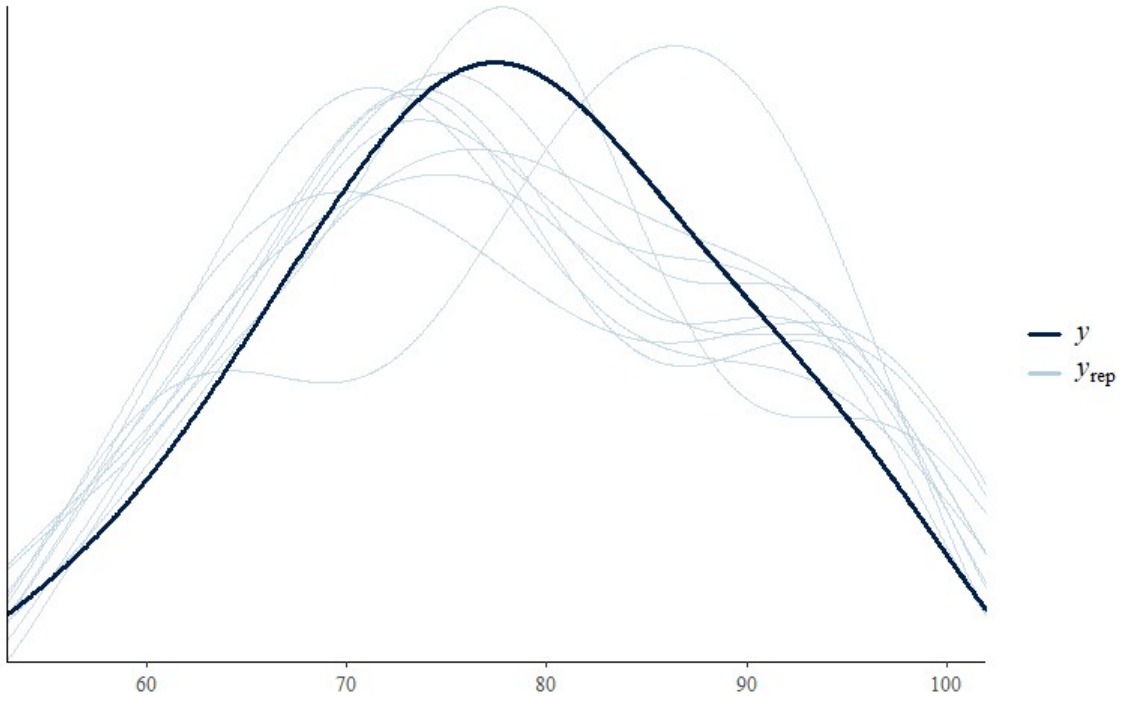

Model 20 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.

### B.10.6 Model 21: Hits.i|trials(Clicks.i) Model: Anna and Drayton Only

```
Model21 <-
brm(data = df.no.glitch.main.trials.only.Anna.and.Drayton,
    family = binomial,
    formula = Hits.i|trials(Clicks.i) ~ Test_Group.f,
    iter = 8000, chains = 4, cores=4, prior = prior("normal(0,1)", class = "b"),
 control = list(adapt_delta = 0.999, max_treedepth=15))
```

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 1.52 | 0.18 | 1.18 | 1.87 | 1.00 | 8460 | 9466 |
| F/D/C* | 0.98 | 0.43 | 0.19 | 1.86 | 1.00 | 10871 | 9767 |
| F/D/ST | 0.32 | 0.23 | -0.14 | 0.78 | 1.00 | 8627 | 10000 |
| M/D/ST* | 0.64 | 0.27 | 0.12 | 1.16 | 1.00 | 9267 | 10388 |
| M/E/C* | 1.39 | 0.44 | 0.57 | 2.31 | 1.00 | 11154 | 10795 |
| M/E/ST* | 0.53 | 0.27 | 0.01 | 1.06 | 1.00 | 9742 | 10539 |

Model 21 Population-level effects. F = Female, M=Male, D = Difficult, E = Easy, C = Control, ST = Stereotype Threat. Stars (*) indicate effects for which the lower and upper 95% CI (Credible Intervals) do not cross zero.




Model 21 Posterior Predictive Check. Y represents the actual representations from the model while $Y_{rep}$ shows the possible model representations.