Essays in Economics

by

Eric Wilken

A dissertation accepted and approved in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Economics

Dissertation Committee:
Michael Kuhn, Chair
Jonathan Davis, Core Member
Benjamin Hansen, Core Member
Leslie Leve, Institutional Representative

University of Oregon
Spring 2026

© 2026 Eric Wilken
All rights reserved.

DISSERTATION ABSTRACT

Eric Wilken

Doctor of Philosophy in Economics

Title: Essays in Economics

This dissertation consists of three essays in applied microeconomics examining how identity shapes economic and social outcomes across a range of contexts. The first chapter provides summaries of all three essays. The second chapter uses a laboratory experiment to study whether leader gender and sexual orientation influence follower effort and stated effectiveness in a nonprofit volunteer setting. I find that women leaders are rated as more effective and more typical than men, though this advantage does not extend to gay women, and leader identity has no detectable effect on actual task performance. The third chapter, co-authored with Mary Kite, introduces a face-morphing methodology to test whether the perceived threshold between young and old differs by gender and race. We find that when accounting for how participants rate the morph anchor faces, results are consistent with the double standard of aging. The fourth chapter estimates the causal effect of sports betting legalization on child maltreatment reports using a staggered difference-in-differences framework. Following legalization, states experience roughly a seven percent increase in screened-in reports, with effects concentrated in states that permit mobile betting and the results are consistent with a financial hardship pathway. This dissertation includes unpublished and co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Eric Wilken

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Ball State University, Muncie, IN, USA
Marietta College, Marietta, OH, USA

DEGREES AWARDED:

Doctor of Philosophy, Economics, 2026, University of Oregon
Master of Science, Economics, 2024, University of Oregon
Master of Science, Statistics, 2021, Ball State University
Master of Arts, Social Psychology, 2021, Ball State University
Bachelor of Science, Mathematics and Environmental Science, 2018, Marietta College
Bachelor of Arts, Economics, 2018, Marietta College

AREAS OF SPECIAL INTEREST:

Behavioral Economics, Labor Economics, Public Policy
Economics of Nonprofits, Leadership Development, Group Dynamics

PROFESSIONAL EXPERIENCE:

Graduate Employee, University of Oregon, 2021–Present
Summer Faculty, Phillips Exeter Academy, 2023–Present
AP Microeconomics Exam Reader, Educational Testing Services, 2024
Graduate Employee, Ball State University, 2019–2021
Business Analyst, Luxottica, 2018–2019

GRANTS, AWARDS AND HONORS:

Graduate Teaching Award – Outstanding Online Instruction, University of Oregon, 2025

Kleinsorge Research Fellowship, University of Oregon, 2025

Department of Economics Travel Grant, University of Oregon, 2023

Edward G. Daniel Scholarship, University of Oregon, 2022

Five Star Society, Omicron Delta Kappa, 2022

Department of Mathematics Merit Fellowship, Ball State University, 2019–2021

Dale Edward Umbach Memorial Scholarship, Ball State University, 2020

Sherrill Cleland Prize in Leadership, Marietta College, 2018

ACKNOWLEDGEMENTS

This dissertation would not exist without the generosity and wisdom of many remarkable people, and it is my privilege to name them here.

I owe an immeasurable debt of gratitude to my advisor, Dr. Michael Kuhn, whose patience through the inevitable setbacks and steady confidence in this work — even when mine wavered — made all the difference. I am a better scholar because of you. To my committee, Dr. Benjamin Hansen, Dr. Jonathan Davis, and Dr. Leslie Leve, thank you for the care and seriousness with which you engaged this project. Your investment in my development as a researcher is a gift I will carry well beyond this degree.

Special thanks to Dr. Mary Kite, who introduced me to the world of research. You gave me both the tools and the confidence to succeed in my career. And to Dr. Tanya Judd, you gave me a clear vision of the kind of educator I aspire to be. I hope to bring what you've taught me into the classroom each and every day. I would also like to thank the late Dr. Gama Perruci, whom I am deeply grateful to have known. Your leadership and integrity left a lasting impression on me, and I hope to embody those same qualities in my own life and work.

To my sister, thank you for the laughter, the honesty, and the kind of support only a sibling can give. To my parents, thank you for the sacrifice that helped me achieve this accomplishment. And to my friends who kept me sane and smiling, thank you for all of the joy that took my mind off of work and made me better for it.

TABLE OF CONTENTS

Chapter Page

I. LEADER IDENTITY, EVALUATIONS, AND PERFORMANCE: EVIDENCE FROM A LABORATORY EXPERIMENT 18

1.1. Introduction 18

1.2. Background 20

1.2.1. Leader Identity 21

1.3. Methods 24

1.3.1. Procedure 24

1.3.2. Measures 27

1.3.2.1. Data-Entry Task 27

1.3.2.2. Stated Leader Effectiveness and Typicality 29

1.3.2.3. Heteronormative Beliefs and Values 29

1.3.2.4. Other Measures of Beliefs 29

1.4. Data 30

1.4.1. Participants 30

1.4.2. Descriptive Statistics Across Groups 31

1.5. Results 34

1.5.1. Leader Identity on Stated Beliefs and Revealed Performance . 34

1.5.2. Page-Level Performance Differences By Leader Identity . . . 38

1.5.3. Heterogeneity By Participant Identity and Beliefs 39

1.6. Discussion 41

II. ARTIFICIAL AGING AND FACE MORPHING: TESTING THE DOUBLE STANDARD OF AGING HYPOTHESIS 45

2.1. Introduction 45

Chapter Page

2.2. Background . . . 48

2.2.1. The Double Standard of Aging . . . 50

2.2.2. Direct Tests of the Double Standard Hypothesis . . . 51

2.3. Methods . . . 53

2.3.1. Participants . . . 53

2.3.2. Procedure . . . 53

2.3.3. Measures . . . 54

2.3.3.1. Morph Task . . . 54

2.3.3.2. Endpoint Age and Attractiveness Ratings . . . 55

2.3.3.3. Task Experience . . . 56

2.3.3.4. Demographics . . . 57

2.4. Results . . . 57

2.4.1. Morph Stop Time . . . 59

2.4.2. Stop Time Versus Perceived Age . . . 62

2.4.3. Participant-Specific Estimated Target Age . . . 63

2.4.4. Covariate-Adjusted Models . . . 65

2.5. Conclusion . . . 67

III. THE IMPACT OF LEGALIZED SPORTS BETTING ON CHILD MALTREATMENT: EVIDENCE FROM CHILD PROTECTIVE SERVICES REPORTS . . . 70

3.1. Introduction . . . 70

3.2. Background . . . 73

3.2.1. History of Legal Sports Betting . . . 73

3.2.2. Background on Child Maltreatment . . . 75

3.2.3. Sports Betting and Maltreatment Risk Factors . . . 80

3.3. Data . . . 80

Chapter Page

3.3.1. Child Protective Services Case Data 80

3.3.2. NFL Game Data 82

3.3.3. Sports Betting Spread and Revenue 84

3.3.4. Population, COVID School Closures, and Debt to Income . . 85

3.4. Empirical Strategy 86

3.5. Results 88

3.5.1. Legalization Increases Reports of Child Maltreatment 88

3.5.2. Pre-Trends and Dynamic Effects 90

3.5.3. Synthetic Controls Method Estimates 94

3.5.4. Financial Health, Abuse, and Emotional Responses to Betting 97

3.5.5. Case and Child Heterogeneity 100

3.6. Conclusion 105

APPENDICES

A. LEADER IDENTITY, EVALUATIONS, AND PERFORMANCE: EVIDENCE FROM A LABORATORY EXPERIMENT APPENDIX 107

A.1. Training Time 107

A.2. Sample Subset: Passed Manipulation Check 107

A.3. Sample Subset: Participant Gender 110

A.4. Sample Subset: Participant Sexuality 110

A.5. Sample Subset: Low vs High Heteronormative Beliefs 111

A.6. Imputation of Missing Values 111

A.6.1. Missing Data Summary 111

Chapter Page

B. ARTIFICIAL AGING AND FACE MORPHING: TESTING THE DOUBLE STANDARD OF AGING HYPOTHESIS APPENDIX 115

B.1. Sensitivity to the Assumed Form of Age Progression 115

B.2. Robustness Check: Classical Split-Plot ANOVA 117

C. THE IMPACT OF LEGALIZED SPORTS BETTING ON CHILD MALTREATMENT: EVIDENCE FROM CHILD PROTECTIVE SERVICES REPORTS APPENDIX 121

Sports Betting Appendix 121

C.1. Child Maltreatment Reporting Trends 121

C.2. Main Specification by Fortnight 122

C.3. Staggered Adoption Robustness Check 123

C.4. Betting Handle As Treatment 125

REFERENCES CITED 129

LIST OF FIGURES

Figure Page

1. Examples of the Leader Vignette 25
2. Examples of the Training and Data Entry Task 26
3. Example of Young Face 56
4. Example of "Old" Face 56
5. Examples of Face Morph End Points 56
6. Estimated Marginal Means of Stop Time By Condition 60
7. Estimated Marginal Means of Stop Time By Condition 61
8. Distribution of Ratings of Anchor Faces for Morphs 62
9. Estimated Marginal Means of Estimated Age By Condition 64
10. Estimated Marginal Means of Estimated Age By Condition 65
11. Trends in Sports Betting Revenue 75
12. Trends in Child Maltreatment Reporting 77
13. Family Stress Model Diagrammed 79
14. Home Team for Each County based on Distance to Nearest Stadium . . 84
15. DiD Imputation Event Studies 93
16. Distribution of Synthetic Control ATTs 95
17. Partially Pooled Cohort Synthetic Control Method 96
B.1. Effect of the non-linearity parameter 115
B.2. Effect of the inflection point parameter 115
B.3. Parameterization of the power function 115

Figure Page

B.4. Significance landscape for the three omnibus tests (Race × Gender interaction, target gender main effect, target race main effect) as a function of the linearity exponent α (x-axis; 1 = linear, 13 = cubic root) and inflection point parameter (y-axis), for participants who initiated the task from the young-face endpoint. Cell color encodes the p-value from the corresponding mixed-effects model (lighter = smaller p). Solid and dashed white contours mark $p = 0.05$ and $p = 0.01$ respectively. 116

B.5. Significance landscape as in Figure B.4, for participants who initiated the task from the old-face endpoint. 118

C.1. Spatial Distribution of Child Abuse Reports 121

C.2. Spatial Distribution of the Change in Child Abuse Reports 122

C.3. Sun & Abraham Event Study Estimates 124

C.4. Cohort Effects (Sun & Abraham) 126

C.5. Treatment Effects by Betting Handle Quintile 127

LIST OF TABLES

Table Page

1. Participant Characteristics and Beliefs by Assigned Leader Condition . . 32
2. Performance, Leader Evaluations, and Time Use by Leader Identity Condition 33
3. Effects of Leader Gender and Sexuality on Item Completion 35
4. Effects of Leader Gender and Sexuality on Accuracy 36
5. Effects of Leader Gender and Sexuality on Leader Evaluations 37
6. Effects of Leader Characteristics on Page-Level Performance 39
7. Effects of Leader Gender and Sexuality on Leader Evaluations by Participant Gender 40
8. Effects of Leader Gender and Sexuality on Leader Evaluations by Participant Sexuality 41
9. Effects of Leader Gender and Sexuality on Leader Evaluations by Heteronormativity 42
10. Participant Characteristics by Condition 54
11. U.S. Sports-Betting Adoption Timeline 74
12. Pre-Treatment Summary Statistics by Case Type, Child Age, and Race/Ethnicity 78
13. OLS and Poisson Estimates of Any Betting 89
14. OLS and Poisson Estimates of Mobile and In-Person Betting 90
15. DiD Imputation Estimates 91
16. Poisson Estimates of Betting Effects by County Debt-to-Income Tercile 98
17. Poisson Estimates for Types of Maltreatment Investigations 99

Table Page

18. Poisson Estimates of Treatment Effects by Local NFL Team Game Outcomes . . . 100

19. Poisson Estimates of Treatment Effects by Race . . . 101

20. Poisson Estimates of Betting Effects by Rurality: County-Level . . . 103

21. Poisson Estimates of Treatment Effects by Child Age Group . . . 104

22. Poisson Estimates of Treatment Effects by Child Sex . . . 104

A.1. Effects of Leader Gender and Sexuality on Training Time . . . 107

A.2. Effects of Leader Gender and Sexuality on Item Completion . . . 108

A.3. Effects of Leader Gender and Sexuality on Accuracy . . . 109

A.4. Effects of Leader Gender and Sexuality on Leader Evaluations . . . 109

A.5. Effects of Leader Gender and Sexuality on Item Completion by Participant Gender . . . 110

A.6. Effects of Leader Gender and Sexuality on Item Completion by Participant Sexuality . . . 110

A.7. Effects of Leader Gender and Sexuality on Accuracy by Participant Sexuality . . . 110

A.8. Effects of Leader Gender and Sexuality on Item Completion by Heteronormativity . . . 111

A.9. Effects of Leader Gender and Sexuality on Accuracy by Heteronormativity 111

A.10 Missing Data Summary for Imputed Control Variables . . . 112

A.11 Effects of Leader Gender and Sexuality on Leader Evaluations . . . 112

A.12 Effects of Leader Gender and Sexuality on Accuracy . . . 112

A.13 Effects of Leader Gender and Sexuality on Item Completion . . . 113

A.14 Effects of Leader Gender and Sexuality on Item Completion . . . 113

A.15 Effects of Leader Gender and Sexuality on Accuracy . . . 113

A.16 Effects of Leader Gender and Sexuality on Leader Evaluations . . . 114

C.1. OLS and Poisson Estimates of Treatment Effects: Any Sports Betting . . . 123

Table Page

C.2. OLS and Poisson Estimates of Treatment Effects: Mobile vs In-Person 123

C.3. Sun and Abraham ATTs 125

C.4. Poisson Estimates of Combined and Separate Betting Treatments 128

This dissertation contains four chapters. This chapter, Chapter I, provides an overview of the other three chapters. Chapters II and III examine questions in the economics of identity, while Chapter IV relates to the economics of crime and child welfare.

Chapter II examines the gap between leadership evaluations and follower behavior using a laboratory experiment in a nonprofit, volunteer setting where motivation is non-pecuniary. Observers often use identity as a proxy for the characteristics they associate with effective leadership, and it remains unclear whether expectations of effectiveness influence how individuals actually behave when working under leaders of differing identities. Participants are randomly assigned to leader vignettes whose gender and sexual orientation vary and are asked to complete an effort-based data entry task and a survey evaluating the effectiveness of their assigned leader. I find that women are rated as more effective and more representative of a typical leader, in the context of nonprofit volunteer coordination, than men. This advantage does not extend to gay women. Despite these differences in evaluations, there are no detectable differences in effort or accuracy across leader types. These findings provide suggestive evidence that stated evaluations of leadership may be a poor signal of actual productivity gains.

Chapter III, co-authored with Mary Kite, introduces a face-morphing methodology to test whether the perceived threshold between young and old differs by gender and race. Undergraduates ($N = 172$) viewed automatically advancing videos which morphed a young face to an artificially aged face by advancing smoothly through a discrete set of slides. Participants were asked to stop the video when the face appeared to cross from young to old. The faces featured four racial groups (White, Asian, Black, Latino) with both men and women. We summarize

the results with two primary outcomes: the slide at which participants stopped the sequence and a participant-specific estimated transition age obtained by linearly interpolating between each participant's own age ratings of the young and old anchors. The two measures produce effects in opposing directions. Participants stopped the slide progression later for female faces across all racial groups, a pattern inconsistent with the double standard of aging. Estimated transition age, by contrast, was lower for female than male faces, with White women judged old at the youngest estimated age and Latino and Black men at the oldest, consistent with the double standard. The divergence traces to systematic differences in anchor ratings: young anchors of female morphs were rated as older than young anchors of male morphs. The gender effect on estimated age was not attenuated by adjustment for rated attractiveness of the aged face.

Chapter IV explores the unintended consequences of sports betting legalization on child maltreatment. Using administrative Child Protective Services case data and a staggered difference-in-differences framework, I estimate the causal effect of sports betting on screened-in reports of child maltreatment and find that following legalization, states experience roughly a seven percent increase. The effect is driven by states that legalize mobile sports betting and is stronger in rural areas. I outline two potential pathways for these effects: an amplification of emotional cues from bad sports betting outcomes, and the family stress of accumulated financial hardship. I do not find that legal betting amplifies the relationship between upset losses and child maltreatment. Instead, the long horizon over which effects emerge suggests a financial hardship pathway.

CHAPTER I

LEADER IDENTITY, EVALUATIONS, AND PERFORMANCE: EVIDENCE FROM A LABORATORY EXPERIMENT

1.1 Introduction

The stereotypical leader is often described as "ambitious, self-sufficient, and dominant" (Powell et al., 2002; Schein, 2001), and as possessing masculine traits or physical features (Giacomin et al., 2022; Koenig et al., 2011). These descriptions reflect widely held beliefs about what makes an effective leader and reflect the image of a leader as consistently portrayed in media from Machiavelli's The Prince to The West Wing. However, it remains unclear whether these stated and portrayed preferences correspond to how individuals truly respond to leader identity.

Most people do not fit this prototypical image, whether due to personality, appearance, or social identity. Prior research shows that women and members of sexual minority groups are often perceived as less prototypical and less desired as leaders (Eagly & Chin, 2010; Liberman & Golom, 2015; Pellegrini et al., 2020). These perceptions are typically measured through stated evaluations, such as survey responses or hypothetical judgments. Yet, such measures may not reflect how followers behave when asked to exert effort on behalf of their leader.

This distinction between stated and revealed preferences is central to understanding leadership effectiveness. If individuals report preferences for certain leader characteristics but do not act on those preferences in real tasks, then commonly used measures of leadership perception may provide a misleading guide to actual performance. This raises the possibility that leadership positions are allocated based on beliefs that do not translate into differences in productivity.

I propose a laboratory experiment to examine whether stated preferences over leader characteristics, specifically social identity, align with revealed measures of follower effort. The study focuses on leadership in the context of nonprofit organizations and volunteerism, where financial incentives are limited and motivation must come from non-pecuniary sources. In this setting, leadership plays a particularly important role in motivating participation.

To test the relationship between stated and revealed preferences, participants are randomly assigned to a fictitious leader whose identity varies along observable dimensions, including gender and sexual orientation. These identity characteristics are used as a tractable way to introduce variation in leader perceptions, as prior research suggests they are associated with systematic differences in stated evaluations. Participants then complete an effort-based task and report their attitudes toward the assigned leader.[1] This design allows for a direct comparison between how individuals say they evaluate leaders and how they behave when working under them.

I find that leader identity shapes how followers evaluate their leaders but has no detectable effect on how hard they work. In the context of nonprofit volunteer coordination, women leaders are rated as more effective and more representative of a typical leader than men. This advantage does not extend to gay women, who are viewed as less prototypical despite the generally favorable ratings received by women leaders overall. When it comes to actual task performance, there are no meaningful differences across any of the four leader types. Followers worked just as hard and just as carefully regardless of whether their assigned leader was a straight man, a straight woman, a gay man, or a gay woman. This disconnect between

[1]Simple effort-based tasks, such as data entry or counting tasks, are commonly used in experimental economics to measure how incentives influence motivation (Charness et al., 2018).

stated evaluations and observed behavior raises important questions about whether the stated assessments organizations use to select and promote leaders are optimal.

1.2 Background

Leadership is not an innate trait but rather a process in which leaders and followers work together toward a shared goal within a particular social and cultural context (McManus & Perruci, 2020). Psychologists, sociologists, and leadership scholars have long studied leadership as a "relational process" (Bass, 1985; Fischer et al., 2017; Gardner et al., 2011; Hooper, 2016; Sheer, 2015), which means that leadership emerges through social interaction and contextual dynamics, even when the "leader" is a title granted by fiat. In contrast, economists and management scholars often focus on how the behavior and strategies of leaders influences task outcomes (Antonakis et al., 2022; Aunga & Masare, 2017; Grossman & Hart, 1992; Kim, 2002; Lazear, 2012; Puni et al., 2014). Within this framework, leadership is frequently conceptualized as a position that motivates subordinate effort through reward and punishment mechanisms.

Historically, these two approaches have remained largely separate. Garretsen, Stoker, and Weber (2020) argue that integrating these perspectives could benefit leadership research. They propose developing a unified theory of leadership that reduces conceptual ambiguity, more frequently identifies causal relationships, and investigates leadership beyond purely individual or group-level analysis. In this paper, I take this approach—combining elements of identity economics, task-motivation, and social perceptions of leadership—to consider the experience of gay and lesbian leaders.

In their foundational work on identity economics, Akerlof and Kranton, 2000 argue that individuals' preferences are influenced by social, cultural, and personal

norms. These norms shape individual identity and affect how individuals interact with economic institutions. Akerlof and Kranton later apply this framework to firms, suggesting that organizations operate more efficiently when employees feel that they share a common identity within the organization (Akerlof & Kranton, 2005). This implies that firms may benefit from investing resources that foster a sense of belonging among employees.

How do leaders do this? There is extensive work that considers the taxonomy and effectiveness of leadership styles. Numerous leadership styles have been theorized, including charismatic leadership and servant leadership (Antonakis et al., 2022; Greenleaf & Spears, 1998). One concept particularly relevant to this study is *G Quotient Leadership*, a theory developed through research on successful gay executives (Snyder, 2006). This theory suggests that there may be systematic differences in leadership styles between straight and gay leaders. While leadership styles are not inherently restricted to particular groups, some scholars argue that certain groups may be more likely to adopt particular styles. Researchers have examined how different leadership styles affect performance both within and outside of firms (Aunga & Masare, 2017; Meslec et al., 2020; Puni et al., 2014). In contrast, the present study holds leadership style constant while varying leader identity. This design allows for a clearer examination of the relationship between leader identity and follower behavior in a controlled setting. To the degree successful leaders create a sense of common identity which affects subordinate performance and leadership perceptions, I examine the role of leader identity, independent of style, directly linking the two threads of Akerlof and Kranton work.

1.2.1 Leader Identity. Research on leadership perception has increasingly focused on how identity-based biases shape who is recognized as a

legitimate leader and who ultimately gains access to leadership roles. Much of this work has centered on gender, offering substantial insight into how men and women are evaluated differently. In contrast, the role of sexual orientation in shaping leadership perceptions remains comparatively underexplored.

Two of the most common terms one reads about women and leadership are the "glass ceiling" and the "glass cliff". The glass ceiling is the term used to describe the persistent difficulty women face in reaching top leadership roles despite comparable qualifications. Similarly, the glass cliff highlights the common occurrence of when women are promoted, they are disproportionately placed into leadership positions during periods of crisis or heightened risk (Cotter et al., 2001; Ryan & Haslam, 2005; Ryan et al., 2016). Evidence suggests that women are less likely to be selected for stable, high-reward leadership roles but more likely to be appointed when the probability of failure is high, exposing them to greater scrutiny and a higher likelihood of negative evaluations (Ryan & Haslam, 2005; Ryan et al., 2016). These dynamics complicate standard interpretations of leadership outcomes, as observed differences in performance may reflect differences in the environments in which leaders are placed rather than differences in underlying ability.

Work on the perceptions of women leaders shows that they are shaped by a mismatch between gender norms and leadership expectations. Role congruity theory in psychology predicts that women are evaluated less favorably in leadership roles because these roles emphasize traits such as dominance and assertiveness that conflict with traditional expectations of femininity (Eagly & Karau, 2002; Heilman, 2001; Ridgeway, 2001). As a result, women may be perceived as less effective leaders even when their behavior is identical to that of men (Eagly & Chin, 2010). These perceptions are reinforced by widely held beliefs about what

leaders "look like." Across multiple research paradigms, leadership traits are consistently associated with masculinity, suggesting that the prototypical leader is implicitly male (Koenig et al., 2011; Powell et al., 2002; Schein, 2001). This pattern persists across contexts and contributes to systematic differences in how men and women are evaluated as potential leaders (Giacomin et al., 2022).

However, when evaluating the performance of leaders rather than leadership perceptions, the evidence is more mixed. A large body of work finds that women are rated as equally or slightly more effective leaders than men across a wide range of settings (Eagly et al., 2003; Paustian-Underdahl et al., 2014). This divergence between perceived leadership and leader effectiveness highlights a gap between beliefs about leadership and beliefs about measurable outcomes. Gender may operate as a diffuse "status characteristic", shaping both selection into leadership and perceptions of those in those roles (Correll, 2004; Ridgeway, 2001) without needing reinforcement of differential effectiveness. These dynamics can lead to inefficient allocation of leadership positions if evaluators rely on biased priors rather than observed performance.

In the case of gay and lesbian leaders, discrimination based on sexual orientation provides one possible explanation for why such biased perceptions could exist. Research suggests that some of the perceived ineffectiveness of queer leaders is explained by evaluators' levels of prejudice, yet research suggests that something more than discrimination—either implicit or explicit—is at play (Morton, 2017). These perceptions also appear to manifest in the real world in the hiring of leadership positions: queer leaders receive fewer formal leadership roles than their straight counterparts, even when they are perceived as equally competent (Aksoy et al., 2019; Fasoli et al., 2017; Frank, 2006; Niedlich & Steffens,

2015). Often referred to as the "lavender ceiling," the scale of this disparity in leadership may be caused by a combination of in-group effects, stereotypes, and discrimination. According to the Williams Institute, in 2021 approximately 50.4% of LGBT employees reported not being out to their current supervisor and 25.8% were not out to any coworkers (Mallory et al., 2024). This likely reflects a strategic response to the lavender ceiling and the continuing presence of discrimination and anti-gay bias in the workplace.

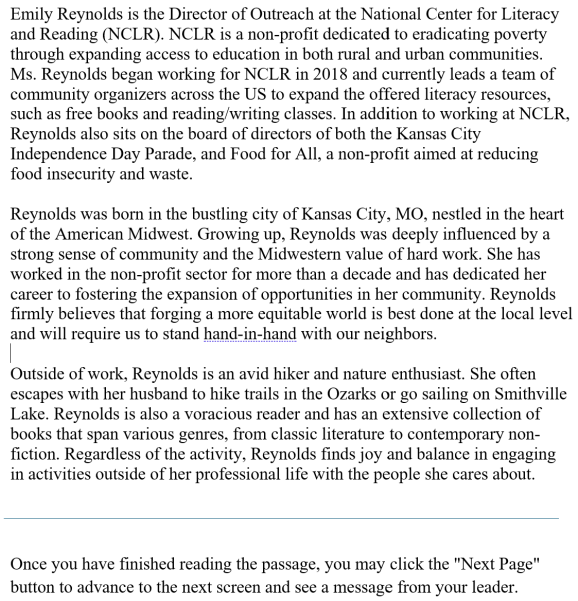
Beyond the motivation of equitable access to leadership roles, discrimination is costly. When societies become more welcoming of queer individuals, higher GDP growth tends to follow, likely due to reduced transaction costs and reduced information loss associated with the avoidance of queer workers at the firm level (Badgett et al., 2019). There may therefore be additional gains from increased inclusivity in leadership if discrimination has led to an inefficient allocation of leadership positions.

1.3 Methods

The experiment addresses three main questions. First, I investigate whether leader identity, specifically gender and sexual orientation, influences the effort their followers exert when completing tasks. Second, I examine whether leader identity influences follower attitudes toward leaders in terms of leader typicality, and perceived success in other roles. Finally, I explore the link between followers' beliefs about gender and sexuality and their stated preferences toward leaders.

1.3.1 Procedure. Participants completed the survey through the online platform Qualtrics. They began by reading and completing an informed consent form.

After consenting, participants were given instructions describing the study procedure and were randomly assigned to one of four fictitious leaders: (1) Gay Woman, (2) Gay Man, (3) Straight Woman, or (4) Straight Man. In all conditions, participants learned that the leader worked for a fictitious nonprofit organization whose mission was to improve adult literacy within the United States. The organization was identical across conditions.



Emily Reynolds is the Director of Outreach at the National Center for Literacy and Reading (NCLR). NCLR is a non-profit dedicated to eradicating poverty through expanding access to education in both rural and urban communities. Ms. Reynolds began working for NCLR in 2018 and currently leads a team of community organizers across the US to expand the offered literacy resources, such as free books and reading/writing classes. In addition to working at NCLR, Reynolds also sits on the board of directors of both the Kansas City Independence Day Parade, and Food for All, a non-profit aimed at reducing food insecurity and waste.

Reynolds was born in the bustling city of Kansas City, MO, nestled in the heart of the American Midwest. Growing up, Reynolds was deeply influenced by a strong sense of community and the Midwestern value of hard work. She has worked in the non-profit sector for more than a decade and has dedicated her career to fostering the expansion of opportunities in her community. Reynolds firmly believes that forging a more equitable world is best done at the local level and will require us to stand hand-in-hand with our neighbors.

Outside of work, Reynolds is an avid hiker and nature enthusiast. She often escapes with her husband to hike trails in the Ozarks or go sailing on Smithville Lake. Reynolds is also a voracious reader and has an extensive collection of books that span various genres, from classic literature to contemporary non-fiction. Regardless of the activity, Reynolds finds joy and balance in engaging in activities outside of her professional life with the people she cares about.

Once you have finished reading the passage, you may click the "Next Page" button to advance to the next screen and see a message from your leader.

Figure 1. Examples of the Leader Vignette

Participants then read a short vignette (see Figure 1) describing the assigned leader. The vignette resembled a short biography containing both professional and personal details about the leader. Participants then received a message from their leader describing the upcoming task. This message was designed to encourage effort based on previous literature on motivating workers (Antonakis et al., 2011, 2022; Fest et al., 2021; Meslec et al., 2020).

The task was introduced as a volunteer opportunity in which participants would enter publication information from books into a database to help the

nonprofit catalog books. The task consists of two stages. The first stage involves training. Participants receive instructions describing how to identify information from a book's publication page (see Figure 2). Participants may spend as much or as little time on the training materials as they wish. The amount of time spent on this training stage is recorded.

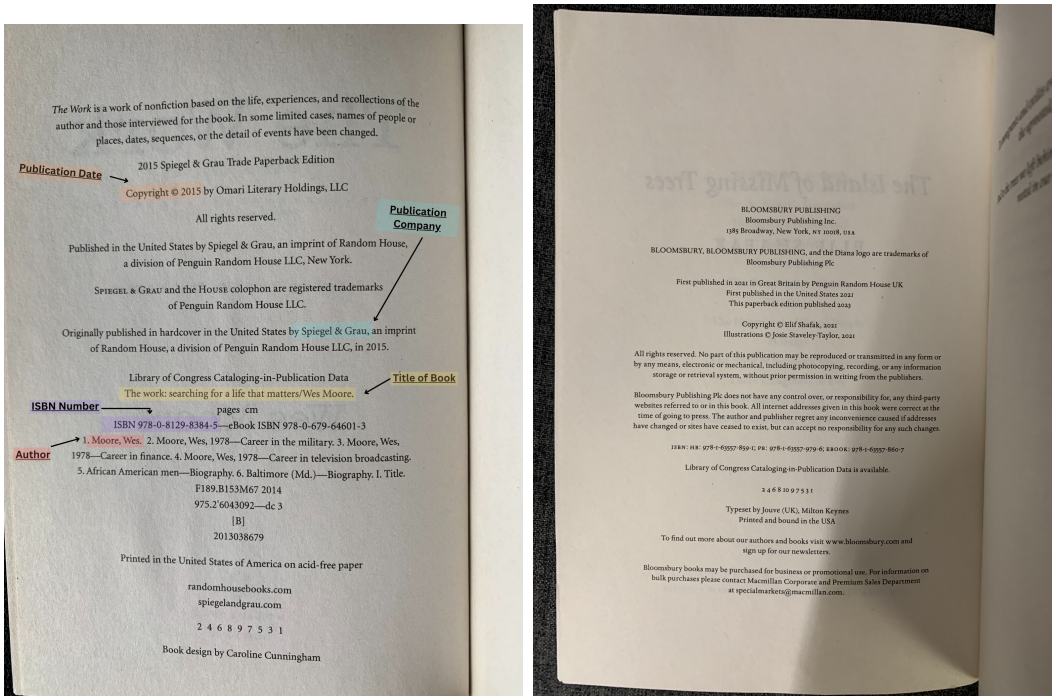


Figure 2. Examples of the Training and Data Entry Task

The second stage consists of a timed effort task. Participants are asked to enter information from as many book publication pages as possible within a 15-minute period. Each page must be displayed for a minimum of 17 seconds before participants can proceed to the next page, and there are enough pages available to ensure that participants remain engaged for the full duration of the task.

From this task, I collect several performance measures, including the time required to complete each entry, the accuracy of each entry, the total number

of entries completed, the total number of entries completed correctly, and the percentage of entries completed correctly.

Participants concluded the study by completing a questionnaire containing the measures described below and demographic questions (age, gender, race, religion, and sexual orientation).

1.3.2 Measures. To test the effect leaders have on followers, I explore followers' self-reported opinions of their assigned leader and their performance on the data-entry task. I highlight below each of the measures used in this experiment to capture underlying participant cultural beliefs, their beliefs about their assigned leader, and their revealed performance.

1.3.2.1 Data-Entry Task. In addition to the stated measure of leader effectiveness, I create several outcomes based on the data-entry task described above. These outcomes cover performance measures such as the number of entries attempted (simple effort-based outcome), the accuracy of entries (attention-based outcome), and an outcome that attempts to quantify the number of outcomes completed correctly.

One of the simplest forms of performance that we can measure is the total number of tasks completed by a volunteer. However, as a measure of effort, this requires that all tasks are completed at the same average level of quality. This is not the case. To capture the quality of each response, I must identify how similar a participant's response is to the correct answer, where the correct answer is the information that is available to each participant on the publication pages.

Determining how similar a participant's response is to the correct answer—or more generally how similar two text strings are to each other—is a common question in computational linguistics and computer science (Navarro, 2001). Edit

distance is a way of quantifying how similar two strings are to each other by counting the minimum number of operations required to transform one string into the other (Navarro, 2001). I use the Jaro-Winkler similarity measure. The Jaro algorithm measures the number of characters in common between two strings while accounting for transpositions within the strings and is normalized between 0 and 1 (Winkler, 1990). A score of 0 means that the strings are identical and a score of 1 indicates that the strings are completely different. Winkler modified Jaro's algorithm by making differences at the beginning of strings more costly than differences near the end of the strings (Winkler, 1990). The Jaro-Winkler algorithm is typically used in the comparison of short strings such as names or, in the case of this paper, publishers, as it is computationally efficient compared to similar measures of edit distance.

Using this algorithm, I calculate the "correctness" of each participant response by transforming the output from the Jaro-Winkler algorithm to instead reflect a measure of similarity rather than distance. When determining what the correct answer should be, I construct and report three different levels of critique for correctness. I report these levels as "generous", "normal", and "strict." Normal correctness is based off of a set of correct answers which require the participant to accurately input the correct title, author, and publisher (spelling matters but capitalization does not) shown on the publication page. However, participants are given flexibility on selecting the correct ISBN number. Strict correctness only permits the exact text string shown on the publication page, including capitalization and punctuation, as completely correct. Only the correct level of publisher and exact ISBN is accepted. Generous correctness allows for the

participant to report either the parent publishing company or the specific imprint and be marked as correct.

1.3.2.2 Stated Leader Effectiveness and Typicality.

Questions measuring leader effectiveness and typicality were adopted from Rosette and Livingston (2012). This measure has high internal consistency ($\alpha = 0.91$) and contains items that can be easily adapted to evaluate leadership effectiveness in this context. It includes four items measuring leader effectiveness and one item measuring leader typicality, which directly asks: "To what extent is NAME typical of a leader?"

1.3.2.3 Heteronormative Beliefs and Values.

To measure participant heteronormativity, I adopt items from the heteronormative attitudes and beliefs scale developed by Habarth (2015). The original measure consists of 16 items divided into two subscales: the Essential Sex and Gender subscale and the Normative Behavior subscale. Because of concerns regarding participant attention when completing a 16-item instrument, I developed a short-form adaptation of the original scale. The shortened measure contains five items—two from the Essential Sex and Gender subscale and three from the Normative Behavior subscale—and demonstrates acceptable internal consistency (Cronbach's alpha $\alpha = 0.77$).

1.3.2.4 Other Measures of Beliefs.

To measure how prosocial a participant may be they are asked to report how many hours they have volunteered in the last 30 days. Additionally, participants are asked to rate their support of the nonprofit listed as the leaders place of work. These two variables are captured to verify that pro-sociality is balanced across conditions or control for it otherwise. These measures are used as controls.

1.4 Data

1.4.1 Participants. Participants were recruited from two university research pools: a marketing research pool and a psychology research pool.[2] They received no monetary compensation for their participation in this research but did receive course credit. All participants had the option to opt out of the research at any time and could complete an alternative assignment for the required course credit.

Overall, 516 participants completed the survey. The final sample consisted of 62.1% women and 31.0% men. Participant age was relatively concentrated, with an average age of $M_{age} = 21.15$. Most participants identified as White (72.5%), and many identified as Christian (43.6%). In terms of sexual orientation, the sample includes participants who self-identified as straight (67.3%) and gay or lesbian (32.7%).

The number of participants targeted for this study was informed by a statistical power analysis. Using a two-sided t-test with $\alpha = 0.05$ and statistical power of 0.8, I estimated that 504 participants would be required to detect a 0.25 standard deviation difference between in group means by leader gender or sexuality.

Table 1 presents participant demographic information separated by treatment condition. Several facets of participant identity varied across groups following random assignment. Any differences in underlying participant

[2]Using a college sample for this study is primarily a matter of convenience. The ability to generalize findings from college subject pools to the broader population is limited (Hanel & Vione, 2016; Wild et al., 2022). This limitation is not a major concern for the present study. Understanding how a leader's identity may influence the effectiveness of college volunteerism is itself an important question, as college students are frequently encouraged by their institutions to volunteer in their local communities. This volunteerism is meaningful for organizations working to accomplish their missions and visions. Additionally, identifying contexts in which leader preferences differ from general expectations allows us to comment more broadly on how strongly employers should rely on stated preferences when making hiring decisions.

characteristics which appear across conditions are controlled for later in the analysis. Participants' ability to correctly recall the leader's identity ranged between a 53% and 82% success rate. The gay man leader condition was associated with the lowest rate of success in the manipulation check. Further examination indicates that this low success rate was primarily due to incorrect identification of the leader's sexuality rather than the leader's gender. In other words, participants often assumed that the gay man leader was straight.

I investigate whether the manipulation check was differentially successful by gender. Across all manipulation checks, women are more accurate than men by approximately 5 percentage points. While I find that only the gender manipulation check shows a statistically significant difference in success by participant gender, the effect warrants some caution in how the conditional-on-passing effects should be interpreted.[3] Any observed differences between men and women in outcome variables could partially reflect differential comprehension of the experimental conditions rather than genuine differences in how they respond to the manipulations themselves.

1.4.2 Descriptive Statistics Across Groups. I calculate at the individual level the percentage of responses completed perfectly and the number of items completed correctly. The average of these values by experimental condition is shown in Table 2.

Table 2 provides a preliminary glance at differences in follower performance and stated opinions of leaders across leader identity conditions. Using the

[3]Using logistic regression models to test the differences across participant gender, I find that both the combined manipulation check (passing both the leader sexuality and gender manipulation checks) and the sexuality manipulation check on their own showed no statistically significant differences in pass rates between men and women ($p = .109$ and $p = .147$, respectively). The gender manipulation check shows that women are significantly more likely than men to correctly identify the leader's gender ($p = .058$).

Table 1. Participant Characteristics and Beliefs by Assigned Leader Condition

	Gay Man	Gay Woman	Straight Man	Straight Woman
Number of Participants	125	120	134	137
% Passed Attention Check	0.560	0.475	0.373	0.263
% Passed Manipulation Check	0.536	0.550	0.828	0.810
Mean Preference for Nonprofit	4.010	4.140	4.070	3.950
Mean Heteronormative Beliefs	2.240	2.300	2.340	2.280
Participant Volunteerism (Mean Monthly Hours)	12.100	16.600	12.200	11.600
Mean Participant Age	21.100	20.800	21.700	21.000
% Women	0.595	0.655	0.623	0.612
% White	0.700	0.716	0.729	0.752
% Straight	0.633	0.655	0.750	0.651
% Christian	0.450	0.447	0.458	0.391

restrictive definition of correctness, differences across groups are modest. Followers assigned to straight woman leaders completed the most tasks completely correctly on average (21.5), while those assigned to gay man leaders completed slightly fewer (20.9). Accuracy under this definition is also similar across conditions, ranging from approximately 53.4% to 54.7%.

With a more relaxed definition of correctness, a similar pattern emerges. Followers assigned to straight woman leaders completed the most items on average (23.8), followed closely by those assigned to straight men (23.6) and gay women (23.5), while those assigned to gay men completed slightly fewer items (22.8). Overall accuracy is again quite similar across conditions, ranging from about 59.3% to 60.0%. Looking at maximum performance, followers assigned to straight woman leaders also achieved the highest number of items completed (32.3), with other groups performing similarly (31.4–32.2). Maximum accuracy is nearly identical across conditions, ranging from 79.6% to 80.6%. Taken together, these performance measures suggest only minor differences in productivity across leader identity conditions.

Stated opinions of leaders are also summarized in Table 2. The effectiveness and leader typicality measures are broadly similar across groups, though some modest differences emerge. Gay women leaders were rated most highly in terms

Table 2. Performance, Leader Evaluations, and Time Use by Leader Identity Condition

	Gay Man	Gay Woman	Straight Man	Straight Woman
Number of Participants	125	120	134	137
Performance (Strict Definition)				
Items Completed Correctly	20.9	21.1	21.4	21.5
	(8.20)	(8.26)	(7.99)	(7.88)
Percent Correct (%)	54.6	53.4	53.9	54.7
	(14.2)	(15.7)	(14.0)	(14.9)
Performance (Relaxed Definition)				
Items Completed	22.8	23.5	23.6	23.8
	(8.15)	(8.14)	(7.77)	(8.05)
Percent Correct (%)	59.3	59.3	59.6	60.0
	(11.2)	(12.9)	(11.3)	(12.4)
Maximum Performance				
Max Items Completed	31.4	31.9	32.2	32.3
	(11.4)	(11.2)	(10.8)	(11.6)
Max Percent Correct (%)	80.6	79.7	80.4	80.2
	(11.9)	(14.1)	(12.8)	(14.1)
Leader Evaluations				
Leader Effectiveness (1–5)	3.61	3.82	3.49	3.61
	(0.83)	(0.83)	(0.89)	(0.86)
Leader Typicality (1–5)	3.51	3.53	3.36	3.60
	(0.76)	(0.88)	(0.79)	(0.81)
Time Measures (seconds)				
Training Time	47.6	39.2	74.5	36.9
	(75.1)	(28.3)	(303.0)	(33.0)
Vignette Reading Time	85.7	62.7	59.1	75.1
	(161.0)	(68.5)	(60.3)	(192.0)
Message Reading Time	33.9	25.9	23.8	25.9
	(82.7)	(23.1)	(25.1)	(48.4)

*Entries are group means with standard deviations in parentheses.

of effectiveness, with an average score of 3.83 out of 5, while straight men received the lowest average rating (3.49). Leader typicality follows a similar pattern, with ratings ranging from 3.36 to 3.60 across conditions. These group averages provide limited evidence of systematic differences in how leaders are perceived across identity conditions.

Turning to the time data at the bottom of Table 2, there are differences in how participants interacted with the materials. The average training time is highest in the straight man leader condition (74.5 seconds), which appears to be driven by an outlier, likely representing a participant who left the experiment running before returning. Participants in the straight woman leader condition spent more time reading the vignette (75.1 seconds) relative to other groups. Excluding extreme observations reduces the average training time in the straight man condition to approximately 67 seconds, suggesting that these differences are driven in part by a small number of unusually long durations.

1.5 Results

1.5.1 Leader Identity on Stated Beliefs and Revealed

Performance. To explore the interaction between leader identity and follower behavior, I estimate a series of linear models where the outcome variables described above are regressed on experimental condition (leader gender, leader sexuality, and their interaction) indicator variables. The model takes the following form:

$$Y_i = \beta_0 + \beta_1 SEX_i + \beta_2 GEN_i + \beta_3 (SEX_i \times GEN_i) + \beta_4 X_i + \epsilon_i \tag{1.1}$$

where Y is the outcome variable of interest, SEX_i is a dummy variable representing the sexual orientation of the assigned leader for subject i (1 = gay, 0 = straight), GEN_i is a dummy variable representing assigned leader gender for

subject i (1 = woman, 0 = man), and X_i is a vector of participant control variables. The comparison group captured by β_0 is a straight man leader.

Across all specifications, there are no statistically significant relationships between leader gender or sexual orientation and follower performance outcomes. This holds for both the number of items completed (Table A.13) and measures of accuracy (Table A.12), and is consistent across models with and without controls. The interaction between leader gender and sexuality is also small and imprecisely estimated in all cases.

That said, the magnitude and direction of several point estimates are informative. In Table A.13, leaders who are women are associated with slightly higher levels of task completion across most specifications, though these effects are small and not statistically distinguishable from zero. Leaders who are gay are generally associated with fewer items completed, particularly under the stricter definition of correctness, but again these estimates are imprecise. Taken together, these results suggest a weak and inconsistent relationship between leader identity and productivity, with no clear directional pattern.

Table 3. Effects of Leader Gender and Sexuality on Item Completion

	Items		Generous Items		Strict Items	
Intercept	23.649***	18.985***	32.231***	28.862***	21.410***	14.394**
	(0.672)	(5.138)	(0.937)	(6.677)	(0.691)	(5.517)
Leader is a Woman	0.117	0.064	0.097	-0.079	0.057	0.066
	(0.962)	(0.938)	(1.361)	(1.286)	(0.965)	(0.976)
Leader is Gay	-0.817	-0.348	-0.839	-0.006	-0.490	0.074
	(0.991)	(1.041)	(1.383)	(1.439)	(1.008)	(1.066)
Woman Leader × Gay Leader	0.576	0.093	0.453	-0.282	0.115	-0.449
	(1.417)	(1.356)	(1.982)	(1.875)	(1.427)	(1.397)
Controls	No	Yes	No	Yes	No	Yes
N	516	458	516	458	516	458

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

A similar story emerges for accuracy. In Table A.12, the estimated effects of leader gender and sexuality fluctuate in sign across specifications and definitions of

correctness. For example, female leaders are associated with slightly higher strict accuracy but slightly lower generous accuracy in some models, while gay leaders are associated with modest decreases in overall accuracy in certain specifications. However, these differences are small relative to baseline accuracy levels and are not statistically significant. This suggests that any relationship between leader identity and accuracy is, at best, limited and highly sensitive to how performance is measured.

Table 4. Effects of Leader Gender and Sexuality on Accuracy

	Pct. Correct		Generous Pct.		Strict Pct.	
Intercept	59.586***	29.158**	80.390***	46.938***	53.897***	20.189.
	(0.972)	(9.404)	(1.102)	(12.838)	(1.211)	(10.554)
Leader is a Woman	0.407	0.087	-0.171	-0.719	0.809	0.643
	(1.436)	(1.499)	(1.634)	(1.644)	(1.760)	(1.902)
Leader is Gay	-0.261	-0.917	0.182	-0.288	0.670	0.153
	(1.399)	(1.466)	(1.534)	(1.675)	(1.758)	(1.869)
Woman Leader × Gay Leader	-0.418	0.475	-0.750	-0.170	-1.970	-1.317
	(2.114)	(2.084)	(2.336)	(2.279)	(2.603)	(2.611)
Controls	No	Yes	No	Yes	No	Yes
N	516	458	516	458	516	458

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

In contrast, leader identity appears to play a more meaningful role in shaping follower evaluations (Table A.11). In the controlled specifications, leaders who are women are associated with higher perceived effectiveness and are more likely to be viewed as typical leaders. The magnitude of these effects is modest but consistently positive, with some reaching conventional levels of statistical significance. By comparison, leader sexuality is not strongly associated with evaluations, as the corresponding coefficients are small and imprecisely estimated.

The interaction between leader gender and sexuality provides some suggestive evidence of an interactive effect. While there is little evidence of an interaction effect for perceived effectiveness, the estimates for leader typicality indicate that gay women leaders may be viewed as less typical than would be

Table 5. Effects of Leader Gender and Sexuality on Leader Evaluations

	Effectiveness		Typicality	
Intercept	3.486***	1.893***	3.362***	2.444***
	(0.080)	(0.412)	(0.070)	(0.695)
Leader is a Woman	0.127	0.238*	0.242*	0.328***
	(0.111)	(0.096)	(0.100)	(0.094)
Leader is Gay	0.126	0.069	0.146	0.107
	(0.111)	(0.110)	(0.099)	(0.095)
Woman Leader × Gay Leader	0.086	-0.015	-0.216	-0.293*
	(0.156)	(0.138)	(0.147)	(0.139)
Controls	No	Yes	No	Yes
N	480	454	490	458

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

implied by the separate effects of gender and sexuality. Although these estimates are only marginally significant, they point to the possibility that combined identities shape perceptions in ways that are not captured by additive effects alone.

Restricting the sample to participants who pass the manipulation check yields a similar overall pattern, though some point estimates become more pronounced. As in the full sample, there are no statistically significant relationships between leader gender or sexual orientation and follower performance outcomes across measures of item completion and accuracy (Tables A.14 and A.15). However, the direction of several coefficients shifts. In contrast to the baseline results, where gay leaders were weakly associated with lower output, the manipulation-check sample shows consistently positive coefficients across most performance measures. For example, gay leaders are associated with higher levels of strict item completion and, in one specification, a statistically significant increase of roughly 4 percentage points in strict accuracy. Although these effects attenuate with controls and remain imprecise overall, they suggest that participants who correctly perceive the treatment may respond more favorably to gay leaders in terms of performance.

Results for leader evaluations remain broadly consistent with the baseline findings but are somewhat stronger in magnitude (Table A.16) in this sample. Female leaders are again rated as more effective and more typical, with larger and more precisely estimated coefficients than in the full sample. There is also some suggestive evidence that gay leaders are viewed as more typical in the uncontrolled specification, though this effect does not persist with controls. The interaction between gender and sexuality remains negative for typicality, indicating that gay women leaders may be perceived as less typical relative to the additive effects of gender and sexuality. Overall, conditioning on passing the manipulation check sharpens some patterns, particularly the positive association between gay leaders and performance and the stronger evaluation effects for women, but the main conclusion remains unchanged: leader identity has limited and inconsistent effects on follower performance, even though it may meaningfully shape perceptions.

1.5.2 Page-Level Performance Differences By Leader Identity.

The participant-level results above aggregate performance across all pages completed by each individual. Since all participants were given the same order of book publication pages, this approach does not account for the fact that task difficulty may vary across pages. As a result, differences in participant accuracy may partly reflect exposure to different segments of the task rather than differences in performance conditional on the same task. Participants who persist longer are more likely to encounter later pages, which may not be identical in realized difficulty. This introduces a potential confound between treatment effects and order effects that cannot be addressed in the participant-level analysis.

To address this concern, Table 6 presents page-level estimates that include page fixed effects and cluster standard errors at the participant level. This

approach compares participants across treatment conditions within the same page, effectively holding constant the difficulty of the task being performed. In this setting, the outcomes capture within-task performance, such as accuracy and speed on a given page, rather than overall persistence or total output across the experiment.

Table 6. Effects of Leader Characteristics on Page-Level Performance

	Output		Accuracy		Speed		Completed	
Leader is a Woman	-0.019	0.000	-0.010	-0.009	-0.046	-0.020	0.051	0.079
	(0.081)	(0.085)	(0.015)	(0.017)	(0.082)	(0.083)	(0.099)	(0.096)
Leader is Gay	-0.059	-0.029	-0.018	-0.013	-0.061	-0.033	0.073	0.042
	(0.079)	(0.073)	(0.016)	(0.014)	(0.081)	(0.077)	(0.079)	(0.071)
Woman Leader × Gay Leader	0.020	-0.038	0.013	0.000	0.050	0.005	-0.096	-0.064
	(0.121)	(0.121)	(0.024)	(0.025)	(0.121)	(0.123)	(0.135)	(0.131)
Page FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clustered SE (Participant)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Participant Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	4,741	4,163	4,515	3,977	4,735	4,158	4,741	4,163

Standard errors clustered at the participant level in parentheses
All models include page fixed effects
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Across all specifications, the page-level results mirror the participant-level findings. There are no statistically significant relationships between leader gender or sexual orientation and any of the performance outcomes, including output, accuracy, speed, or completion. The estimated coefficients are small in magnitude and unstable in sign across specifications. These findings reinforce the conclusion that leader identity has limited and inconsistent effects on follower behavior, both in aggregate performance and in within-task execution.

1.5.3 Heterogeneity By Participant Identity and Beliefs. To explore potential mechanisms, I re-estimate the main specifications separately by participant gender and sexuality. I focus my discussion on leader evaluations rather than performance outcomes given that there is little evidence that the null effects of identity on performance mask interesting heterogeneity (see Appendix Tables A.14, A.15, A.6, and A.7 for the performance heterogeneity results).

Across subgroups, the most consistent patterns show up in leader effectiveness and typicality rather than in any of the performance measures. When I split by participant gender, female leaders tend to be rated more favorably, especially by female participants. Among women, female leaders receive higher ratings on both effectiveness and typicality, with the latter showing clearer statistical support. In contrast, male participants do not show much of a systematic response to leader gender on either dimension. This suggests that the positive evaluation effects in the main results are largely being driven by women's responses rather than reflecting a general preference across the full sample.

Table 7. Effects of Leader Gender and Sexuality on Leader Evaluations by Participant Gender

	Effectiveness		Typicality	
	Men Only	Women Only	Men Only	Women Only
Leader is a Woman	0.090	0.237.	0.202	0.421**
	(0.179)	(0.138)	(0.202)	(0.131)
Leader is Gay	-0.030	0.038	-0.231	0.206
	(0.333)	(0.166)	(0.227)	(0.145)
Woman Leader × Gay Leader	0.064	0.088	0.224	-0.456*
	(0.410)	(0.202)	(0.353)	(0.221)
Controls	Yes	Yes	Yes	Yes
N	97	207	97	211

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

A similar pattern emerges when splitting by participant sexuality. Among straight participants, female leaders are again rated as more effective and more typical, lining up with the main results. Among gay participants, the estimates are noisier and less consistent, although there is some suggestive evidence that gay leaders may be viewed as more typical. The interaction between leader gender and sexuality tends to pull typicality ratings down, particularly among women, which

suggests that gay women leaders may be seen as less prototypical in the nonprofit setting.

Table 8. Effects of Leader Gender and Sexuality on Leader Evaluations by Participant Sexuality

	Effectiveness		Typicality	
	Straight Only	Gay Only	Straight Only	Gay Only
Leader is a Woman	0.286*	0.023	0.290*	0.515.
	(0.122)	(0.239)	(0.122)	(0.268)
Leader is Gay	-0.010	0.207	0.022	0.403
	(0.180)	(0.309)	(0.142)	(0.284)
Woman Leader × Gay Leader	-0.046	0.279	-0.221	-0.404
	(0.241)	(0.305)	(0.239)	(0.349)
Controls	Yes	Yes	Yes	Yes
N	217	87	218	90

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

I also investigate these effects by the underlying heteronormative beliefs of participants. Low scores of heteronormativity would indicate beliefs of gender egalitarianism and positive views of non-straight sexualities. I find in Table 9 under a median split of the distribution of participant beliefs that those that fall in the low category drive the higher typicality rating of women leaders that does not extend to gay women leaders. While the results are noisy, there is evidence in Appendix Tables A.8 and A.9 that followers with high level of heteronormative beliefs perform more slowly and less accurately for gay male leaders. However, this sample skews toward low levels of heteronormative belief and as such these findings should be treated as suggestive pending replication in more ideologically diverse samples.

1.6 Discussion

A central question of this study is whether leader identity shapes how followers allocate effort and perform on assigned tasks. The evidence provides

Table 9. Effects of Leader Gender and Sexuality on Leader Evaluations by Heteronormativity

	Effectiveness		Typicality	
	Low Hetnorm	High Hetnorm	Low Hetnorm	High Hetnorm
Leader is a Woman	0.172	0.236.	0.459**	0.226
	(0.139)	(0.142)	(0.141)	(0.141)
Leader is Gay	0.112	0.127	0.203	-0.024
	(0.152)	(0.143)	(0.145)	(0.141)
Woman Leader × Gay Leader	0.013	0.040	-0.526**	-0.012
	(0.195)	(0.208)	(0.200)	(0.212)
Controls	Yes	Yes	Yes	Yes
N	211	212	217	210

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

little support for this relationship. Across specifications, differences in both task completion and accuracy by leader gender are small and statistically indistinguishable from zero. Likewise, leader sexuality has no meaningful effect on follower performance, with coefficients that are both economically small and statistically insignificant. Evaluations of leaders, however, reveal differences in how leaders are perceived based on their identity. In particular, women leaders tend to receive higher ratings in effectiveness and typicality than male leaders, with this pattern driven primarily by female participants. Strikingly, this boost in ratings does not extend to gay women.

These results suggest a disconnect between stated beliefs and observed behavior. While leader identity has little measurable impact on follower performance, it does influence how leaders are evaluated. This divergence suggests that perceptions of leadership may not translate into meaningful differences in task execution, even when participants express systematic differences in how they view leaders across identity dimensions.

Interpreting this disconnect requires some consideration of what the experimental setting can and cannot capture. The use of a biography and a

personal message from the leader to convey identity grounds the manipulation in a richer and more believable portrayal than simply stating a leader's gender and sexual orientation outright. Yet, leadership is fundamentally a relationship that is built through repeated human interaction. Even a carefully constructed vignette cannot fully replicate the experience of being led, and so while these findings speak to how people respond to perceived leader identity in a controlled setting, they do not capture the ways in which identity intersects with leadership style and interpersonal dynamics in real organizational contexts.

In many organizations, promotion and advancement decisions rely heavily on written evaluations, peer feedback, and subjective assessments of leadership ability. If these stated evaluations are systematically shaped by identity-based perceptions that are only weakly related to actual performance, then they may provide a noisy or even misleading signal of underlying effectiveness. As a result, organizations that rely heavily on such feedback may inadvertently reward perceived leadership qualities rather than realized outcomes. These findings suggest value in incorporating objective performance measures alongside subjective evaluations when making leadership decisions, as well as carefully considering how evaluation systems may amplify or dampen identity-based biases.

An important direction for future research is to directly compare subjective evaluations, such as written feedback or peer ratings, with objective performance metrics in both experimental and field settings. Doing so would help identify when and why these measures diverge, and whether certain evaluation frameworks systematically amplify identity-based perceptions. More importantly, this comparison allows for an assessment of the economic consequences of these differences. If leadership positions are allocated based on subjective evaluations

that do not align with true performance, organizations may incur efficiency losses by misallocating talent. Disentangling these relationships is critical for quantifying the gains or losses associated with how leadership roles are assigned and for designing evaluation systems that better promote efficient outcomes.

CHAPTER II

ARTIFICIAL AGING AND FACE MORPHING: TESTING THE DOUBLE STANDARD OF AGING HYPOTHESIS

This chapter was co-authored with Mary Kite. Mary developed the original project idea, and we both contributed to the data collection, analysis, and writing.

2.1 Introduction

Older adults are often perceived as warm but incompetent, a pattern that holds across cultures (Chan et al., 2012; Cuddy et al., 2005). At the same time, aging is strongly associated with declines in perceived physical attractiveness, with older faces described using explicitly negative language (Henss, 1991; Kite et al., 1991). These perceptions matter. A large literature shows that physical attractiveness shapes social and economic outcomes, influencing everything from earnings to perceived competence (Hamermesh, 2011; Mobius & Rosenblat, 2006). Taken together, this suggests that age-based declines in attractiveness are not merely aesthetic, but may translate into meaningful social and economic disadvantages.

Aging is universally experienced, yet its consequences are profoundly influenced by gender. Older women experience more disadvantages than older men, both economically and socially. For example, they face higher rates of poverty and financial insecurity, reflecting the accumulation of gender disparities in earnings and career advancement (Calasanti & Slevin, 2003). In addition, evidence from hiring suggests that older women are more likely to face discrimination, particularly as they approach retirement age, ostensibly because they are perceived as less competent or less desirable by employers (Neumark et al., 2017). These disparities are not confined to the labor market. Older women also face disadvantages

in marriage markets (England & McClintock, 2009; Fileborn et al., 2015), are underrepresented or narrowly portrayed in media (Lemish & Muhlbauer, 2012), and encounter biases in healthcare settings (Chrisler et al., 2016). Sontag (1972) coined the phrase "double standard of aging" to describe these observations suggesting that men gain status and dignity with age, whereas women become perceived as less attractive and are consequently devalued.

Evidence for the double standard of aging is most consistent when individuals directly evaluate physical appearance. Across a range of studies, signs of aging such as wrinkles and gray hair are viewed as less attractive in women than in men (Berman et al., 1981; Deutsch et al., 1986; Harris, 1994; Henss, 1991; Hummert et al., 1997). These differences extend beyond external evaluations. Women also report experiencing aging differently, frequently describing feelings of social invisibility and increased anxiety tied to societal standards of youthful appearance (Clarke & Korotchenko, 2010; Grabe et al., 2008; Ward & Holland, 2011; Wilcox, 1997). In response, many engage in appearance management, such as dyeing their hair, even when doing so conflicts with their sense of authenticity (Cecil et al., 2022). Taken together, these patterns suggest that the consequences of aging are shaped not only by age itself, but by when individuals come to be perceived as old. To the extent that this transition occurs earlier for women, the social and economic costs associated with aging are likely to be disproportionately borne by them.

We study the transition between adulthood and elderhood to offer new insights into the double standard of aging. Existing work shows that women are perceived as reaching middle and elder age sooner than men (Kogan, 1979) and are thought to enter the prime of life at younger ages compared to men (Harris,

1994; Seccombe & Ishii-Kuntz, 1991; Zepelin et al., 1986). While researchers have tried to pin down a typical age at which an individual perceived as old, this goal is fraught with complications. For example, Wettstein et al. (2024) find that older adults place the onset of old age later in life, while younger individuals report earlier thresholds. Researchers that have used facial features to establish the elderhood threshold have discovered that raters are mixed in their accuracy when assessing age (Burt & Perrett, 1995; Christensen, 2009; Dehon & Brédart, 2001; Norja, 2021; Willner & Rowe, 2001). Specifically, raters are most accurate for middle-aged faces, tend to underestimate older ages, and overestimate the age of young women. These studies use static photographs or surveys to estimate when individuals are perceived as old. They also typically used sets of photographs the researchers selected or had created for their study and so reflect the investigators' ideas about what representative older and younger people look like (Deutsch et al., 1991; Hummert et al., 1997). To address these limitations, we developed a new methodology; specifically, we take a more active approach by requiring participants to stop an automatically advancing video to identify when a target face "looks old." We call this approach "face morphing", and use it to compare how the perceived transition to elderhood occurs differentially by race and gender.

We find that when participants start with a young face ("target") they stop the video later for female targets versus male targets. Considered in isolation, this result is inconsistent with the double standard of aging. However, participants' ratings of the morph anchors (the static young and old starting points of each morph sequence) varied systematically by target gender and race: young anchors of female morphs were rated as older than young anchors of male morphs, and old-anchor ratings differed by gender and varied across racial groups. The morph

sequences therefore did not span an equivalent perceived age range across faces within participants, and thus stop time cannot be interpreted as a direct measure of perceived age at the point of stopping.[1]

We use these anchor ages to interpolate a participant-specific estimated age for each face at every point in time during a morph sequence. Using this measure, female faces were judged old at a younger estimated age than male faces, with White women judged old at the youngest estimated age and Latino and Black men at the oldest. This pattern is consistent with the double standard of aging. The gender effect on estimated age was not attenuated by adjustment for rated attractiveness of the aged face, self-reported tiredness, or self-reported focus, in contrast to suggested pathways for the double-standard of aging in the literature. We report results for both stopping time and estimated transition age, and discuss the implications of their divergence for the interpretation of morph-based age-perception paradigms.

2.2 Background

People's beliefs about older adults reflect a combination of positive and negative characteristics; older adults are seen as warm, likable, and experienced, but they are also thought to be incompetent, rigid, and complaining (Cuddy & Fiske, 2002; Narayan, 2008; Rupp et al., 2005). The view that older adults are "weak but wonderful" (Bugental & Hehman, 2007, p. 184) emerges across a number of cultures, including both collectivist and individualist societies. For example, respondents from 26 countries reported that older adults are warmer, but less active, than younger people (Chan et al., 2012), and cross-cultural research shows

[1]While the heterogeneity here is less systematic, this also means that the morph sequences did not necessarily span an equivalent perceived age range across participants within faces, highlighting the value of a methodology where the perceived ages of each anchor are measured.

that older adults are consistently viewed as warm but incompetent (Cuddy et al., 2005). When people evaluate older adults' physical appearance, they are less generous. Older adults are generally rated as less attractive than younger adults (Canetto et al., 1995; Henss, 1991; Kissler & Bäuml, 2000), and their appearance is described by terms such as wrinkled, hard of hearing, slow, and in decline (Kite et al., 1991). Terms commonly associated with old age reinforce this perceived unattractiveness, including crone, fossil, withered, wizened, and wrinkled (Palmore, 1999).

The negative consequences of being perceived as physically unattractive are well documented. A meta-analysis of 76 studies on the "what is beautiful is good" stereotype showed that, on average, physically unattractive people are viewed as less socially competent than physically attractive people (Eagly et al., 1991). Physically unattractive people also report fewer friends, poorer social skills, and a less active sex life (Rhodes et al., 2005). In the labor market, attractive workers' performance is more highly rewarded than equivalent performance by unattractive workers (Mobius & Rosenblat, 2006), and attractive workers earn a measurable wage premium over their less attractive counterparts (Hamermesh, 2011; Scholz & Sicinski, 2014). Notably, attractiveness is interconnected with age: unattractive faces are judged to be older than attractive ones (Wernick & Manaster, 1984), and unattractive people in their 60s and beyond are judged to have less favorable personalities, fewer successful life experiences, and fewer occupational achievements than more attractive people of similar ages (Hummert et al., 1997; Johnson & Pittenger, 1984). Because perceived attractiveness and perceived age are tightly coupled in raters' own judgments, any study of how people categorize faces as old has to take seriously the visual information on which that categorization is based.

2.2.1 The Double Standard of Aging.

The cost of aging is not borne equally. Physical attractiveness is perceived to be more important for women than for men (Chrisler et al., 2015), and Hamermesh (2011) found that judgments of women's appearance were more polarized than judgments of men's appearance. Women, more so than men, are expected to attend to their appearance by doing things such as wearing make-up and staying thin (Saucier, 2004). If these expectations are met, women are seen as more attractive, competent, and likable than otherwise (Etcoff et al., 2011). Harris (1994) found that most visible signs of aging, such as wrinkles and gray hair, were viewed more negatively for women than men, while baldness was viewed as less attractive for men. These findings suggest that women who age might be judged particularly harshly. Sontag (1972) coined the phrase "double standard of aging" to refer to this asymmetry, arguing that women's success is measured by how they look while men's is measured by what they do, so that men gain status and dignity with age whereas women are devalued.

When people are asked to name the specific ages at which middle- and old-age begin, their responses line up with Sontag's account. Women are perceived to attain middle age and elder status sooner than men (Kogan, 1979) and to reach the prime of life earlier (Harris, 1994; Seccombe & Ishii-Kuntz, 1991; Zepelin et al., 1986). A U.S. national survey found women were categorized as old at 55 compared to 60 for men (Seccombe & Ishii-Kuntz, 1991). A more recent survey of older U.S. adults showed that male and female respondents agreed men were old at 70, but disagreed about the corresponding threshold for women. Men placed it at 68, a double standard favoring men, while women placed it at 75, a double standard favoring women (AARP, 2014). The divergence between respondents is

itself a reminder that these estimates of when elderhood begins are self-reported judgments, and may reflect the respondent's position as much as the target's.

Actual behavior is more consistent. In a Canadian sample, 67 percent of women reported using anti-aging products to maintain a youthful appearance even though most also reported that doing so was probably ineffective (Muise & Desmarais, 2010), and in a qualitative study most of the 50- to 70-year-old Canadian women interviewed engaged in beauty work, such as make-up, hair color, and surgery, to push back against society's views of older women (Clark & Griffin, 2008). Harris (1994) found that older women in the U.S. were much more likely than older men to engage in such behaviors. These are non-trivial investments of time and money, and they are consistent with women facing a steeper appearance-related penalty as they age. Evidence from media representation points the same direction. Older women are underrepresented in U.S. movies relative to older men (Bazzini et al., 1997; Lauzen & Dozier, 2005), and when they do appear the roles skew white, high-income, and attractive (Lemish & Muhlbauer, 2012). Older male actors appear more frequently, including in romantic roles, and are routinely paired with much younger actresses (Saucier, 2004). Older women are likewise underrepresented in advertisements in Japan (Prieler et al., 2011), Great Britain (Kay & Furnham, 2013), India (Raman et al., 2008), and the United States (Lee et al., 2007; Raman et al., 2008).

2.2.2 Direct Tests of the Double Standard Hypothesis. Despite this indirect support, direct tests of the double standard have produced mixed results, and the pattern of that mix is informative. When photographs of actual faces are assessed, older women are usually judged more harshly than older men (Berman et al., 1981; Deutsch et al., 1986; Foos & Clark, 2011; Henss,

1991; Hummert et al., 1997), though some studies find the effect only among younger raters (Foos & Clark, 2011) or among men (Berman et al., 1981; Henss, 1991). When ratings are instead based on a verbal target label such as "75-year-old woman," rather than on a photograph, the double standard favoring men disappears and people rate women's and men's physical appearance similarly (Canetto et al., 1995; Sherman, 1986). In one case, the ratings even tilt in women's favor (Kite et al., 1991). A similar visual-versus-verbal pattern appears on gender-associated traits. Photographs of older women are rated as less feminine than photographs of younger women while men's masculinity does not vary with perceived age (Deutsch et al., 1986), but when age is conveyed by label the gender differences attenuate or disappear (Canetto et al., 1995; Kite et al., 1991; Sherman, 1986).

This gap between photograph-based and label-based findings is important for interpretation. It implies that the double standard is carried by the visual features of an aging face rather than by abstract category knowledge about older women and men, and it means that any credible test of whether women are judged old sooner than men requires visual rather than verbal stimuli. The existing photograph-based literature either compares different younger and older targets or uses personal photographs over time of the person. This introduces possible confounds. Specifically, the measured age effect could be due to between-target differences in identity, expression, and imaging conditions or it could be due to other cultural markers of age such as choice of clothing or hairstyle. Separating the effect of age from those other sources of variation requires stimuli in which a single face is observed across a range of apparent ages, holding everything else constant. Additionally, past studies have used photographs that represent distinct periods in

a target's life (20s, 40s, 60s) which does not as accurately represent the continuous measure that is age.

We develop a method, described in Section 3.3, to address both of these concerns: artificial aging and face morphing. Participants view a continuous young-to-old morph of the same identity and stop the sequence at the point at which they judge the face to have crossed from young to old, permitting a within-identity estimate of the perceptual threshold for elderhood and a direct test of whether that threshold differs by target gender and across White, Asian, Black, and Latino faces.

2.3 Methods

2.3.1 Participants. Participants were a convenience sample of 172 students at a midsize Midwestern public university who took part in the research in partial fulfillment of an introductory psychology course requirement. The average age of participants was 20.9 years (SD = 2.58) and the sample was predominantly White, with 69.6% of participants identifying as such. In terms of gender, 75.6% of participants self-identified as woman, 11.6% as men, and 10.5% as non-binary, with 2.3% of participants not reporting their gender. On average, participants reported moderate levels of tiredness (M = 3.73, SD = 1.61) but with an ability to stay focused throughout the task (M = 3.46, SD = 1.37).

2.3.2 Procedure. Data were collected in person in the laboratory. Upon arrival, participants were randomly assigned to a target race condition (White, Asian, Black, or Latino) and to a starting-point condition (young-to-old or old-to-young), determining which faces they would view and the direction in which the morph sequences would advance. Before beginning the experimental trials, participants completed a practice trial using a cat-to-dog morph sequence

Table 10. Participant Characteristics by Condition

	Asian	Black	Latino	White
N	43	46	35	48
Age	20.5	21.5	20.3	21.0
	(1.04)	(4.42)	(1.19)	(1.73)
% White	0.674	0.682	0.735	0.702
% Man	0.093	0.130	0.143	0.104
% Woman	0.767	0.717	0.771	0.771
% Non-Binary	0.140	0.109	0.057	0.104
Tired	3.23	3.77	3.82	4.06
	(1.59)	(1.48)	(1.53)	(1.75)
Focus	3.60	3.30	3.50	3.45
	(1.50)	(1.32)	(1.31)	(1.38)

Notes: Standard deviations in parentheses.

to familiarize themselves with the task of stopping a sequence at a predetermined threshold.

Participants then completed six morph sequences — three male and three female faces from their assigned racial group — presented in a random order. Each sequence advanced automatically at one frame per second. A researcher recorded the slide number at which the participant stopped each sequence. Following the morph task, participants completed a questionnaire containing the endpoint rating items, task experience items, and demographic questions described in the Measures section.

2.3.3 Measures.

2.3.3.1 Morph Task. The primary outcome measures were collected via the face morphing task. Each morph sequence consisted of 50 images created by linearly interpolating between a young face and an artificially aged version of the same face, with each step incrementing the blend by 2%, such that the sequence

spanned the full range from young to old appearance over 50 frames advancing at one frame per second. Young faces were drawn from the Chicago Face Database (Ma et al., 2015). Artificial aging was applied using the FaceApp application, which transforms a face to approximate an aged appearance; because FaceApp provides no numeric age metric of its own, the aged endpoint faces were pretested to confirm that participants perceived them as aged and to verify that attractiveness ratings were equivalent across young and old versions of each face. Morphing of the young and artificially aged endpoints was performed using the MorphMan application. Examples of the young and aged endpoints used in the task are shown in Figure 5.

Participants were shown six morph sequences — three male and three female faces, all of the same racial group to which they were assigned — presented in a random order. Half of participants were shown sequences that began with the young face and progressed toward the aged face; the remaining half were shown sequences beginning with the aged face and progressed toward the young face. Participants were instructed to stop the sequence at the point at which they judged the face to have crossed the threshold from young to old (or, in the reverse condition, from old to young). Prior to the experimental trials, participants completed a practice trial using a cat-to-dog morph sequence to familiarize themselves with the stopping procedure. A researcher recorded the slide number at which each participant stopped the sequence for each face. The slide number at which the participant stopped the sequence constitutes the first primary outcome measure, referred to throughout as *stop time*.

2.3.3.2 Endpoint Age and Attractiveness Ratings. Following the morph task, participants were shown a static image of each morph endpoint — that is, the young anchor face and the aged anchor face for each sequence they



Figure 3. Example of Young Face



Figure 4. Example of "Old" Face

Figure 5. Examples of Face Morph End Points

had viewed — and asked to provide age and attractiveness ratings for each. Age was assessed with a single open-ended item: *"Estimate the exact age of this person (in years)."* Participants also indicated the perceived race/ethnicity and gender of each face. Attractiveness was assessed by asking participants to rate the person relative to others of the same race, age, and gender in the United States (e.g., relative to other 40-year-old males of a given ethnicity), providing a within-category comparison that controls for the confound between age and attractiveness that would arise from an absolute rating. The age ratings assigned to the young and old endpoints were subsequently used to construct a participant-specific estimate of the perceived age of the face at the point of stopping, via linear interpolation between each participant's own endpoint ratings (see Equation 2.2).

2.3.3.3 Task Experience. Two items assessed participants' subjective experience during the morph task. Focus was measured with the item *"How easy or difficult was it for you to focus on the faces as they changed during the study?"* and tiredness was measured with the item *"How tired did you feel during the study?"* Both items were included as covariates in sensitivity analyses to account for the

possibility that variation in task engagement influenced stop time or perceived age judgments (see Section 2.4.4).[2]

2.3.3.4 Demographics. Following the task, participants completed a brief demographic questionnaire. Race/ethnicity was assessed with a multi-select item allowing participants to select all applicable categories or to self-describe. Gender was assessed with a multi-select item. Age was assessed with a single open-ended numeric item.

2.4 Results

The experimental design comprised one main randomized between-subjects factor — the racial group of the target face (White, Asian, Black, Latino) — and one within-subjects factor — target gender (male, female). Young-to-old vs. old-to-young progression order was also randomized between-subjects for robustness. This combination of between- and within-subjects variation makes the study well-suited for mixed-design analysis of variance modeling (also termed a "split-plot ANOVA"), in which between- and within-subjects variance components are partitioned and allowed to have separate error terms.

All models were estimated as linear mixed-effects models via restricted maximum likelihood (REML) using the `lme4` package in R, with F-ratios and p-values derived from Satterthwaite-approximated degrees of freedom (Bates et al., 2015; Luke, 2017). This specification is formally equivalent to a traditional mixed-design ANOVA while accommodating unbalanced cell sizes and partially missing

[2]Participants were also asked an open-ended item regarding their decision strategy: *"How did you determine when the faces became [old/young]?"* where the bracketed term reflected the participant's assigned condition. Responses to this item were not analyzed quantitatively but provide qualitative context for interpreting individual differences in stopping behavior.

data.[3] All models included two random effects: a participant random intercept and a participant-by-target gender random intercept-shifter.[4] Each experimental factor — target race, target gender, and their interaction — was tested using Type III sums of squares, such that each factor is evaluated after accounting for all remaining factors in the model. The effect sizes of these experimental conditions are reported as partial eta-squared ((η_p^2)), which expresses the variance attributable to each factor as a proportion of the variance remaining after all other factors are accounted for, placing the between-subjects factor (target race) and within-subjects factor (target gender) on a comparable scale.

The decision to frame the analysis as a mixed-design ANOVA rather than as a regression model reflects the structure of the research question and norms of the existing literature on this topic rather than any substantive difference in the underlying model assumptions. The two methods are cases of the general linear model

$$Y = XB + U \tag{2.1}$$

[5] and, with appropriate coding of the categorical explanatory variables, produce the same underlying model fit and equivalent test statistics. The approaches differ in what they emphasize as output. A regression specification emphasizes prediction. Its primary output is a set of coefficients that describe the direction

[3] We re-estimate all the main models from this section using classical split-plot ANOVA in Appendix B.2.

[4] Rather than one intercept per participant, each participant gets a family of intercepts — one per target gender level — and the by-participant-by-target-gender term captures how much those shift relative to the participant's overall intercept.

[5] Where $\mathbf{Y}$ is a matrix of multivariate measurements (each column representing a dependent variable), $\mathbf{X}$ is a design matrix of observations on independent variables (each column representing an independent variable), $\mathbf{B}$ is a matrix of parameters to be estimated, and $\mathbf{U}$ is a matrix of errors assumed to be uncorrelated across measurements and following a multivariate normal distribution, $\mathbf{U} \sim \mathcal{MN}(0, \Sigma)$.

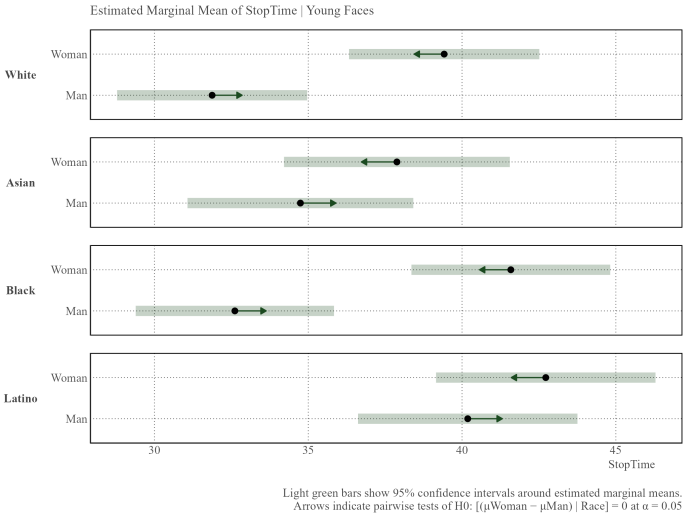
and magnitude of each predictor's relationship with the outcome, representing the estimated difference between each non-reference level and the reference category, holding other predictors constant. The ANOVA framing, by contrast, emphasizes hypothesis testing. It partitions the total variance in the outcome into components attributable to each factor (target race, target gender, their interaction, and residual error) and uses an F-statistic to determine whether each component is larger than would be expected by chance. This produces a single test per factor, asking whether that factor accounts for meaningful variance in age evaluations across all its levels simultaneously, rather than reporting differences level by level. The individual group means, and the distances between them, are background quantities in this framework; they become the object of analysis only in subsequent post-hoc procedures such as Tukey's HSD or Scheffé's method, which conduct pairwise comparisons of the factor-level means. The ANOVA framing is, in our view, more readily interpretable given the categorical structure of the design.

2.4.1 Morph Stop Time. The first set of analyses examines whether the point in the morph sequence at which observers stopped the face — hereafter *stop time* — differed as a function of target race and gender. Models were estimated separately for participants who initiated the task from the young-face endpoint and those who initiated from the old-face endpoint. Each sequence was 50 images.

We observe a large main effect of target gender ($F(1,\ 77.24) = 106.59$, $p < 0.001$, $\eta_p^2 = 0.58$), indicating that stop time differed substantially between male and female faces. This gender effect is heterogeneous by race, with a significant Race × Gender interaction ($F(3,\ 77.28) = 8.81$, $p < 0.001$, $\eta_p^2 = 0.25$), indicating that the magnitude of the gender difference in stop time varied across racial groups.

Figure 6 contains the point estimate of the marginal mean (dot), a 95% confidence interval (shaded region), and pairwise comparisons of the mean stop time of male versus female targets within a given target race using Tukey's HSD adjustment (arrows). The arrows represent a visual representation of these pairwise comparisons and can be interpreted such that any overlay implies we fail to reject the null hypothesis at the 5% level. The pairwise comparisons reveal that female faces were stopped later in the morph sequence (measured in terms of images, 1-50) than male faces across all racial groups and that this represents a statistically significant difference at the $\alpha = 0.05$ level. The gender gap was most pronounced for White faces (men: mean = 31.9, 95% CI [28.8, 35.0]; women: mean = 39.4, 95% CI [36.3, 42.5]; $\Delta = 7.5$) and Black faces (men: mean = 32.6, 95% CI [29.4, 35.8]; women: mean = 41.6, 95% CI [38.4, 44.8]; $\Delta = 9.0$).

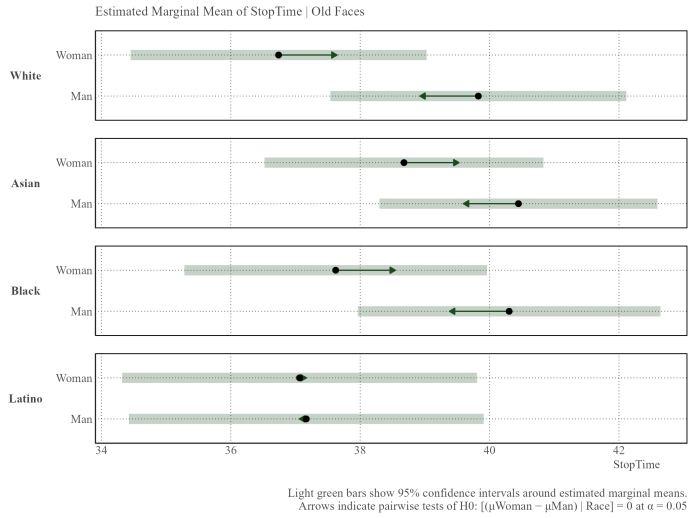
Figure 6. Estimated Marginal Means of Stop Time By Condition



Among participants who initiated the task from the old-face endpoint, we again estimate a significant main effect of gender, though it is attenuated relative to the young-face condition ($F(1,\ 82.99) = 16.23$, $p < 0.001$, $\eta_p^2 = 0.16$).

Neither the main effect of race ($F(3,\ 83.00)\ =\ 0.83$, $p\ =\ 0.481$, $\eta_p^2\ =\ 0.03$) nor the Race × Gender interaction ($F(3,\ 83.00)\ =\ 1.69$, $p\ =\ 0.175$, $\eta_p^2\ =\ 0.06$) are statistically significant at the 5% level. Figure 7 shows the estimated marginal means when participants see an old-face first. We see that male faces were stopped at a later point in the morph sequence than female faces across most racial groups (e.g., White men: $M = 39.8$, 95% CI [37.5, 42.1]; White women: $M = 36.7$, 95% CI [34.5, 39.0]). Since participants are now stopping the videos when they view the target as looking young, this implies the same result as the previous group: that women experience the transition between old and young age at a later point in the morph sequence than men.

Figure 7. Estimated Marginal Means of Stop Time By Condition



These findings are inconsistent with predictions derived from the double standard of aging (Kite et al., 1991; Sontag, 1972). Under that framework, female faces would be expected to cross the threshold for "old" appearance at an earlier point in the young-old morph sequence than male faces, and a later point in the old-young morph sequence, reflecting a harsher standard applied to visible aging

in women. The present data demonstrate the opposite pattern: across all racial groups, observers permitted female (male) faces to advance further along the aging (de-aging) continuum before stopping the morph.

2.4.2 Stop Time Versus Perceived Age. While stop time provides a direct observation of the point in facial feature space at which observers judged a face to have crossed some threshold of "old", it is unclear exactly what it means for age. Stop time is a perfect proxy for target age only under the assumption all participants agree on how they perceive the age of the start and end faces within the morphing sequence.[6]

Figure 8. Distribution of Ratings of Anchor Faces for Morphs

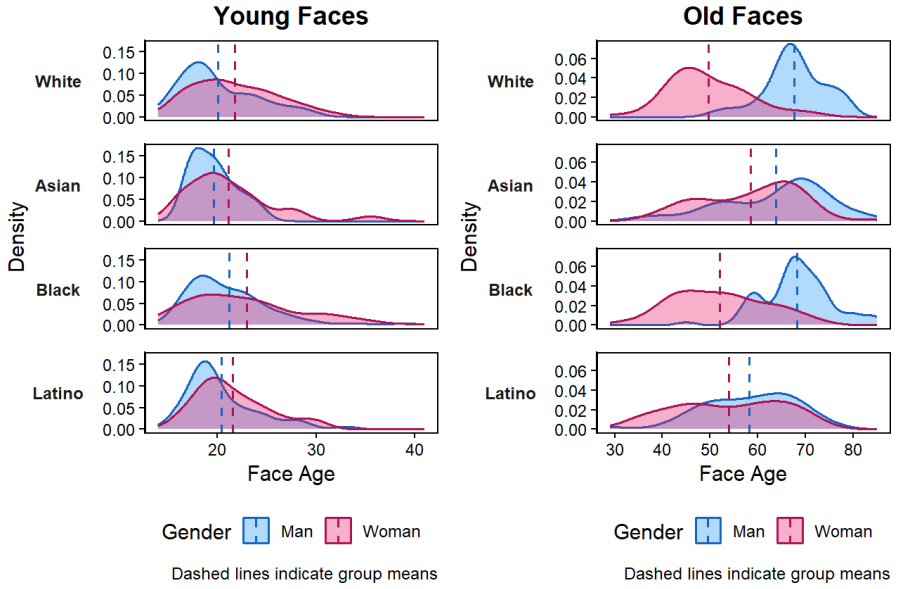


Figure 8 illustrates that this assumption does not hold in the present data. The left panel displays the distribution of age ratings assigned to the *young* anchor of each morph sequence—that is, how old participants judged the youngest face

[6]While this is a necessary condition, it is not sufficient. Sufficiency requires that all participants agree on the time-path of face ages throughout the morph. We explore this issue further in Appendix C.

in the sequence to be—while the right panel displays the equivalent ratings for the *old* anchor. Critically, this variability is not random noise: the distributions differ systematically as a function of the conditions under investigation, namely the gender and racial group of the target face. For the young anchor (left panel), female faces received consistently higher age ratings than male faces across all racial groups, indicating that participants perceived the youngest point of the female morph as older than the equivalent point of the male morph. For the old anchor (right panel), the distributions are far wider, spanning approximately 30 to 85 years, and the between-gender gap is considerably more pronounced and variable across racial groups—White faces in particular show marked separation, with male old anchors rated as substantially older than female old anchors.

This is a key observation for the interpretation of *stop time* as an outcome measure. If the anchor ages themselves vary as a function of target gender and race, then group differences in stop time may reflect, at least in part, differences in how the morph sequence was perceptually scaled rather than differences in the age at which observers judged a face to appear old. Disentangling these two sources of variance requires an outcome measure that accounts for each participant's own perception of the endpoints, which is the motivation for the participant-specific age estimation approach adopted in the following section.

2.4.3 Participant-Specific Estimated Target Age. To obtain a measure of perceived age, we create a participant-specific estimate for each morph by interpolation between their ratings of the age of the faces which constitute the youngest and oldest within a morph. For a participant who assigned age y to the young endpoint and age o to the old endpoint of a given face, and who stopped the

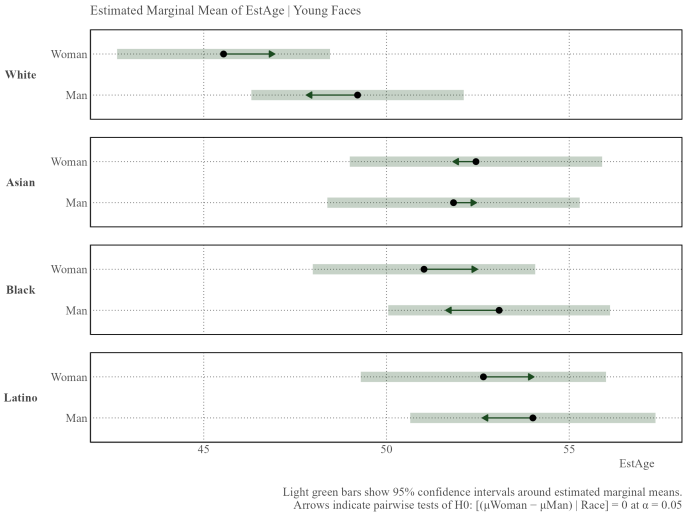
morph at image t, the estimated age at the point of stopping is computed as

$$\widehat{\text{Age}} = y + t\,(\frac{o - y}{50}) \tag{2.2}$$

Under this formulation, perceived age is modeled as advancing in equal increments across the full stimulus range.

Applying the same mixed-effects model structure to the estimated age outcome, significant main effects of both race ($F(3,\ 76.72) = 3.83$, $p = 0.013$, $\eta_p^2 = 0.13$) and gender ($F(1,\ 77.27) = 4.40$, $p = 0.039$, $\eta_p^2 = 0.05$) remain in the young-face-starting condition. However, Figure 9 shows that among the estimated marginal means, White women are estimated as youngest at the point of stopping (mean = 45.5, 95% CI [42.6, 48.5]) and Latino men as oldest (mean = 54.0, 95% CI [50.7, 57.4]). When age is estimated as described above, the only target gender difference that remains statistically significant is for white targets.

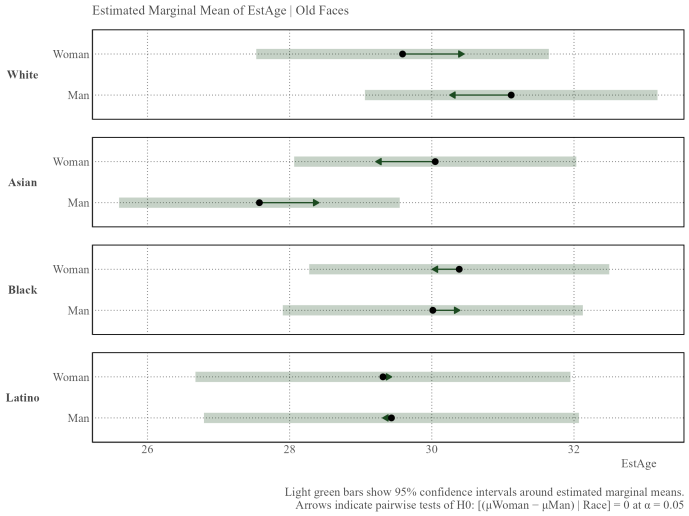
Figure 9. Estimated Marginal Means of Estimated Age By Condition



In the old-face-starting condition, neither race ($F(3,\ 79.98) = 0.60$, $p = 0.620$, $\eta_p^2 = 0.02$) nor gender ($F(1,\ 412.28) = 0.40$, $p = 0.528$, $\eta_p^2 < 0.01$) featured significant main effects. However there is a significant Race × Gender interaction

($F(3,\ 412.30) = 3.66$, $p = 0.013$, $\eta_p^2 = 0.03$), reflecting divergence in estimated age across racial groups as a function of target gender. The Race × Gender interaction, as seen in Figure 10, is driven by Asian Man targets being rated as much younger when they reach the threshold between adulthood and elderhood than other races.[7]

Figure 10. Estimated Marginal Means of Estimated Age By Condition



2.4.4 Covariate-Adjusted Models. To assess whether the primary effects of target race and gender on *stop time* and $\widehat{\text{Age}}$ are robust to individual differences in task engagement and stimulus perception, we re-estimate each model with three participant-level covariates: rated attractiveness of the target face when old, self-reported tiredness at the time of participation, and self-reported focus during the task.

In the young-face-starting condition, the inclusion of covariates meaningfully alters the pattern of results relative to the unadjusted model for *stop time*.
The main effect of target race is more precisely estimated with covariates

[7]The old-face estimated age model converged with a singular fit, indicating that the variance of at least one random effect was estimated at the boundary of the parameter space. Estimates are reported for completeness but should be interpreted with caution.

($F(3,\ 66.25) = 3.39$, $p = 0.023$, $\eta_p^2 = 0.13$). The main effect of gender remains large and statistically significant ($F(1,\ 78.80) = 98.70$, $p < 0.001$, $\eta_p^2 = 0.56$), as does the Race × Gender interaction ($F(3,\ 78.04) = 9.40$, $p < 0.001$, $\eta_p^2 = 0.27$). The adjusted estimated marginal means indicate that the gender gap in *stop time* persisted across all racial groups, with the largest disparities observed for Black faces (men: mean = 34.9, 95% CI [30.4, 39.3]; women: mean = 44.1, 95% CI [39.7, 48.6]; $\Delta = 9.2$) and Latino faces (men: mean = 42.8, 95% CI [38.0, 47.6]; women: mean = 45.2, 95% CI [40.4, 50.0]; $\Delta = 2.4$). Notably, the covariate-adjusted means for Latino men shift upward relative to the unadjusted estimates, accounting for the emergence of a significant race main effect. In the old-face-starting condition, the effects are unchanged from the model without covariates.

When considering how the inclusion of covariates impacted group differences in $\widehat{\text{Age}}$, the effects are largely preserved. The main effect of race remains statistically significant ($F(3,\ 64.99) = 4.56$, $p = 0.006$, $\eta_p^2 = 0.17$) and is somewhat larger than in the unadjusted model ($\eta_p^2 = 0.13$), and the main effect of gender remains statistically significant as well ($F(1,\ 78.53) = 3.98$, $p = 0.049$, $\eta_p^2 = 0.05$). The adjusted estimated marginal means reveal that White women were assigned the youngest $\widehat{\text{Age}}$ at the point of stopping (mean = 47.4, 95% CI [43.5, 51.4]) while Black men and Latino men were assigned the oldest (mean = 56.4, 95% CI [52.2, 60.6] and mean = 56.4, 95% CI [51.9, 60.9], respectively). Compared with the unadjusted model, the covariate-adjusted means shifted upward across all cells, for White men in particular, as a result of adjustment for attractiveness and tiredness. As with *stop time*, the old-face-starting condition results for $\widehat{\text{Age}}$ were unchanged by the inclusion of covariates.

Across all four models, controlling for face attractiveness did not attenuate the gender effect on either stop time or estimated age. Effect sizes for the gender main effect are virtually unchanged before and after adjustment, with no reduction in precision. These findings are inconsistent with the theoretical prediction that gender differences in perceived aging are driven by the attractiveness penalty associated with aging in women. Instead, the results suggest that observers' age-related judgments of female faces reflect something distinct from attractiveness evaluations, a pattern that complicates straightforward attractiveness-mediation accounts of the double standard of aging.

2.5 Conclusion

This study examines the perceived threshold between young and old across gender and race using a face-morphing paradigm with two outcome measures: stop time, the morph position at which participants judged the face to have become old, and estimated age, a participant-specific measure derived by linear interpolation between each participant's ratings of the morph anchors. The two measures produced effects in opposing directions. In the young-face-starting condition, stop time was later for female than male faces across all four racial groups, whereas estimated age was lower for female than male faces, with White women judged old at the youngest estimated age and Latino and Black men at the oldest. The estimated-age pattern is consistent with the double standard of aging (Kite et al., 1991; Seccombe & Ishii-Kuntz, 1991; Sontag, 1972) and with prior evidence from verbal age judgments (AARP, 2014; Seccombe & Ishii-Kuntz, 1991) and evaluative ratings of photographs (Berman et al., 1981; Deutsch et al., 1986; Henss, 1991).

The divergence between the two measures is attributable to systematic variation in participant ratings of the morph anchors. Young anchors of female

morphs were rated as older than young anchors of male morphs, and old-anchor ratings varied by gender and race. Consequently, the morph sequence did not span an equivalent age range across conditions: a later stop time on a female morph could correspond to a lower interpolated age than an earlier stop time on a male morph. The source of this variation in anchor ratings cannot be identified from the present data and may reflect properties of the source photographs, the aging transformation applied by the stimulus-generation software, the perception of identical images under different target attributions, or some combination of these.

Adjustment for rated attractiveness of the aged face did not attenuate the gender effect on estimated age. This result is inconsistent with a strong form of the hypothesis that gendered ageism in age categorization is mediated by an attractiveness penalty (Chrisler et al., 2015; Harris, 1994). Two interpretations are compatible with the data: perceptual age categorization and attractiveness evaluation may be partially dissociable components of gendered ageism, or the single post-task attractiveness item employed here may be insufficiently sensitive to detect a mediating relationship. Designs incorporating more extensive attractiveness measurement, or experimental manipulation of attractiveness, would be required to distinguish these accounts.

Several limitations constrain the generalizability of these findings. The sample was drawn from a single Midwestern university and was predominantly composed of young women; prior work indicates that the magnitude and direction of double-standard effects can vary with observer age and gender (AARP, 2014; Berman et al., 1981; Foos & Clark, 2011). The stimulus set comprised three male and three female faces per racial group, which does not rule out stimulus-specific contributions to the observed group-level effects. The aged endpoints

were generated by a commercial application whose underlying aging model is not publicly documented and may encode gender-specific aging trajectories; the anchor-rating data are consistent with this possibility but do not establish it. Finally, the old-face-starting condition produced smaller effects, which may not be robust to broader assumptions about the functional form of the age-morph mapping, as documented in Appendix C. Future work could consider whether reverse aging stimuli could serve to inoculate individuals against biases in perceived aging.

CHAPTER III

THE IMPACT OF LEGALIZED SPORTS BETTING ON CHILD MALTREATMENT: EVIDENCE FROM CHILD PROTECTIVE SERVICES REPORTS

3.1 Introduction

Child maltreatment remains a pervasive public health challenge in the United States, with nearly four million referrals and 550,000 confirmed victims each year (U.S. Department of Health & Human Services, 2024). The social costs of adverse childhood experiences are immense: victims face elevated risks of mental illness, chronic disease, substance abuse, and premature death (Hughes et al., 2017; Vinnerljung et al., 2015; Widom et al., 2012). Maltreatment also disrupts education, increases contact with the criminal justice system (Berger et al., 2016), and can perpetuate cycles of disadvantage across generations (Currie & Spatz Widom, 2010; Font & Maguire-Jack, 2020). Economic hardship and parental conflict, including intimate partner violence, are among the strongest predictors of abuse and neglect (Renner & Slack, 2006; Sedlak et al., 2010). These factors magnify stress and instability within households, reducing parents' capacity to provide consistent care and supervision (Conger & Conger, 2002; Conger & Donnellan, 2007). Given these established risk factors, policies that affect household finances and parental relationships may have unintended consequences for child welfare. This paper investigates one such policy —the recent legalization of sports betting in much of the United States.

In 2018, the United States Supreme Court effectively struck down the Professional and Amateur Sports Act, paving the way for states to legalize sports betting for the first time since 1992. Since then, 40 states have legalized some

form of sports betting and 34 have allowed sports betting via the internet. Since then, the market has grown rapidly, with more than $100 billion wagered in 2024 (American Gaming Association, 2023; Baker et al., 2024; Hoffer, 2024; Hollenbeck et al., 2024).[1] The growing popularity of sports betting raises concerns about the prevalence of "problem gambling" and its impact on households.

For most, betting is a harmless form of entertainment that leads to small financial losses on average. However, problem gamblers may experience far more severe outcomes, such as household financial stress and the disruption of family dynamics (Dowling et al., 2016; Gabellini et al., 2022; Griffiths et al., 2009; Shaffer & Korn, 2002). Previous work on sports betting legalization indicates the size of the population that experiences these more severe betting-related outcomes is increasing. For example, states that legalize sports betting, particularly mobile sports betting, exhibit a greater number of help-seeking posts on online betting support groups (Van Der Maas et al., 2022), worsening consumer financial health specifically through lower credit scores, increased rates of bankruptcy, and lower rates of investing (Baker et al., 2024; Hollenbeck et al., 2024), and the amplification of sports outcome-related intimate partner violence (Matsuzawa & Arnesen, 2024). Both financial stress and intimate partner violence are known risk factors for child maltreatment, and they both appear to increase in response to the legalization of sports betting, yet the impacts of this policy on child welfare remain unexplored.

In this paper, I explore the causal effect of legal sports betting on the reporting of child maltreatment. To estimate this relationship, I use Child Protective Services (CPS) case data retrieved from the National Child Abuse and Neglect Data System (NCANDS), and leverage variation from the staggered

[1]In the same year, sports books brought in $13.7 billion in revenue of which more than 90% was generated by online betting (American Gaming Association, 2025)

adoption of legal sports betting across states. Using two-way fixed effects Poisson models, and pooling all law changes that legalize sport betting in some form, I find that on average states experience between a 5-7% increase in the number of screened-in reports of child maltreatment (more details on the report handling and screening process are in Section 3.3). When I split out the legalization of mobile (online) sports betting and in-person sport betting, I find that the overall effects are driven almost entirely by the former. I explore the heterogeneity of this effect in terms of county characteristics, child demographics, report source, and type of allegation.

Other work suggests two likely pathways through which these effects could manifest. One is an amplification of the impact of sports-related emotional cues following upset losses. Looking at quarterly reports of abuse, rather than any reports of maltreatment, I find that the isolated effect of mobile sports betting legalization is large (13% increase) and precisely estimated. Though, when I consider the effect of upset National Football League losses, following Card and Dahl (2011), I find that while upset losses lead to an increase in reports, the effect is not exacerbated by legal sports betting. Another pathway is long-run financial hardship, which the family stress model suggests creates frictions in family dynamics leading to increased chance for the occurrence of abuse or neglect (Conger et al., 1994). Indeed, the effect of these policies is slow to emerge and grows over time, showing very similar patterns to existing findings on sport betting and consumer financial health (Baker et al., 2024; Hollenbeck et al., 2024).

The rest of this paper is structured as follows. Section 3.2 provides a brief history of sports betting in the United States and background on child abuse

and neglect. Section 3.3 outlines each data source used. Section 3.4 the empirical strategy for the paper. Section 3.5 provides results and Section 3.6 concludes.

3.2 Background

3.2.1 History of Legal Sports Betting. In 1992, the Professional and Amateur Sports Protection Act (PAPSA), or "Bradley Act" effectively banned sports betting in most of the United States. Oregon, Delaware, and Montana were exempt due to their existing sports lottery systems, along with Nevada's licensed sports pools.[2] The challenge that ultimately led to the repeal of the federal sports betting ban began when New Jersey passed legislation to remove its own state-level bans. The NCAA and several professional sports leagues quickly sued, arguing that New Jersey's actions violated the Professional and Amateur Sports Protection Act (PASPA). The case reached the U.S. Supreme Court, which in May 2018 issued its decision in Murphy v. National Collegiate Athletic Association, striking down PASPA and overturning the federal prohibition on sports betting. The Court ruled that PASPA's ban on state sports betting violated the anticommandeering doctrine of the 10th Amendment by preventing states from making their own laws. Thus, this decision didn't legalize sports betting nationwide, rather, it gave each state the power to decide for itself.

Delaware became the first state to legalize sports betting following this decision, soon followed by New Jersey. As of November 2025, 39 states and the District of Columbia have legalized some form of sports betting. There are two main ways that such laws permit people to bet on sports. First, retail or in-person betting takes place in a licensed physical space within the state, often at casinos, bars, or stadiums. Second, people can bet using the internet, either on a computer

[2]The law also allowed states with established betting industries a one year window to legalize sports betting and become exempt from the ban, but no states elected to do so.

or mobile device. Often called mobile or online betting, this form of betting can happen anywhere within a state.[3] Sports betting laws govern the location where a bet is placed, rather than the state residency of the bettor. Online sportsbooks enforce this by checking the GPS location of the device used to access the platform. While there are a sophisticated set of tools available to sportbooks to prevent illegal use (e.g. via a virtual private network), there is still some potential for spillovers via bettors taking their phones across state lines, or betting through proxy individuals in other states.

Table 11. U.S. Sports-Betting Adoption Timeline

State	Online	Retail	Online Launch	Retail Launch	State	Online	Retail	Online Launch	Retail Launch
Arizona	YES	YES	9/9/2021	9/9/2021	Nevada	YES	YES	1/1/2010	1/1/1949
Arkansas	YES	YES	3/6/2022	7/1/2019	New Hampshire	YES	YES	12/30/2019	8/12/2020
Colorado	YES	YES	5/1/2020	6/17/2020	New Jersey	YES	YES	8/1/2018	6/14/2018
Connecticut	YES	YES	10/19/2021	9/30/2021	New Mexico	NO	YES	—	10/16/2018
Delaware	YES	YES	12/27/2023	6/5/2018	New York	YES	YES	1/8/2022	7/16/2019
Florida	YES	YES	11/7/2023	12/7/2023	North Carolina	YES	YES	3/11/2024	3/18/2021
Illinois	YES	YES	3/5/2022	3/9/2020	North Dakota	NO	YES	—	6/23/2021
Indiana	YES	YES	10/3/2019	9/1/2019	Ohio	YES	YES	1/1/2023	1/1/2023
Iowa	YES	YES	8/15/2019	8/15/2019	Oregon	YES	YES	8/27/2019	8/27/2019
Kansas	YES	YES	9/1/2022	9/1/2022	Pennsylvania	YES	YES	5/1/2019	11/15/2018
Kentucky	YES	YES	9/28/2023	9/7/2023	Rhode Island	YES	YES	9/14/2019	11/26/2018
Louisiana	YES	YES	1/28/2022	10/6/2021	South Dakota	NO	YES	—	9/9/2021
Maine	YES	YES	5/2/2022	5/2/2022	Tennessee	YES	NO	11/1/2020	—
Maryland	YES	YES	11/23/2022	12/9/2021	Vermont	YES	NO	1/11/2024	—
Massachusetts	YES	YES	3/10/2023	1/31/2023	Virginia	YES	YES	1/21/2021	1/21/2021
Michigan	YES	YES	1/22/2021	3/11/2020	Washington	NO	YES	—	9/9/2021
Mississippi	NO	YES	—	8/1/2018	District of Columbia	YES	YES	5/28/2020	8/31/2020
Montana	NO	YES	—	3/11/2020	West Virginia	YES	YES	12/1/2018	9/1/2018
Missouri	YES	YES	12/1/2025	12/1/2025	Wisconsin	NO	YES	—	11/31/2021
Nebraska	NO	YES	6/22/2023		Wyoming	YES	YES	9/1/2021	9/1/2021

Table 11 shows the market open date for both in-person and/or mobile betting in each state. Typically, states legalize in-person betting prior to mobile betting, and the timing between opening the markets for these different modalities is often close. However, 8 states have legalized in-person betting with no mobile betting during the sample period January 2012 - December 2022. Several of these

[3]Some states restrict online betting to specific areas. For example, Washington only allows mobile betting while on tribal lands. Such laws are considered in-person betting for the sake of this study as they do not enable betting at all times and locations for the vast majority of a state's residents.

states—New Mexico, North Dakota, Washington, and Wisconsin—only legalized sports betting on tribal lands.

Figure 11. Trends in Sports Betting Revenue



Regardless of the type of legalization, a key question for policy makers is: does the legalization of sports betting actually change how much people gamble or is it just a new substitute for other gambling? Figure 11 shows the trends in sports betting revenue (green) versus total revenue for all gambling industries (brown) collected from Legal Sports Report and the Federal Reserve Bank of St. Louis, respectively. This figure shows that as sports betting legalization advanced across the country, the total amount of money flowing into sports-books, casinos, and other gambling activities has increased. I take this increase as evidence that the introduction of sports betting didn't just reallocate betting from casinos to sports books. Rather, it brought additional people into the gambling market, led to increased expenditures of existing bettors, or both.

3.2.2 Background on Child Maltreatment. Child abuse and neglect are critical public health issues in the United States. In 2023,

approximately 4.3 million child maltreatment referrals were made, involving about 7.7 million children. Of these, 546,159 children were confirmed victims of abuse or neglect, and 1,968 children died as a result (U.S. Department of Health & Human Services, 2024). The majority of confirmed reports are of neglect, but physical, sexual, and emotional abuse reports are also prominent (Wildeman et al., 2014). Official reports likely under-count the actual prevalence of abuse, as many incidents go unreported. By age 18, 25% to 37% of U.S. children will experience a CPS investigation for abuse or neglect, with higher rates for some groups, such as Black children (Kim et al., 2017).

The prevalence of reports varies spatially and over time within the United States. Panel A of Figure 12 shows how the number of children investigated through reports has changed over the last 11 years differentially across states.[4] The variation across states does not appear to be explained by regional factors. Panel B shows the overall incidence of reports over the same 11 year window. Child populations in the U.S. are falling, which coincides with a falling number of investigations.[5] The bottom panel shows the rate of investigations per 10,000 children, which also indicates a return to a rate of investigations similar to 2010.

Table 12 shows the quarterly pre-sports betting legalization prevalence of child maltreatment across states. Additionally, it includes information about the distribution of report targets (alleged victims) in terms of their race and age. On

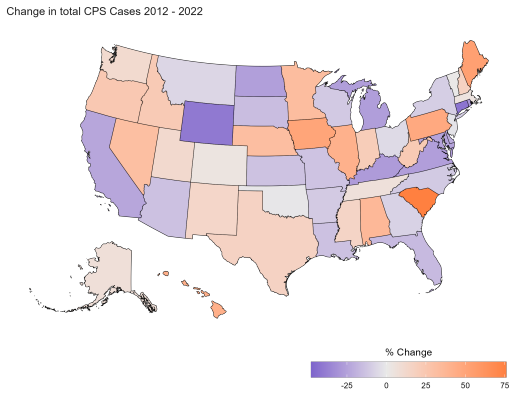
[4]These figures show child-report pairs, meaning that each child within report is counted as a separate occurrence. Children that are reported repeatedly after an investigation is complete show up as multiple entries. Repeat victimization is included in these statistics as it represents continued harm to children and captures the impact on case load of CPS workers.

[5]Maps similar to Panel A for 2012 - 2018 and 2018 - 2022 are in the Appendix, providing more context for the periods of expansion and contraction of caseloads shown in Panel B.

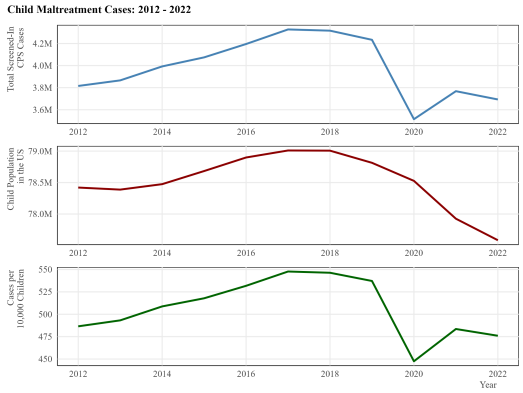
average, more than 20,000 reports make it into the hands of CPS professionals in each state each quarter of the year.

Figure 12. Trends in Child Maltreatment Reporting

Panel A: Change in CPS Reports by State 2012 - 2022



Panel B: CPS Reports over Time 2012 - 2022



The social costs of adverse childhood experiences are immense. Children who are found to be victims of abuse or neglect are more likely to experience developmental delays, mental health disorders (such as depression, anxiety, and Post-Traumatic Stress Disorder), chronic physical conditions (like heart disease and diabetes), and substance abuse (Hughes et al., 2017; Widom et al., 2012). Victims of abuse are more likely to receive disability pensions, be incarcerated, and experience preventable deaths (Berger et al., 2016; Vinnerljung et al., 2015). They also face disruptions in education, and juvenile offenders with abuse histories are at greater risk of continued involvement with the criminal justice system (Currie & Spatz Widom, 2010; Font & Maguire-Jack, 2020). Beyond individual impacts, maltreatment often sticks with the victim throughout their life which in many reports can impact their success in raising their own kids leading to generational consequences.

Table 12. Pre-Treatment Summary Statistics by Case Type, Child Age, and Race/Ethnicity

Category	**Mean**	**Min**	**1st Qu.**	**Median**	**3rd Qu.**	**Max**
By Case Type						
All Reports	20,258	818	4,574	11,053	24,776	122,303
Abuse Reports	4,448	66	941	2,676	4,528	27,787
Neglect Reports	12,600	33	2,452	6,462	15,757	74,123
By Child Age						
<1 year old	1,637	66	382	968	2,086	9,021
1–5 years old	6,192	240	1,426	3,462	7,170	35,874
6–11 years old	7,024	235	1,649	3,848	8,674	43,252
12–17 years old	5,258	148	1,197	2,748	6,750	35,138
By Child Race						
Asian	192	1	19	62	159	4,216
Black	4,393	1	271	1,976	5,006	31,063
Other	2,986	7	377	927	2,081	66,343
White	11,607	1	2,987	7,127	14,486	63,811

The Family Stress Model (FSM) developed by Conger et al. (1994) theorizes the mechanisms which cause child abuse and neglect. The FSM posits that economic hardship undermines family functioning through a cascading sequence of stress processes (see Figure 13 for a diagrammatic view of this model). Economic hardship constrains a family's ability to meet their basic needs and creates emotional distress among one or both parents (Slack et al., 2011). This can destabilize the couple's relationship via inter-parental conflict, and erodes the quality of parenting—often manifesting as harsher, less responsive interactions with children (Conger & Conger, 2002; Conger & Donnellan, 2007). Harsh, inconsistent parenting undermines the development of children's non-cognitive skills like self-control, emotional regulation, and cooperation (Neppl et al., 2016), which increases the likelihood of conflict between parents and children. Originally formulated

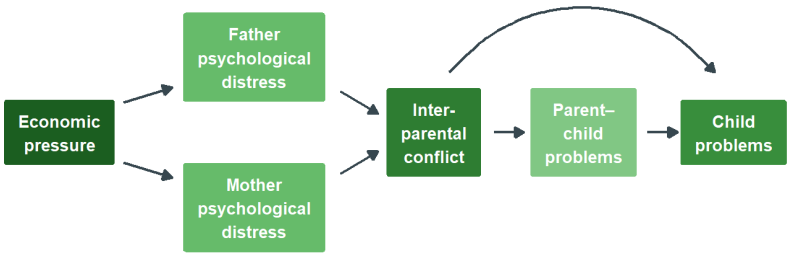


Figure 13. Family Stress Model Diagrammed

during the 1980s farm crisis (Conger & Conger, 2002), there is now a large body of evidence across different cultures and income levels supporting the causal chain of events described by the FSM (Brooks-Gunn & Duncan, 1997; Masarik & Conger, 2017; Repetti et al., 2002; Zietz et al., 2022). Overall, children from low socioeconomic status (SES) families face maltreatment rates five times higher than their more affluent peers (Sedlak et al., 2010), and aggregate changes in child poverty are associated with changes in maltreatment reports (Kim & Drake, 2023).

Intimate partner violence (IPV) can amplify the risk factors outlined by the FSM. Economic hardship escalates partner conflict, and abusive relationships can exacerbate financial instability (Stylianou, 2018). Families facing both poverty and IPV are often at the highest risk of maltreatment, as multiple stressors converge to undermine family stability. Part of this relationship is mechanical since exposing children to domestic violence among adults in their household is considered a form of emotional abuse in its own right (Kitzmann et al., 2003). Approximately 30-60% of child maltreatment reports involve domestic violence in the home (Edleson, 1999; Renner & Slack, 2006). IPV can directly harm children or lead to emotional and physical abuse as a result of the trauma experienced by the non-abusive parent. It also impairs parental capacity, as a parent experiencing abuse may struggle to care for or protect their children (Hamby et al., 2010). Chronic IPV often co-occurs

with child abuse, and the dynamics of power and control in these situations can exacerbate maltreatment (Renner & Slack, 2006).

3.2.3 Sports Betting and Maltreatment Risk Factors. Together, financial hardship and violence among adults are the two significant risk factors for child maltreatment that motivate this study. This is because past literature has identified direct effects of sports betting on both. First, sports betting can create the economic hardship central to the FSM. One-time large losses can create moments of economic hardship, particularly those that occur immediately following a payday, thus causing households to more strictly budget until the next. In the longer-run, repeated losses lead households to have worsening credit scores, and increased rates of bankruptcy (Baker et al., 2024; Hollenbeck et al., 2024). Second, emotional shocks in high-stakes situations often translate into real-world violence, and this pathway is heightened with the addition of sports betting. Using National Football League data, Card and Dahl (2011) found that unexpected (upset) home-team losses increased male-on-female partner assaults in the hours immediately after the game.[6] The results match the frustration–aggression mechanism that links anger from violated expectations to violence (Berkowitz, 1989). Matsuzawa and Arnesen (2024) find that the legalization of sports betting amplifies the effects of game outcomes on IPV following NFL games, driven by states that legalized mobile betting.

3.3 Data

3.3.1 Child Protective Services Case Data. I obtain child protective services (CPS) case-level data from the National Child Abuse and

[6]Data from England following soccer matches shows a similar pattern; during the world cup domestic-abuse calls climb roughly 25% after national-team wins or draws and 38% after losses (Kirby et al., 2014).

Neglect Data System (NCANDS). NCANDS is a federally-sponsored collection and analysis of state child abuse and neglect data mandated under the 1988 amendments to the Child Abuse Prevention and Treatment Act (CAPTA). Data are submitted voluntarily by all 50 states, Washington D.C. and Puerto Rico. NCANDS reporting follows the federal fiscal calendar (October - September) and culminates into two files each year: the agency file and the child file. The agency file provides aggregate information at the state level regarding the total number of reports, the number of screened in reports, and aggregated information about services rendered. I use data from the FY2012 to FY2023 child files to measure child maltreatment reports from January 2012 to December 2022.[7]

The child file provides biweekly case-level data which includes information about the characteristics of screened-in reports. The exact report date for each report is suppressed for privacy and recoded as either the 8th or the 23rd depending on which date is closer. In my main specification, I aggregate the data up to the quarter level.

The file also contains information about the county of the offense, types of alleged maltreatment, the final result (known as "disposition"), and characteristics about the children, caregivers, and perpetrators of the report. Only counties which have more than 700 reports in a fiscal year are identified. Otherwise, data from a county with fewer than 700 reports in a fiscal year is identified only at the state level for that entire year. When conducting analysis at the county level, I remove counties which are not consistent reporters during my sample period; counties

[7]Each row in the data is a child-report pair. This means that every time a child is reported it contributes towards the overall count. Reports featuring multiple children are already disaggregated for proper counting. Children can be linked across multiple reports during the sample period.

which report for less than two years in the pre-period and less than two years in the post-period are dropped from the sample.

Exact statutes regarding valid reasons for reporting and how to report cases of child abuse and neglect vary across states. In general, the process of reporting starts with a concerned adult or mandatory reporter (e.g. schoolteacher) encountering information that leads them to suspect that a child is experiencing maltreatment. When a report comes in, an intake worker first gathers basic facts and determines whether the allegations fall within a state's statutory definition of maltreatment and whether the child's caregiver is the alleged perpetrator. This triage step is called screening. If the report lacks jurisdiction or alleges concerns that do not meet legal criteria, it is "screened out" and closed or referred to other services. If it meets all basic criteria, it is screened in and assigned a response priority.[8] The case is then either routed to a traditional investigation or in some states to a family-assessment/alternative-response track. Next, a caseworker begins an investigation and determines appropriate course of actions. Once the investigation is complete and the child is deemed safe the report is given a disposition or outcome. Substantiated and indicated dispositions mean that a CPS investigator determined that a child in the report is a victim of abuse or neglect. Unsubstantiated and alternative response dispositions mean that either the alleged act did not occur, or the evidence did not meet the legal definition required for a formal intervention.

3.3.2 NFL Game Data. To determine the role of emotional cues, I use spread (*ex-ante* expected difference in points scored between the teams in a game) data on the predicted outcome of each regular season NFL game during

[8]Usually "immediate" (within 24 hours), or "non-emergency" (within 72 hours).

my sample period from NFLOddsHistory.com. The sign of the spread indicates which team is favored (negative indicating the favored team) and the magnitude of the spread provides information about how close the game should be. A larger magnitude spread indicates less uncertainty about the winner.

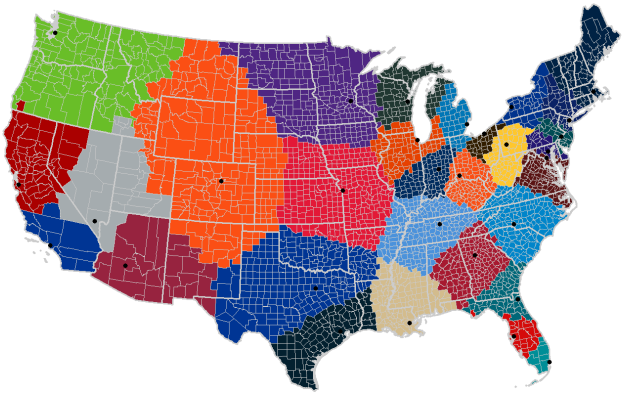
Akin to Card and Dahl (2011) and Matsuzawa and Arnesen (2024), I match spread data to the outcome of each game to create an indicator for games where the favored team ended up losing —an event likely to cause a negative emotional response. Again following the above authors, I define an upset loss for a team that lost but was favored by four or more points according to the spread. Upset wins are defined symmetrically.

Determining where sports fans reside and which counties favor specific teams is a challenge. Some states have multiple NFL teams, during the sample period three teams move, and in two areas of the country, Los Angeles and New York, two teams play in the same stadium or relatively close to one another. I adopt an approach similar to Matsuzawa and Arnesen (2024), who determine a county's home team as the team whose stadium is closest by linear distance to that county's centroid.

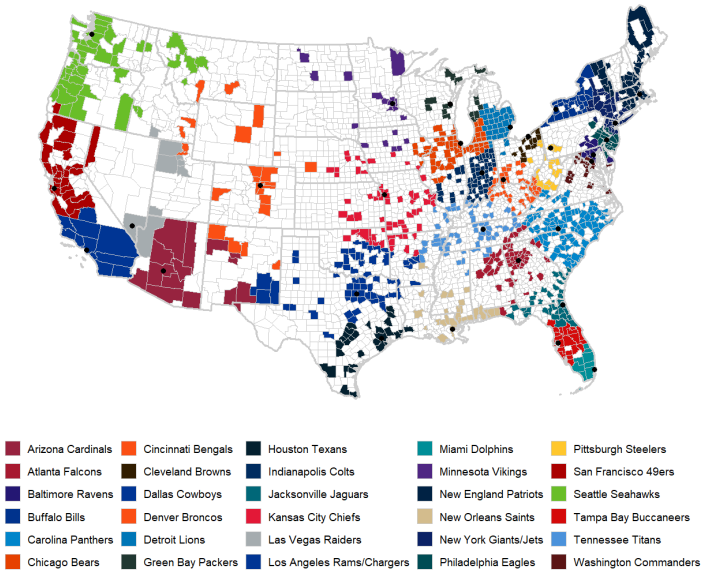
I use the linear distance from a stadium to a county's nearest edge. Because there are several large, oddly shaped counties which are between two cities with NFL teams but touch only one of those cities, the county edge approach better deals with this. Panel A of Figure 14 shows the team to which each county in the contiguous U.S. is assigned. Panel B shows the counties in which NCANDS data is available at some point during the sample period.

Figure 14. Home Team for Each County based on Distance to Nearest Stadium

(a) All U.S. Counties



(b) U.S. Counties Identifiable in NCANDS Data



Notes: (A) Map of the contiguous United States showing the "home team" that each county should be assigned based on the linear distance to the nearest stadium. (B) takes the map of all counties and subsets it to only include the counties which are consistent reporters in the data.

3.3.3 Sports Betting Spread and Revenue. To measure the expansion and intensity of legalized sports wagering across the United States, I collect state-level data from Legal Sports Report (2025), which compiles official regulatory filings from each state's gaming commission. The dataset provides monthly observations beginning with the first month of legal operation in each

state and includes both "handle" and revenue. Handle refers to the total dollar amount wagered by consumers within a given period, while revenue represents the amount retained by operators after paying out winnings but before deducting taxes, licensing fees, or administrative expenses. As such, handle captures the overall scale of betting activity, whereas revenue reflects the realized earnings to sportsbooks.

To contextualize sports wagering within broader gambling activity, I supplement these data with quarterly measures of total gambling receipts from the U.S. Census Bureau's *Quarterly Services Survey* (U.S. Census Bureau, 2025). The QSS collects national estimates of gross receipts from establishments classified under NAICS 7132, which includes casinos (except online), lotteries, and other gaming operations. These data provide a consistent benchmark for tracking trends in overall gambling expenditures, allowing for comparison between the rise of sports betting and aggregate gaming revenue.

3.3.4 Population, COVID School Closures, and Debt to Income.

In addition to the above sources of data, I retrieve county-level population data from the Surveillance, Epidemiology, and End Results (SEER) Program. The SEER population files provide annual July 1 mid-year estimates for every U.S. county stratified by single year of age, sex, bridged race, and Hispanic origin. Using these data, I find the size of the population in each state or county which would be eligible for any type of CPS investigation.

COVID-19 impacted many of the institutions which were most likely to identify child maltreatment, such as schools and daycare facilities (Brown et al., 2022; Rapoport et al., 2021). In order to control for the time varying state specific policies enacted to reduce the spread of COVID, I collect school closure policy severity information from the Oxford Covid-19 Government Response Tracker

(OxCGRT). School closure severity ranges from 0 (normal operation) to 3 (schools are prevented from operating in-person) and is reported at the onset of COVID through 2022.

To proxy for a households financial health, I use Debt-to-Income ratios collected from the Federal Bank of New York Consumer Credit Panel. Data is collected at the county level and represents a DTI ratio for each quarter of the year.

3.4 Empirical Strategy

My empirical strategy exploits the staggered adoption of sports betting between 2018 and 2022. My baseline specification uses state-level counts of child protective services reports to estimate a two-way fixed effects (TWFE) difference-in-difference model. To adjust for the use of count data as the dependent variable, I estimate a Poisson pseudo-maximum likelihood model with two-way fixed effects:

$$\mathbb{E}[\text{Reports}_{st} \mid Policy_{st}, X_{st}, \alpha_s, \lambda_t] = \exp\left(\tau\, Policy_{st} + \alpha_s + \lambda_t + \beta' X_{st} + \ln(\text{Population}_{st})\right) \tag{3.1}$$

where $Reports_{st}$ is the count of reports of child maltreatment started in state s in time period t. $Policy_{st}$ is a dummy variable indicating whether sports betting was legalized for more than 50% of the time period t in state s and X_{st} represents co-variates such as COVID school closure controls. (α_s) denotes state fixed effect and (λ_t) time fixed effects. When estimating rates, I include $\ln(\text{Population}_{st})$ as an "offset" term with its coefficient fixed at one, allowing the model to account for differences in exposure across observations without estimating its effect. In this case, the offset converts the model from counts to rates by scaling reports by the child population in each state and year. Under this specification, $\exp(\tau) - 1$ is the

proportional change in the expected number of reports per child associated with the implementation of $Policy_{st}$.

Recent literature shows that in the presence of a staggered treatment with heterogeneous effects, TWFE regressions can result in biased estimates. To address this concern, I also estimate my main results using methods proposed by Sun and Abraham (2021) and Borusyak et al. (2024). Additionally, I report OLS TWFE estimates comparable to the Poisson specification from equation 3.1. The TWFE OLS specification uses logged number of investigated reports as the outcome variable.

Equation 3.1 identifies a causal effect of sports betting legalization on child maltreatment under the assumption that in the absence of a policy change, treated states would have experienced similar time trends in child maltreatment reports as non-treated states. In Section 3.5 I present event-study figures that lend support to this parallel trends assumption. This assumption could be violated if policy adoption is driven by unobserved, time-varying factors affecting reporting trends. News reporting and commentary from state legislatures suggests legalization timing is not correlated with trends in child maltreatment. Most reporting around sports betting legalization focuses on tax revenues for states and the expansion of individual rights.

I consider two possible treatments regarding sports betting legalization. In the first, I consider the first quarter in which any market (online or in-person) opens as the moment of treatment. In the second, I consider mobile betting as an independent treatment from in-person betting which is determined by the first quarter in which the respective market first begins operation.

To test the mechanisms underlying the relationship between sports betting and child maltreatment, I use an alternative specification to consider the impact of sports-related emotional cues on child maltreatment (Card & Dahl, 2011; Matsuzawa & Arnesen, 2024). To take advantage of within state variation of sports team loyalty and the higher frequency of games, I create a panel of child maltreatment data at the county–biweekly level. I specifically estimate:

$$\begin{aligned}\mathbb{E}[\text{Reports}_{ct} \mid \text{Policy}_{ct}, X_{ct}, \alpha_c, \lambda_t] = \exp\Big(&\gamma_1 \,\text{ExpWin}_{ct} + \gamma_2 \,\text{ExpLoss}_{ct} + \gamma_3 \,\text{UpsetLoss}_{ct} \\ &+\gamma_4 \,\text{UpsetWin}_{ct} + \tau_0 \,\text{Policy}_{ct} + \tau_1(\text{ExpWin}_{ct} \times \text{Policy}_{ct}) + \tau_2(\text{ExpLoss}_{ct} \times \text{Policy}_{ct}) \\ &+\tau_3(\text{UpsetLoss}_{ct} \times \text{Policy}_{ct}) + \tau_4(\text{UpsetWin}_{ct} \times \text{Policy}_{ct}) + \alpha_c + \lambda_t + \beta' X_{ct} + \log(\text{Population}_{ct})\Big)\end{aligned} \tag{3.2}$$

where $ExpWin_{ct}$ is a variable indicating the number of games a county's home team won as expected —with a pregame spread of -4 or better— within two-week period t. $ExpLoss_{ct}$ is the equivalent for games lost as expected —a pregame spread of +4 or worse— $UpsetWin_{ct}$ is the equivalent for games unexpectedly won —with a pregame spread of +4 or worse— $UpsetLoss_{ct}$ is the equivalent for games unexpectedly lost —with a pregame spread of -4 or better— and the omitted game outcome category is thus both close games and weeks with no games. Equation 3.2 identifies the causal effect of wins vs. losses under the assumption that holding fixed the pregame expectation, the realization of the outcome is as-good-as-random. This assumption underpins a substantial literature on sports outcomes in economics, and has been supported in past work (Card & Dahl, 2011; Gandar et al., 1988).

3.5 Results

3.5.1 Legalization Increases Reports of Child Maltreatment.

To analyze the impact of sports betting on child maltreatment investigations, I

estimate Equation 3.1 using a state-by-quarter panel of child protective services reports. Table 13 shows the DiD estimates of Equation 1 when treatment is considered to be the legalization of any type of sports betting (in-person or online). Column (3) shows the baseline Poisson specification from equation 3.1 without covariates. The coefficient suggests that the legalization of sports betting results in a $\sim 7\%$ increase in the rate of reports of child maltreatment. The estimates of the effect of any betting legalization are noisy, statistically different than zero at only the 10% level. The inclusion of COVID School Control and the use of OLS instead of Poisson attenuates the size of this effect slightly, as shown in columns (1), (2), and (4). The smallest estimate still suggests that the rate of investigations increases by 6.1% in response to legalized sports betting.

Table 13. OLS and Poisson Estimates of Any Betting

	(1) OLS	(2) OLS	(3) Poisson	(4) Poisson
Any Betting	0.0652	0.0612	0.0701$^{\cdot}$	0.0668
	(0.0399)	(0.0380)	(0.0431)	(0.0415)
N	2,200	2,200	2,200	2,200
State FE	X	X	X	X
Quarter FE	X	X	X	X
COVID School Control		X		X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Notes: Columns (1) and (3) exclude COVID School Control. Columns (2) and (4) include them. OLS estimates are weighted by child population and include log(`child population`) as a covariate. Poisson estimates include log(`child population`) as an offset. All specifications include state and quarter fixed effects. Standard errors are clustered at the state level.

Table 14 separates treatment into the legalization of mobile and in-person betting. Similar to other research on the impact of legalized gambling, I find that access to mobile betting has a larger effect than in-person betting on the rate of

investigations.[9] The legalization of mobile betting alone contributes to a 6.8% in the rate of reports (using the baseline Poisson model). The decomposition of any betting into retail and mobile improves the precision of the estimates somewhat, such that I can reject the null hypothesis that the mobile betting coefficient is zero at the 10% level across all models.

Table 14. OLS and Poisson Estimates of Mobile and In-Person Betting

	(1) OLS	(2) OLS	(3) Poisson	(4) Poisson
Mobile Betting	0.0654$^{\cdot}$	0.0606$^{\cdot}$	0.0677$^{\cdot}$	0.0630$^{\cdot}$
	(0.0344)	(0.0325)	(0.0346)	(0.0325)
In-Person Betting	0.0308	0.0300	0.0340	0.0340
	(0.0332)	(0.0327)	(0.0333)	(0.0333)
N	2,200	2,200	2,200	2,200
State FE	X	X	X	X
Quarter FE	X	X	X	X
COVID School Control		X		X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Notes: Columns (1) and (3) exclude COVID School Control. Columns (2) and (4) include them. OLS estimates are weighted by child population and include log(`child population`) as a covariate. Poisson estimates include log(`child population`) as an offset. All specifications include state and quarter fixed effects. Standard errors are clustered at the state level.

3.5.2 Pre-Trends and Dynamic Effects. Before further investigating the estimated effects of the policy change, it is important to verify that treated and control units were on similar trajectories prior to adoption and to understand how outcomes evolve over time following treatment. Because TWFE estimates can be biased in the presence of heterogeneous treatment effects (Callaway & Sant'Anna, 2021; De Chaisemartin & D'Haultfœuille, 2020; Goodman-

[9]While the point estimates consistently suggest a larger effect for mobile than in-person betting, I formally test whether the two coefficients are statistically distinguishable using a Wald test of the restriction $\beta_{Mobile} = \beta_{In-Person}$. I fail to reject this hypothesis in all specifications, reflecting insufficient precision to statistically distinguish the two treatment effects.

Bacon, 2021; Sun & Abraham, 2021) I re-estimate the main specification of the paper using methods developed by Sun and Abraham (2021) and Borusyak et al. (2024). Since Borusyak et al. (2024)'s estimator uses not-yet-treated units as the comparison, I argue that in the context of sports betting it is the more applicable method.[10]

Borusyak et al. (2024) develop a two-step imputation method that predicts counterfactual untreated and not-yet-treated outcomes using unit and time fixed effects and then computes treatment effects as the difference between actual and imputed outcomes. The results in Table 15 and Figure 15 show the DiD Imputation estimates and event study, respectively.

Table 15. DiD Imputation Estimates

	Without COVID Controls		With COVID Controls	
	(1)	(2)	(3)	(4)
Any Betting	0.0526$^{\cdot}$		0.0485$^{\cdot}$	
	(0.0300)		(0.0292)	
Mobile Betting		0.0673***		0.0626***
		(0.0206)		(0.0200)
In-Person Betting		0.0289		0.0288
		(0.0251)		(0.0246)
N	2,200	2,200	2,200	2,200
State FE	X	X	X	X
Quarter FE	X	X	X	X

$^{\cdot}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Notes: Imputation-based DiD estimator (Borusyak et al.). Entries are treatment effects with state-clustered SEs in parentheses. Columns (1)–(2) exclude COVID school-related controls, while Columns (3)–(4) include them. Columns (1) and (3) report "Any" legalization; Columns (2) and (4) stack Mobile and In-Person estimates in the same specification. All use the same state-by-quarter panel (2,200 obs.).

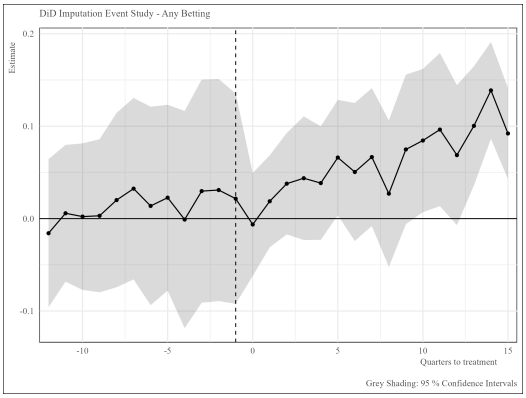
[10]Results for the Sun and Abraham (2021) estimator are in the Appendix.

The estimates in Table 15 are slightly smaller in magnitude than those in Table 13 and Table 14, but more precisely estimated. Notably, the effect of mobile betting becomes statistically significant at the 1% level. Figure 15 shows the dynamic effects of the policy broken down by legalization type. The pre-treatment differences are near zero, not statistically significant, and don't show evidence of a trend. In the post period, the effect of mobile betting manifests more quickly while the effect of in-person betting is more delayed and relatively small. Taken together, the DiD Imputation estimates and the lack of a concerning pre-trend provides suggests the TWFE results are not driven by heterogeneous treatment effect bias.

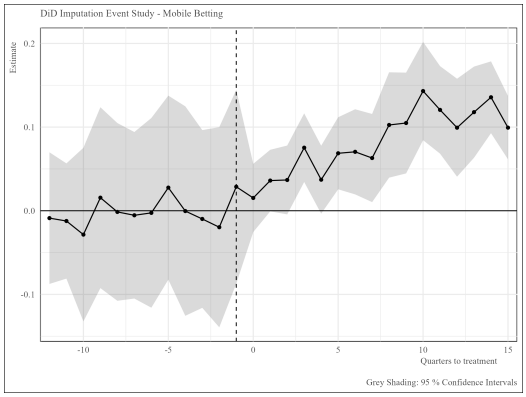
The effect of legalized betting does appear to grow over time. There are two possible explanations for this. The first is that uptake of betting is slow. When betting markets first opened they were a fraction of the size they are today. Either fewer people were betting when markets first opened, they were betting smaller amounts, or both. But as more people engaged in this activity and spent more money, the externalities began to appear. The second explanation is the dynamics of the Family Stress Model. When a family member is a problem gambler and engages in sports betting, they are putting income at risk that is needed to pay for rent, food, and other child expenses. Each consecutive loss of income has the possibility to lead to arguments between partners and the use of savings or loans to make up the difference. As resources get depleted and conversations remain unresolved, the FSM suggests that these problems increasingly trickle down to children (Conger et al., 1994). It takes time for this to occur and then be detected by the reporting adult. Both of these explanations suggest a financial motive as the

Figure 15. DiD Imputation Event Studies

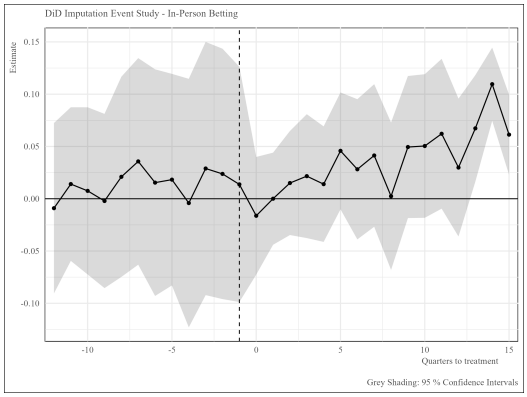
Panel A: Overall



Panel B: Mobile Betting



Panel C: In-Person Betting



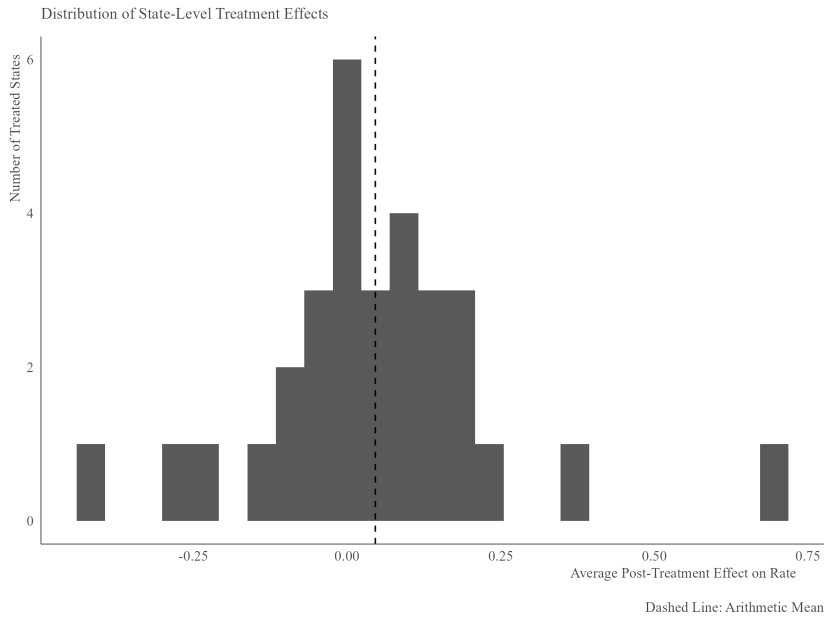
Notes: This figure plots the estimated dynamic treatment effects of sports betting legalization relative to the quarter of adoption (period 0). Estimates prior to treatment indicate pre-trends are likely not present. The imputation approach ensures that estimates are not contaminated by inappropriate comparisons across units with different treatment timings. Shaded areas denote 95% confidence intervals.

driving explanation of the increase in child maltreatment reports following sports betting legalization.

3.5.3 Synthetic Controls Method Estimates. As an alternative approach to using staggered difference-in-difference estimators, I use a synthetic control method (SCM) approach to characterize the full distribution of effect sizes across treated states. SCM, originally introduced by Abadie et al. (2010), constructs a weighted average of control units to approximate the counterfactual outcome trajectory of a treated unit.

Figure 16 shows the distribution of SCM-estimated average treatment effects (ATTs) of sports betting on the log rate of child maltreatment reports. The ATTs are calculated as the mean difference in outcomes between the treated and synthetic unit in the post-period. The dashed-line at 0.045 in the figure shows the average ATT across all units when equal weight is applied to each treated unit. This estimate aligns closely with the DiD Imputation estimate. Because these results do not include placebos or confidence intervals, the figure should be interpreted as descriptive evidence of the cross-state distribution of estimated effects rather than a formal inference result.

Figure 16. Distribution of Synthetic Control ATTs

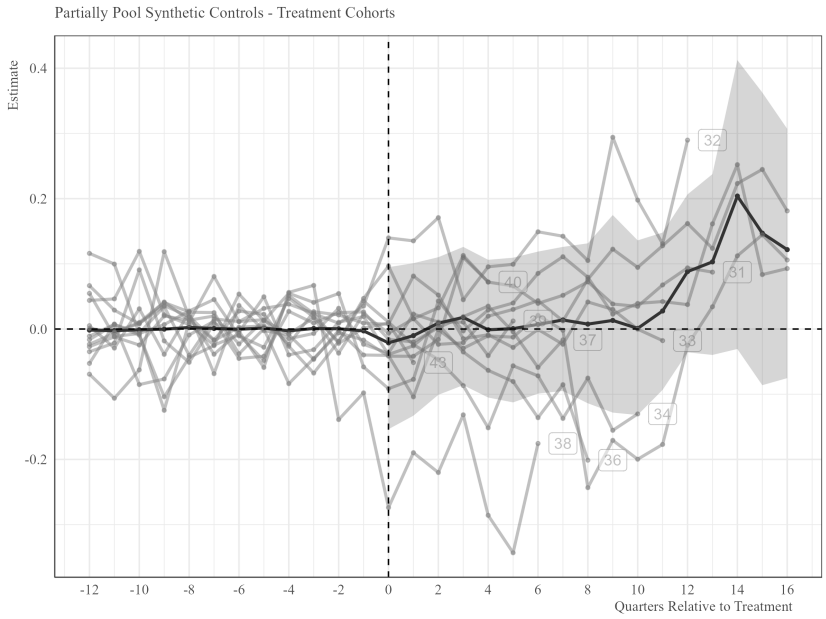


Notes: The histogram summarizes state-level average treatment effects (ATTs) from separate synthetic control estimates. For each treated state i, a synthetic counterfactual of the log rate of reports is constructed using pre-treatment weights determined by rolling means of the log rate of reporting for the three half years prior to adoption. The ATT for state i is calculated as the mean post-t_0 difference $\overline{(y^{\text{real}} - y^{\text{synth}})}$ across all post-treatment periods, where positive values indicate higher post-treatment outcomes relative to the synthetic counterfactual. The dashed vertical line denotes the arithmetic mean ATT across treated states, while the dispersion of the histogram illustrates heterogeneity in treatment effects.

One limitation of this approach, generally, is that the fit of SCM varies by adoption state and it doesn't use any information across units. The SCM was developed for a single treated unit and does not directly accommodate multiple units with staggered adoption. While one strategy for multiple treated units involves computing separate synthetic controls for each treated unit and then averaging their estimated effects (Donohue et al., 2019; Dube & Zipperer, 2015), this may not be appropriate when some treated units have poorer pre-treatment fits than others. This could lead to a high variance in unit-level ATT estimates and a potentially biased across-unit average (Ben-Michael et al., 2021). At the other extreme, pooled synthetic control approaches a single set of weights that minimizes

the average pre-treatment imbalance across treated units. When all treated units share one common set of weights, the method targets the average treated trajectory rather than each unit. This reduces variance but risks bias if units differ in their underlying outcome dynamics or treatment timing. Instead, I follow the partially-pooled synthetic control approach introduced by Ben-Michael et al. (2021) which minimizes a weighted average of the pooled and unit-specific imbalance measures. By tuning the degree of pooling, the partially pooled estimator helps mitigate the bias–variance trade-off that arises when estimating heterogeneous treatment effects across staggered adoption settings.

Figure 17. Partially Pooled Cohort Synthetic Control Method



Notes: This figure presents dynamic treatment effects estimated using a partially pooled synthetic control framework. The pooling parameter nu is selected flexibly to minimize variance and bias. SCs are estimated for each legalization cohort.

Figure 17 shows the evolution of both the average and legalization cohort-specific effects of sports betting legalization on the log rate of child maltreatment reports using the partially-pooled method. Cohorts which adopted sports betting

later appear feature a poorer fit, while early appear to drive the overall increase in reports.

3.5.4 Financial Health, Abuse, and Emotional Responses to Betting.

Investigating how sports betting impacts child maltreatment through household financial constraints poses interesting challenges. Due to the sensitive nature of child maltreatment data, it is not possible to try to identify or link any household specific information to individual incidences of abuse or neglect. For this reason, I use the 2018 county-level Debt-to-Income ratio as a proxy for the average financial health of households.

Table **??** shows treatment effects by county debt-to-income terciles. Column 1 consists of counties with the lowest debt-to-income ratio and column 3 consists of the highest. Effects appear to be more pronounced in areas outside the highest debt-to-income counties. One possible explanation is that this grouping includes households with enough disposable income to engage in high-stakes and frequent sports betting.

Table 16. Poisson Estimates of Betting Effects by County Debt-to-Income Tercile

	(1) (Lowest)	(2) (Middle)	(3) (Highest)
Mobile Betting	0.0699$^{\cdot}$	0.0550*	0.0393
	(0.0418)	(0.0273)	(0.0400)
In-Person Betting	0.0154	0.0574	0.0428
	(0.0380)	(0.0437)	(0.0299)
Observations	14,370	14,370	14,339
County FE	X	X	X
Quarter FE	X	X	X

$^{\cdot}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$

Notes: Each column reports results from a separate Poisson regression within terciles of county debt-to-income ratio. All models include an offset for child population, county (FIPS) and quarter fixed effects. Standard errors are clustered at the state level.

Table 17 looks at the impact of sports betting legalization on reports of abuse and neglect separately. Both reports of neglect and abuse appear to increase on average following legalization. Mobile betting increase reports of abuse especially.

Beyond the financial pathway outlined above, the literature suggests another possible explanation for the connection between child maltreatment and sports betting: increases in intimate partner violence rates (IPV) following sports matches with unexpectedly poor outcomes. In fact, in most jurisdictions, the act of a child witnessing IPV is in-and-of-itself an act of maltreatment. For this reason, increased rates of IPV may lead to increased rates of child maltreatment if the increases in IPV occur in households with children. On the other hand, most reports of child abuse and neglect are from trusted adults (teachers, doctors, police, etc.) in the community. This means that many acts of maltreatment go unreported as they

Table 17. Poisson Estimates for Types of Maltreatment Investigations

	(1) Abuse	(2) Abuse	(3) Neglect	(4) Neglect
Mobile Betting	0.1276**		0.0563	
	(0.0633)		(0.0503)	
In-Person Betting	−0.0466		0.0394	
	(0.0754)		(0.0492)	
Any Betting		0.0649		0.0595
		(0.0687)		(0.0596)
N	2,200	2,200	2,200	2,200
State FE	X	X	X	X
Quarter FE	X	X	X	X
COVID School Control	X	X	X	X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Notes: All models are Poisson regressions with an offset for child population. Specifications include state and quarter fixed effects and standard errors clustered at the state level. Columns (1)–(2) present results for **abuse**, while Columns (3)–(4) present results for **neglect**. Models (2) and (4) use a combined treatment indicator. Models include a control for COVID-19 school closures.

are never directly or indirectly observed by a reporter. This could especially be true for child maltreatment via witnessed IPV or co-incident with IPV. To explore this channel, I limit the sample to abuse reports and estimate a model of upset losses akin to Card and Dahl, 2011. Specifically, I estimate Equation 3.2 leveraging the variation in NFL game outcome conditional on the pregame expectations to measure emotional responses following unexpected losses.

In Table 18, I find little evidence that emotional cues from sports games influence child maltreatment investigations. Most estimates are close to zero, and the coefficient on the interaction of legalized betting with an upset loss is negative. One important limitation is that my analysis is aggregated at the biweekly level because of how the NCANDS data are structured. In contrast, earlier studies (e.g., Card and Dahl, 2011; Matsuzawa and Arnesen, 2024) looked at outcomes just hours

after a game, when emotional effects are likely to be strongest. By aggregating over two weeks, where multiple games occurred of varying outcomes, any short-term spikes in reports are likely to be smoothed out, making them harder to detect.

Table 18. Poisson Estimates of Treatment Effects by Local NFL Team Game Outcomes

	(1)	(2)
Upset Loss	0.0061	0.0081
	(0.0065)	(0.0067)
Exp. Loss	0.0095$^{\cdot}$	0.0076
	(0.0053)	(0.0052)
Upset Win	-0.0085	-0.0068
	(0.0054)	(0.0060)
Exp. Win	0.0004	-0.0016
	(0.0040)	(0.0037)
Mobile Betting		0.0925*
		(0.0469)
Upset Loss × Mobile Betting		-0.0255$^{\cdot}$
		(0.0154)
Exp. Loss × Mobile Betting		0.0039
		(0.0156)
Upset Win × Mobile Betting		-0.0187
		(0.0196)
Exp. Win × Mobile Betting		0.0198
		(0.0168)
Observations	71,024	71,024
County FE	X	X
Fortnight FE	X	X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Notes: Poisson regression with offset for child population. Standard errors are clustered at the state level.

3.5.5 Case and Child Heterogeneity. Baseline reporting rate differences by race are well documented. For example, black children are persistently more likely to be reported than white children as a percentage of the population (Kim et al., 2017). For this reason, I conduct heterogeneity analysis by child race. Table 19 shows the estimates of Equation 3.1 subset by child race.

These estimates suggest that sports betting is impacting all communities, but that some may be impacted more than others. The largest estimated increase in investigations is among Asian households, who ex-ante have the lowest investigation rate of any group.

Table 19. Poisson Estimates of Treatment Effects by Race

	White	Black	Latino	Asian
Mobile Betting	0.0721$^{\cdot}$	0.0871	0.0592	0.0948
	(0.0401)	(0.0580)	(0.0605)	(0.0590)
In-Person Betting	0.0411	0.0683	0.0555	0.1211**
	(0.0504)	(0.0646)	(0.0475)	(0.0613)
Observations	42,928	41,100	41,840	22,700
State FE	X	X	X	X
Quarter FE	X	X	X	X
COVID School Control	X	X	X	X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$

Notes: Each column reports estimates from a separate Poisson regression subset by child race. All models include an offset for child population, county (FIPS) fixed effects, fortnight fixed effects, and a control for COVID school closures. Standard errors are clustered at the state level. Sample period: 2012–2022.

Cities may have more resources to detect, address, and prevent child abuse and neglect (Maguire-Jack et al., 2022). For this reason, I investigate whether treatment effects are heterogeneous for urban vs. rural counties. Table 20 shows the differential effects of legalization for urban and rural counties. Since this analysis only includes counties which have at least 700 cases of reported maltreatment due to data suppression these results represent the comparison between reporting urban area and reporting rural areas. Column (1) uses a binary indicator variable for whether a county is rural based on a simplified version of the USDA Rural-Urban Continuum Code for the year the data was collected. This binary breakdown shows limited heterogeneity, with a gambling effect in reporting rural counties that is two

percentage points higher than in urban counties. Column (2) breaks down counties into finer groups by population size to better understand which types of rural counties are most impacted by betting. I find suggestive evidence that the most populated ($>$ 20k) and least populated ($<$ 5k) rural areas experience significantly larger effects of legalized betting than urban areas. Column (3) looks at rurality in yet another way—whether a county is adjacent to a metro area or not. I find that rural areas not adjacent to metro areas experience the larger impact from sports betting legalization. These results together suggest that rural counties which are large and/or not next to a metro areas experience the greatest increase in the rates of child abuse and neglect following sports betting legalization.

Table 21 provides estimates by age group and suggests that the relationship between betting legalization and child maltreatment may vary with children's developmental stage. The data are subset to represent a rough categorization into infants, pre-school age, elementary school, and middle/high school. For mobile betting, the largest and most precisely estimated increase appears among children ages 1–5. The estimates are also positive and large for children ages 6–11 and 12–17. In-person betting appears to increase reports of younger groups and essentially no effect for adolescents. These results suggest that younger children —but not infants— experience the greatest increase in investigated reports following the legalization of sports betting.

Table 20. Poisson Estimates of Betting Effects by Rurality: County-Level

	(1)	(2)	(3)
Any Sports Betting	0.038092	0.064274	0.064294
	(0.045369)	(0.041705)	(0.041697)
Rural	0.006532		
	(0.014848)		
Rural pop $\geq$ 20k		−0.014108	
		(0.048239)	
Rural pop 5k–20k		−0.016104	
		(0.046922)	
Rural pop $<$ 5k		−0.037986	
		(0.076162)	
Rural Area Adj.			−0.025531
			(0.044100)
Rural Area Not Adj.			0.021788
			(0.051147)
Betting × Rural	0.015368**		
	(0.006018)		
Betting × Rural pop $\geq$ 20k		0.063986***	
		(0.014129)	
Betting × Rural pop 5k–20k		0.011368	
		(0.020059)	
Betting × Rural pop $<$ 5k		0.102920·	
		(0.059712)	
Betting × Rural Adj.			0.033951**
			(0.016102)
Betting × Rural Not Adj.			0.083693***
			(0.027094)
Observations	43,156	43,156	43,156
County FE	X	X	X
Quarter FE	X	X	X
COVID School Control	X	X	X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Notes: All models are Poisson regressions with an offset for child population. Column (1) includes a binary rural indicator interacted with treatment. Column (2) includes three rural population categories: $\geq$ 20,000 people, 5,000–20,000 people, $<$ 5,000 people, each interacted with sports betting legalization. Column (3) includes indicators for *Rural Area Adjacent* to Metro Areas and *Rural Area Not Adjacent* to Metro Areas. All specifications include county (FIPS) and time (fortnight) fixed effects, and a control for covid school closures. Standard errors are clustered at the state level.

Table 21. Poisson Estimates of Treatment Effects by Child Age Group

	(1) <1 year old	(2) 1–5 years old	(3) 6–11 years old	(4) 12–17 years old
Mobile Betting	0.0396	0.0764*	0.0638·	0.0635·
	(0.0300)	(0.0358)	(0.0345)	(0.0363)
In-Person Betting	0.0314	0.0495·	0.0510	−0.0011
	(0.0361)	(0.0299)	(0.0339)	(0.0447)
Observations	2,200	2,200	2,200	2,200
State FE	X	X	X	X
Quarter FE	X	X	X	X
COVID School Control	X	X	X	X

·$p < 0.1$, *$p < 0.05$

Notes: Each column reports a separate Poisson regression with an offset for child population log(child population). Models include state and quarter fixed effects and a COVID school-closure control. Standard errors are clustered at the state level.

Table 22. Poisson Estimates of Treatment Effects by Child Sex

	(1) Boys	(2) Girls
Mobile Betting	0.0602·	0.0621·
	(0.0339)	(0.0320)
In-Person Betting	0.0387	0.0323
	(0.0331)	(0.0338)
Observations	2,200	2,200
State FE	X	X
Quarter FE	X	X
COVID School Control	X	X

·$p < 0.1$

Notes: Each column reports a separate Poisson regression with an offset for child population log(child population). Models include state and quarter fixed effects and a COVID school-closure control. Standard errors are clustered at the state level.

The estimates by child sex in Table 22 show small and marginally significant increases for both boys and girls, with little evidence of differences between the two groups.

3.6 Conclusion

Legislation regarding the legal status and regulation of sports betting is still active in many states. Missouri, Georgia, and Minnesota among others have recently proposed bills to legalize sports betting. Bills proposed in North Carolina and New York look to add regulation on the age at which people can bet and what types of bets people can engage in and the SAFE Bet Act introduced into the U.S. congress proposes regulations on advertising and a ban on college sports prop bets. New Jersey, which played a central role in overturning PASPA, is now advancing new regulations to address problem gambling. The proposals would require operators to designate a Responsible Gaming Lead, implement monitoring systems to identify potentially risky behavior, and use a tiered intervention process ranging from education to direct outreach. In addition, pending legislation seeks to prohibit micro-betting, reflecting concern over the potential for impulsive wagering in real time. It is crucial to policymakers in this evolving environment to have ample information regarding the negative externalities caused by sports betting.

This paper provides causal estimates of the impact of sports betting legalization on reports of child maltreatment. Overall, I find that sports betting increases the rate of reports of maltreatment by 5-7%. A 6% increase corresponds to a 923 case increase per quarter per state that legalizes sports betting. Since thirty-two states, and the District of Columbia, legalized during the sample period of this study, this amounts to a 30,468 case increase in the United States each quarter. This number does not represent the number of victims, but about 13% of all reports are substantiated as abuse. A back of the envelope calculation, assuming the reports generated by the policy follow the substantiation rate of the overall sample, puts the current child maltreatment externality costs of sports betting

legalization at roughly $830 million per quarter.[11]. The increase in reports also has implications on the constraints of case workers and their ability to adequately investigate reports as well as municipal budget planning.

My paper adds to the existing literature of the negative externalities caused by sports betting legalization (Baker et al., 2024; Hollenbeck et al., 2024; Matsuzawa & Arnesen, 2024; Van Der Maas et al., 2022). While many states have chosen to legalize sports betting for the possible benefits related to state tax revenues or personal freedoms, these benefits may be offset by the growing number of unintended costs now including child abuse/neglect.

An important avenue for future research concerns the ways in which legalized sports betting, especially through mobile apps, integrates gambling into the home environment. Unlike casinos or racetracks, mobile platforms allow betting to occur during family time, often while parents are simultaneously watching sports with their children. This shift raises questions about how gambling in the household may compete with time, attention, and financial resources devoted to children. Future studies could examine whether parental engagement decreases when gambling is embedded in leisure activities at home, and whether children's exposure to betting behaviors shapes their own attitudes toward risk and gambling. Investigating these inter-generational and household-level effects would provide valuable insights into how sports betting legalization influences family well-being beyond direct financial strain.

[11]Fang et al. (2012) estimate of the cost of nonfatal incidents of child maltreatment at $210,012 per victim

APPENDIX A

LEADER IDENTITY, EVALUATIONS, AND PERFORMANCE: EVIDENCE FROM A LABORATORY EXPERIMENT APPENDIX

A.1 Training Time

This appendix presents estimates of the effect of leader identity on the time participants spent on the training materials. Across specifications, neither leader gender nor sexual orientation is associated with statistically significant differences in training time. The coefficient on female leader is negative and moderately sized but imprecisely estimated, and the gay leader coefficient is near zero. These results suggest that leader identity does not influence on average how much effort participants invest in preparing for the task, consistent with the null performance results in the main text.

Table A.1. Effects of Leader Gender and Sexuality on Training Time

	Training Time
Intercept	-62.317
	(54.808)
Leader is a Woman	-22.214
	(20.604)
Leader is Gay	2.291
	(14.613)
Woman Leader × Gay Leader	9.485
	(18.879)
Controls	Yes
N	454

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

A.2 Sample Subset: Passed Manipulation Check

Tables A.14 through A.16 replicate the main analyses restricting the sample to participants who correctly identified their assigned leader's gender and sexual orientation.

For item completion and accuracy (Tables A.14 and A.15), the pattern of null results from the main analysis is broadly preserved. The coefficient on gay leader becomes positive across most performance specifications. This pattern is in the opposite direction from what a discrimination-based account would predict, and may reflect that participants who attended carefully to their leader's identity responded more favorably to gay leaders in terms of task effort. These estimates are imprecise, however, and should be interpreted cautiously.

Leader evaluation results (Table A.16) are larger in magnitude in this subsample than in the full sample. Female leaders are rated as more effective and more typical with larger and more precisely estimated coefficients, consistent with the main results. The interaction between gay and female leader remains negative for typicality, reinforcing the finding that gay women leaders are perceived as less prototypical than would be implied by the additive effects of gender and sexuality alone.

Table A.2. Effects of Leader Gender and Sexuality on Item Completion

	Items		Max Items		Strict Items	
Intercept	23.658***	22.163***	32.342***	31.070***	21.423***	19.255**
	(0.668)	(6.070)	(0.971)	(7.785)	(0.706)	(6.119)
Leader is a Woman	0.117	-0.243	-0.090	-0.448	0.135	-0.145
	(1.025)	(0.955)	(1.488)	(1.333)	(1.022)	(0.992)
Leader is Gay	0.417	0.659	0.389	1.050	1.293	1.395
	(1.058)	(1.073)	(1.535)	(1.527)	(1.094)	(1.102)
Woman Leader × Gay Leader	0.975	0.551	1.071	0.278	-0.018	-0.517
	(1.635)	(1.509)	(2.346)	(2.138)	(1.628)	(1.544)
Controls	No	Yes	No	Yes	No	Yes
Passed Manipulation Check	Yes	Yes	Yes	Yes	Yes	Yes
N	355	330	355	330	355	330

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.3. Effects of Leader Gender and Sexuality on Accuracy

	Pct. Correct		Max Pct.		Strict Pct.	
Intercept	59.949***	34.813**	80.935***	49.410**	54.209***	27.932*
	(1.005)	(11.464)	(1.135)	(15.811)	(1.308)	(12.422)
Leader is a Woman	0.418	-0.088	-0.502	-0.980	1.081	0.685
	(1.539)	(1.610)	(1.760)	(1.766)	(1.898)	(2.009)
Leader is Gay	1.875	-0.126	2.266	0.147	4.002*	1.535
	(1.568)	(1.645)	(1.747)	(1.868)	(1.940)	(2.027)
Woman Leader × Gay Leader	1.061	1.956	0.856	1.445	-1.578	-1.008
	(2.346)	(2.421)	(2.463)	(2.509)	(2.873)	(2.964)
Controls	No	Yes	No	Yes	No	Yes
Passed Manipulation Check	Yes	Yes	Yes	Yes	Yes	Yes
N	355	330	355	330	355	330

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.4. Effects of Leader Gender and Sexuality on Leader Evaluations

	Effectiveness		Typicality	
Intercept	3.502***	1.522***	3.383***	2.287**
	(0.086)	(0.400)	(0.074)	(0.860)
Leader is a Woman	0.210.	0.268*	0.299**	0.342***
	(0.117)	(0.104)	(0.106)	(0.103)
Leader is Gay	0.223	0.043	0.252*	0.095
	(0.142)	(0.145)	(0.121)	(0.113)
Woman Leader × Gay Leader	-0.025	-0.011	-0.325.	-0.332.
	(0.195)	(0.176)	(0.185)	(0.175)
Controls	No	Yes	No	Yes
Passed Manipulation Check	Yes	Yes	Yes	Yes
N	337	326	341	330

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

A.3 Sample Subset: Participant Gender

Table A.5. Effects of Leader Gender and Sexuality on Item Completion by Participant Gender

	Items		Max Items		Strict Items	
	Men Only	Women Only	Men Only	Women Only	Men Only	Women Only
Leader is a Woman	2.012	-0.979	2.929	-1.954	2.538	-0.948
	(1.665)	(1.173)	(2.399)	(1.605)	(1.695)	(1.246)
Leader is Gay	-0.010	1.369	-1.312	1.954	1.880	1.987
	(1.954)	(1.334)	(2.497)	(1.935)	(1.973)	(1.366)
Woman Leader × Gay Leader	-1.061	0.622	-1.378	0.480	-3.542	-0.007
	(2.678)	(1.853)	(3.593)	(2.559)	(2.801)	(1.894)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	97	212	97	212	97	212

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

A.4 Sample Subset: Participant Sexuality

Table A.6. Effects of Leader Gender and Sexuality on Item Completion by Participant Sexuality

	Items		Max Items		Strict Items	
	Straight Only	Gay Only	Straight Only	Gay Only	Straight Only	Gay Only
Leader is a Woman	-0.074	0.444	-0.556	0.914	0.580	0.301
	(1.119)	(2.067)	(1.534)	(2.985)	(1.139)	(2.260)
Leader is Gay	0.825	1.338	1.218	1.309	1.344	3.488
	(1.194)	(2.624)	(1.605)	(4.024)	(1.176)	(2.774)
Woman Leader × Gay Leader	1.135	-2.500	0.519	-2.980	-0.177	-4.133
	(1.745)	(3.235)	(2.367)	(4.785)	(1.806)	(3.320)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	219	90	219	90	219	90

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.7. Effects of Leader Gender and Sexuality on Accuracy by Participant Sexuality

	Pct. Correct		Max Pct.		Strict Pct.	
	Straight Only	Gay Only	Straight Only	Gay Only	Straight Only	Gay Only
Leader is a Woman	-0.008	0.464	-1.603	1.526	2.062	0.540
	(1.944)	(2.846)	(2.124)	(3.467)	(2.333)	(4.098)
Leader is Gay	-0.570	2.757	-0.340	2.872	0.652	7.767
	(1.955)	(3.743)	(2.349)	(4.332)	(2.200)	(4.867)
Woman Leader × Gay Leader	3.680	-1.743	2.609	-2.495	-0.271	-5.762
	(2.888)	(4.291)	(3.223)	(4.691)	(3.506)	(5.884)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	219	90	219	90	219	90

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

A.5 Sample Subset: Low vs High Heteronormative Beliefs

Table A.8. Effects of Leader Gender and Sexuality on Item Completion by Heteronormativity

	Items		Max Items		Strict Items	
	Low Hetnorm	High Hetnorm	Low Hetnorm	High Hetnorm	Low Hetnorm	High Hetnorm
Leader is a Woman	0.850	-0.048	0.478	-0.045	0.639	0.666
	(1.277)	(1.382)	(1.780)	(1.879)	(1.352)	(1.426)
Leader is Gay	-0.209	-1.586	-0.176	-1.021	0.571	-1.729
	(1.379)	(1.420)	(1.836)	(1.989)	(1.513)	(1.458)
Woman Leader × Gay Leader	-1.185	0.114	-1.138	-0.877	-1.872	-0.684
	(1.873)	(2.071)	(2.600)	(2.819)	(1.980)	(2.120)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	217	212	217	212	217	212

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.9. Effects of Leader Gender and Sexuality on Accuracy by Heteronormativity

	Pct. Correct		Max Pct.		Strict Pct.	
	Low Hetnorm	High Hetnorm	Low Hetnorm	High Hetnorm	Low Hetnorm	High Hetnorm
Leader is a Woman	1.606	0.496	0.835	-0.259	1.472	2.538
	(1.857)	(2.271)	(1.889)	(2.578)	(2.601)	(2.678)
Leader is Gay	-1.743	-2.251	-1.324	-0.632	-0.047	-2.295
	(2.121)	(2.038)	(2.226)	(2.475)	(2.698)	(2.564)
Woman Leader × Gay Leader	-1.719	0.634	-2.228	-0.270	-3.732	-1.951
	(2.805)	(3.179)	(2.885)	(3.679)	(3.703)	(3.923)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	217	212	217	212	217	212

Heteroskedasticity-robust standard errors in parentheses
Control coefficients omitted for brevity
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

A.6 Imputation of Missing Values

A.6.1 Missing Data Summary. Covariate missingness was low across all control variables (Table A.10). Missing values for continuous controls were replaced with variable means; missing values for categorical controls were replaced with the mode. Missingness indicators were included in all models with controls. The below tables establish that treatment effect estimates are substantively unchanged between listwise deletion and mean/mode imputation.

Table A.10. Missing Data Summary for Imputed Control Variables

Variable	Type	N Missing	% Missing
Continuous			
Volunteer Hours	Numeric	24	4.7%
Heteronormativity	Numeric	28	5.4%
Leader Typicality	Numeric	26	5.0%
Leader Effectiveness	Numeric	36	7.0%
Categorical			
Nonprofit Preference	Categorical	33	6.4%
Participant Sex	Categorical	23	4.5%
Participant Race	Categorical	22	4.3%
Participant Gender	Categorical	20	3.9%
Total observations		516	
Complete cases (with controls)		458	88.8%

Full sample. Treatment variables have no missing data.

Table A.11. Effects of Leader Gender and Sexuality on Leader Evaluations

	Effectiveness		Typicality	
Intercept	3.497***	1.888***	3.370***	2.510***
	(0.074)	(0.410)	(0.067)	(0.699)
Leader is a Woman	0.118	0.191**	0.229**	0.266***
	(0.102)	(0.087)	(0.095)	(0.089)
Leader is Gay	0.117	0.112	0.139.	0.140.
	(0.103)	(0.100)	(0.094)	(0.090)
Woman Leader × Gay Leader	0.088	-0.020	-0.204.	-0.264**
	(0.145)	(0.127)	(0.140)	(0.131)
Controls	No	Yes	No	Yes
N	516	516	516	516

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.12. Effects of Leader Gender and Sexuality on Accuracy

	Pct. Correct		Generous Pct.		Strict Pct.	
Intercept	59.586***	28.862***	80.390***	45.611***	53.897***	20.431*
	(0.972)	(9.369)	(1.102)	(12.781)	(1.211)	(10.423)
Leader is a Woman	0.407	0.283	-0.171	0.051	0.809	0.535
	(1.436)	(1.444)	(1.634)	(1.603)	(1.760)	(1.812)
Leader is Gay	-0.261	-0.044	0.182	0.051	0.670	0.733
	(1.399)	(1.420)	(1.534)	(1.599)	(1.758)	(1.767)
Woman Leader × Gay Leader	-0.418	-0.386	-0.750	-1.249	-1.970	-1.758
	(2.114)	(2.087)	(2.336)	(2.300)	(2.603)	(2.542)
Controls	No	Yes	No	Yes	No	Yes
N	516	516	516	516	516	516

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.13. Effects of Leader Gender and Sexuality on Item Completion

	Items		Generous Items		Strict Items	
Intercept	23.649***	17.519***	32.231***	26.673***	21.410***	13.323**
	(0.672)	(4.950)	(0.937)	(6.427)	(0.691)	(5.330)
Leader is a Woman	0.117	-0.197	0.097	-0.209	0.057	-0.291
	(0.962)	(0.932)	(1.361)	(1.274)	(0.965)	(0.962)
Leader is Gay	-0.817	-0.705	-0.839	-0.771	-0.490	-0.415
	(0.991)	(1.023)	(1.383)	(1.388)	(1.008)	(1.042)
Woman Leader × Gay Leader	0.576	0.794	0.453	0.598	0.115	0.364
	(1.417)	(1.349)	(1.982)	(1.846)	(1.427)	(1.384)
Controls	No	Yes	No	Yes	No	Yes
N	516	516	516	516	516	516

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.14. Effects of Leader Gender and Sexuality on Item Completion

	Items		Max Items		Strict Items	
Intercept	23.658***	21.561***	32.342***	30.114***	21.423***	18.776***
	(0.668)	(5.864)	(0.971)	(7.490)	(0.706)	(5.925)
Leader is a Woman	0.117	-0.173	-0.090	-0.264	0.135	-0.183
	(1.025)	(0.957)	(1.488)	(1.347)	(1.022)	(0.995)
Leader is Gay	0.417	0.595	0.389	0.864	1.293	1.219
	(1.058)	(1.075)	(1.535)	(1.524)	(1.094)	(1.103)
Woman Leader × Gay Leader	0.975	0.605	1.071	0.263	-0.018	-0.309
	(1.635)	(1.514)	(2.346)	(2.141)	(1.628)	(1.537)
Controls	No	Yes	No	Yes	No	Yes
Passed Manipulation Check	Yes	Yes	Yes	Yes	Yes	Yes
N	355	355	355	355	355	355

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.15. Effects of Leader Gender and Sexuality on Accuracy

	Pct. Correct		Max Pct.		Strict Pct.	
Intercept	59.949***	35.466***	80.935***	49.662***	54.209***	28.431**
	(1.005)	(11.416)	(1.135)	(15.666)	(1.308)	(12.347)
Leader is a Woman	0.418	-0.069	-0.502	-0.623	1.081	0.549
	(1.539)	(1.580)	(1.760)	(1.742)	(1.898)	(1.978)
Leader is Gay	1.875	0.356	2.266	0.672	4.002**	1.885
	(1.568)	(1.600)	(1.747)	(1.782)	(1.940)	(1.954)
Woman Leader × Gay Leader	1.061	1.572	0.856	0.670	-1.578	-1.079
	(2.346)	(2.405)	(2.463)	(2.483)	(2.873)	(2.926)
Controls	No	Yes	No	Yes	No	Yes
Passed Manipulation Check	Yes	Yes	Yes	Yes	Yes	Yes
N	355	355	355	355	355	355

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

Table A.16. Effects of Leader Gender and Sexuality on Leader Evaluations

	Effectiveness		Typicality	
Intercept	3.508***	1.524***	3.387***	2.386***
	(0.082)	(0.398)	(0.071)	(0.867)
Leader is a Woman	0.199*	0.263***	0.288***	0.321***
	(0.111)	(0.100)	(0.102)	(0.100)
Leader is Gay	0.211.	0.059	0.240**	0.143
	(0.133)	(0.135)	(0.115)	(0.113)
Woman Leader × Gay Leader	-0.016	-0.030	-0.309*	-0.357**
	(0.185)	(0.166)	(0.177)	(0.171)
Controls	No	Yes	No	Yes
Passed Manipulation Check	Yes	Yes	Yes	Yes
N	355	355	355	355

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, . $p < 0.15$

APPENDIX B

ARTIFICIAL AGING AND FACE MORPHING: TESTING THE DOUBLE STANDARD OF AGING HYPOTHESIS APPENDIX

B.1 Sensitivity to the Assumed Form of Age Progression

Equation 2.2 assumes that perceived age increases as a linear function of morph progression. This assumption may not capture the dynamics of facial age perception. For example, perceptual sensitivity to age-related facial change may be relatively low during early adulthood and increase suddenly in middle age. A power function with exponent 13 (i.e., a cubic root transformation) represents the limiting case of such concavity within the parameter range examined here.

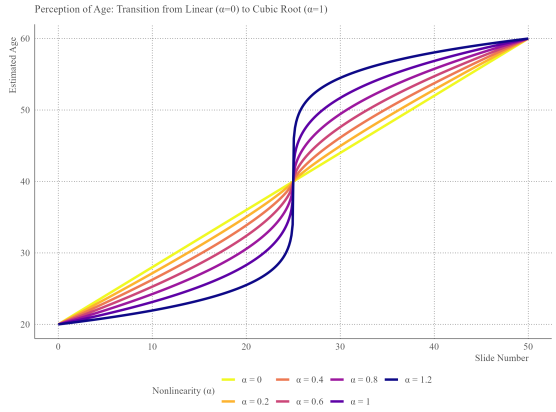


Figure B.1. Effect of the non-linearity parameter

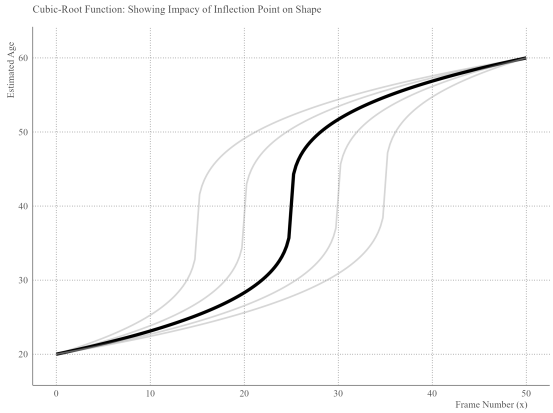


Figure B.2. Effect of the inflection point parameter

Figure B.3. Parameterization of the power function

To assess the robustness of the reported findings to this modeling assumption, we re-compute estimated age across a grid of functional forms parameterized by an exponent $\alpha \in [13,\ 1]$, where $\alpha = 1$ recovers the linear model and $\alpha = 13$ corresponds to the cubic root, and an inflection point parameter governing the locus of maximum curvature within the morph sequence. The full

mixed-effects model was refit at each grid point and the resulting p-values for the Race $\times$ Gender interaction, target gender main effect, and target race main effect are shown as contours in Figures B.4 and B.5 for young-face and old-face starting conditions respectively, with lighter hues encoding smaller p-values and white contours demarcating the $p = 0.05$ (solid) and $p = 0.01$ (dashed) boundaries.

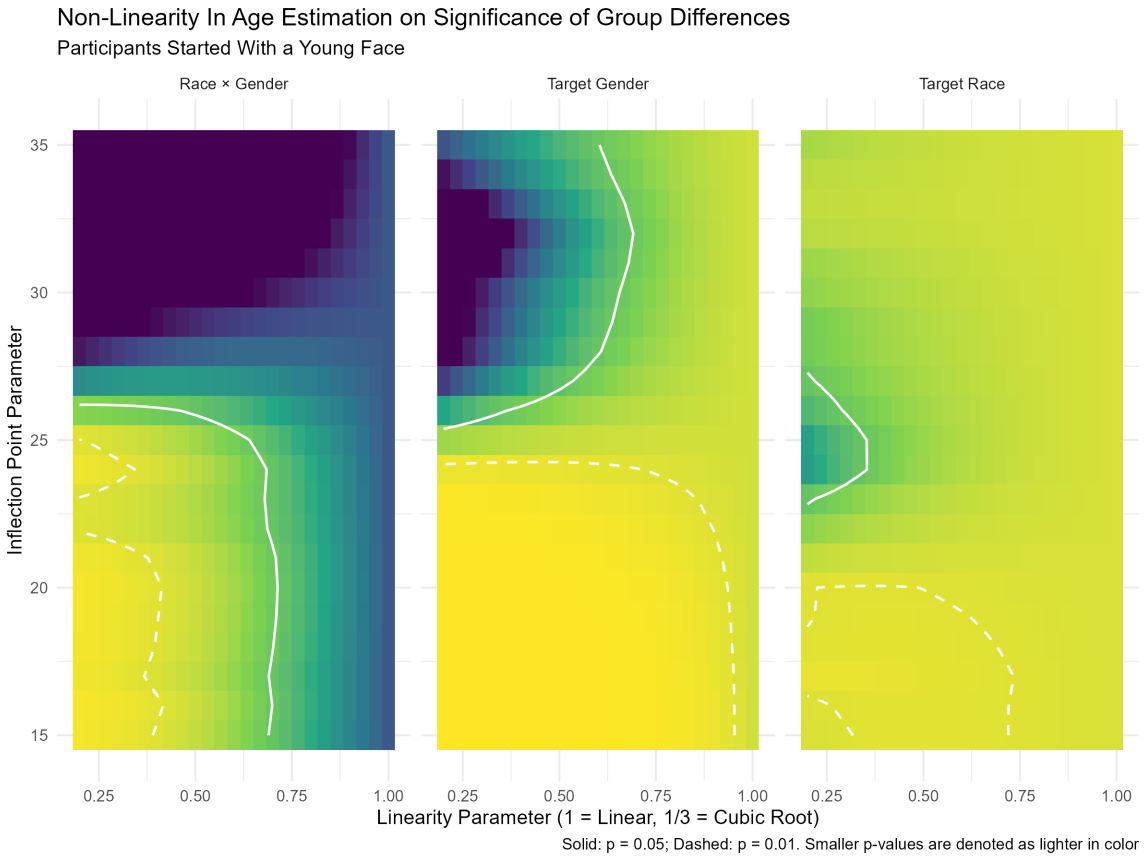


Figure B.4. Significance landscape for the three omnibus tests (Race $\times$ Gender interaction, target gender main effect, target race main effect) as a function of the linearity exponent α (x-axis; 1 = linear, 13 = cubic root) and inflection point parameter (y-axis), for participants who initiated the task from the young-face endpoint. Cell color encodes the p-value from the corresponding mixed-effects model (lighter = smaller p). Solid and dashed white contours mark $p = 0.05$ and $p = 0.01$ respectively.

As shown in Figure B.4, the target gender main effect in the young-face condition is robust across the majority of the parameter space, retaining statistical

significance at $p < 0.05$ across most combinations of α and inflection point in the Target Gender panel. The Race $\times$ Gender interaction attains statistical significance only within a circumscribed region corresponding to lower exponents and moderate inflection points, indicating that this effect is contingent on the assumed functional form of the age-morph mapping. The target race main effect was near uniformly statistically significant across parameterizations. In the old-face condition (Figure B.5), the contours are different: the Race $\times$ Gender interaction is nearly always statistically significant, with neither gender nor race on their own being statistically significant across most of the space examined (an exception is for highly non-linear, late-inflection models for gender). The old-face interaction reported under the linear model should accordingly be treated as hypothesis-generating rather than confirmatory.

B.2 Robustness Check: Classical Split-Plot ANOVA

The primary analyses are estimated as linear mixed-effects models via REML, a specification that offers several advantages — accommodation of unbalanced cell sizes, partial missingness handled through model-based estimation rather than listwise deletion, and Satterthwaite-approximated degrees of freedom that account for the random-effect structure. As a robustness check, all four models were re-estimated using a classical split-plot ANOVA, which differs from the primary specification in three consequential ways.

First, the classical approach computes F-ratios using a fixed error term structure derived from the balanced-design assumption: between-subjects effects are tested against a between-subjects error mean square, and within-subjects effects are tested against a pooled within-subjects error mean square. The mixed-effects approach, by contrast, constructs separate variance components for each random

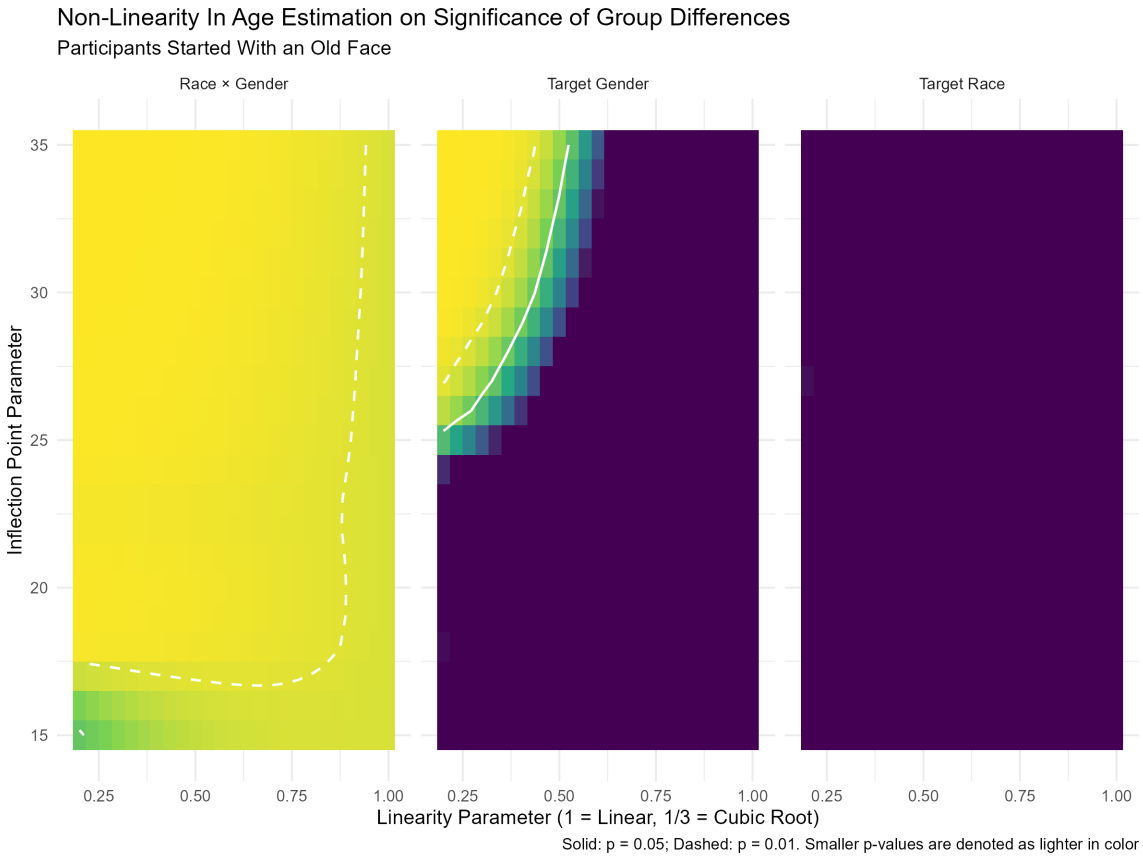


Figure B.5. Significance landscape as in Figure B.4, for participants who initiated the task from the old-face endpoint.

effect and derives denominator degrees of freedom from the estimated covariance structure via the Satterthwaite approximation. In practice, this means the two approaches can yield meaningfully different F-ratios and p-values when random-effect variance is poorly identified or when the data are unbalanced.

Second, the classical split-plot ANOVA requires complete data across all within-subjects cells: any participant missing a response for even one condition is excluded entirely from the analysis. The mixed-effects specification retains partial observations, estimating cell means from whatever data are available under the assumption of missing at random. This distinction is consequential in the present data, where several participants are missing responses.

Third, the classical specification does not include random slopes — it partitions variance into a participant random intercept and a single pooled within-subjects error term. Where that random slope accounts for meaningful individual variation in the gender effect, the classical approach will absorb this variance into the error term, potentially altering power and Type I error rates.

For participants who initiated the task from the young-face endpoint, the classical split-plot results for stop time are consistent with the primary analysis. The main effect of gender is large and highly significant ($F(1,\ 76) = 104.42$, $p < 0.001$, $\eta_p^2 = 0.58$), closely matching the primary estimate ($\eta_p^2 = 0.58$). We replictae the Race $\times$ Gender interaction ($F(3,\ 76) = 8.64$, $p < 0.001$, $\eta_p^2 = 0.25$), while the main effect of race did not reach the 5% threshold of statistical significance ($F(3,\ 76) = 2.55$, $p = 0.062$, $\eta_p^2 = 0.09$). One participant was excluded due to a missing data; under the mixed-effects specification this participant was retained via partial-data estimation.

For participants who initiated from the old-face endpoint, the classical results again closely reproduce the primary analysis. The gender main effect is significant ($F(1,\ 83) = 16.23$, $p < 0.001$, $\eta_p^2 = 0.16$), with an effect size identical to the primary estimate. Neither the race main effect ($F(3,\ 83) = 0.84$, $p = 0.474$, $\eta_p^2 = 0.03$) nor the interaction ($F(3,\ 83) = 1.69$, $p = 0.175$, $\eta_p^2 = 0.06$) was statistically significant.

When looking at $\widehat{\text{Age}}$, the young-face-starting condition yields significant main effects of both race ($F(3,\ 75) = 4.23$, $p = 0.008$, $\eta_p^2 = 0.15$) and gender ($F(1,\ 75) = 4.33$, $p = 0.041$, $\eta_p^2 = 0.06$), consistent with the primary results ($\eta_p^2 = 0.13$ and $\eta_p^2 = 0.05$, respectively). The interaction is not statistically significant ($F(3,\ 75) = 1.35$, $p = 0.266$, $\eta_p^2 = 0.05$). Two participants were excluded

due to missing data; the modest upward shift in the race effect size ($\eta_p^2 = 0.15$ vs. 0.13) most plausibly reflects the composition of the analytic sample under the non-missing restriction rather than a substantive difference between specifications.

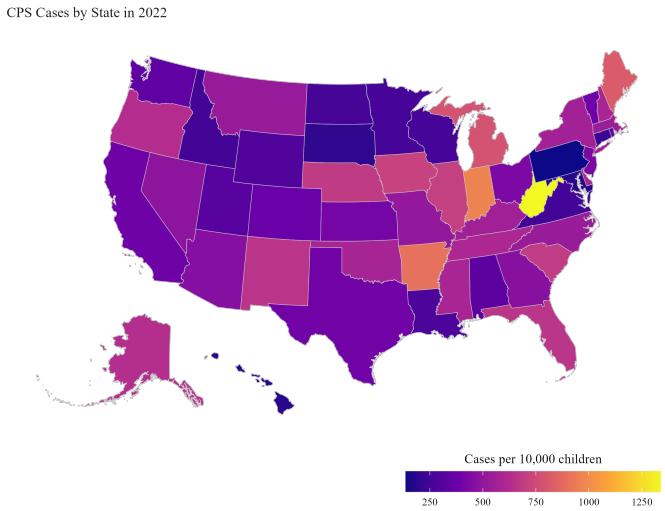
For the old-face-starting condition, the two approaches diverge. As in the primary model, neither race ($F(3,\ 76) = 0.41$, $p = 0.748$, $\eta_p^2 = 0.02$) nor gender ($F(1,\ 76) = 0.67$, $p = 0.415$, $\eta_p^2 = 0.01$) are significant. The Race $\times$ Gender interaction, however, is significant under the classical specification ($F(3,\ 76) = 4.41$, $p = 0.006$, $\eta_p^2 = 0.15$), with a considerably larger effect size than in the primary model ($\eta_p^2 = 0.03$). This divergence is most plausibly attributable to the combined influence of exclusion of seven participants with missing data and the singular fit of the primary `lme4` model, which indicated that the participant-by-gender random slope is estimated at the boundary of the parameter space. When the random slope contributes negligible variance — as the singular fit implies — the classical approach's pooled error term may be more efficient, but the non-missing restriction simultaneously alters the sample. Given the uncertainty introduced by both the singular fit and the differential exclusions, neither specification should be treated as definitive for this model, and the interaction effect in the old-face estimated age condition warrants cautious interpretation regardless of which approach is preferred.

APPENDIX C

THE IMPACT OF LEGALIZED SPORTS BETTING ON CHILD MALTREATMENT: EVIDENCE FROM CHILD PROTECTIVE SERVICES REPORTS APPENDIX

C.1 Child Maltreatment Reporting Trends

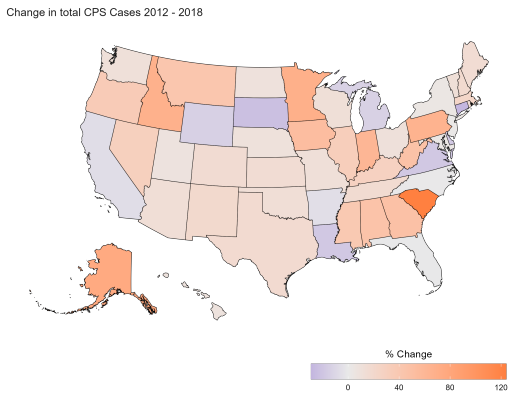
Figure C.1. Spatial Distribution of Child Abuse Reports



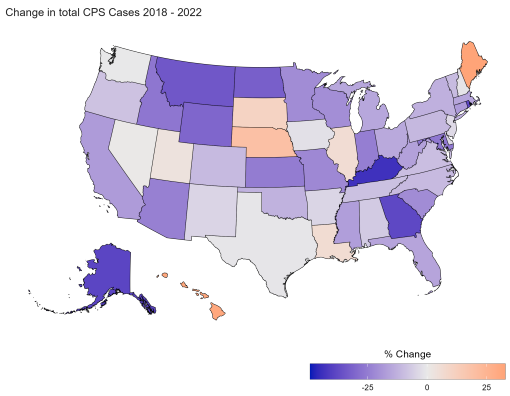
Notes: This map shows the number of reports of child abuse and neglect reports per 10,000 children across the United States in 2022. It shows variation in the severity of child maltreatment and that the prevalence is not directly connected to the change in reports.

Figure C.2. Spatial Distribution of the Change in Child Abuse Reports

Panel A: Percent Change in Reports from 2012 to 2018



Panel B: Percent Change in Reports from 2018 to 2022



Notes: Panel A and B aim to decompose the percentage change in reports of abuse and neglect across the U.S. shown in Figure 12. Panel A shows that from 2012 to 2018 most states saw an increase in the number of reports of child maltreatment. Panel B shows that most states saw a decrease in the number of reports from 2018 to 2022 largely driven by the height of the COVID pandemic and it's impacts on reporting. Though, this decrease in case volume is less homogeneous than the increases seen prior to 2018.

C.2 Main Specification by Fortnight

The main specification of this paper uses data aggregated to the quarter-level. This aggregation helps to compare the results to the literature and to reduce noise and seasonality seen at lower aggregations. To show that this decision does not dramatically impact the results, Table C.1 and C.2 provide the results to my main results if a fortnight panel is used instead of a quarterly panel.

Table C.1. OLS and Poisson Estimates of Treatment Effects: Any Sports Betting

	(1) OLS	(2) OLS	(3) Poisson	(4) Poisson
Any Sports Betting	0.0683$^{\cdot}$	0.0658$^{\cdot}$	0.0724$^{\cdot}$	0.0702$^{\cdot}$
	(0.0392)	(0.0378)	(0.0420)	(0.0407)
N	13,200	13,200	13,200	13,200
State FE	X	X	X	X
Biweekly FE	X	X	X	X
COVID School Closure Control		X		X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Notes: Models (1) and (2) are OLS regressions weighted by state population under 18 years old. Models (3) and (4) are Poisson regressions with an offset for child population. All specifications include state and time fixed effects. Standard errors are clustered at the state level. Model (2) and Model (4) include a COVID-19 school closure indicator. All models control for child population either as a covariate (OLS) or as an offset (Poisson).

Table C.2. OLS and Poisson Estimates of Treatment Effects: Mobile vs In-Person

	(1) OLS	(2) OLS	(3) Poisson	(4) Poisson
Mobile Betting	0.0604$^{\cdot}$	0.0568$^{\cdot}$	0.0648$^{\cdot}$	0.0614$^{\cdot}$
	(0.0336)	(0.0324)	(0.0345)	(0.0331)
In-Person Betting	0.0362	0.0361	0.0371	0.0372
	(0.0328)	(0.0325)	(0.0314)	(0.0314)
N	13,200	13,200	13,200	13,200
State FE	X	X	X	X
Biweekly FE	X	X	X	X
COVID School Closure Control		X		X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

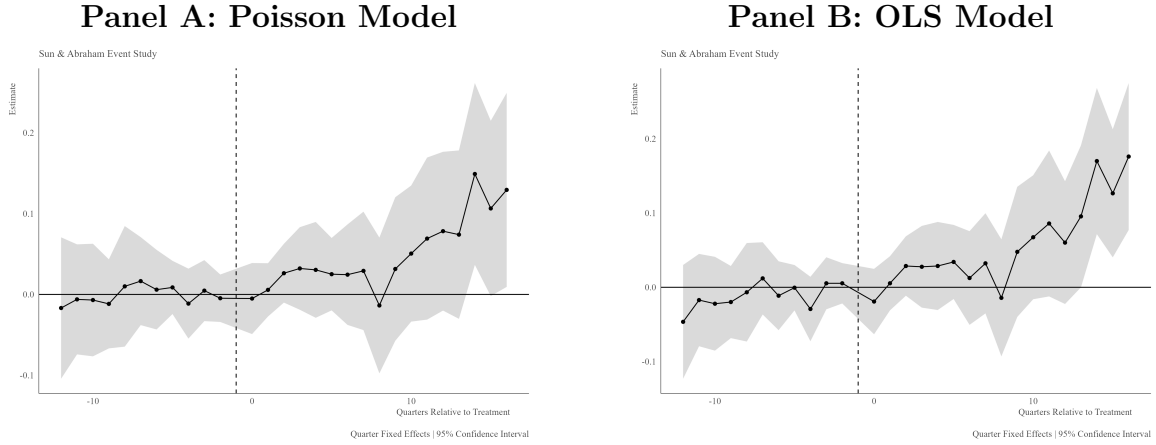
Notes: Models (1) and (2) are OLS regressions weighted by state population under 18 years old. Models (3) and (4) are Poisson regressions with an offset for child population. All specifications include state and time fixed effects. Standard errors are clustered at the state level. Model (2) and Model (4) include a COVID-19 school closure indicator. All models control for child population either as a covariates (OLS) or as an offset (Poisson).

C.3 Staggered Adoption Robustness Check

The Sun and Abraham (2021) estimator identifies treatment effects by interacting cohort and relative-time indicators, estimating cohort-specific event-study coefficients, and aggregating them using non-negative weights based on

untreated comparison groups. This approach corrects the bias and negative-weight issues inherent in traditional two-way fixed effects estimators, but requires sufficient untreated comparisons in each period. Panels A and B of Figure C.3 show the event study estimates for the Sun & Abraham estimator. It shows that pre-treatment coefficients remain near and are not statistically different from zero, supporting the parallel trends assumption. The effect of treatment is delayed in these figures becoming statistically significant in the third year following legalization.

Figure C.3. Sun & Abraham Event Study Estimates



Notes: The plots show Sun and Abraham (2021)-style event-study estimates for the effect of any form of sports-betting legalization on child maltreatment reports.

Since the effects evolve over time, this may bias the pooled TWFE coefficients reported in the main table. To address this concern, I re-create the main effects of any gambling legalization treatment using the average treatment effect on treated units. Table C.3 shows qualitatively similar effects to Table 13, with an increase in the number of investigations post-legalization, but the effects are smaller. The Sun & Abraham estimator reported uses only never-treated units

as the comparison group. This may be unreasonable given the large number of adopters leaving only about twelve never treated units in the sample.

Table C.3. Sun and Abraham ATTs

	Poisson		OLS	
	(1)	(2)	(3)	(4)
Any Sports Betting	0.0352	0.0328	0.0365	0.0334
	(0.0274)	(0.0260)	(0.0259)	(0.0246)
N	2,200	2,200	2,200	2,200
State FE	X	X	X	X
Quarter FE	X	X	X	X
COVID School Controls		X		X

$^{\cdot}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

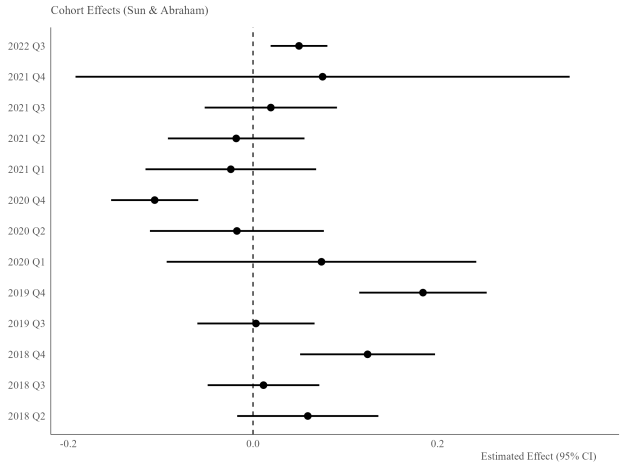
Notes: Columns (1)–(2) report Poisson estimates; Columns (3)–(4) report OLS estimates. Each pair contrasts models with and without COVID school-related controls. Poisson specifications include an offset for the child population. All models include state and quarter fixed effects. Standard errors are clustered at the state level.

Heterogeneous effects may in-part explain why the Sun and Abraham (2021) ATT effects are smaller in magnitude than the TWFE main effects. Figure **??** shows the Sun and Abraham (2021) ATT estimate separately for each policy change cohort. These results suggest that some states are experiencing disproportionate impacts from sports betting. The cohort which legalized in the fourth quarter of 2020 is a notable standout. It is the only cohort to shows a negative, statistically significant impact of sports gambling legalization on child maltreatment investigations. This estimate is likely heavily influenced by reporting difficulties brought on by the COVID-19 pandemic and not fully captured by controls. Additionally, this only consists of one state, Tennessee.

C.4 Betting Handle As Treatment

Sports betting legalization was a binary treatment in accordance to the law, but its popularity has grown over time. This popularity can be gauged by the

Figure C.4. Cohort Effects (Sun & Abraham)



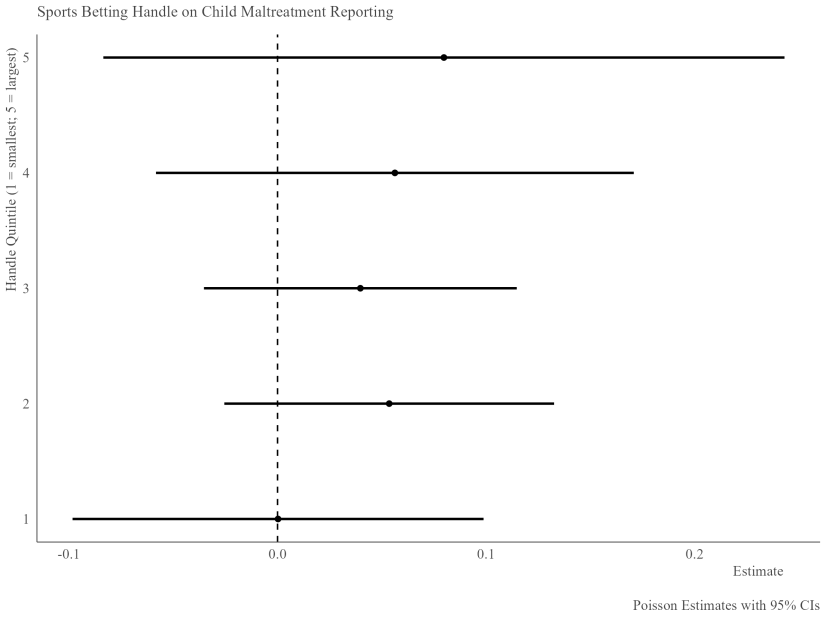
Notes: Points show cohort-level ATT estimates; bars are 95% confidence intervals. The horizontal dashed line denotes zero.

amount of money that people collectively wager (handle) each quarter. Specifically, per capita handle is one way to view how states that legalized sports betting may experience heterogeneous treatments. States with higher per capita handle may experience broader social or economic effects if betting activity leads to problem gambling and negative outcomes such as child maltreatment.

Figure C.5 plots estimates from a Poisson model which breaks per capita handle into quintiles and interacts it with treatment status. In the smallest quintile, the coefficient is near zero. The magnitude of the coefficients, more or less, increase as states spend more on sports betting. This is suggestive evidence that as more people engage in sports betting the more likely that negative outcomes may occur.

Table C.4 reports results from a Poisson model which adds per capita betting as a continuous treatment alongside the dummy treatment indicators. The estimated coefficient on per capita betting handle accounts for about a three percent increase in the number of reports per $100 spent per person per quarter

Figure C.5. Treatment Effects by Betting Handle Quintile



Notes: Bars are 95% confidence intervals. The horizontal dashed line denotes zero.

on sports betting. More interestingly, it appears that the availability of betting in any capacity may have an impact on child maltreatment reports. One possible explanation for this relationship is a change in time use following sports betting legalization that reallocates time away from parenting towards sports betting.

Table C.4. Poisson Estimates of Combined and Separate Betting Treatments

	(1)	(2)
Per Capita Betting Handle	0.0003	0.0002
	(0.0003)	(0.0003)
Any Betting	0.0480	
	(0.0358)	
Mobile Betting		0.0494$^{\cdot}$
		(0.0254)
In-Person Betting		0.0329
		(0.0328)
N	2,200	2,200
State FE	X	X
Quarter FE	X	X

$^{\cdot}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Notes: All specifications show Poisson estimates with a child population offset which include state and quarter fixed effects.

REFERENCES CITED

AARP. (2014). You're old, i'm not: How americans really feel about aging. *AARP: The Magazine*, *40*, 40–43.

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Akerlof, G. A., & Kranton, R. E. (2000). Economics and Identity*. *Quarterly Journal of Economics*, *115*(3), 715–753. https://doi.org/10.1162/003355300554881

Akerlof, G. A., & Kranton, R. E. (2005). Identity and the Economics of Organizations. *Journal of Economic Perspectives*, *19*(1), 9–32. https://doi.org/10.1257/0895330053147930

Aksoy, C. G., Carpenter, C. S., Frank, J., & Huffman, M. L. (2019). Gay glass ceilings: Sexual orientation and workplace authority in the UK. *Journal of Economic Behavior & Organization*, *159*, 167–180. https://doi.org/10.1016/j.jebo.2019.01.013

American Gaming Association. (2023). Assessing shifts in the sports betting market 5 years post-paspa. https://www.americangaming.org/wp-content/uploads/2023/05/AGA_PASPA_LSBResearch.pdf

American Gaming Association. (2025, May). State of the states 2025 [Accessed: 2025-11-10].

Antonakis, J., d'Adda, G., Weber, R. A., & Zehnder, C. (2022). "Just Words? Just Speeches?" On the Economic Value of Charismatic Leadership. *Management Science*, *68*(9), 6355–6381. https://doi.org/10.1287/mnsc.2021.4219

Antonakis, J., Fenley, M., & Liechti, S. (2011). Can Charisma Be Taught? Tests of Two Interventions. *Academy of Management Learning & Education*, *10*(3), 374–396. https://doi.org/10.5465/amle.2010.0012

Aunga, D., & Masare, O. (2017). Effect of leadership styles on teacher's performance in primary schools of Arusha District Tanzania. *International Journal of Educational Policy Research and Review*, *4*(4).

Badgett, M. L., Waaldijk, K., & Rodgers, Y. V. D. M. (2019). The relationship between LGBT inclusion and economic development: Macro-level evidence. *World Development*, *120*, 1–14. https://doi.org/10.1016/j.worlddev.2019.03.011

Baker, S., Balthrop, J., Johnson, M., Kotter, J., & Pisciotta, K. (2024, November). *Gambling Away Stability: Sports Betting's Impact on Vulnerable Households* (tech. rep. No. w33108). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w33108

Bass, B. M. (1985). Leadership: Good, better, best. *Organizational Dynamics*, *13*(3), 26–40. https://doi.org/10.1016/0090-2616(85)90028-2

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bazzini, D. G., McIntosh, W. D., Smith, S. M., Cook, S., & Harris, C. (1997). The aging woman in popular film: Underrepresented, unattractive, unfriendly, and unintelligent. *Sex Roles*, *36*, 531–543. https://doi.org/10.1007/BF02766689

Ben-Michael, E., Feller, A., & Rothstein, J. (2021, June). *Synthetic Controls with Staggered Adoption* (tech. rep. No. w28886). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w28886

Berger, L. M., Cancian, M., Cuesta, L., & Noyes, J. L. (2016). Families at the Intersection of the Criminal Justice and Child Protective Services Systems. *The ANNALS of the American Academy of Political and Social Science*, *665*(1), 171–194. https://doi.org/10.1177/0002716216633058

Berkowitz, L. (1989). Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin*, *106*(1), 59–73. https://doi.org/10.1037/0033-2909.106.1.59

Berman, P. W., O'Nan, B. A., & Floyd, W. (1981). The double standard of aging and the social situation: Judgments of attractiveness of the middle-aged woman. *Sex Roles*, *7*, 87–96. https://doi.org/10.1007/BF00287795

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *Review of Economic Studies*, *91*(6), 3253–3285. https://doi.org/10.1093/restud/rdae007

Brooks-Gunn, J., & Duncan, G. J. (1997). The Effects of Poverty on Children. *The Future of Children*, *7*(2), 55. https://doi.org/10.2307/1602387

Brown, S. M., Orsi, R., Chen, P. C. B., Everson, C. L., & Fluke, J. (2022). The Impact of the COVID-19 Pandemic on Child Protection System Referrals and Responses in Colorado, USA. *Child Maltreatment*, *27*(1), 3–11. https://doi.org/10.1177/10775595211012476

Bugental, D. B., & Hehman, J. A. (2007). Ageism: A review of research and policy implications. *Social Issues and Policy Review*, *1*(1), 173–216. https://doi.org/10.1111/j.1751-2409.2007.00007.x

Burt, D. M., & Perrett, D. I. (1995). Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *259*(1355), 137–143. https://doi.org/10.1098/rspb.1995.0021

Calasanti, T. M., & Slevin, K. F. (2003). *Gender, social inequalities, and aging*. AltaMira Press.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Canetto, S. S., Kaminski, P. L., & Felicio, D. M. (1995). Typical and optimal aging in women and men: Is there a double standard? *International Journal of Aging and Human Development*, *40*, 187–207. https://doi.org/10.2190/RX0U-T56B-1G0F-266U

Card, D., & Dahl, G. B. (2011). Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior*. *The Quarterly Journal of Economics*, *126*(1), 103–143. https://doi.org/10.1093/qje/qjr001

Cecil, V., Pendry, L. F., Salvatore, J., Mycroft, H., & Kurz, T. (2022). Gendered ageism and gray hair: Must older women choose between feeling authentic and looking competent? *Journal of Women & Aging*, *34*(2), 210–225. https://doi.org/10.1080/08952841.2021.1899744

Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., Löckenhoff, C. E., De Bolle, M., & Terracciano, A. (2012). Stereotypes of age differences in personality traits: Universal and accurate? *Journal of Personality and Social Psychology*, *103*, 1050–1066. https://doi.org/10.1037/a0029712

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, *149*, 74–87. https://doi.org/10.1016/j.jebo.2018.02.024

Chrisler, J. C., Barney, A., & Palatino, B. (2016). Ageism and women's health. *Journal of Social Issues*, *72*, 86–104.

Chrisler, J. C., Rossini, M., & Newton, J. R. (2015). Older women, power, and the body. In V. Muhlbauer, J. C. Chrisler & F. L. Denmark (Eds.), *Women and aging: An international, intersectional power perspective* (pp. 9–30). Springer International Publishing.

Christensen, K. e. a. (2009). Perceived age as biomarker. *BMJ*, *339*, b5262.

Clark, L. H., & Griffin, M. (2008). Visible and invisible ageing: Beauty work as a response to ageism. *Ageing & Society*, *28*, 653–674. https://doi.org/10.1017/S0144686X07007003

Clarke, L. H., & Korotchenko, A. (2010). Shades of grey. *Ageing and Society*, *30*, 1011–1026.

Conger, R. D., Ge, X., Elder, G. H., Lorenz, F. O., & Simons, R. L. (1994). Economic stress, coercive family process, and developmental problems of adolescents. *Child Development*, *65*(2 Spec No), 541–561.

Conger, R. D., & Conger, K. J. (2002). Resilience in Midwestern Families: Selected Findings from the First Decade of a Prospective, Longitudinal Study. *Journal of Marriage and Family*, *64*(2), 361–373. https://doi.org/10.1111/j.1741-3737.2002.00361.x

Conger, R. D., & Donnellan, M. B. (2007). An Interactionist Perspective on the Socioeconomic Context of Human Development. *Annual Review of Psychology*, *58*(1), 175–199. https://doi.org/10.1146/annurev.psych.58.110405.085551

Correll, S. J. (2004). Constraints into Preferences: Gender, Status, and Emerging Career Aspirations. *American Sociological Review*, *69*(1), 93–113. https://doi.org/10.1177/000312240406900106

Cotter, D. A., Hermsen, J. M., Ovadia, S., & Vanneman, R. (2001). The Glass Ceiling Effect. *Social Forces*, *80*(2), 655–681. https://doi.org/10.1353/sof.2001.0091

Cuddy, A. J. C., & Fiske, S. T. (2002). Doddering but dear: Process, content, and function in stereotyping of older persons. In T. D. Nelson (Ed.), *Ageism: Stereotyping and prejudice against older persons* (pp. 3–26). MIT Press.

Cuddy, A. J. C., Norton, M. I., & Fiske, S. T. (2005). This old stereotype: The pervasiveness and persistence of the elderly stereotype. *Journal of Social Issues*, *61*, 267–285. https://doi.org/10.1111/j.1540-4560.2005.00405.x

Currie, J., & Spatz Widom, C. (2010). Long-Term Consequences of Child Abuse and Neglect on Adult Economic Well-Being. *Child Maltreatment*, *15*(2), 111–120. https://doi.org/10.1177/1077559509355316

De Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, *110*(9), 2964–2996. https://doi.org/10.1257/aer.20181169

Dehon, H., & Brédart, S. (2001). An 'other-race' effect in age estimation from faces. *Perception*, *30*(9), 1107–1113. https://doi.org/10.1068/p3122

Deutsch, F. M., Zalenski, C. M., & Clark, M. E. (1986). Is there a double standard of aging? *Journal of Applied Social Psychology*, *16*, 771–785. https://doi.org/10.1111/j.1559-1816.1986.tb01167.x

Deutsch, F. M., Zalenski, C. M., & Clark, M. E. (1991). Is there a double standard of aging? *Journal of Applied Social Psychology*, *21*(13), 1096–1108.

Donohue, J. J., Aneja, A., & Weber, K. D. (2019). Right-to-Carry Laws and Violent Crime: A Comprehensive Assessment Using Panel Data and a State-Level Synthetic Control Analysis. *Journal of Empirical Legal Studies*, *16*(2), 198–247. https://doi.org/10.1111/jels.12219

Dowling, N., Suomi, A., Jackson, A., Lavis, T., Patford, J., Cockman, S., Thomas, S., Bellringer, M., Koziol-Mclain, J., Battersby, M., Harvey, P., & Abbott, M. (2016). Problem Gambling and Intimate Partner Violence: A Systematic Review and Meta-Analysis. *Trauma, Violence, & Abuse*, *17*(1), 43–61. https://doi.org/10.1177/1524838014561269

Dube, A., & Zipperer, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies [Available at SSRN: https://ssrn.com/abstract=2589786]. *IZA Discussion Paper*, (8944).

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . . : A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*, 109–128. https://doi.org/10.1037/0033-2909.110.1.109

Eagly, A. H., & Chin, J. L. (2010). Diversity and leadership in a changing world. *American Psychologist*, *65*(3), 216–224. https://doi.org/10.1037/a0018957

Eagly, A. H., Johannesen-Schmidt, M. C., & Van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin*, *129*(4), 569–591. https://doi.org/10.1037/0033-2909.129.4.569

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Edleson, J. L. (1999). The Overlap Between Child Maltreatment and Woman Battering. *Violence Against Women*, *5*(2), 134–154. https://doi.org/10.1177/107780129952003

England, P., & McClintock, E. (2009). The gendered double standard of aging in u.s. marriage markets. *Population and Development Review*, *35*, 797–816. https://doi.org/10.1111/j.1728-4457.2009.00309.x

Etcoff, N. L., Stock, S., Haley, L. E., Vickery, S. A., & House, D. M. (2011). Cosmetics as a feature of the extended human phenotype: Modulation of the perception of biologically important facial signals. *PLoS ONE*, *6*, 1–9. https://doi.org/10.1371/journal.pone.0025656

Fang, X., Brown, D. S., Florence, C. S., & Mercy, J. A. (2012). The economic burden of child maltreatment in the United States and implications for prevention. *Child Abuse & Neglect*, *36*(2), 156–165. https://doi.org/10.1016/j.chiabu.2011.10.006

Fasoli, F., Maass, A., Paladino, M. P., & Sulpizio, S. (2017). Gay- and Lesbian-Sounding Auditory Cues Elicit Stereotyping and Discrimination. *Archives of Sexual Behavior*, *46*(5), 1261–1277. https://doi.org/10.1007/s10508-017-0962-0

Fest, S., Kvaløy, O., Nieken, P., & Schöttner, A. (2021). How (not) to motivate online workers: Two controlled field experiments on leadership in the gig economy. *The Leadership Quarterly*, *32*(6), 101514. https://doi.org/10.1016/j.leaqua.2021.101514

Fileborn, B., Thorpe, R., Hawkes, G., Minichiello, V., Pitts, M., & Dune, T. (2015). Sex, desire and pleasure: Considering the experiences of older Australian women. *Sexual and Relationship Therapy*, *30*(1), 117–130. https://doi.org/10.1080/14681994.2014.936722

Fischer, T., Dietz, J., & Antonakis, J. (2017). Leadership Process Models: A Review and Synthesis. *Journal of Management*, *43*(6), 1726–1753. https://doi.org/10.1177/0149206316682830

Font, S. A., & Maguire-Jack, K. (2020). It's not "Just poverty": Educational, social, and economic functioning among young adults exposed to childhood neglect, abuse, and poverty. *Child Abuse & Neglect*, *101*, 104356. https://doi.org/10.1016/j.chiabu.2020.104356

Foos, P. W., & Clark, M. C. (2011). Adult age and gender differences in perceptions of facial attractiveness: Beauty is in the eye of the older beholder. *Journal of Genetic Psychology*, *172*, 162–175. https://doi.org/10.1080/00221325.2010.526154

Frank, J. (2006). Gay Glass Ceilings. *Economica*, *73*(291), 485–508. https://doi.org/10.1111/j.1468-0335.2006.00516.x

Gabellini, E., Lucchini, F., & Gattoni, M. E. (2022). Prevalence of Problem Gambling: A Meta-analysis of Recent Empirical Research (2016–2022). *Journal of Gambling Studies*, *39*(3), 1027–1057. https://doi.org/10.1007/s10899-022-10180-0

Gandar, J., Zuber, R., O'Brien, T., & Russo, B. (1988). Testing Rationality in the Point Spread Betting Market. *The Journal of Finance*, *43*(4), 995–1008. https://doi.org/10.1111/j.1540-6261.1988.tb02617.x

Gardner, W. L., Cogliser, C. C., Davis, K. M., & Dickens, M. P. (2011). Authentic leadership: A review of the literature and research agenda. *The Leadership Quarterly*, *22*(6), 1120–1145. https://doi.org/10.1016/j.leaqua.2011.09.007

Giacomin, M., Tskhay, K. O., & Rule, N. O. (2022). Gender stereotypes explain different mental prototypes of male and female leaders. *The Leadership Quarterly*, *33*(6), 101578. https://doi.org/10.1016/j.leaqua.2021.101578

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Grabe, S., Ward, L. M., & Hyde, J. S. (2008). The role of the media in body image concerns. *Psychological Bulletin*, *134*, 460–476.

Greenleaf, R. K., & Spears, L. C. (1998). *The power of servant-leadership: Essays*. Berrett-Koehler Publishers.

Griffiths, M., Hayer, T., & Meyer, G. (Eds.). (2009). *Problem Gambling in Europe*. Springer New York. https://doi.org/10.1007/978-0-387-09486-1

Grossman, S. J., & Hart, O. D. (1992). An Analysis of the Principal-Agent Problem [Series Title: Huebner International Series on Risk, Insurance and Economic Security]. In *Foundations of Insurance Economics* (pp. 302–340, Vol. 14). Springer Netherlands. https://doi.org/10.1007/978-94-015-7957-5_16

Habarth, J. M. (2015). Development of the heteronormative attitudes and beliefs scale. *Psychology & Sexuality*, *6*(2), 166–188. https://doi.org/10.1080/19419899.2013.876444

Hamby, S., Finkelhor, D., Turner, H., & Ormrod, R. (2010). The overlap of witnessing partner violence with child maltreatment and other victimizations in a nationally representative survey of youth. *Child Abuse & Neglect*, *34*(10), 734–741. https://doi.org/10.1016/j.chiabu.2010.03.001

Hamermesh, D. S. (2011). *Beauty pays: Why attractive people are more successful*. Princeton University Press.

Hanel, P. H. P., & Vione, K. C. (2016). Do Student Samples Provide an Accurate Estimate of the General Public? (M. Voracek, Ed.). *PLOS ONE*, *11*(12), e0168354. https://doi.org/10.1371/journal.pone.0168354

Harris, M. B. (1994). Growing old gracefully: Age concealment and gender. *Journal of Gerontology: Psychological Sciences*, *49*, P149–P158. https://doi.org/10.1093/geronj/49.4.P149

Heilman, M. E. (2001). Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder. *Journal of Social Issues*, *57*(4), 657–674. https://doi.org/10.1111/0022-4537.00234

Henss, R. (1991). Perceiving age and attractiveness in facial photographs. *Journal of Applied Social Psychology*, *21*(11), 933–946. https://doi.org/10.1111/j.1559-1816.1991.tb00451.x

Hoffer, A. (2024). *Bets on Legal Sports Markets Pay Off Big for States, Sportsbooks, and Consumers* (tech. rep.). Tax Foundation. https://taxfoundation.org/wp-content/uploads/2024/12/FF854-3.pdf

Hollenbeck, B., Larsen, P., & Proserpio, D. (2024). The Financial Consequences of Legalized Sports Gambling. *SSRN*. https://ssrn.com/abstract=4903302

Hooper, A. (Ed.). (2016). *Leadership perspectives* [OCLC: 988029023]. Routledge, Taylor & Francis Group.

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., & Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: A systematic review and meta-analysis. *The Lancet Public Health*, *2*(8), e356–e366. https://doi.org/10.1016/S2468-2667(17)30118-4

Hummert, M. L., Garstka, T. A., & Shaner, J. L. (1997). Stereotyping of older adults: The role of target facial cues and perceiver characteristics. *Psychology and Aging*, *12*, 107–114. https://doi.org/10.1037/0882-7974.12.1.107

Johnson, D. F., & Pittenger, J. B. (1984). Attribution, the attractiveness stereotype, and the elderly. *Developmental Psychology*, *20*, 1168–1172. https://doi.org/10.1037/0012-1649.20.6.1168

Kay, A., & Furnham, A. (2013). Age and sex stereotypes in british television advertisements. *Psychology of Popular Media Culture*, *2*(3), 171–186. https://doi.org/10.1037/a0033083

Kim, H., & Drake, B. (2023). Has the relationship between community poverty and child maltreatment report rates become stronger or weaker over time? *Child Abuse & Neglect*, *143*, 106333. https://doi.org/10.1016/j.chiabu.2023.106333

Kim, H., Wildeman, C., Jonson-Reid, M., & Drake, B. (2017). Lifetime Prevalence of Investigating Child Maltreatment Among US Children. *American Journal of Public Health*, *107*(2), 274–280. https://doi.org/10.2105/AJPH.2016.303545

Kim, S. (2002). Participative Management and Job Satisfaction: Lessons for Management Leadership. *Public Administration Review*, *62*(2), 231–241. https://doi.org/10.1111/0033-3352.00173

Kirby, S., Francis, B., & O'Flaherty, R. (2014). Can the FIFA World Cup Football (Soccer) Tournament Be Associated with an Increase in Domestic Abuse? *Journal of Research in Crime and Delinquency*, *51*(3), 259–276. https://doi.org/10.1177/0022427813494843

Kissler, J., & Bäuml, K.-H. (2000). Effects of beholder's age on the perception of facial attractiveness. *Acta Psychologica*, *104*, 145–166. https://doi.org/10.1016/S0001-6918(00)00018-4

Kite, M. E., Deaux, K., & Miele, M. (1991). Stereotypes of young and old: Does age outweigh gender? *Psychology and Aging*, *6*, 19–27. https://doi.org/10.1037/0882-7974.6.1.19

Kitzmann, K. M., Gaylord, N. K., Holt, A. R., & Kenny, E. D. (2003). Child witnesses to domestic violence: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *71*(2), 339–352. https://doi.org/10.1037/0022-006X.71.2.339

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin*, *137*(4), 616–642. https://doi.org/10.1037/a0023557

Kogan, N. (1979). A study of age categorization. *Journal of Gerontology*, *34*, 358–367. https://doi.org/10.1093/geronj/34.3.358

Lauzen, M., & Dozier, D. (2005). Maintaining the double standard: Portrayals of age and gender in popular films. *Sex Roles*, *52*(7-8), 437–446. https://doi.org/10.1007/s11199-005-3710-1

Lazear, E. P. (2012). Leadership: A personnel economics approach. *Labour Economics*, *19*(1), 92–101. https://doi.org/10.1016/j.labeco.2011.08.005

Lee, M. M., Carpenter, B., & Meyers, L. S. (2007). Representations of older adults in television advertisements. *Journal of Aging Studies*, *21*, 23–30. https://doi.org/10.1016/j.jaging.2006.04.001

Legal Sports Report. (2025). Sports betting revenue tracker [Accessed November 10, 2025]. https://www.legalsportsreport.com/sports-betting-states/revenue

Lemish, D., & Muhlbauer, V. (2012). "can't have it all": Representations of older women in popular culture. *Women & Therapy*, *35*(3-4), 165–180. https://doi.org/10.1080/02703149.2012.684541

Liberman, B. E., & Golom, F. D. (2015). Think manager, think male? Heterosexuals' stereotypes of gay and lesbian managers. *Equality, Diversity and Inclusion: An International Journal*, *34*(7), 566–578. https://doi.org/10.1108/EDI-01-2015-0005

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behaviour Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Maguire-Jack, K., Smith, B. D., & Spilsbury, J. C. (2022). Rural and Urban Child Maltreatment Considersations in the United States. In *Handbook of child maltreatment* (Second edition, pp. 295–301). Springer.

Mallory, C., Sears, B., & Moran, R. (2024, August). *LGBTQ People's Experiences of Workplace Discrimination and Harassment* (tech. rep.). The Williams Institute, UCLA School of Law. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Workplace-Discrimination-Aug-2024.pdf

Masarik, A. S., & Conger, R. D. (2017). Stress and child development: A review of the Family Stress Model. *Current Opinion in Psychology*, *13*, 85–90. https://doi.org/10.1016/j.copsyc.2016.05.008

Matsuzawa, K., & Arnesen, E. (2024). Sports Betting Legalization Amplifies Emotional Cues & Intimate Partner Violence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4938642

McManus, R. M., & Perruci, G. (Eds.). (2020). *Understanding leadership: An arts and humanities perspective* (Second Edition). Routledge.
Revised edition of Understanding leadership, 2015.

Meslec, N., Curseu, P. L., Fodor, O. C., & Kenda, R. (2020). Effects of charismatic leadership and rewards on individual performance. *The Leadership Quarterly*, *31*(6), 101423. https://doi.org/10.1016/j.leaqua.2020.101423

Mobius, M. M., & Rosenblat, T. S. (2006). Why beauty matters. *American Economic Review*, *96*, 222–235. https://doi.org/10.1257/000282806776157515

Morton, J. W. (2017). Think leader, think heterosexual male? The perceived leadership effectiveness of gay male leaders. *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration*, *34*(2), 159–169. https://doi.org/10.1002/cjas.1434

Muise, A., & Desmarais, S. (2010). Women's perceptions and use of "anti-aging" products. *Sex Roles*, *63*, 126–137. https://doi.org/10.1007/s11199-010-9791-5

Narayan, C. (2008). Is there a double standard of aging? older men and women and ageism. *Educational Gerontology*, *34*(9), 782–787. https://doi.org/10.1080/03601270802042123

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, *33*(1), 31–88. https://doi.org/10.1145/375360.375365

Neppl, T. K., Senia, J. M., & Donnellan, M. B. (2016). Effects of economic hardship: Testing the family stress model over time. *Journal of family psychology: JFP: journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, *30*(1), 12–21. https://doi.org/10.1037/fam0000168

Neumark, D., Burn, I., & Button, P. (2017). Age discrimination and hiring of older workers. *FRBSF Economic Letter*.

Niedlich, C., & Steffens, M. C. (2015). On the interplay of (positive) stereotypes and prejudice: Impressions of lesbian and gay applicants for leadership positions. *Sensoria: A Journal of Mind, Brain & Culture*, *11*(1). https://doi.org/10.7790/sa.v11i1.408

Norja, R. e. a. (2021). Accuracy of assessing age from faces. *Nordic Psychology*, *74*, 70–85.

Palmore, E. B. (1999). *Ageism: Negative and positive*. Springer.

Paustian-Underdahl, S. C., Walker, L. S., & Woehr, D. J. (2014). Gender and perceptions of leadership effectiveness: A meta-analysis of contextual moderators. *Journal of Applied Psychology*, *99*(6), 1129–1145. https://doi.org/10.1037/a0036751

Pellegrini, V., De Cristofaro, V., Giacomantonio, M., & Salvati, M. (2020). Why are gay leaders perceived as ineffective? The role of the type of organization, sexual prejudice and gender stereotypes. *Personality and Individual Differences*, *157*, 109817. https://doi.org/10.1016/j.paid.2020.109817

Powell, G. N., Butterfield, D. A., & Parent, J. D. (2002). Gender and Managerial Stereotypes: Have the Times Changed? *Journal of Management*, *28*(2), 177–193. https://doi.org/10.1177/014920630202800203

Prieler, M., Kohlbacher, F., Hagiwara, S., & Arima, A. (2011). Gender representation of older people in japanese television advertisments. *Sex Roles*, *64*, 405–415. https://doi.org/10.1007/s11199-010-9923-y

Puni, A., Ofei, S. B., & Okoe, A. (2014). The Effect of Leadership Styles on Firm Performance in Ghana. *International Journal of Marketing Studies*, *6*(1), p177. https://doi.org/10.5539/ijms.v6n1p177

Raman, P., Harwood, J., Weis, D., Anderson, J. L., & Miller, G. (2008). Portrayals of older adults in u.s. and indian magazine advertisements: A cross-cultural comparison. *Howard Journal of Communications*, *19*(3), 221–240. https://doi.org/10.1080/10646170802218214

Rapoport, E., Reisert, H., Schoeman, E., & Adesman, A. (2021). Reporting of child maltreatment during the SARS-CoV-2 pandemic in New York City from March to May 2020. *Child Abuse & Neglect*, *116*, 104719. https://doi.org/10.1016/j.chiabu.2020.104719

Renner, L. M., & Slack, K. S. (2006). Intimate partner violence and child maltreatment: Understanding intra- and intergenerational connections. *Child Abuse & Neglect*, *30*(6), 599–617. https://doi.org/10.1016/j.chiabu.2005.12.005

Repetti, R. L., Taylor, S. E., & Seeman, T. E. (2002). Risky families: Family social environments and the mental and physical health of offspring. *Psychological Bulletin*, *128*(2), 330–366. https://doi.org/10.1037/0033-2909.128.2.330

Rhodes, G., Simmons, L. W., & Peters, M. (2005). Attractiveness and sexual behavior: Does attractiveness enhance mating success? *Evolution and Human Behavior*, *26*, 186–201. https://doi.org/10.1016/j.evolhumbehav.2004.08.014

Ridgeway, C. L. (2001). Gender, Status, and Leadership. *Journal of Social Issues*, *57*(4), 637–655. https://doi.org/10.1111/0022-4537.00233

Rosette, A. S., & Livingston, R. W. (2012). Failure is not an option for Black women: Effects of organizational performance on leaders with single versus dual-subordinate identities. *Journal of Experimental Social Psychology*, *48*(5), 1162–1167. https://doi.org/10.1016/j.jesp.2012.05.002

Rupp, D. E., Vodanovich, S. J., & Credé, M. (2005). The multidimensional nature of ageism: Construct validity and group differences. *Journal of Social Psychology*, *145*(3), 335–362. https://doi.org/10.3200/SOCP.145.3.335-362

Ryan, M. K., & Haslam, S. A. (2005). The Glass Cliff: Evidence that Women are Over-Represented in Precarious Leadership Positions. *British Journal of Management*, *16*(2), 81–90. https://doi.org/10.1111/j.1467-8551.2005.00433.x

Ryan, M. K., Haslam, S. A., Morgenroth, T., Rink, F., Stoker, J., & Peters, K. (2016). Getting on top of the glass cliff: Reviewing a decade of evidence, explanations, and impact. *The Leadership Quarterly*, *27*(3), 446–455. https://doi.org/10.1016/j.leaqua.2015.10.008

Saucier, M. G. (2004). Midlife and beyond: Issues for aging women. *Journal of Counseling & Development*, *82*, 420–425.

Schein, V. E. (2001). A Global Look at Psychological Barriers to Women's Progress in Management. *Journal of Social Issues*, *57*(4), 675–688. https://doi.org/10.1111/0022-4537.00235

Scholz, J. K., & Sicinski, K. (2014). Facial attractiveness and lifetime earnings: Evidence from a cohort study. *Review of Economics and Statistics*, *97*, 14–28. https://doi.org/10.1162/REST_a_00435

Seccombe, K., & Ishii-Kuntz, M. (1991). Perceptions of problems associated with aging: Comparisons among four older age cohorts. *The Gerontologist*, *31*, 527–533. https://doi.org/10.1093/geront/31.4.527

Sedlak, A. J., Mettenburg, J., Basena, M., Petta, I., McPherson, K., Greene, A., & Li, S. (2010). *Fourth National Incidence Study of Child Abuse and Neglect (NIS–4): Report to Congress* (Technical Report). U.S. Department of Health and Human Services. Washington, DC.

Shaffer, H. J., & Korn, D. A. (2002). Gambling and Related Mental Disorders: A Public Health Analysis. *Annual Review of Public Health*, *23*(1), 171–212. https://doi.org/10.1146/annurev.publhealth.23.100901.140532

Sheer, V. C. (2015). "Exchange lost" in leader–member exchange theory and research: A critique and a reconceptualization. *Leadership*, *11*(2), 213–229. https://doi.org/10.1177/1742715014530935

Sherman, S. R. (1986). Sex role stereotypes of middle and old age. *Journal of Social Service Research*, *8*, 23–35. https://doi.org/10.1300/J079v08n04_02

Slack, K. S., Berger, L. M., DuMont, K., Yang, M.-Y., Kim, B., Ehrhard-Dietzel, S., & Holl, J. L. (2011). Risk and protective factors for child neglect during early childhood: A cross-study comparison. *Children and Youth Services Review*, *33*(8), 1354–1363. https://doi.org/10.1016/j.childyouth.2011.04.024

Snyder, K. (2006). *The G quotient: Why gay executives are excelling as leaders–and what every manager needs to know* (1st ed) [OCLC: ocm65340952]. Jossey-Bass.

Sontag, S. (1972). The double standard of aging.

Stylianou, A. M. (2018). Economic Abuse Within Intimate Partner Violence: A Review of the Literature. *Violence and Victims*, *33*(1), 3–22. https://doi.org/10.1891/0886-6708.33.1.3

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

U.S. Census Bureau. (2025). Quarterly services survey (qss): Gambling industries, naics 7132 [Accessed November 10, 2025]. https://www.census.gov/programs-surveys/qss.html

U.S. Department of Health & Human Services. (2024). *Child Maltreatment 2022* (tech. rep.). U.S. Department of Health & Human Services. https://www.acf.hhs.gov/cb/data-research/child-maltreatment

Van Der Maas, M., Cho, S. R., & Nower, L. (2022). Problem gambling message board activity and the legalization of sports betting in the US: A mixed methods approach. *Computers in Human Behavior*, *128*, 107133. https://doi.org/10.1016/j.chb.2021.107133

Vinnerljung, B., Brännström, L., & Hjern, A. (2015). Disability pension among adult former child welfare clients: A Swedish national cohort study. *Children and Youth Services Review*, *56*, 169–176. https://doi.org/10.1016/j.childyouth.2015.07.001

Ward, R., & Holland, C. (2011). Hair and later-life image dilemmas. *Ageing and Society*, *31*, 288–307.

Wernick, M., & Manaster, G. J. (1984). Age and the perception of age and attractiveness. *The Gerontologist*, *24*, 408–414. https://doi.org/10.1093/geront/24.4.408

Wettstein, M., Park, R., Kornadt, A. E., Wurm, S., Ram, N., & Gerstorf, D. (2024). Postponing old age: Evidence for historical change toward a later perceived onset of old age. *Psychology and Aging*, *39*(6), 625–636. https://doi.org/10.1037/pag0000812

Widom, C. S., Czaja, S. J., Bentley, T., & Johnson, M. S. (2012). A Prospective Investigation of Physical Health Outcomes in Abused and Neglected Children: New Findings From a 30-Year Follow-Up. *American Journal of Public Health*, *102*(6), 1135–1144. https://doi.org/10.2105/AJPH.2011.300636

Wilcox, S. (1997). Age and gender in relation to body attitudes. *Psychology of Women Quarterly*, *21*, 549–565.

Wild, H., Kyröläinen, A.-J., & Kuperman, V. (2022). How representative are student convenience samples? A study of literacy and numeracy skills in 32 countries (S. Frisson, Ed.). *PLOS ONE*, *17*(7), e0271191. https://doi.org/10.1371/journal.pone.0271191

Wildeman, C., Emanuel, N., Leventhal, J. M., Putnam-Hornstein, E., Waldfogel, J., & Lee, H. (2014). The Prevalence of Confirmed Maltreatment Among US Children, 2004 to 2011. *JAMA Pediatrics*, *168*(8), 706. https://doi.org/10.1001/jamapediatrics.2014.410

Willner, P., & Rowe, G. (2001). Alcohol servers' estimates of young people's ages. *Drugs: Education, Prevention and Policy*, *8*(4), 375–383. https://doi.org/10.1080/09687630110063960

Winkler, W. (1990). *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage.* (tech. rep.). US Census.

Zepelin, H., Sills, R. A., & Heath, M. W. (1986). Is age becoming irrelevant? an exploratory study of perceived age norms. *International Journal of Aging and Human Development*, *24*, 241–256. https://doi.org/10.2190/1RAF-8YEW-QKU8-RTF8

Zietz, S., Lansford, J. E., Liu, Q., Long, Q., Oburu, P., Pastorelli, C., Sorbring, E., Skinner, A. T., Steinberg, L., Tapanya, S., Tirado, L. M. U., Yotanyamaneewong, S., Alampay, L. P., Al-Hassan, S. M., Bacchini, D., Bornstein, M. H., Chang, L., Deater-Deckard, K., Di Giunta, L., ... Gurdal, S. (2022). A longitudinal examination of the family stress model of economic hardship in seven countries. *Children and Youth Services Review*, *143*, 106661. https://doi.org/10.1016/j.childyouth.2022.106661