

Essays on the Role of Identity in Economic and Political Behavior

by

Hannah K. Ruebeck

B.A, Wellesley College (2016)

Submitted to the Department of Economics in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

©Hannah Ruebeck 2024. All rights reserved. The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Hannah K. Ruebeck
Department of Economics
May 15, 2024

Certified by: Frank Schilbach
Associate Professor of Economics
Thesis Supervisor

Certified by: Esther Duflo
Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics
Thesis Supervisor

Certified by: Parag Pathak
Class of 1922 Professor of Economics
Thesis Supervisor

Accepted by: Isaiah Andrews
Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Essays on the Role of Identity in Economic and Political Behavior

by

Hannah K. Ruebeck

Submitted to the Department of Economics
on May 15, 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract. These essays consider the role of various personal and social identities and resulting decision-making in the domains of education, work, and political participation. The first essay studies beliefs about experiencing racial or gender discrimination, or perceived discrimination, and its consequences for worker behavior. Using a large randomized controlled trial (RCT, N=5,000) in a constructed online labor market, I show that perceived racial and gender discrimination has large negative effects on worker retention, future labor supply, and cooperation with managers and that these effects are driven by large psychological costs to interacting with a biased manager. Firms can therefore improve both equity and efficiency by reducing perceived discrimination. I then test whether implementing hiring procedures that reduce the potential for actual discrimination are effective at reducing perceived discrimination. The procedures I test—blinding hiring managers to demographics and using unbiased algorithms—at best moderately reduce rates of perceived discrimination when members of minority groups remain highly under-represented.

The second essay studies childhood confidence, a potential determinant of educational and labor-market behavior when ability is imperfectly observed. This essay documents two main facts in a large, national sample of children whose outcomes are followed for 20 years. First, childhood confidence in math and reading is starkly gendered along stereotypical lines: girls are more likely to be under-confident in math and over-confident in reading, and vice-versa for boys. Second, childhood over- and under-confidence in math strongly predicts adolescent test scores, educational attainment, and majoring or working in STEM.

The final essay studies political efficacy, or beliefs about government responsiveness to citizen preferences and action in an RCT with 6,000 participants. In the context of US climate policy, we test how these beliefs and preferences for government action change when citizens learn about the recent, largest climate bill in US history. Learning about policy progress has small positive effects on political efficacy and small negative effects on preferences for the government to focus on climate policy. These countervailing effects may be why we see no effect of this treatment on citizen climate action. On the other hand, additionally watching a short, fictional narrative about a young, initially apathetic woman who goes on to organize a climate march has large effects on political efficacy and subsequently large effects on donations to climate lobbying groups and revealed interest in climate marches.

JEL Codes: D91, J7, J16

Thesis Supervisor: Frank Schilbach
Title: Associate Professor of Economics

Thesis Supervisor: Esther Duflo
Title: Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics

Thesis Supervisor: Parag Pathak
Title: Class of 1992 Professor of Economics

Acknowledgements

I am extremely grateful for the support I have received from advisors and mentors, classmates, friends, and family over the course of my academic career.

I benefitted tremendously from the guidance of my advisors, Frank Schilbach, Esther Duflo, and Parag Pathak. Their advice has been instrumental in developing the research I have undertaken during the PhD and my ways of thinking about the research I hope to tackle going forward.

Frank's mentorship was invaluable. His commitment to, patience with, and care for my projects and myself as a researcher and a person were fundamental to the last six years and I hope to emulate his generosity of time and spirit with students in the future. He carefully considered every question I asked and provided thorough and thoughtful feedback on research designs and many drafts of papers and slides. Frank was also instrumental in acquiring the external funding needed for my projects and so helped make this research possible in that very logistical way as well. His sheep and hoptimists had their intended effect on my mental health (and so, surely, my productivity!)

I am so grateful to have had Esther as an advisor. I so admire her decisiveness and engagement with such a broad scope of research, and have learned so much from her during the PhD along both of these dimensions. Esther also supported my research logistically, funding the pilots necessary to win external funding (and needed to move forward with research in the meantime), and was just as generous with her time and engagement with my research.

Parag's support and advising was crucial to the development of my projects, and he often helped me see the forest for the trees and pointed me to connections I had not considered. I am so grateful for our conversations and the precision with which he engaged with my writing and thinking.

I also benefitted immensely from feedback and support from Abhijit Banerjee throughout the PhD. Abhijit's enthusiasm and fresh perspectives were so helpful at times when they were very needed. Sendhil Mullainathan was extremely generous with his time and insight this past year. I am also so thankful for other members of the Economics department's engagement with my work, individually and in the labor and behavioral lunches.

Lucy Page has been my coauthor, friend, and teacher for the last six years. I am awed by and grateful for her commitment to research and teaching, precise and detail-oriented thinking and writing, and willingness to help with anything she can at a moment's notice. Lisa Ho has been an incredible sounding board, co-TA, and friend, and Lisa and Lucy's camaraderie throughout the job market was key to keeping sane. I could always rely on Abby Ostriker and Maya Bidanda to read a quick email or abstract, and would not have made it through the first year (and probably the rest) without them and Lucy in our booth in the Sloan cafeteria.

I would not have pursued this PhD without the incredible economics faculty at Wellesley College. Kartini Shastry has been my advisor, coauthor, and friend and I cannot express how much I appreciate the role she has played in my life, from introducing me to economics research to helping me navigate the job market. I also want to thank Grace and Yeye for their friendship and advice over so many years. I have loved navigating our cross-disciplinary PhD's together.

My parents instilled in me a desire to try to answer complicated questions and a commitment to serving others, particularly as a teacher. I have learned so much from both of them and am so thankful for their unending support. My siblings help keep me grounded and mindful of what matters, and I respect them both so much.

Finally, *thank you* to my wife, Hannah Rhodenhiser, who has seen me through it all (and, tangentially, thanks to MIT and the Boston Fed for bringing us together). Your confidence in me at my lowest and joyfulness with me at my highest kept me going. My admiration for you as a person and economist cannot be overstated, and I am so happy to share this accomplishment with you.

Contents

1	Perceived Discrimination at Work	18
1.1	Introduction	19
1.2	Research design	27
1.2.1	Setting and sample	27
1.2.2	The promotion experiment	29
1.2.2.1	The job	30
1.2.2.2	Design overview and background	31
1.2.2.3	Experimental design	32
1.2.3	The hiring experiment	38
1.2.4	Experimental fidelity: Attention and comprehension	39
1.3	The effects of perceived discrimination	40
1.3.1	Estimation	40
1.3.2	Manipulation check: Treatment affects perceptions	42
1.3.3	Results: Effects of perceived discrimination	43
1.3.3.1	Retention and performance	43
1.3.3.2	Future labor supply	45
1.3.3.3	Mechanisms	47
1.4	The effects of anti-bias policies on perceptions	51
1.4.1	Estimation	51
1.4.2	Results	53

1.5	The effects of minority-group representation	55
1.5.1	Estimation	55
1.5.2	Results	56
1.6	The effects of perceived algorithmic discrimination	58
1.6.1	Estimation	58
1.6.2	Manipulation check: Treatment affects perceptions	58
1.6.3	Results: Effects of perceived algorithmic discrimination	59
1.7	Robustness	60
1.8	Conclusion	63
2	Childhood Confidence, Schooling, and the Labor Market: Evidence from the PSID	74
2.1	Introduction	75
2.2	How might childhood math confidence affect economic outcomes?	81
2.3	Measuring confidence and later-life outcomes in the PSID	83
2.3.1	Sample and survey design	83
2.3.2	Measuring over- and under- confidence in math	86
2.3.3	Biased beliefs or measurement error?	89
2.4	Patterns of over- and under-confidence in the population	93
2.4.1	Prevalence of biased beliefs	93
2.4.2	Biased beliefs and other child characteristics	94
2.5	Confidence and long-term outcomes: Empirical strategy	95
2.6	Confidence and long-term outcomes: Results	97
2.6.1	Medium-term educational achievement	97
2.6.2	Educational attainment	98
2.6.3	College quality, college major choice, and graduate education	99
2.6.4	Employment outcomes	100
2.7	Robustness	102

2.7.1	Key confounders: Personality, adult investment, and school quality	102
2.7.2	Alternate definitions of biased beliefs	104
2.8	Snowballing investment or persistent over- and under-confidence?	105
2.8.1	The persistence of childhood confidence in math	106
2.8.2	Gaps in intermediate outcomes do not fully explain results	108
2.9	Conclusion	109
3	The Narrative of Policy Change: Fiction Builds Political Efficacy and Climate Action	120
3.1	Introduction	121
3.2	Research design	124
3.2.1	Sample selection	124
3.2.2	Experimental survey	125
3.2.2.1	IRA information randomization	126
3.2.2.2	Fictional climate-advocacy story	127
3.2.3	Experimental fidelity	128
3.2.4	Main outcomes	129
3.2.4.1	Political efficacy	129
3.2.4.2	Climate action	129
3.3	Results	131
3.3.1	Specifications	131
3.3.2	Political efficacy	131
3.3.3	Climate action	132
3.3.4	Mechanisms	133
3.3.4.1	Emotions	134
3.3.4.2	Desire for climate policy	135
3.3.4.3	Beliefs about others	136
3.3.4.4	Memory	136

3.3.4.5	Combining mechanisms	137
3.4	Conclusion	137
A	<i>Appendix to Perceived Discrimination at Work</i>	156
A.1	Supplementary figures and tables	157
A.2	Manager recruitment and manager task	196
A.3	Training the algorithm	198
A.4	Variable definitions	199
A.4.1	Outcome variables	199
A.4.2	Control variables	206
A.4.3	Variables used in heterogeneity	209
A.5	Coding the text data	213
A.5.1	The main measure of perceived discrimination	213
A.5.2	Secondary measure of perceived discrimination (complaints about bias)	214
A.5.3	Study topic	215
A.6	Instrumental variables (IV) specification results	216
A.7	Other heterogeneity	221
A.8	Effects on future labor supply in the hiring experiment	228
A.9	Robustness analyses	231
A.9.1	Alternative inference methods	231
A.9.2	Attrition and attention	232
A.9.3	Specification choices	232
A.9.3.1	Sample definitions	232
A.9.3.2	Included controls	232
A.9.4	Demand effects	233
A.9.5	Multiple price list elicitations	235
A.9.6	Deviations from the Pre-Analysis Plan:	235

B	<i>Appendix to Childhood Confidence, Schooling, and the Labor Market</i>	266
B.1	Supplementary Figures and Tables	267
B.2	Constructing Indices	324
B.3	Biased Beliefs and Other Attitudes Towards School	331
B.4	Over- versus under-confidence	332
B.5	Measuring key confounders	333
B.6	Alternate definitions of childhood biased beliefs	337
C	<i>Appendix to The Narrative of Policy Change</i>	341
C.1	Supplementary tables and figures	342
C.2	Study recruitment	368
C.3	Obfuscating the follow-up survey	370
C.4	Story production	371
C.5	Filler questions	372
	C.5.1 Open-ended filler questions	372
	C.5.2 Multiple choice filler questions	374
C.6	Comprehension questions	376
C.7	Variable definitions	378
	C.7.1 Outcome variables	378
	C.7.2 Control variables	392

List of Figures

1.1	Perceived discrimination in the manager arms of the promotion experiment	65
1.2	Treatment effects on retention and performance by gender \times race	66
1.3	Decomposing the effects on reservation wages in the promotion experiment	66
1.4	Perceived discrimination in the hiring experiment	67
1.5	Effects of seeing one previously-promoted minority-group worker	67
1.6	Perceived discrimination in the algorithm arm of the promotion experiment	68
1.7	Effects of perceived algorithmic discrimination on worker behavior	69
2.1	Distributions of self-assessed and demonstrated ability	111
2.2	Controlling for intermediate outcomes	112
3.1	Impacts on emotions	138
3.2	Impacts of the story on climate action: Controlling for mediating emotions and beliefs	139
A.1	Information workers received about the job(s)	157
A.2	Experimental design of the promotion experiment	158
A.3	Timeline of study	159
A.4	Design of promotion experiment	160
A.5	Timeline of experimental survey (promotion experiment)	161
A.6	Treatment variation in manager groups	162
A.7	Treatment variation in algorithm groups (promotion experiment)	163

A.8	Correlation of perceived discrimination measures	164
A.9	Design of hiring experiment	165
A.10	Perceived discrimination by race \times gender	166
A.11	Effects on retention and overall earnings in the promotion experiment	167
A.12	Effects on psychological well-being, gender and race heterogeneity	167
A.13	Perceived discrimination in the hiring sample, secondary measures of perceived discrimination	168
A.14	Comprehension of hiring procedure inputs	169
A.15	Comprehension of hiring procedure incentives (manager arms) and design (algorithm arms)	170
A.16	Effects of seeing one previously-promoted minority-group worker of a same or different demographic group	171
A.17	Reported past experiences of discrimination at work in the experimental sample	172
A.18	All predictors of perceived discrimination in the promotion experiment	173
A.19	Avatar-making procedure	174
A.20	Avatar characteristics by self-reported race and gender	175
A.21	Perceived discrimination by quiz-score quintile group and quintile	176
A.22	CDFs of workers' beliefs about the likelihood of future promotion	177
A.23	Perceived discrimination in the hiring experiment (full sample)	177
A.24	Making information about decision-making inputs visually salient in the hiring experiment	178
A.25	Perceived discrimination in the hiring experiment, heterogeneity by race \times gender	179
A.26	Effects on beliefs, reservation wages, and affect in the hiring experiment, separately by treatment arm	230
A.27	Randomization inference and multiple hypothesis testing (effects of perceived discrimination)	238
A.28	Randomization inference and multiple hypothesis testing (hiring experiment)	238
A.29	Randomization inference and multiple hypothesis testing (historical representation)	239

A.30	Randomization inference and multiple hypothesis testing (perceived alg. discrimination)	239
A.31	Alternative specifications: Effect of being in non-blind manager arm on perceived discrimination	240
A.32	Alternative specifications: Effect of being in non-blind manager arm on retention	241
A.33	Alternative specifications: Effect of being in non-blind manager arm on effort	242
A.34	Alternative specifications: Effect of being in non-blind manager arm on performance	243
A.35	Alternative specifications: Effect of being in non-blind manager arm on future labor supply for the same job	244
A.36	Alternative specifications: Effect of being in non-blind manager arm on future labor supply (cooperative task) and generosity	245
A.37	Alternative specifications: Effect of being in non-blind manager arm on beliefs about promotion	246
A.38	Alternative specifications: Effect of being in non-blind manager arm on future labor supply, accounting for beliefs	247
A.39	Alternative specifications: Hiring experiment (if previous hires are all white men)	248
A.40	Alternative specifications: Hiring experiment (regardless of demographic makeup of previous hires)	249
A.41	Alternative specifications: Effects of minority-group representation	250
A.42	Alternative specifications: Effects of seeing avatars on perceived algorithmic discrimination (promotion experiment)	251
A.43	Alternative specifications: Effects of seeing avatars in the algorithm arm on retention	252
A.44	Alternative specifications: Effects of seeing avatars in the algorithm arm on effort	253
A.45	Alternative specifications: Effects of seeing avatars in the algorithm arm on performance	254
A.46	Alternative specifications: Effects of seeing avatars in the algorithm arm on future labor supply	255
A.47	Attrition and attention in the hiring experiment	256
A.48	Knowledge of study topic in the hiring experiment	256

A.49 Accounting for possible demand effects in the manager arms of the promotion experiment: effort and future labor supply	257
A.50 Accounting for possible demand effects in the manager arms of the promotion experiment: secondary outcomes	258
A.51 Accounting for possible demand effects in the algorithm arm of the promotion experiment: effort and future labor supply	259
A.52 Accounting for possible demand effects in the algorithm arm of the promotion experiment: secondary outcomes	260
A.53 Accounting for possible demand effects among workers who saw three versus two white men previously promoted	261
A.54 Accounting for possible demand effects in the hiring experiment	262
B.1 Distribution of the degrees of over- and under-confidence measure	267
B.2 Patterns in over- and under-confidence by age	268
B.3 Differences in over- or under-confidence classification using 2 subtests of the WJ-R	269
B.4 Specification chart for persistence into adolescence	270
B.5 Specification chart for adolescent math scores	272
B.6 Specification chart for adolescent reading scores	274
B.7 Specification chart for graduating from high school	276
B.8 Specification chart for graduating from college	278
B.9 Specification chart for college quality index	280
B.10 Specification chart for college's 75th percentile math SAT score	282
B.11 Specification chart for majoring in STEM	284
B.12 Specification chart for having a graduate school degree	286
B.13 Specification chart for working in STEM	288
B.14 Specification chart for working in non-STEM high-education occupation	290
B.15 Specification chart for ln(earnings)	292

B.16	Specification chart for unemployment	294
B.17	Specification chart for gender differences	296
B.18	Coefficients on each confidence level fixed effect (medium-term educational achievement and attainment)	297
B.19	Coefficients on each confidence level fixed effect (college quality, college major, and post-college schooling)	298
B.20	Coefficients on each confidence level fixed effect (employment outcomes)	299
C.1	Political efficacy among those who want more climate policy	342
C.2	Research Design	343
C.3	Baseline desire for government climate action	344
C.4	Specification chart: Standardized political-efficacy index, main survey	345
C.5	Specification chart: Gradient of bill passage with respect to citizen calls	346
C.6	Specification chart: Standardized political-efficacy index, follow-up survey	347
C.7	Specification chart: Started process of writing to Congress	348
C.8	Specification chart: Wrote custom text for letter to Congress	349
C.9	Specification chart: Clicked to send letter to Congress	350
C.10	Specification chart: Clicked link for climate marches	351
C.11	Specification chart: Downloaded guide for contacting Congress	352
C.12	Specification chart: Whether donated to climate organization in main survey	353
C.13	Specification chart: Amount donated to climate organization in main survey	354
C.14	Specification chart: Whether donated to climate organization in follow-up survey	355
C.15	Specification chart: Amount donated to climate organization in follow-up survey	356

List of Tables

1.1	Summary statistics and comparison to ACS sample	70
1.2	Effects on retention	71
1.3	Effects on effort and performance in the first six (required) paragraphs	71
1.4	Effects on future labor supply	72
1.5	Effects on sharing with and avoidance of manager	73
1.6	Effects on beliefs about future promotion	73
2.1	The persistence of math over- and under-confidence	113
2.2	Demographic predictors of over- and under-confidence	114
2.2	Demographic predictors of over- and under-confidence (continued)	115
2.3	Childhood math confidence and medium-term educational achievement and attainment	116
2.4	Childhood math confidence and college quality, college major choice, and post-college schooling	117
2.5	Childhood math confidence and employment outcomes	118
2.6	Childhood math confidence and young adult confidence outcomes	119
3.1	Impacts of treatments on political efficacy	140
3.2	Impacts of treatments on climate donations and citizen advocacy	141
A.1	Balance table, promotion experiment	180
A.2	Balance table, hiring experiment	181

A.3	Who is selected by each promotion/hiring procedure	182
A.4	Balance table, never versus ever promoted/hired	183
A.5	Effects on secondary measures of perceptions of discrimination in the manager sample	184
A.6	Effects on retention, gender and racial heterogeneity	185
A.7	Effects on effort and performance in the manager sample of the promotion experi- ment, gender and racial heterogeneity	186
A.8	Effects on self-reported interest in future work	187
A.9	Balance table by whether see two or three white men previously promoted	188
A.10	Effects on other reasons for non-promotion in manager sample of the promotion experiment	189
A.11	Effects on psychological outcomes in the promotion experiment	190
A.12	Effects on psychological outcomes, gender and racial heterogeneity	191
A.13	Effects on future labor supply, gender and racial heterogeneity	192
A.14	Correlations of manager characteristics with previously selected workers' charac- teristics	193
A.15	Effects on secondary measures of perceptions of discrimination in the algorithm sample of the promotion experiment	194
A.16	Effects on other reasons for non-promotion in algorithm sample of the promotion experiment	195
A.17	The effect of perceived discrimination (IV estimates) on retention	217
A.18	The effect of perceived discrimination (IV estimates) on effort and performance . .	218
A.19	The effect of perceived discrimination (IV estimates) on future labor supply	219
A.20	The effect of perceived discrimination (IV estimates) on beliefs about promotion .	220
A.21	The effect of perceived discrimination (IV estimates) on sharing with and avoid- ance of manager	220
A.22	Effects on retention, all measures of heterogeneity	222

A.23	Effects on effort and performance, all measures of heterogeneity	223
A.24	Effects on future labor supply, all measures of heterogeneity	224
A.25	Effects on retention, heterogeneity by employment outside Prolific	225
A.26	Effects on effort and performance, heterogeneity by employment outside Prolific	226
A.27	Effects on future labor supply, heterogeneity by employment outside Prolific	227
A.28	Effects on future labor supply and beliefs in the hiring experiment	229
A.29	Attrition	263
A.30	Knowledge of study topic	264
A.31	Comparing workers with sensical and non-sensical MPL elicitations	265
B.1	WJ-R Applied Problems section scores predict long-run outcomes	300
B.2	The persistence of reading over- and under-confidence	301
B.3	Childhood reading confidence and medium-term educational achievement and attainment	302
B.4	Childhood reading confidence and college quality, college major choice, and post-college schooling	303
B.5	Childhood reading confidence and employment outcomes	304
B.6	Summary statistics	305
B.7	The persistence of math over- and under-confidence (weighted)	307
B.8	Childhood math confidence and medium-term educational achievement and attainment (weighted)	308
B.9	Childhood math confidence and college quality, college major choice, and post-college schooling (weighted)	309
B.10	Childhood math confidence and employment outcomes (weighted)	310
B.11	Demographic predictors of over- and under-confidence (decile coefficients)	311
B.12	Benchmarking the relationships between confidence and long-run outcomes and test scores	312

B.13 Robustness to potential confounders	313
B.14 Correlations between childhood confidence and personality measures	314
B.15 Parent and teacher predictors of math over- and under-confidence	315
B.16 Robustness to definitions of confidence	316
B.17 Sample means by whether missing confidence variables in the CDS	318
B.18 Correlations between math confidence and other attitudes	319
B.19 Childhood math confidence and average employment outcomes from age 28-33 . .	320
B.20 Comparing predictiveness of biased beliefs and Big-Five traits	321
B.21 Math confidence and young adult social outcomes	322
B.22 Heterogeneity by over- and under-confidence using the degrees of confidence mea- sure	323
C.1 Descriptive statistics and sample balance	357
C.2 Impacts of treatments on policy knowledge	358
C.3 Attrition by treatment groups	359
C.4 Impacts of treatments on political efficacy: Lower Lee (2009) bounds	360
C.5 Impacts of treatments on climate donations and citizen advocacy: Lower Lee (2009) bounds	361
C.6 Effects on donations to each cause in the follow-up survey	362
C.7 Correlations between action index and emotions in the control group	363
C.8 Impacts of treatments on climate worry and desire for action	364
C.9 Effects on beliefs about support and advocacy for climate policy	365
C.10 Effects of on political efficacy by wave	366
C.11 Effects of on climate action by wave	367
C.12 Effects on probabilistic beliefs about passing climate policy	368

Chapter 1

Perceived Discrimination at Work

Abstract

Beliefs about experiencing discrimination are widespread but understudied. In an online experiment ($N \approx 5000$), I randomly assign workers to be evaluated by promotion procedures with varied potential to discriminate and provide information about the procedure. Learning that managers knew workers' race and gender and previously promoted mostly white men increases perceived discrimination rates from 3-34%, lowers retention by 3-6%, and increases reservation wages by 9%. Reducing perceived discrimination is therefore important for equity and efficiency. However, increasingly-common anti-bias procedures—blinding managers to demographics or using unbiased algorithms—are unlikely to alone eliminate perceived discrimination when minority groups remain under-represented.

I am immeasurably grateful for the support of Esther Duflo, Parag Pathak, and Frank Schilbach, who have advised this project since its earliest days. I would also like to thank Abi Adams-Prassl, David Autor, Abhijit Banerjee, Lisa Ho, Sendhil Mullainathan, Lucy Page, Ashesh Rambachan, Nina Roussile, Kartini Shastry, Lise Vesterlund, and the participants in MIT's labor and behavioral lunches for insightful feedback. This project was supported by the Social Policy and Research Initiative at the Abdul Latif Jameel Poverty Action Lab, the George and Obie Shultz Fund, Esther Duflo and Frank Schilbach's research funds, and the Administration for Children and Families (ACF) of the United States (U.S.) Department of Health and Human Services (HHS). The latter was as part of a financial assistance award (Grant #: 90PD0314) totaling \$25,000 (28 percent) funded by ACF/HHS and \$65,178 (72 percent) funded by non-government sources. The contents are those of the author and do not necessarily represent the official views of, nor an endorsement, by ACF/HHS, or the U.S. Government. I was also supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1745302. The pre-registration for the two experiments can be found here: <https://www.socialsciceregistry.org/trials/9592> and here: <https://www.socialsciceregistry.org/trials/11806>. IRB approval for both studies was obtained under MIT's Committee on the Use of Humans as Experimental Subjects as protocols 2201000547 and 2307001048.

1.1 Introduction

Minority-group workers know discrimination is pervasive, have experiences that could be explained by discrimination, and often infer they have been discriminated against. In the last four decades, social scientists have documented widespread workplace discrimination (Goldin and Rouse, 2000; Bertrand and Mullainathan, 2004; Bertrand and Duflo, 2017; Neumark, 2018; Small and Pager, 2020). We know little, however, about either the causes or consequences of *perceived* discrimination because it is difficult to measure and its effects are difficult to separate causally from those of discrimination itself. Perceived discrimination is widespread, so it is important for understanding the impacts of discrimination in the labor market.¹ Whether or not discrimination is present, perceived discrimination may exacerbate race and gender gaps if it affects worker behavior via distrust, retaliation, or anticipated further discrimination.

This paper describes two large, pre-registered experiments ($N \approx 5000$). The first experiment tests whether perceived discrimination has quantitative importance for workers and firms; it also provides some evidence on how to reduce perceptions of discrimination. The second experiment tests the effects of common anti-bias hiring policies on *perceived* discrimination. Both experiments take place in a labor market with multiple rounds of worker evaluation and work opportunities that I constructed on Prolific, an online platform where workers take surveys for pay. I focus on gender and racial discrimination in a white- and male-dominated domain (science), and oversample women and racial minority men in the US. To my knowledge, my study is unique in the quasi-lab style experimental literature on discrimination in that it studies racial discrimination in addition to gender and is powered to detect heterogeneity along these lines; racial minorities most systematically experience discrimination in the workplace (Kline et al., 2021).

The novel design element of both experiments is that workers are randomly assigned to be

¹40 percent of women in the US report ever experiencing gender discrimination at work, and 25 percent of Black and Hispanic workers report experiencing discrimination at work *in the last year* (Pew Research, 2017; Gallup Inc, 2021). Among the 3,240 workers screened into my first experiment, 20, 35, and 50 percent of Hispanic, Asian, and Black men and women report ever experiencing racial discrimination at work before the survey, and around 35 percent of white and racial minority women report ever experiencing gender discrimination (Appendix Figure A.17).

evaluated under procedures with varied potential to discriminate. When non-promoted workers learn how they were evaluated, I also truthfully, randomly vary what workers know about who was previously promoted under the same procedure. This design generates exogenous variation in whether workers infer they have been discriminated against when not promoted and identifies how workers respond to different evaluation procedures and their outcomes. Then, I measure perceived discrimination and observe subsequent work behavior. I test whether perceived discrimination is quantitatively important by focusing on the effects of treatment on worker performance, retention, and future labor supply. Specifically, I observe worker performance in required scientific proof-reading tasks and the number of optional paragraphs they proofread—which I call retention—and elicit incentivized reservation wages for future job opportunities and cooperation with managers.

I also introduce a new measure of perceived discrimination that circumvents concerns with other methods. Prior surveys prime workers by asking directly about discrimination or differential treatment (e.g. [Kessler et al., 1999](#); [Goldsmith et al., 2004](#); [Mukerjee, 2014](#)). Administrative data relies on filed complaints that may be under-reported. Instead, I ask workers what they think needed to be different about their profile for them to be promoted. I define perceived discrimination as a worker’s open-ended response mentioning demographic information, which correlates strongly with more explicit questions asked at the end of the survey.

Specifically, in the first experiment ($N \approx 2,400$), workers are recruited for an initial work task and one of three randomly-assigned procedures determines whether they are “promoted”—offered a harder, higher-paying job—or not, in which case they are offered an easier, lower-paying one. The procedures are (i) a *demographic-blind manager* who sees prior performance and education, (ii) a *non-blind manager* who also sees avatars, race, and gender, and (iii) an algorithm that predicts performance using prior performance and education. I implement the randomly-assigned procedure’s decision for each worker. The analysis sample restricts to the 90 percent of workers who would not be promoted under *any* procedure so the only differences between treatment arms are workers’ perceptions of why they were not promoted.²

²There was no observable discrimination. The algorithms selected workers at similar rates regardless of race and gender, and race and gender are not predictive of performance conditional on the algorithms’ prediction. I could not

I focus on the two manager arms to study the effects of perceived discrimination. Before starting the easier job, non-promoted workers see how they were evaluated and this generates differences in perceived discrimination. Workers see demographic and work history information for their manager and the profiles of the three workers their manager promoted in the past, which communicates the information their manager had when making decisions. These previously-promoted workers are primarily white men due to their over-representation in earlier cohorts.

Perceived discrimination affects retention and performance in ways that would exacerbate racial and gender gaps. Learning that one's manager knew race and gender and previously promoted mostly white men—i.e. being in the non-blind manager arm—increases perceived discrimination by 31pp compared to the blind manager arm. There are subsequently large effects on worker behavior. Being in the non-blind manager arm reduces the number of optional tasks completed by 3 percent ($p=0.05$); this mechanically reduces the earnings of those induced to leave. On average, however, these negative effects on total earnings are canceled out by an improvement in performance among those who stay. The retention and performance effects are strongly gendered and more negative for racial minority women than white women.

Perceived discrimination also reduces future labor supply, primarily due to direct disutility from future interactions rather than lower expected wages due to future discrimination. Workers in the non-blind manager arm have 9 percent ($p=0.03$) higher reservation wages for a job opportunity with a chance for promotion that is determined by the same manager. I measure beliefs about the likelihood of future promotion and elicit reservation wages when there is no scope for additional discrimination, allowing me to disentangle these overall effects into: (i) lower expected wages due to anticipated discrimination and (ii) psychological mechanisms, e.g. disutility from interacting with a biased manager. Only 25 percent of the 9 percent increase in reservation wages is due to lower beliefs about the likelihood of promotion in the future. Workers are also 9 percent ($p=0.09$) less likely to share with their manager in a dictator game and have 18 percent ($p=0.04$) higher willingness to pay to be able to choose a different manager to work with in a collaborative task.

control managers' decisions, but they did not appear to discriminate against women or racial minorities.

Perceived discrimination therefore has quantitatively important implications for workers and firms, so firms should care about how their hiring procedures are perceived. Reducing actual bias may not reduce *perceived* discrimination, as reasoning about whether one has experienced discrimination is a complex problem. Individuals only receive few, noisy signals about why they do not receive certain opportunities and must draw their own conclusions about whether discrimination played a role (Jones et al., 2016; Doering et al., 2023); on the other hand, they get many signals about the prevalence of discrimination at scale.

In the second part of the paper, I test how firms might reduce perceived discrimination and its effects. First, I test whether anti-bias hiring procedures—using unbiased algorithms to rank candidates and blinding decision-makers to demographic information—affect perceived discrimination. In my setting, workers receive a strong signal that there may be racial or gender bias since the majority of previously-hired workers are white men. I also (truthfully) randomly vary whether workers see previously-hired workers who are all or two-thirds white men to understand how changes in minority-group representation affect worker perceptions. The experimental manipulations mimic ongoing structural changes in the labor market: the growing ubiquity of algorithms in managerial decision-making (Cowgill, 2020; Jarrahi et al., 2021) and pressure on companies to de-bias and diversify their workforces (Chang et al., 2019; Brecheisen, 2023; Fath, 2023; Gallup Inc, 2023).³ Finally, I consider whether algorithmic evaluation can mitigate the *effects* of perceived discrimination. Algorithmic bias is generated by statistical processes, not explicit prejudice, which may change how perceived discrimination affects behavior (Bigman et al., 2023).

The second experiment ($N \approx 2,700$ new workers) tests how anti-bias hiring policies affect perceived discrimination. I randomly assign workers to be evaluated by a manager or an algorithm

³Algorithms can embed human biases but are easier to regulate and audit, perhaps leading to less discrimination (e.g. Kleinberg et al., 2020) and thus less perceived discrimination, but are unavoidably opaque, undermining this potential. Companies' increasing emphasis on diversity, equity, and inclusion may successfully change perceptions, or fail or backfire if they seem inauthentic or ineffective. Flory et al. (2021) show that job ads emphasizing company diversity increase minority-group application rates, regardless of whether the statement is evidence-backed, and Baker et al. (2023) show that firms opportunistically publicize diversity commitments and that these firms are *more* likely to incur EEOC penalties and *less* likely to hire diverse candidates. Blinding decision-makers to demographic information can effectively improve representation of minority groups (e.g. Goldin and Rouse, 2000; Fath, 2023), but a recent survey of HR practitioners suggests that less than 20 percent of firms are employing this strategy (Fath and Zhu, 2021).

and cross-randomize whether the decision-maker knows their race and gender. Among workers evaluated by a demographic-blind decision-maker, I further cross-randomize whether they learn previous hires' demographics; they are all white men. Perceived discrimination is the primary outcome, and workers do not do an effortful task—the sample is workers who are not “hired,” rather than not “promoted.” Everything else about the labor market is the same as the first experiment.

I test whether minority-group representation—i.e. the *outcomes* of these procedures rather than the *inputs*—affects perceptions using additional random variation from the first experiment. Some workers are randomly assigned to managers who previously promoted three white men while others are assigned to managers who promoted two white men and someone else. Workers evaluated by the demographic-blind algorithm are cross-randomized to learn previously-promoted workers' race and gender or not; there is similar variation in whether these are all or two-thirds white men.

Changing hiring procedures to mitigate the direct effects of human biases—using algorithms or blinding managers to race and gender—does little to mitigate common *perceptions* of discrimination when past decisions “seem biased,” but increasing minority-group representation effectively does so—an example of the representativeness heuristic possibly affecting belief formation (e.g. [Kahneman and Tversky, 1972](#); [Bordalo et al., 2021](#)).⁴ Specifically, when workers see that only white men were previously hired and know their manager knew their race and gender, 52 percent perceive discrimination.⁵ Blinding managers to race and gender only reduces perceived discrimination by 11pp (20 percent, $p=0.04$) when white men are so over-represented. Regardless of whether decision-makers can use demographic information or not, using an algorithm *increases* perceived discrimination by 9-11pp (20 percent, $p<0.03$). These results suggest that using algorithmic decision-making or de-biasing managers can, at best, only moderately reduce perceived discrimination if the outcomes of those procedures do not change. If these procedures increase representation of minority groups, however, they may be more effective at reducing perceived

⁴This is conditional on the hiring policies used in the experiment—increasing representation via other policies, like affirmative action, may have other effects.

⁵This is higher than above because all previously-hired workers are white men and women (who are more likely to perceive discrimination) make up a higher fraction of the second experimental sample. In the second experiment all workers see that three white men were previously promoted which also raises perceived discrimination rates.

discrimination: seeing one woman or racial minority man among the three previously-selected workers reduces perceived manager and algorithmic discrimination by 40-50 percent ($p < 0.01$).

Since workers do perceive algorithmic discrimination, a final question is whether algorithmic decision-making could mitigate the effects of perceived discrimination on worker behavior. In the first experiment, workers evaluated by the algorithm are cross-randomized to see previously-promoted workers' profiles with or without demographics. This generates differences in perceived algorithmic discrimination in the sample in which I observe workers' subsequent performance in the work task. Seeing that a demographic-blind algorithm previously promoted mostly white men—which increases perceived discrimination by 17pp—reduces performance by 5-6 percent ($p = 0.07, 0.09$) but does not affect retention. When workers anticipate future discrimination, perceived algorithmic discrimination also has similarly negative effects on future labor supply as perceived manager discrimination.

In sum, this paper shows that rectifying the disparate impacts of discrimination will require addressing worker perceptions due to both material and psychological costs. Furthermore, employers could improve equity and efficiency by reducing perceived discrimination, but eliminating such perceptions will likely be unattainable as long as minority groups remain under-represented. Implementing objectively race- and gender-neutral procedures cannot eliminate perceived discrimination when decisions seem intuitively biased. These concerns are likely to become even more central as opaque algorithms play an increasingly important role in rendering high-stakes decisions that affect access to opportunity.

This paper contributes to literatures in labor and behavioral economics and cross-disciplinary work on algorithmic bias. First, racial and gender discrimination in the workplace is well-documented (e.g. Pager and Shepherd, 2008; Bartoš et al., 2016; Bertrand and Duflo, 2017; Blau and Kahn, 2017; Glover et al., 2017; Neumark, 2018; Bohren et al., 2019, 2022; Small and Pager, 2020; Kline et al., 2021). I show that perceived discrimination affects worker behavior and therefore exacerbates gaps caused by discrimination itself, including affecting future opportunities even if future evaluation is objectively neutral. Prior work therefore understates the overall effects of a

discriminatory labor market.

This paper's second contribution is to disentangle the effects of perceived discrimination into (i) changes in returns to effort or the wage value of work opportunities versus (ii) psychological mechanisms like retaliation or direct disutility from anticipated interactions with a biased manager. Psychological costs play a substantial role. Thus, I build on work showing the importance of integrating non-classical preferences into models of worker effort provision (Akerlof, 1982; Fehr et al., 2009; Card et al., 2012; Cohn et al., 2014; Breza et al., 2018; Dube et al., 2019; DellaVigna et al., 2022; Fehr and Charness, 2023). I also build on a theoretical and experimental literature on the effects of *anticipated* discrimination on human capital investment due to changes in incentives (Lundberg and Startz, 1983; Coate and Loury, 1993; Fryer et al., 2005; de Haan et al., 2017; Dianat et al., 2020) and recent experiments showing that anticipated discrimination affects job search behaviors (Alston, 2019; Charness et al., 2020; Lepage et al., 2022; Ridley, 2022; Agüero et al., 2023; Aksoy et al., 2023; Angeli et al., 2023; Avery et al., 2023).⁶ These papers do not identify the mechanisms behind these effects and infer that workers' responses are entirely (and sometimes incorrectly) strategic, whereas I show that this avoidance may be due to psychological costs. Other work shows that workers value related concepts like dignity or lack of sexual harassment at work using vignette experiments and amenity valuation (Sockin, 2021; Dube et al., 2022; Folke and Rickne, 2022; Adams-Prassl et al., Forthcoming). I provide experimental evidence that perceived discrimination affects worker sorting in and out of jobs and increases willingness to take a pay cut to avoid future discrimination, due to both a direct utility cost and lower expected wages.

Most closely related is contemporaneous work by Gagnon et al. (2024). In an experiment with Prolific workers in the UK, they show that explicitly attributing wage inequality to gender discrimination further reduces lower-wage workers' labor supply (analogous to the behavior I call retention) relative to the effects of wage inequality alone.⁷ Since I study a situation in which

⁶Other work shows that Black workers reporting previously experiencing discrimination search for jobs more widely and that controlling for perceived discrimination explains racial differences in labor supply and job satisfaction (Goldsmith et al., 2004; Mukerjee, 2014; Pager and Pedulla, 2015).

⁷They also study inferred perceived discrimination by informing workers of the other worker's gender and wage and find that this makes workers (especially women) more likely to perceive discrimination as measured on a 7-point Likert scale in response to the question, "During the task, did you believe that gender discrimination was used to

workers learn about hiring procedures and their results and possibly infer discrimination, we learn that perceived discrimination reduces retention in a more realistic setting and among those who naturally perceive discrimination; I additionally show that perceived discrimination affects future labor supply and cooperation. They provide suggestive evidence that reduced morale drives their results; using various reservation wage and quantitative belief elicitation I show that there are large psychological costs to perceived discrimination using revealed-preference measures. Finally, by introducing variation in the degree to which evaluation procedures can discriminate, we also learn in my setting that implementing anti-bias policies alone is unlikely to reduce perceived discrimination unless minority-group representation improves.⁸

Finally, an extensive literature in computer science, law, and economics shows that algorithms can either exacerbate or mitigate human bias (e.g. Kleinberg et al., 2018; Obermeyer et al., 2019; Kleinberg et al., 2020; Raghavan et al., 2020). Consistently, there is mixed evidence on how people perceive algorithmic versus human fairness. There is little work on perceived race and gender bias specifically.⁹ Two recent papers study revealed preferences for algorithmic versus manager decision-making and also find mixed results. Women are 35 percent more likely to apply to tech jobs when they will be evaluated by an algorithm rather than managers because they anticipate less discrimination (Avery et al., 2023), but workers in a lab setting are about equally split when they can choose between manager and algorithmic evaluation (Dargnies et al., 2022). I experimentally compare perceptions of discrimination by managers and algorithms when workers can see that they previously promoted mostly white men. In this setting, workers are *more* likely to perceive algorithmic than manager discrimination and both have negative effects on subsequent behavior.

The paper proceeds as follows: Section 1.2 describes the setting and research design. Sec-

determine your payment per line.” This treatment has no effect on workers’ labor supply.

⁸I also introduce a novel measure of perceived discrimination to document these effects. There are two other notable differences. First, I consider the role of race in addition to gender and find substantial heterogeneity by race conditional on gender, as well as by gender. Second, my experiment tests these effects in a context in which there is stereotypically gender and racial discrimination (proofreading and summarizing scientific articles), versus their neutral context (copying randomly-generated strings of numbers and letters).

⁹People find depictions of algorithmic or human decision-making more fair in different small-sample studies depending on the definition of fairness (Lee, 2018; Kaibel et al., 2019; Acikgoz et al., 2020; Newman et al., 2020; Noble et al., 2021; Zhang and Yench, 2022). People are less morally outraged by descriptions of algorithmic discrimination than human discrimination due to inferred differences in underlying prejudice (Bigman et al., 2023).

tion 1.3 provides evidence on the first research question—the effects of perceived discrimination. Sections 1.4, 1.5, and 1.6 provide evidence on the effect of anti-bias policies on perceived discrimination, the role of minority-group representation, and whether algorithms can mitigate the effects of perceived discrimination, respectively. Section 1.7 discusses robustness and Section 1.8 concludes.

1.2 Research design

The key challenge in studying the effects of perceived discrimination is that those who perceive discrimination are more likely to have experienced the direct effects of discrimination and to be from oft-discriminated groups, who are—observably and non-observably—different from others. This experiment overcomes these challenges by creating random variation in perceived discrimination. Importantly, it does so while minimizing actual discrimination, being truthful with workers, and proxying a realistic work scenario in which workers are likely to perceive discrimination—circumventing new challenges introduced in an experimental context.

This section describes how I address these challenges. In summary, workers are randomly assigned to be evaluated by procedures that vary in their potential to discriminate (though in this setting, none observably did so). Workers who are not selected to do a difficult job learn about their evaluation and see workers that were previously selected under the same procedure, only sometimes learning these workers' demographics—they are primarily white men due to overrepresentation in past cohorts. Finally, the analysis sample restricts to workers who would not have been selected under any procedure, ensuring that differences across treatment arms come only from differences in perceptions of the procedures, and not the procedures' decisions themselves.

1.2.1 Setting and sample

I recruit workers on Prolific, an online platform commonly used in economics research where workers complete surveys for pay. Prolific is a suitable setting for my experiment for several

reasons. First, manipulating perceived discrimination requires experimental variation and it is costly and difficult to randomly manipulate worker-selection procedures in the field. My approach on Prolific, a functional labor market, allows me to vary perceived discrimination while holding fixed: (i) any effects of actual discrimination (though there is none), (ii) all information workers have about the jobs, selection procedures, and other workers, and (iii) the effect of not being promoted. The main limitation is that, unlike most workplaces, workers and managers have short, impersonal “interactions.” The implications for generalizability are ambiguous: with extended interactions, workers may perceive more discrimination if they take it more personally or less discrimination if they know more context. Effects on behavior may be larger in a more interactive setting if they arise due to retaliation or distrust, or smaller if there are higher stakes.

In both experiments, I over-sample workers likely to experience and perceive racial and gender discrimination. 2,080 workers make up the main analysis sample in the first experiment, of which 50 percent are white women, 21 percent each are racial minority women and men, and 8 percent are white men. 2,527 workers make up the main analysis sample in the second experiment, of which 72 percent are white women, 15 percent are racial minority women, and 13 percent are racial minority men.¹⁰ In both samples, the racial minority men and women are one-third each Asian, Black, and Hispanic. For comparison, the population of US workers *excluding white men* is 41 percent white women, 30 percent racial minority women, and 29 percent racial minority men (calculated from the 2021 American Community Survey, ACS).

The experimental sample includes people representative of large swaths of the US population and 96 percent have non-Prolific work experience (Table 1.1). Seventy percent are currently employed outside Prolific. The sample is well-educated—50 percent have a 4-year college degree or more and 35 percent have some college experience but no degree or a two-year degree—and there is a large spread in income—about 50 percent have household income less than \$60,000 per year,

¹⁰3,240 [3,960] workers were recruited to the baseline survey for the first [second] experiment. 120 [360] are dropped from the experimental sample because the list of possible managers to evaluate them was exhausted (note that randomization occurred within these groups of 120 [360]). Of these 2,886 [3,510] are not promoted [hired] by their randomly-assigned evaluation procedure and thus offered the chance to participate in the experiment, which 80 [76] percent of them complete. Of those, 90 [94] percent would not have been promoted [hired] under any procedure regardless of their random assignment and make up the final analysis sample.

25 percent earn between \$60,000 and \$100,000, and the rest earn above \$100,000. The sample is on the young side of working age, with 70 percent between ages 18 and 44.

The experimental sample is on average more educated, but poorer and younger than a nationally representative sample. The fraction with a job outside Prolific is the same as the fraction of the US population in this age range that is in the labor force. Table 1.1 compares summary statistics from the experimental samples to their analogues from the nationally-representative 2021 ACS. Column 3 restricts the ACS sample to individuals ages 18-71, the minimum and 99th percentile of age in the experimental sample, and column 4 further restricts to employed women and racial minority men. The experimental sample is more likely to have a college degree, but also more likely to have household income between \$20,000-\$60,000 and less likely to earn above \$160,000 than either representative sample.

1.2.2 The promotion experiment

The first experiment is referred to as the “promotion experiment” because workers are evaluated and assigned to one of two jobs, one of which represents “promotion” and the other “non-promotion.” Non-promoted workers are the sample of interest and their performance in the non-promotion job is a key outcome. This experiment primarily tests whether perceived discrimination affects worker behavior.

I refer to the second experiment as the “hiring experiment” because when workers are evaluated, they are either selected to do a job (the same as the “promotion job”), and thus “hired,” or not. The non-hired workers are the sample of interest, and I do *not* observe their performance in an effortful job. It primarily follows the same structure as the promotion experiment with the deviations described in Section 1.2.3; the primary goal of this experiment is to test whether hiring policies that reduce *actual* discrimination affect *perceptions* of discrimination.

1.2.2.1 The job

Workers are offered scientific proofreading jobs. In the lower-paying, easier job (“non-promotion”), they proofread up to eighteen paragraphs from elementary-school-level articles in the *Science Journal for Kids* and can earn up to \$4.50 in bonuses. In the higher-paying, harder job (“promotion”), they proofread up to twelve paragraphs from articles in leading scientific journals (*Nature*, *Science*, etc.) and write a short summary of each, and can earn up to \$9.00 in bonuses. Each paragraph has around 100 words and four inserted errors. Most workers use almost the entire allotted minute to proofread each paragraph. After the first six paragraphs, workers choose after each paragraph whether to continue proofreading.¹¹

To proofread a paragraph, workers highlight words they think contain errors. They click on a word to highlight it, or again to un-highlight it if they change their mind. Their bonus is determined by the number of correctly-highlighted mistakes and incorrectly-highlighted non-mistakes when they submit their work: they earn the bonus for a given paragraph if two times the number of correct highlights minus the number of incorrect highlights is at least the median of the same statistic for that paragraph in a pilot. Workers, however, are only told they are paid for each paragraph in which they do “a good job” and that on average, workers could expect to be paid for half of the paragraphs.

The scientific proofreading job was chosen for several reasons. First, women and racial minority men are under-represented and stereotypically under-perform in most scientific fields. Second, of the effortful tasks that can be embedded in a survey experiment, it requires relatively high skill levels and effort and can be long without becoming tedious. Third, unlike other science- or math-based survey tasks, performance might be subjective, and as just described, workers knew only that they would be paid per paragraph that they did a “good job” proofreading. The effects of perceived discrimination could therefore have operated through workers anticipating further discrimination in payment decisions. Empirically, however, workers anticipated the use of objective performance

¹¹Workers who quit *before* finishing the first six paragraphs forgo their participation payment; 1.5 percent of potential participants who consented to taking the survey did so.

metrics when asked at the end of the survey how they thought “a good job” was determined.

What did workers know about the jobs? At the beginning of the experimental survey, workers are told they are working on a team that proofreads scientific articles, and that a [manager/algorithm] assigned workers to one of two jobs. They learn what makes one job harder and the other easier—the complexity—and that the harder job pays two to three times as much per paragraph (see Appendix Figure A.1).¹² They then learn they were assigned to the easier job.

1.2.2.2 Design overview and background

Appendix Figure A.2 illustrates the overall experimental design. Workers are randomly assigned to be evaluated by one of three promotion procedures with varied potential to discriminate. These procedures—managers who do or do not know workers’ race and gender and a demographic-blind algorithm—know workers’ prior performance and education and decide who to promote. Before recruiting the experimental sample, the same managers and algorithm evaluate groups of workers in which white men were over-represented (a timeline is in Appendix Figure A.3). The key to inducing variation in perceptions of discrimination in the experimental sample is that workers learn about the procedure under which they were not promoted and who was previously promoted under that procedure.

The historical sample. The experiment aims to proxy a setting in which a worker from an under-represented group is not promoted and knows that mostly white men were previously promoted. To generate these dynamics online, 1,800 predominantly (85 percent) white male “historical workers” are initially evaluated by the *same managers* who later evaluate the experimental sample, and by the same algorithm. Then, the experimental workers see who their manager or the algorithm previously promoted when they learn about their own evaluation (see Section 1.2.2.3).

The managers. The managers, all white men,¹³ are recruited from Prolific and tasked with

¹²At the end of the survey, 68 percent of workers report preferring the harder job, 6 percent were ambivalent, and 20 percent were unsure but didn’t think they would prefer the harder job. This does not differ by treatment group.

¹³This serves two purposes: to best represent cases in which a women or racial minority man might be the most likely to feel discriminated against and to minimize noise in the experiment caused by variation in manager characteristics (only 54 participated in the study).

choosing three of twenty-four historical-sample workers to promote. They know their decisions will be implemented and the performance of the workers they promote will determine their bonus (most of their total payment). Half of the managers are randomly assigned to be “demographic-blind” (they see only baseline performance and education levels) while the other half are “non-blind” and also see workers’ self-identified race, gender, and an avatar that workers created in the baseline survey (see below). Baseline performance is shown as 1-5 stars indicating the quintile of workers’ average quiz scores and education is in three broad categories: no college degree, college degree, and more than a college degree.

The same managers return to evaluate workers in the experimental sample and maintain the same “demographic-blind” or “non-blind” status as when they evaluated the historical sample. They also evaluate workers with similar quiz scores as those they evaluated in the historical sample (see below). Each manager evaluates three sets of forty workers independently, promoting three from each set. Their decisions are implemented for one randomly-chosen set, and the performance of workers they promote in that set determines their bonus payment. Appendix A.2 provides details about manager recruitment and randomization.

The algorithm. The algorithm uses a random forest regression-based model with workers’ quintile of average quiz scores and education as inputs to predict performance in the harder job. It promotes the three workers with the highest predicted performance in each group of twenty-four (in the historical sample) or forty (in the experimental sample). Ties are broken randomly. It was trained on a sample of 500 workers recruited from MTurk using Cloud Research’s pre-approved participant pool. Like the experimental sample, women and racial minority men are over-represented in the training sample. Appendix A.3 provides more detail.

1.2.2.3 Experimental design

This section describes the promotion experiment design, using the vocabulary from the previous section (i.e. *algorithm*, *demographic-blind manager*, *non-blind manager*, and *previously-promoted workers*, which refers to promoted workers in the historical sample). Appendix Figure A.4 illus-

trates the treatment arms.

Baseline survey. Workers are recruited on Prolific for an initial survey, during which they take three timed quizzes that are predictive of their ability in the proofreading jobs and report how well they thought they did on each.¹⁴ Workers know their performance on the quizzes could affect their future work opportunities and are paid a small bonus for each correct answer. Workers then answer demographic questions, build an avatar that “looks like them,”¹⁵ report a detailed work history, including past experiences of discrimination at work, and report incentivized beliefs about the prevalence of discrimination against different groups.

Randomization and worker evaluation. Workers are randomly assigned to groups of 120 with average baseline quiz scores within one quintile of each other. Each group is evaluated by two randomly-assigned managers, one demographic-blind and one non-blind, and the algorithm. Both managers had previously been randomly assigned to evaluate workers in the same two quiz-score quintiles in the historical sample, so workers with low scores see previously-promoted workers with similar scores to themselves and can infer they had a chance at getting the harder job. Otherwise, low-scoring workers would mainly see top-scoring workers being promoted and would be unlikely to attribute decisions to discrimination.¹⁶

Within each group, each worker has an equal probability of being randomly assigned to have the decision of the demographic-blind manager, non-blind manager, or algorithm implemented. Among workers assigned to be evaluated by the algorithm, half are randomized to see previously-promoted workers’ avatars, race, and gender in their profiles (like the non-blind manager arm)

¹⁴A spelling quiz of 10 scientific words, 15 grammar questions, and 11 science questions. The difference between workers’ performance and their (incentivized) beliefs about performance is a measure of confidence, which could predict perceived discrimination if motivated reasoning causes some workers to attribute rejection to discrimination rather than their performance (Heidhues et al., 2019). This does not seem to be the case (Appendix Figure A.18).

¹⁵Workers made their avatars following the procedure shown in Appendix Figure A.19. They do not know what the avatars will be used for or that they will be seen by other participants. They therefore should *not* have chosen their avatars strategically, indeed, workers largely choose avatars that align with their self-reported race and gender in expected ways (Appendix Figure A.20).

¹⁶Workers were told “[your manager/the algorithm] evaluated workers with similar quiz scores as you.” As intended, workers at all performance levels were similarly likely to perceive discrimination, conditional on being in the higher or lower quintile in their group, though workers in the fifth quintile in the 4-5 quintile group were more likely to perceive discrimination than those in the higher quintile in the 3-4 and 2-3 quintile groups (Appendix Figure A.21, Panel A). Unsurprisingly, workers in the higher of the two quintiles of their group are more likely to perceive discrimination. Workers with more education were also more likely to perceive discrimination (Panel B).

whereas the other half do not (like the demographic-blind manager arm). Workers are balanced across treatment arms on baseline characteristics (Appendix Table A.1).

Workers that are not promoted under their randomly-assigned promotion procedure are offered the easier job and the 80 percent who take it up become the experimental sample (N=2,397). They are all offered the same Prolific task (i.e. same wage and job description) and only learn about their evaluation *after* opting in. Attrition after starting the survey was uncommon, and similar across all treatment arms (Section 1.7). Promoted workers are offered the harder job.

Information about non-promotion and evaluation procedures. The timeline of the experimental survey is shown in Appendix Figure A.5. Workers are invited to take the experimental survey at least three weeks after taking the baseline survey. Upon starting the survey, workers learn about the two jobs and that they were not promoted (i.e., they were assigned to the easier, lower-paying job). Framed as context for the next question, “*what do you think needed to be different about your profile to be assigned the harder job?*,” they then learn how they were evaluated.

In the manager sample, workers learn that managers was recruited from Prolific, reviewed example paragraphs, evaluated worker profiles, and are paid based on the performance of the workers they promote and so should be trying to promote the best workers. They also see their manager’s profile (avatar, age, education, gender, race, and tenure on Prolific). Thus, differences in perceived discrimination between the two manager arms control for having a white male manager.

The differences in what workers know about their evaluation (and thus perceived discrimination) comes next: they are told their manager previously evaluated workers with similar quiz scores as them, then see their own and the previously-promoted workers’ profiles. The profiles communicate what the manager knew when making their decision. Appendix Figure A.6 shows the implementation.

In the non-blind manager arm, workers see profiles that include avatars, average quiz-score quintiles as 1-5 stars, education (no college degree, college degree, or more than college degree), self-identified gender (Man, Woman, or Non-Binary/Other), and self-identified race/ethnicity.¹⁷

¹⁷Only 6 of 2397 participants identified as non-binary; they are grouped with women in results that split by gender as the “not male” category.

In the demographic-blind manager arm, the profiles show only quiz-score stars and education. This controls for any effect of feeling misjudged for not having a college degree, misunderstanding what the stars represented, or not being promoted.

Thus, “treatment” in the manager sample is learning that one’s manager knew race and gender and seeing the avatars, race, and gender of previously-promoted workers (mostly white men), relative to the “control group” whose manager did not know race and gender and who does not know that the previously-promoted workers are mostly white men.

In the algorithm arm, all workers learn that an algorithm used data from previous iterations of the survey to predict who would do the best at the harder job. They are explicitly told that the algorithm only uses information on average quiz scores and education to predict performance. Next, workers see previously-promoted workers’ profiles. Half are randomly assigned to see profiles in the same format as the non-blind manager arm: they include avatars, stars, education, gender, and race. The other half see profiles in the same format as the demographic-blind manager arm: just stars and education (Appendix Figure A.7). Thus, “treatment” in the algorithm arm is seeing the avatars, race, and gender of previously-promoted workers (mostly white men), relative to the “control group” who does not know that the previously-promoted workers are mostly white men.

Finally, when workers see previously-promoted workers’ avatars, race, and gender, 40 percent of workers see that three white men were previously promoted. The rest see two white men and someone from another demographic group. This is random due to the random pairing of workers with managers or groups of historical-sample workers jointly evaluated by the algorithm.

Eliciting perceived discrimination. After learning about their evaluation procedure and who was previously promoted, workers are asked, “*what do you think would have needed to be different about your profile for you to be assigned to the harder job? For example, would it have helped if you scored higher on the quizzes, or had more education?*” and answer in a text box. The main measure of perceived discrimination is an indicator for whether workers’ responses suggest that they think their demographics (age, race, gender, etc.) played a role in their evaluation.¹⁸

¹⁸Age was not communicated directly but was somewhat observable through avatar hair color and baldness. This was pre-registered as the main measure of perceived discrimination. The free response was coded as mentioning demo-

To my knowledge, no other study has attempted to identify individuals who felt that they had been discriminated against in a recent interaction.¹⁹ This measure aims to capture that feeling without priming workers to think about discrimination or their own race and gender. Other, more explicit questions that identify perceived discrimination—asked at the end of the survey and described below—correlate highly with the main measure (Appendix Figure A.8). This elicitation and coding, all other outcomes, and control variables are described in detail in Appendices A.5 and A.4).

Observed work outcomes. Workers are required to proofread six paragraphs (spending a maximum of six minutes) to earn their participation payment; after the sixth and each subsequent paragraph they are explicitly asked if they want to proofread another paragraph or skip to the end of the survey. The primary outcome is how many paragraphs workers choose to proofread, including indicators for whether they did more than six or all eighteen.

In each paragraph, workers highlight (or unhighlight) identified errors by clicking on a word. For each paragraph, I observe the time spent, the number of clicks, and how many final highlights correctly identify mistakes or incorrectly identify non-mistakes. A summary measure of performance is their earned bonus. After the first, third, or fifth paragraph—timing randomly assigned—workers also complete a standard psychological scale that elicits current emotional states.

Incentivized survey outcomes. I elicit reservation wages via a multiple price list: workers indicate for various wage schedules whether they would like to participate in a future round of the survey at the given wage schedule for the harder and easier job; the wage in the harder job is always twice the wage in the easier job. “Participation in a future round” involves being evaluated again and either being promoted or not promoted. One wage schedule is randomly chosen and workers’ choices are implemented for a random subset of workers.

Workers answer these questions under two conditions: (1) if they would be evaluated by the

graphics independently by two MIT undergraduates and a PhD psychologist who professionally codes qualitative data and was hired on Upwork. The externally-generated variables used in the analysis are highly correlated with variables generated by the author. The full coding scheme for this and other text-based variables is in Appendix A.5.

¹⁹Surveys (e.g. [Pew Research, 2017](#); [Gallup Inc, 2021](#)) ask whether someone has experienced discrimination or been treated differently because of their gender or race, “ever” or in the past year.

same procedure as in the experiment and (2) if their assignment would depend only on whether they had the highest baseline quiz scores. After each multiple price list, they report what they think is the probability that they will be promoted if they participate in the future round with the given evaluation procedure.

In the manager arms only, I similarly elicit incentivized reservation wages to participate in a collaborative task with their manager from the experiment and their willingness to pay to be able to choose a different manager to work with in that task.²⁰ Workers in the manager arms also play a dictator game. They know they may be randomly chosen to receive a \$20 thank-you bonus for their participation and choose how much they would want to share with their manager if they are selected, in which case their choice will be implemented. These outcomes measure workers' willingness to share and work cooperatively with their manager, proxying for e.g. willingness to stay late at work or take on extra tasks.

Self-reported outcomes. Workers next report their interest in future work, job satisfaction, and whether they would prefer the harder or easier job in the future, and complete self-efficacy scales for the proofreading jobs and related skills.

Secondary measures of perceived discrimination. Finally, they answer a series of questions that validate the main measure of perceived discrimination. First, they are asked if they have a complaint about the promotion procedure and if so, they can describe their complaint in an open-ended response, which is coded for complaints about discrimination using similar methods as the main measure (Appendix A.5). Second, they answer multiple-choice questions about whether they would have been promoted if they were a different race, or a different gender (masked with similar questions about education and quiz scores); if they say “Yes, I think so” or “Yes, definitely” they are coded as perceiving discrimination. These more direct measures of perceived discrimination are highly correlated with the main measure (Appendix Figure A.8). Finally, they answer incentivized questions about whether they thought various groups of workers were over- or under-represented

²⁰Workers summarize the paragraphs used in the harder version of the proofreading job, managers provide edits, and workers make revisions. They earn a base wage that the manager cannot influence, but worker-manager pairs who provide some of the best summaries will earn a bonus, and the manager has discretion over how it is split.

(or neither) among promoted workers, a measure of perceived discrimination in *general* rather than against themselves specifically.²¹

1.2.3 The hiring experiment

The hiring experiment tests how firms' hiring policies affect whether workers perceive discrimination. Specifically, it identifies the effects of using an algorithm versus a manager and the effect of blinding decision-makers to race and gender when workers see that primarily white men were previously hired. It takes place in the same constructed online labor market with newly-recruited but similar workers (Section 1.2.1); the only sampling difference is that white women are more heavily over-represented, due to constraints on the number of racial minority participants available who had not already participated in the promotion experiment. The not-hired workers take an experimental survey without any proofreading job, and the key outcome is perceived discrimination.

3,960 *new* workers are recruited to take a shortened baseline survey and then are evaluated by various procedures. There are six treatment arms; four replicate the arms of the promotion experiment (Appendix Figure A.9). Workers are randomly assigned to be evaluated by a manager or an algorithm and cross-randomized to arms where the decision-maker can use race and gender or not. Workers evaluated by a demographic-blind decision-maker are cross-randomized again to either see the avatars, race, and gender of previous hires or not. Again, workers are evaluated by all four procedures, but the decision of their randomly-assigned procedure is implemented. Workers are balanced across treatment arms (Appendix Table A.2).

The *same* managers evaluate these workers as in the promotion experiment. They maintain their earlier random assignment as demographic-blind or non-blind, and now choose one worker from groups of forty to do the harder proofreading job. The rest are offered the experimental survey without any proofreading job.

²¹All of these “explicit” measures of perceived discrimination require the worker to think about or describe an experience of discrimination. I also observe implicit measures—differences in the number of stars that workers report thinking they would have needed to be promoted compared to other demographic groups. The signs of the treatment effects on these variables are consistent with the explicit measures, but generally imprecise (available upon request).

The demographic-blind algorithm is the same as in the promotion experiment. The non-blind algorithm interacts each baseline performance and education variable in the demographic-blind algorithm with the eight race \times gender groups and predicts performance. Most algorithms would not explicitly use race and gender information, but a realistic hiring algorithm would use sufficient predictors to be able to approximate race and gender. I aim to proxy this type of black-box algorithm in a setting without large amounts of data. The non-blind algorithm does not observably discriminate—it meets the same calibration criteria as the demographic-blind algorithm (Appendix A.3)—and in fact, would hire more racial minorities than the demographic-blind algorithm (Appendix Table A.3).

When non-hired workers return for the experimental survey, they learn about the proofreading job, are told they were not chosen for it and see the analogous information about the managers and algorithms as in the promotion experiment. They then answer the question about what needed to be different about their profile in order to be hired. They do *not* subsequently do a proofreading job, but continue directly to reservation wage elicitation. Finally, they answer questions measuring memory and comprehension of the evaluation procedures and the questions that underly the secondary measures of perceived discrimination.

1.2.4 Experimental fidelity: Attention and comprehension

The mechanism ensuring attention to the information about evaluation procedures was the open-ended question, “*what do you think needed to be different about your profile to be assigned to the harder task?*” Workers had to answer this question; fewer than 5 percent gave a reason outside the standard codes (Appendix A.5).

In the hiring experiment, workers also answered comprehension questions about the procedures at the end of the survey. Approximately 80 percent of workers in all manager arms correctly identified the managers’ payment structure and approximately 70 percent of workers in all algorithm arms correctly identified that the algorithm predicts performance. 88-98 percent of workers correctly identified that the decision-makers did use education and average quiz-score quintiles

represented by stars and did not use age, work history, or how long they’ve worked on Prolific; this did not vary by treatment group.

In both experiments, workers answered an “attention check” during the final un-incentivized survey section. 75 and 80 percent of workers answered the question correctly in the promotion and hiring experiments, respectively.²² There is no differential attention across treatment arms and the main results are robust to dropping the workers who failed the attention check (Section 1.7).

1.3 The effects of perceived discrimination

1.3.1 Estimation

To understand the effects of perceived discrimination, I focus on intent-to-treat (ITT) specifications that compare outcomes between workers assigned to the demographic-blind manager arm and the non-blind manager arm in the promotion experiment. I estimate the following:

$$Y_i = \alpha + \beta T_i + X' \delta + H' \lambda + \gamma_g + \varepsilon_{ig} \quad (1.1)$$

among workers evaluated by a manager, where T_i is one if worker i is in the non-blind manager arm and zero otherwise; β is the effect of learning that one’s manager saw avatars, race, and gender and previously promoted mostly white men. γ_g is fixed effects for workers’ quiz-score quintile group, X is a vector of baseline worker characteristics and H is a vector describing the education and quiz scores of the previously-promoted workers seen by worker i .²³ Appendix A.4 provides details on all outcome and control variables. Standard errors are robust to heteroskedasticity.

One threat to the causal interpretation of this specification as the effects of interventions that change only *perceptions* of discrimination is that different evaluation procedures could have pro-

²²Workers were asked to select one column for both rows in a matrix with two Likert-scale questions.

²³In the main specification, X includes baseline quiz scores, education, income, age, family status, gender, and race. H includes indicators for whether the worker saw one, two, or three previously-promoted workers with the maximum number of stars possible in their quiz-score group (the rest had one fewer star) and ten exhaustive and mutually-exclusive indicators for how the workers’ education compared to the education of the previously-promoted workers they saw. The controls in H were not pre-registered; the results are the same excluding these controls (Section 1.7).

moted different types of workers. Since their real decisions were implemented, this could generate differences between the treatment arms among workers who are not promoted. To eliminate this threat, the experimental design builds in the counterfactual evaluation of every worker by every procedure. In the main analysis, the sample restricts to workers who would not have been selected regardless of their random assignment—i.e. the workers not promoted or hired under any procedure. In the promotion experiment, this is 2,080 workers of 2,317. In the hiring experiment, this is 2,527 workers of 2,680. In practice, the different procedures do not appear to select systematically different workers (Appendix Table A.3) and the workers who are selected under at least one procedure appear similar to those who are never selected (Appendix Table A.4).

The choice of controls and analysis sample do not affect the main results. They are also robust to multiple hypothesis testing and using randomization inference (Section 1.7).

An instrumental variables (IV) specification. I also present IV specifications in Appendix A.6, instrumenting for the main measure of perceived discrimination with random assignment. These effectively divide the ITT estimates of the effects on behavior by the “first stage” effect on perceived discrimination rates. The IV estimates can be interpreted as the effect of perceived discrimination, but to do so, random assignment must only affect outcomes via its effects on *the instrumented measure of perceived discrimination*. This assumption may be violated—for example, there may be a certain degree of perceived discrimination that has effects, which may be more or less frequent than the measures of perceived discrimination I collect. Of those I collect, the main measure has the largest first stage, and thus implies the lowest IV estimates. There may be workers who perceive discrimination and are affected that are not picked up by this measure, in which case the IV estimate may still be an over-estimate. The ITT estimates are thus the focus in the paper and represent a lower bound on the average effects of perceived discrimination caused by a particular treatment.

1.3.2 Manipulation check: Treatment affects perceptions

Workers' perception of discrimination strongly depends on whether they learn that their manager knew demographic information and previously promoted mostly white men. Figure 1.1 plots rates of perceived discrimination in each manager arm estimated using Equation 1.1 for the full sample and separately for women, racial minority men, and white men.²⁴

Perceived discrimination is uncommon in the demographic-blind manager arm. Only 3 percent of workers mention demographics when asked about what needed to be different about their profile in order to be assigned the higher-status job; these workers think the manager had more information than was shown in the profiles.

Overall, being in the non-blind manager arm increases the share of workers perceiving discrimination by 31pp (se=2pp).²⁵ Being in the non-blind manager arm increases perceived discrimination by 37pp (se=3pp) for women and 20pp (se=4pp) for racial minority men. Appendix Figure A.10 plots rates in each treatment arm for each of eight race \times gender cells; women generally perceive discrimination at similar rates regardless of race as do racial minority men.²⁶

Appendix Table A.5 shows parallel results using the secondary measures of perceived discrimination. There is always a large, significant effect, and patterns across demographic groups are the same for all measures, though the prevalence in the demographic-blind manager group and

²⁴Heterogeneity by race and gender was pre-registered, along with heterogeneity by past experiences of discrimination, beliefs about the prevalence of discrimination, and confidence. I focus on gender and race as they most systematically predict perceived discrimination, but multivariate regressions that include all pre-registered characteristics are in Appendix Figure A.18.

²⁵Almost all workers who do not perceive discrimination cite needing more education or higher quiz scores to be promoted. Being in the non-blind manager arm reduces the share who cite the *primary* reason being their education or quiz scores by 11pp and 6pp (27 and 13 percent, se=2.4pp and 2.6pp), respectively (Appendix Table A.10). Note that the main measure of perceived discrimination is an indicator for mentioning demographics *at all*; being in the non-blind manager arm increases the share of workers who mention demographics as the *primary* reason they think they were not promoted by 18pp (se=2pp, Appendix Table A.5).

²⁶Why are there such stark gender differences? Gender discrimination could have been more salient than racial discrimination in this experiment, as the avatars' shirt colors differed by gender but race was only partially communicated in the avatar via skin tone. However, among racial minority women, the majority of those who perceive discrimination report that they think they would have been hired if they had a different race *or* if they had a different gender; of those who only mention one or the other, about twice as many perceive only racial discrimination. Aksoy et al. (2023) similarly find that women are more likely to anticipate discrimination on the basis of some other negatively-discriminated characteristic than men (in their case, sexual orientation).

magnitude of treatment effects vary.

1.3.3 Results: Effects of perceived discrimination

1.3.3.1 Retention and performance

Learning that one's manager knew race and gender and previously promoted mostly white men lowers retention—persistence in the work task—and performance, particularly for women, though it also improves some workers' performance.

Retention. Workers must proofread six paragraphs to be paid their participation wage, and can then choose to proofread up to eighteen. Learning that one's manager knew race and gender and previously promoted mostly white men—i.e. being in the non-blind manager arm—causes workers to proofread 0.5 fewer paragraphs (3 percent, $se=0.25$, $p=0.05$) and reduces the probability that they proofread all eighteen paragraphs by 4.9pp (6 percent, $se=2.5pp$, $p=0.06$) (Table 1.2).

Intensive effort and performance. I identify causal effects of treatment on effort and performance in the first six required paragraphs, before treatment causes differential selection in or out of the sample. Being in the non-blind manager arm does not affect measures of how hard workers choose to work: time spent per paragraph (capped at 1 minute) or the number of times they click on the page (a proxy for whether they check their answers). It also does not affect performance—the number of correct or incorrect highlights, or earned bonus (Table 1.3). This masks substantial heterogeneity, however, which is discussed below.

Total earnings. If the bonus was a piece-rate per paragraph, the negative retention effect would correspond to the same reduction in total earnings. Instead, the bonus depended on passing a performance threshold in each paragraph, so the change in retention may not correspond to the effect on total earnings if there is an effect on performance among those who do not quit.

Indeed, perceived discrimination decreases total earnings for some but increases total earnings for others. The overall effect of being in the non-blind manager arm on total earnings is zero (Appendix Figure A.11). Combined with the negative effect on retention, this implies that those who

stay experience a *positive* treatment effect of perceived discrimination on earnings in the later paragraphs, canceling out the negative earnings effect on those induced to quit early. While this effect cannot be quantitatively separated from selection effects in who chooses to stay, heterogeneous treatment effects by race and gender provide a sense of magnitudes.

Heterogeneity by race and gender. The negative effects on retention are driven by women, and racial minority women experience a negative performance effect in the required paragraphs, where selection cannot play a role. In contrast, racial minority men experience positive performance effects. Figure 1.2 plots the effect of being in the non-blind manager arm on the cumulative percent quit, bonus earned, and correct and incorrect highlights separately by race and gender.

White and racial minority women drive the negative effects on retention. White women proofread 0.65 (4 percent, $se=0.32$) fewer paragraphs and are 3.9pp (70 percent, $se=2.1pp$, $p=0.07$) more likely to immediately quit. Racial minority women proofread one (6 percent, $se=0.56$, $p=0.07$) fewer paragraph and are 12pp (16 percent, $se=5.7pp$) less likely to complete all eighteen paragraphs. In contrast, racial minority men proofread 0.81 *more* paragraphs, though the estimate is not statistically significant. The effects on retention for men and women are significantly different from each other at a less than 5 percent level (Appendix Table A.6).

Racial minority men and women's performance is also affected. Racial minority men *correctly* highlight 1.3 (10 percent, $se=0.47$) more mistakes in the required paragraphs, whereas racial minority women *incorrectly* highlight 0.8 (30 percent, $se=0.41$) more non-mistakes. They also highlight about 5 percent more words correctly, but the effect is not statistically significant (Appendix Table A.7, Panel A). Thus, women and men respond differently when they perceive discrimination—on average, white and racial minority women disengage and do worse, while men try to prove themselves, in this case, successfully. Racial minority women may similarly try to prove themselves by highlighting more words, but do so unsuccessfully, or disengage and simply highlight more words less thoughtfully. The positive performance effect for racial minority men in the required paragraphs is significantly higher ($p=0.03$) from the zero effect for women when white and racial

minority women are pooled together (Panel B).²⁷

Gender differences in confidence and competitiveness (Niederle and Vesterlund, 2007; Coffman, 2014; Bordalo et al., 2019) could underly this result, though other factors may also be at play. Consistently, though not statistically significant, there is a weakly positive effect of being in the non-blind manager arm on racial minority men’s psychological well-being—they are less upset, discouraged, annoyed, and anxious—and a weakly negative effect for racial minority women—they are more annoyed and less motivated, though also less anxious (Appendix Figure A.12).²⁸

In summary, perceived discrimination overall negatively affects retention, with important heterogeneous effects on performance. Some workers may react by trying to prove themselves to biased managers, but only some do so successfully. Even if future evaluations are race- and gender-neutral, perceived discrimination may affect the outcomes of future evaluation via workers’ prior performance. The results suggest that perceived discrimination may exacerbate racial gaps among women and gender gaps conditional on race.²⁹

1.3.3.2 Future labor supply

Perceived discrimination also reduces future labor supply. These effects can be partially explained by beliefs about lower expected wages but derive primarily from disutility from additional interactions with a biased manager or employer.³⁰

After workers complete the proofreading job, I remind them of how they were assigned, explain that some workers will be offered additional work, and elicit reservation wages for these opportunities to be promoted or not promoted. The reservation wages are measured in terms of both the

²⁷This plays out as one would expect in terms of total earnings. There is a negative but insignificant effect on total earnings for women and a positive effect for men (Figure 1.2).

²⁸There are no effects on self-efficacy or job satisfaction (Appendix Tables A.11 and A.12).

²⁹I focus on heterogeneity by race and gender as it yields the most consistent patterns of the pre-registered heterogeneity analysis. Appendix A.7 looks at other pre-registered dimensions of heterogeneity. The negative effect of perceived discrimination on retention is most negative for workers with below-median confidence, after accounting for racial and gender heterogeneity. The stark gender heterogeneity remains after accounting for confidence.

³⁰The primary analysis of these outcomes comes from the promotion experiment, so that the sample is the same as in the previous section. A subset of the labor supply outcomes were also measured in the hiring experiment; this was pre-registered as a replication of the promotion experiment results. Indeed the results are very similar (Appendix A.8).

non-promoted piece rate *and* the promoted piece rate simultaneously, with the higher piece rate always twice the lower. In other words, each row of the multiple price list reflected two wages, e.g. “Earn \$0.20 per paragraph if not promoted and \$0.40 per paragraph if promoted.” Table 1.4 shows effects on reservation wages calculated as if all workers assume the *same expected wages* in each row, i.e. expected the same probability of promotion.³¹ I use a 50 percent probability because the average worker in the control group believes this is the likelihood of future promotion.³²

Thus, the effect on this formulation of reservation wages could be driven either by beliefs about future discrimination or psychological mechanisms like anticipated disutility. I decompose these channels in Section 1.3.3.3.

Perceived discrimination increases reservation wages for future work and willingness to pay for a transparent, unbiased mechanism. Being in the non-blind manager arm increases reservation wages from 26.5 cents to 28.8 cents per paragraph (2.3 cents, 9 percent, $se=1$ cent) if workers would be evaluated by the same manager (Table 1.4). Those in the non-blind manager arm are also willing to pay an additional 1 cent per paragraph ($se=0.5$ cents) to be evaluated by a cutoff rule in baseline performance in which there is no potential for future discrimination or interaction with your manager, measured as the difference in reservation wages if they would be evaluated by their manager or the cutoff rule. This willingness to pay is economically meaningful: it is 4 percent of the average reservation wage if one anticipates not being promoted in the next round.

Perceived discrimination also reduces workers’ willingness to interact with and generosity towards their manager (Table 1.5). Being in the non-blind manager arm increases workers’ willingness to pay to be able to choose a different manager in a cooperative job by 4.1 cents per paragraph (18 percent, $se=2$ cents). Workers in the non-blind manager arm are 5pp (10 percent, $se=3pp$, $p=0.09$) more likely to share \$0 with their manager of their possible \$20 bonus, and overall share

³¹Sample sizes are slightly smaller for these variables as I drop observations in which workers make inconsistent choices indicating they do not understand the MPL elicitation method, as is standard in this type of analysis. Appendix A.9 shows that the effects on retention and performance are similar in this restricted sample.

³²This puts this overall reservation wage on the same scale as the variable that accounts for workers’ beliefs about future promotion in the next section. 50 percent is much higher than the actual promotion rate, (which workers had no information about). Appendix Figure A.22 plots the CDFs of beliefs by treatment group; a large share of workers guess 50 percent but there is wide variation.

about 26 cents (9 percent, $se=23$ cents, $p\text{-value}=0.28$) less with their manager (Table 1.5).³³

Applied to a work context where employees and managers interact on a regular basis, this response to perceived discrimination could be self-fulfilling—if workers react to perceived discrimination by avoiding their managers or refusing their requests, managers may be less likely to reward these workers in the future.³⁴

1.3.3.3 Mechanisms

The effects on willingness to interact with and generosity towards managers suggest that animus or avoidance may underly the effects of perceived discrimination on worker behavior. The effects on reservation wages, however, could also be attributed to higher beliefs about the likelihood of future discrimination and thus lower expected wages. Past work on anticipated discrimination either focuses on the second channel—lower returns to effort—or does not separate the two.³⁵

Understanding how each channel contributes to the effects of perceived discrimination sheds light on whether models of anticipated discrimination that focus on changing returns to effort may be missing a key, psychologically-driven element. Next, I describe a framework to disentangle these effects, map the framework into the experimental design, and then describe results.

Simple framework. In a standard model, perceived discrimination affects reservation wages only through beliefs about the probability of future promotion. Reservation wages equalize ex-

³³This “dictator game” measures distributional preferences and retaliatory preferences if one has already interacted with the recipient. Workers whose managers did not see demographic information shared on average \$2.80 (14 percent of the total) with their manager. A meta-analysis suggests that workers would have shared around \$6 (30 percent) with strangers (Engel, 2011), so even workers in the demographic-blind manager arm may be retaliating against their managers for not promoting them, or the hierarchical nature of the manager-employee framing affects distributional preferences. The effect of being in the non-blind manager arm is on top of any effect of non-promotion or the manager-employee framing.

³⁴In terms of race and gender, being in the non-blind manager arm increases all workers’ willingness to pay to avoid their manager and causes retaliation (Appendix Table A.13). That said, managers may react more negatively to women or racial minorities who are less cooperative or generous. Appendix A.7 looks at other pre-registered dimensions of heterogeneity. The positive effect of perceived discrimination on willingness to pay to choose one’s own manager is largest for workers who report experiencing discrimination at work in the past in the baseline survey.

³⁵Theory focuses on returns to effort and lab experiments testing these theories randomly assign workers to the discriminated “identity,” potentially making psychological costs less relevant (Lundberg and Startz, 1983; Coate and Loury, 1993; Fryer et al., 2005; de Haan et al., 2017; Dianat et al., 2020). Experiments studying identity-based (gender, age, sexual orientation, caste) discrimination do not separate (incorrect) beliefs about discrimination from avoidance of psychological costs (Alston, 2019; Charness et al., 2020; Lepage et al., 2022; Ridley, 2022; Agüero et al., 2023; Aksoy et al., 2023; Avery et al., 2023; Angeli et al., 2023).

pected utility from work and an outside option, where

$$\text{Expected utility} = \hat{\pi}U(w_H) + (1 - \hat{\pi})U(w_L),$$

$\hat{\pi}$ is beliefs about the likelihood of future promotion, w_H is the higher wage if you are promoted, w_L is the lower wage if you are not, and $U(w)$ is an indirect utility function representing utility at a given wage w when a worker chooses optimal effort. Utility is weakly increasing in the wage.

Let D indicate perceived discrimination in a previous encounter and $\hat{\sigma}$ be beliefs about the likelihood of future discrimination. In the standard model, perceived discrimination can only lower expected utility and thus reservation wages if beliefs about future discrimination are increasing in perceived discrimination: $\hat{\sigma} = \sigma(D)$, with $\frac{\partial \hat{\sigma}}{\partial D} > 0$, and if beliefs about future promotion are a decreasing function of anticipated discrimination: $\hat{\pi} = \pi(\hat{\sigma})$, with $\frac{\partial \hat{\pi}}{\partial \hat{\sigma}} < 0$. This simple framework has two implications: (i) when future discrimination is impossible, perceived discrimination should not affect beliefs about promotion, and (ii) after taking into account any effects on beliefs about future promotion, there should be no effect of perceived discrimination on reservation wages.

In a richer framework, perceived discrimination could affect expected utility and thus reservation wages through two other channels. First, beliefs about future promotion may depend on D independently of $\hat{\sigma}$: $\hat{\pi} = \pi(\hat{\sigma}, D)$. This could be because perceived discrimination changes e.g. self-confidence or trust in the employer. Second, workers may experience or anticipate a cost $C(\hat{\sigma}, D)$ which can affect expected utility independently of $\hat{\pi}$. In words, workers may dislike discrimination or dislike interacting with a biased manager or an employer who hires biased managers, even if they do not anticipate future discrimination. The anticipation cost is increasing in both $\hat{\sigma}$ and D . Altogether, now:

$$\text{Expected utility} = \pi(\hat{\sigma}, D)U(w_H) + (1 - \pi(\hat{\sigma}, D))U(w_L) - C(\hat{\sigma}, D).$$

Results. The evidence from my setting rejects both implications of the standard model. First, anticipated discrimination and changes in e.g. trust or self-confidence each explain about half of

the effect of perceived discrimination on beliefs about the likelihood of promotion. Second, the resulting change in expected utility from wages explains only 25 percent of the overall effect on reservation wages. The effect on disutility $C(\hat{\sigma}, D)$ explains the rest. It seems to derive about equally from (i) disutility from anticipated discrimination or interactions with biased managers and (ii) disutility from continuing to work for the employer.

First, I estimate the effects on the likelihood of future promotion, separately when future discrimination is possible (when one will be evaluated by the same manager, i.e. $\hat{\sigma} \geq 0$) and when it is not (when one will be promoted based on a cutoff rule in baseline performance, i.e. $\hat{\sigma} = 0$). In the standard model there may be an effect when $\hat{\sigma} \geq 0$, but not when $\hat{\sigma} = 0$.

Being in the non-blind manager arm increases anticipated discrimination *and* affects beliefs about promotion when no future discrimination is possible. Being in the non-blind manager arm lowers beliefs about the likelihood of future promotion by 2.7pp (6 percent, se=1.2pp). About 40 percent of this effect, however, can be explained by changes in beliefs about the likelihood of promotion when no future discrimination is possible. 1.5pp (se=0.7pp) of the effect of perceived discrimination on beliefs about the likelihood of future promotion comes from increased anticipated discrimination (Table 1.6).

Next, I estimate the effects of perceived discrimination on reservation wages, constructed to account for changes in beliefs about the likelihood of promotion. Any effect on this construction of reservation wages isolates the role of $C(\hat{\sigma}, D)$. Recall that the reservation wages described earlier (Table 1.4) were constructed from the multiple price list assuming that workers had the *average* worker's belief about the likelihood of promotion. Now, the high and low wages in the multiple price list are multiplied by workers' *own* belief about the probability that they would be assigned to each type of job.³⁶ This adjusts for any effect of perceived discrimination on beliefs about the probability of promotion, but average reservation wages in the demographic-blind manager arm

³⁶E.g. if a worker thinks there is a 20 percent chance of being promoted, and they switch from not wanting and wanting to be evaluated between the piece rates 30 and 35 cents per paragraph for the lower-paying job (and correspondingly 60 cents and 70 cents per paragraph for the higher-paying job), then their reservation wage would be calculated as $0.8 \times 32.5 + 0.2 \times 65 = 39$ cents per paragraph, whereas a worker with a higher expected probability of promotion but the same switching point would have a higher estimated reservation wage.

using both measures are the same and the treatment effects can be interpreted on the same scale. When future discrimination is possible, effects on reservation wages constructed in this way can come from $\frac{\partial C}{\partial \sigma}$ or from $\frac{\partial C}{\partial D}$; when future discrimination is not, the effects isolate $\frac{\partial C}{\partial D}$.

Being in the non-blind manager arm increases reservation wages for future evaluation by the same manager by 1.7 cents per paragraph (a 6.5 percent increase, $se=1$ cent, $p=0.10$) using this construction of reservation wages. This is 74 percent of the total effect on reservation wages (2.3 cents per paragraph) when beliefs about promotion can also play a role (Figure 1.3, top panel). When workers will be evaluated by the cutoff rule, treatment increases reservation wages by 0.9 cents per paragraph after accounting for beliefs (though this is statistically insignificant). Taking the point estimates at face value, this means that about half of the overall effect on $C(D, \sigma)$ is therefore due to disutility from anticipated discrimination or continued interactions, and the other half is due to disutility from continued work for the employer in general. Consistently, being in the non-blind manager arm reduces whether workers strongly agree that they are interested in future work for the employer by 7.9pp (12 percent, $se=2.7pp$), are interested in more jobs with tasks assigned like this one by 5.3pp (9 percent, $se=2.9pp$, $p=0.07$), and are interested in doing the harder job in the future by 6.3pp (12 percent, $se=2.9pp$) (Appendix Table A.8).

Altogether, these results suggest that the standard model cannot explain the effects of perceived discrimination on future labor supply. Thus, models of anticipated discrimination that focus on changing returns to effort may be missing an important, psychologically-driven element. As such, perceived and anticipated discrimination may have even more pernicious effects than previously considered because they have a direct utility cost. Even if workers face no monetary consequences, perceived or anticipated discrimination makes them worse off.

In fact, these costs are also the most likely explanation for the negative effects on retention discussed in the previous section. While proofreading, workers were not aware of the specifics of the future work opportunities or that they would possibly interact with their manager again, and they did not think that their manager would be involved in determining their bonus payment.³⁷

³⁷This was intentionally left ambiguous, but workers did not think this was the case. Towards the end of the survey, workers were asked the open-ended question, “*how do you think it is determined whether you did a “good job”*”

Thus, it is unlikely that perceived discrimination affected worker behavior through the channel of changes in expected wages, so the large negative effects of perceived discrimination on retention for women are most likely due to similar retaliation or avoidance mechanisms.

One psychological channel that could explain the effects of perceived discrimination on performance I find is by activating *stereotype threat* (e.g. [Spencer et al., 2016](#); [Liu et al., 2021](#)) for women or *stereotype challenge* (e.g. [Alter et al., 2010](#)) for men. This is not something the experiment was designed to isolate from other psychological channels but could be considered an underlying channel through which perceived discrimination affected effort costs and thus performance.

Together, the results from the promotion experiment show that perceived discrimination has quantitatively important effects on worker behavior and subsequent outcomes. These effects would exacerbate gender gaps conditional on race and racial gaps among women. Addressing the disparate impacts of discrimination therefore requires addressing worker perceptions alongside discrimination itself. In addition to these equity concerns, these results show that employers face an efficiency rationale to reduce perceived discrimination. They could improve retention, some workers' performance, and worker-manager cooperation by doing so. The next sections test how firms might achieve this goal.

1.4 The effects of anti-bias policies on perceptions

1.4.1 Estimation

The hiring experiment tests whether procedural changes that reduce the likelihood of discrimination—blinding managers to demographics and using unbiased algorithms—mitigate *perceptions* of dis-

[and thus are paid for] proofreading these paragraphs? Fewer than 1.5 percent of workers in either manager arm mentioned “manager” or “person/people reviewing” in their answer and fewer than 5 percent of workers in the algorithm arm mentioned “algorithm” or “AI.” The most common hypothesis was that there were pre-determined answers to compare to, and many workers also mentioned a cutoff rule in the number of correct answers.

crimination. I estimate the following:

$$Y_i = \alpha + \eta Alg_i + \theta_1 BA_i + \theta_2 NB_i + \chi_1 BA \times Alg_i + \chi_2 NB \times Alg_i + X' \delta + H' \lambda + \gamma_g + \varepsilon_{ig},$$

where Alg_i is an indicator for being randomly assigned to be evaluated by an algorithm rather than a manager, BA_i is an indicator for being randomly assigned to be evaluated by a demographic-blind decision-maker and to see previous hires' avatars, race, and gender, and NB_i is an indicator for being evaluated by a non-blind decision-maker and seeing previous hires' avatars, race, and gender. All other variables are the same as in the previous analysis (Section 1.3.1). Now, I focus on one outcome: Y_i is an indicator for whether workers perceive discrimination.

The effect of being evaluated by an algorithm is η when decision-makers are demographic-blind and workers do not know previous hires' race and gender, χ_1 (χ_2) when the decision-maker is demographic-blind (non-blind) and workers do know that previous hires were mostly white men. The effect of learning that a demographic-blind manager previously hired mostly white men is θ_1 and the effect of learning that a manager knew race and gender is $\theta_2 - \theta_1$. The effect of learning that a demographic-blind algorithm previously hired mostly white men is $\theta_1 + \chi_1$ and the effect of learning that an algorithm used race and gender is $\theta_2 - \theta_1 + \chi_2 - \chi_1$.

The hiring experiment specifically tests whether introducing unbiased evaluation procedures is sufficient to reduce perceptions of discrimination *when minority-group workers still do not see themselves represented among previously-promoted workers*. That is, the unbiased procedures have not led to any observable changes in decision-making—a realistic outcome when minority-groups are severely under-represented. In the hiring experiment, almost all workers saw exclusively white men being hired. In the main analysis, the above regression is estimated only on those workers (a random subset, as in the promotion experiment).³⁸

³⁸Results for the full sample are similar but conflate perceptions of procedures with minority-group representation. (Appendix Figure A.23). The procedures are differentially likely to hire entirely white men in the historical sample. This exacerbates differences in perceived discrimination between the algorithm and manager arms compared to the results in the main text (by depressing perceived discrimination in the manager arms) and reduces differences between the demographic-blind with avatars arm and non-blind manager arm (by depressing perceived discrimination by more in the non-blind manager arm than in the demographic-blind manager arm).

Like the promotion experiment, the hiring experiment facilitates the counterfactual evaluation of each worker by every procedure and the sample restricts to workers who would not have been hired regardless of their random assignment (2,527 workers of 2,680). Again, in practice, this does not seem to be a concern (Appendix Tables A.3 and A.4). Standard errors are robust to heteroskedasticity. The results do not depend on any of these specification decisions (Section 1.7).

1.4.2 Results

Changing hiring procedures in a way that reduces or eliminates the scope for human bias to affect decisions cannot eliminate perceptions of discrimination, *when workers see only white men were previously promoted*. Figure 1.4 plots rates of perceived discrimination in each treatment arm in the hiring experiment estimated using the main specification. No workers perceive discrimination in the demographic-blind manager or algorithm arms where workers do not see previously-promoted workers' demographics, so I focus on the other arms.

Workers perceive *more* discrimination by algorithms than managers. When decision-makers see workers' avatars, race, and gender and workers know all previous hires were white men, 52 percent of workers evaluated by a manager and 63 percent evaluated by an algorithm perceive discrimination, an 11pp difference (se=5pp). This difference is similar when decision-makers do not know demographics (a 9pp difference, se=4pp, Figure 1.4).

Blinding decision-makers to demographics reduces perceived discrimination, but not nearly to zero when workers know that all previous hires were white men. Specifically, blinding decision-makers reduces perceived discrimination by 11pp (21 percent, se=5pp) and 12pp (19 percent, se=3pp) when workers are evaluated by managers and algorithms, respectively, but 42 percent and 51 percent of workers still perceive discrimination. The differences are similar using the secondary measures of perceived discrimination (Appendix Figure A.13).³⁹

³⁹As in the promotion experiment, in each treatment arm, white and racial minority women perceive similar rates of discrimination and these rates are about twice as high as those for racial minority men (Appendix Figure A.25). Treatment effects are similar for white and racial minority women. Blinding managers to demographics is more effective at reducing perceived discrimination for racial minority men; but using a demographic-blind algorithm rather than managers when workers know that previous hires were all white men increases perceived discrimination

The high rates of perceived discrimination that persist when decision-makers are demographic-blind are due to workers not attending to or believing this information—in other words, they see previous decisions that “look like” discrimination, and convincing them otherwise is hard. In fact, while 50-60 percent of workers in the other four arms correctly identify the information that their manager or the algorithm knew when making hiring decisions, only 7 and 16 percent of workers do so when they are evaluated by a demographic-blind manager and algorithm, respectively, if they see previous hires’ avatars, race, and gender (Appendix Figure A.14). This is *not* due to (differential) inattention or general confusion (Section 1.2.4).

Instead, they are more likely to be incorrect precisely because they are highly likely to think the decision-maker knew their race and gender (Appendix Figure A.14). 95 percent of workers correctly identify that the non-blind manager and algorithm knew race and gender, but 87 and 70 percent of workers *incorrectly* think so when they are evaluated by a demographic-blind manager and algorithm, respectively but see that the previously-promoted workers are all white men. That said, I cannot separate disbelief or motivated reasoning from differential forgetfulness or confusion about this particular decision input.⁴⁰

At least some workers perceive algorithmic discrimination even if they know the algorithm only uses past performance and education as inputs. In the demographic-blind arms that did not see avatars, race, and gender for previous hires, 21 and 7 percent of workers thought that the manager and algorithm used demographic information, respectively. The difference is likely due to workers’ seeing *managers’* demographics in all manager treatment arms. Thus, there is no significant difference in the effect of seeing the previous hires’ avatars, race, and gender on whether workers think demographic-blind decision-makers used demographic information between the manager arms and the algorithm arms (66pp versus 64pp, Appendix Figure A.14). Recall that workers are 9pp (21 percent) *more* likely to perceive discrimination by an algorithm than a manager when both are demographic-blind. Together, the results suggests that workers are more likely to know that

by 5 times.

⁴⁰The information about decision-makers’ inputs and who was previously promoted were both communicated visually, so should have been similarly salient (Appendix Figure A.24).

the algorithm didn't use demographic information but perceive discrimination all the same. One explanation is that they understand that algorithms can discriminate without using demographics as inputs but managers cannot. In general, however, workers do not understand the implications of what they are told about the algorithm's design, whereas they do understand the manager's incentives (Appendix Figure A.15). So instead, general distrust or misunderstanding of the algorithm may make them more likely to jump to conclusions when they see what "looks like discrimination."

1.5 The effects of minority-group representation

1.5.1 Estimation

I estimate the effects of minority-group representation by comparing workers who saw that three white men were previously promoted to those who saw two white men and someone else. There is truthful, random variation in the demographic composition of the previously-promoted workers that each worker in the experimental sample saw (Appendix Figure A.4).

In the manager arms, workers were randomly assigned to managers who had made different choices in the historical sample. In the non-blind manager arm of the promotion experiment, 46 percent of workers were randomly assigned to managers who had previously promoted three white men and 54 percent were assigned to managers who had promoted two white men and someone else. Workers who see that two versus three white men were previously promoted have similar observable characteristics (Appendix Table A.9).⁴¹ In the demographic-blind arm, this was 40 percent and 60 percent, respectively, but note that these two groups of workers did not actually see anything different about the previously-promoted workers.

In the algorithm arm, workers were randomly assigned to groups of historical workers who had been jointly evaluated by the algorithm; in some groups three white men were promoted and in others it was two white men and someone else. Recall that workers in the algorithm arm were

⁴¹One concern is that the fraction of white men among previously-promoted workers could co-vary with other manager characteristics, but this is not reflected in the data and there was minimal heterogeneity in manager profiles (Appendix Table A.14).

cross-randomized to either learn the demographics of previously-promoted workers or not. When workers learned previously-promoted workers' demographic information, 33 percent saw three white men and 67 percent saw two white men and someone else. Workers who see that two versus three white men were previously promoted have similar observable characteristics (Appendix Table A.9). I estimate the following:

$$Y_i = \alpha + \delta_1 3WM_i + \gamma_1 T_i^M + \beta_1 T^M \times 3WM_i + \theta_1 A_i + \delta_2 A \times 3WM_i + \gamma_2 T_i^A + \beta_2 T^A \times 3WM_i + X' \delta + H' \lambda + \gamma_g + \varepsilon_{ig},$$

where $3WM_i$ is an indicator for seeing three white men previously promoted, T_i^M is an indicator for being evaluated by a non-blind manager, A_i is an indicator for being evaluated by the demographic-blind algorithm and not learning previously-promoted workers' demographics, and T_i^A is an indicator for being evaluated by the demographic-blind algorithm and learning previously-promoted workers' demographics. All other variables are the same as previously (see Section 1.3.1), and Y_i indicates perceived discrimination.

The effect of seeing that only white men were previously promoted is β_1 when workers were evaluated by a non-blind manager and β_2 when workers were evaluated by a demographic-blind algorithm. These are the primary coefficients of interest. δ_1 and δ_2 should both be zero, as perceived discrimination rates among workers who did not see previously-promoted workers' demographics should not depend on the demographics of those workers.

As in all previous analysis, the sample is restricted to workers who would not have been promoted under any procedure, and standard errors are robust to heteroskedasticity. Note that this analysis was not pre-registered (see Section 1.7).

1.5.2 Results

More representation of women and racial minority men among previously-promoted workers substantially reduces perceived discrimination. Seeing one minority-group member among the three

previously-promoted workers reduces rates of perceived discrimination by 17pp (40 percent, $se=3pp$) in the non-blind manager arm and by 15pp (50 percent, $se=5pp$) in the algorithm arm when workers see previously-promoted workers' demographics (Figure 1.5).

Whether the previously-promoted minority-group worker is of the worker's same demographic group also matters. I add indicators for seeing one previously-promoted worker from a workers' own demographic group and the corresponding interactions to the above regression. In the non-blind manager arm, seeing one previously-promoted minority-group worker from a *different* demographic group as oneself reduces perceived discrimination from 44 to 34 percent relative to seeing three white men (a 10pp difference, $se=4pp$), and seeing one minority-group member from one's *own group* reduces perceived discrimination by an *additional* 18pp ($se=5pp$). The pattern is the same in the algorithm arm (Appendix Figure A.16).

Thus, worker beliefs about discrimination are much more responsive to signals from decision outputs—who was selected—than they are to signals about inputs—information about how the decision was made. When a decision “looks like discrimination,” workers pay little attention to decision inputs—a potential application of the representativeness heuristic affecting belief formation (e.g. Kahneman and Tversky, 1972; Bordalo et al., 2021). When previously-promoted workers include members of minority groups, especially one's own group, workers are much less likely to perceive discrimination even when there could have been discrimination.

Taken together, these results have several implications for settings where some demographic groups are severely under-represented. First, anti-bias policies—like transitioning to algorithmic screening or censoring demographic identifiers from resumes—may slightly reduce perceptions of discrimination, but workers may still be likely to perceive that discrimination played a role if there is no change in the demographics of workers who are hired. On the other hand, if they improve minority-group representation, they may be more effective at reducing perceived discrimination. These results also suggest that settings with more equal representation may have less perceived discrimination. Future study of other policies concerned with minority-group representation—e.g. affirmative action—should consider the channel of perceived discrimination and perceptions of the

policy more broadly.

1.6 The effects of perceived algorithmic discrimination

Since implementing algorithmic hiring procedures cannot fully eliminate perceived discrimination (Section 1.4), even if it improves minority-group representation (Section 1.5), a final related question is whether algorithmic hiring procedures can mitigate the *effects* of perceived discrimination. People are less morally outraged by depictions of algorithmic bias than human bias and are less likely to attribute algorithmic bias to prejudice when it does occur (Bigman et al., 2023); if the effects documented in Section 1.3 are due to perceived *prejudice* rather than perceived *discrimination*, they may not arise when workers perceive algorithmic discrimination.

1.6.1 Estimation

To understand the effects of perceived *algorithmic* discrimination, I focus on intent-to-treat (ITT) specifications that are analogous to those in Section 1.3.1. Restricting to workers randomly assigned to be evaluated by the algorithm in the promotion experiment, I now compare outcomes between workers who were cross-randomized to see previously-promoted workers' race, gender, and avatar versus those who did not. Now, T_i in Equation 1.1 indicates seeing the demographics of previously-promoted workers; β is the effect of learning that the demographic-blind algorithm previously promoted mostly white men.

1.6.2 Manipulation check: Treatment affects perceptions

Again, worker perceptions depend on whether they learn that the (demographic-blind) algorithm previously promoted mostly white men. No workers perceive discrimination when they do not see the avatars, race, and gender of previously-promoted workers and are evaluated by a demographic-

blind algorithm.⁴² Learning that the algorithm previously promoted mostly white men increases perceived discrimination rates from 0 to 17 percent (se=2.6pp), even though workers know the algorithm did not access demographic information (Figure 1.6). The effect is 22pp for women (se=3.4pp) and 8.5pp (se=5pp) for racial minority men.⁴³

1.6.3 Results: Effects of perceived algorithmic discrimination

The effects of perceived algorithmic discrimination seem to be negative, but differ in some ways from the effects of perceived manager discrimination.

Retention. Learning that the demographic-blind algorithm previously promoted mostly white men does not statistically significantly affect retention. The point estimates across all three outcomes (number of paragraphs proofread, proofreading more than the required six paragraphs, and finishing all eighteen paragraphs) are positive, but the confidence intervals are wide and the ITT effect of the algorithm sub-treatment cannot be statistically distinguished from the negative ITT effect of being in the non-blind manager arm on retention (Figure 1.7, Row 1).

Performance and earnings. Learning that the demographic-blind algorithm previously promoted mostly white men negatively affects performance in the first six required paragraphs (the number of words correctly highlighted and bonuses earned), and these effects are statistically significantly different from the ITT effect of being in the non-blind manager arm relative to the demographic-blind manager arm (Figure 1.7, Row 3). Learning that the demographic-blind algorithm previously promoted mostly white men lowers performance by about 5 percent.

Future labor supply. Learning that the demographic-blind algorithm previously promoted mostly white men has no significant effect on reservation wages or willingness to pay for the unbiased cutoff, but these are not statistically distinguishable from the positive effects in the manager

⁴²Given high-profile articles about algorithmic discrimination (e.g. ‘Amazon scraps secret AI recruiting tool that showed bias against women’, Reuters), this was not obvious *ex ante*. It was possible that individuals would infer bias just from the word “algorithm.”

⁴³Results are similar using the secondary measures of perceived discrimination (Appendix Table A.15). In turn, it reduces the share who cite the primary reason for their non-promotion being their education by 7pp (21 percent, se=3.4pp, Appendix Table A.16) while it increases the share who cite discrimination as the *primary* reason by 9pp (se=2pp, Appendix Table A.15).

arms (Figure 1.7, Row 4). These smaller, less distinguishable effects may be due to the fact that there is no effect of learning that the demographic-blind algorithm previously promoted white men on beliefs about future promotion or anticipated discrimination (Figure 1.7, Row 5). This did not, however, replicate in the hiring experiment, where these beliefs and reservation wages were the only outcomes collected that related to worker behavior. Appendix A.8 describes these results, which were pre-registered as a replication exercise. In the hiring experiment, perceived algorithmic discrimination did significantly increase anticipated discrimination in the future and reservation wages for future work (as did perceived manager discrimination). Thus, the anticipated psychological costs of algorithmic discrimination and manager discrimination may be similar.

In summary, the use of algorithmic hiring procedures changes the effects of perceived discrimination but the effects are still generally negative. These results are relatively imprecisely estimated (the sample is half the size, and random assignment generates less of a difference in perceived discrimination rates). Taken at face value, they are consistent with the idea that workers react differently to perceived *prejudice* than they do to perceived *discrimination* (Bigman et al., 2023), though many other things are different about perceived human versus algorithmic discrimination that may drive these differences. Other potential explanations are differences in *who* perceives discrimination and heterogeneous treatment effects, the fact that algorithms cannot themselves “benefit” from a worker’s performance, or the fact that an algorithm cannot be “proven wrong.”

1.7 Robustness

The results are broadly quite robust to accounting for standard concerns with experiments in general and survey experiments in particular. This section provides a broad overview of these analyses; Appendix A.9 provides the details and results.

Multiple hypothesis testing and randomization inference. The statistical significance of my main results does not change when I (1) account for multiple hypothesis testing using a Romano-Wolf (2005; 2016) correction or (2) when I relax the normality assumptions associated with asymp-

otic standard errors and instead use treatment randomization to estimate the distribution of treatment effects under a null hypothesis of no treatment effect (Fisher, 1935; Rosenbaum, 2002).

Attrition and attention. There is no differential attrition in either experiment that can explain my results for the effects of perceived manager discrimination on worker behavior, the effects of anti-bias hiring procedures or minority-group representation on perceived discrimination rates, or the effects of perceived algorithmic discrimination. There is one difference in the rate at which workers failed the attention check (between workers who saw three white men versus two white men previously promoted in the demographic-blind algorithm arm of the promoted experiment); the specification charts described next show that the corresponding effect on perceived discrimination is robust to dropping workers who failed the attention check. Section 1.2.4 discusses overall experimental fidelity and shows that worker attention and comprehension was high overall.

Specification choices. The main results largely do not depend on the choice of control variables or sample definition. The effects of perceived manager discrimination and the effects of anti-bias hiring procedures and minority-group representation on perceived discrimination rates are unchanged if I further restrict the sample to workers who passed the attention check near the end of the survey or expand the sample from workers who would not have been promoted or hired under any procedure to the full sample. The only non-robust result is that the effects of perceived algorithmic discrimination on performance are less negative and are statistically insignificant when restricting to those who passed the attention check.

All four sets of results are very stable regardless of the control variables used, consistent with effective randomization and the lack of observable differences between the treatment arms. The results are robust to replacing the specifications with the exact pre-registered specification along two dimensions, using lasso-selected control variables (Chernozhukov et al., 2018), or dropping all control variables.

Experimenter demand effects. The experimental surveys asked workers to indicate what they thought the survey was about at multiple points throughout the survey (timing was randomly-determined) in an open-ended text response. I use coded measures of whether they think the study

is about discrimination (Appendix A.5) to understand the extent to which experimenter demand effects (differential knowledge of the study purpose and resultant changes in worker behavior) could explain my main results. While there were differences across treatment arms in the fraction of workers who believed the study was about discrimination, the rates at which workers did so remained low until after the more explicit measures of perceived discrimination were elicited at the very end of the survey. To test whether these differences affected worker behavior in a way that explains my treatment effects, I bound my results by replacing outcome variables with higher or lower values for workers who believed the topic of the study was discrimination. I replace the worker behavior outcomes with plus or minus 0.2sd of the variable in the control group (following the results in de Quidt et al., 2018), and I replace the perceived discrimination outcome with 0 or 1. I use the more conservative bound for the latter since this outcome was not incentivized.

The effects of perceived discrimination on worker behavior are unchanged by this exercise.⁴⁴ The effects of blinding hiring procedures to demographics and the effect of minority-group representation on perceived discrimination all fall by about half when replacing the perceived discrimination measure with zero for workers who believe the study is about discrimination, but remain statistically significant.

Multiple price list elicitations. The four measures of future labor supply were based on multiple price list-style elicitations of reservation wages and willingness to pay in terms of foregone wages (Appendix A.4 shows exact implementations). These were incentivized following the literature by telling workers I would randomly choose one row of the list and implement randomly-chosen workers' choices for that row, though the incentives were relatively low-powered as workers could report low reservation wages and then choose to not take the job if offered later. However, only 10-15 percent of workers answered "I would want the job at this wage" for all wages and the results are robust to dropping these workers.

In the main results, I follow the literature by dropping any worker who does not respond to the MPL questions in a sensible way; Appendix A.9 discusses how these workers compare to those

⁴⁴This is consistent with Mummolo and Peterson's (2019) result that workers do not change their behavior when they are told the purpose of a study.

who did answer sensibly. The main results are robust to dropping these workers from all analyses.

Deviations from Pre-Analysis Plan. I posted detailed pre-analysis plans before each experiment, and follow them closely. Here, I outline the main deviations from the pre-analysis plans; again, see Appendix A.9 for details and smaller deviations.

The primary deviation from the pre-analysis plan was to exploit random variation in the fraction of previously-promoted workers that were white men, which was built into the original experimental design by randomly assigning workers to managers and jointly-evaluated groups of historical workers. I did not anticipate the amount of variation in the composition of previously-promoted workers. Given that (i) there was substantial variation and (ii) the effect of minority-group representation on perceived discrimination is an important comparison to the effects of procedural changes, I include these results prominently in the paper.

There were also two deviations related to recruitment. First, in the promotion experiment I planned to more substantially over-sample racial minorities. I was unable to recruit the pre-registered number of racial minorities on Prolific. That said, I am still able to test for racial and gender heterogeneity and obtain moderately precise results. Second, in both experiments, the sample size of workers offered the experiment is 91-93 percent of pre-registered sample size. This is due to difficulty in recruiting managers and in bringing them back to participate in the second and third rounds of worker evaluations.

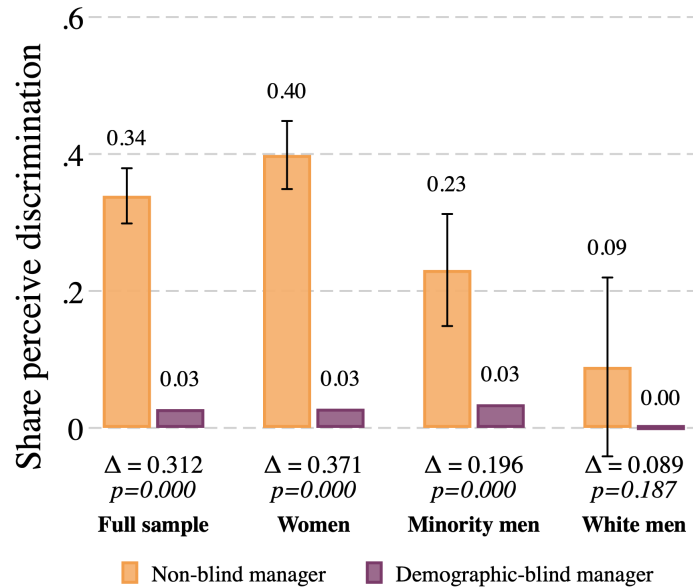
1.8 Conclusion

This paper shows that addressing the disparate impacts of discrimination will require addressing worker perceptions in addition to discrimination itself. Perceived discrimination negatively affects retention and future labor supply, which would exacerbate gender and racial gaps caused by discrimination itself. Workers are also willing to pay to avoid discrimination—beyond just its expected economic consequences—implying that a discriminatory labor market has disparate psychological costs as well as material ones.

While employers stand to benefit from higher retention and easier or cheaper recruitment by reducing perceptions of discrimination, I show that it is going to be difficult to do so purely by implementing anti-bias procedures in contexts where minority-group workers are severely under-represented. Reducing biased decision-making itself is of course the right place to start. But just changing procedures so that they have less potential for bias—in this case, blinding decision-makers to race and gender or using unbiased algorithms—at best only mildly reduces perceived discrimination if the outcomes of those procedures—who is hired or promoted—remain the same and members of under-represented groups continue to see decisions that “look like discrimination.” Increased diversity may be a more promising tool, and perceptions of discrimination after affirmative action or other diversity initiatives warrant additional future study. Implicit bias trainings seem not to be very effective at reducing bias (Chang et al., 2019), but an effect on perceptions could help explain their popularity.

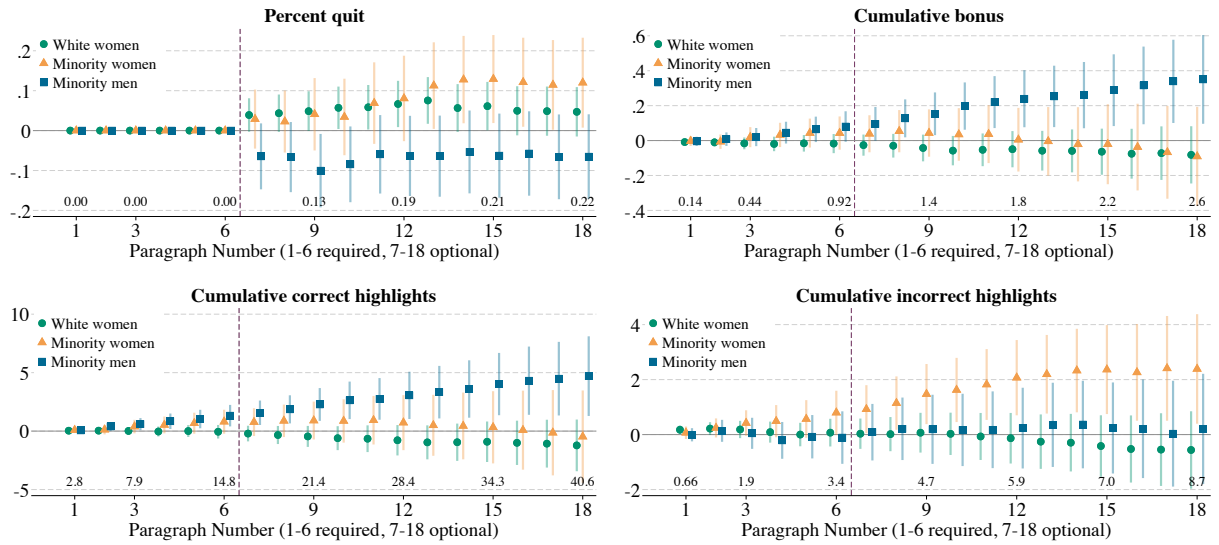
These results offer other interesting avenues for future research on perceptions of discrimination as well. I studied particular anti-bias tools and communicated with workers about them in a particular way, but this leaves open the question of whether workers can be encouraged to trust information about de-biased decision-making—unbiased algorithms in particular. On the other hand, my results suggest that beliefs about discrimination may be very difficult to change. If that is the case, it will be important to study whether firms can mitigate the negative effects of perceived discrimination by encouraging workers to report such concerns and taking them seriously, as well as whether there are psychological or behavioral interventions that might help workers overcome the effects of perceived discrimination that later put them at a disadvantage.

Figure 1.1: Perceived discrimination in the manager arms of the promotion experiment



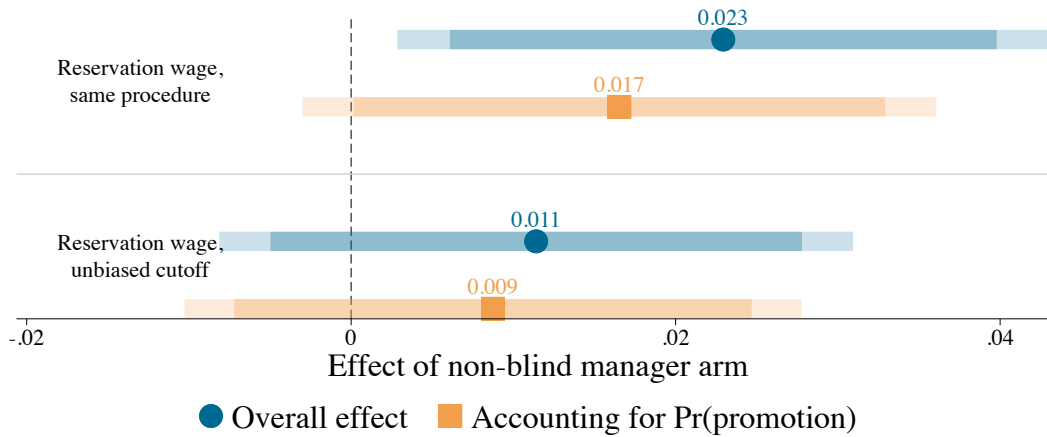
Note: This figure plots the share of workers in the experimental sample of the promotion experiment who perceive discrimination using the main measure (whether a worker mentions demographics in their response to the open-ended question, “*what needed to be different about your profile in order to be assigned to the harder task?*”) first for all workers and then separately for women (of all races), racial minority men, and white men in the two manager treatment arms. The plotted means are estimated from the coefficients estimated by Equation 1.1 separately for each subgroup. The figure restricts to workers who would not have been promoted under any promotion procedure. For breakdowns by race \times gender, see Appendix Figure A.10. 95 percent confidence intervals calculated with standard errors robust to heteroskedasticity are indicated by the black bars. The corresponding effects on the secondary measures of perceived discrimination are in Appendix Table A.5.

Figure 1.2: Treatment effects on retention and performance by gender \times race



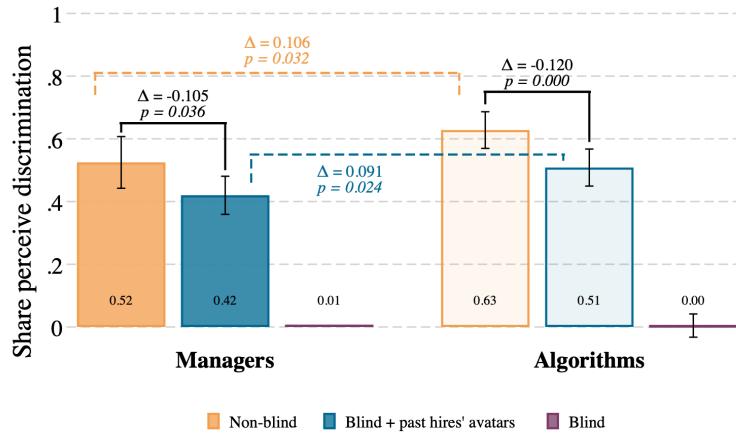
Note: This figure plots the effect of being in the non-blind manager arm of the promotion experiment on retention and performance measures, each cumulatively by the given paragraph. Equation 1.1 is estimated interacting the treatment indicator with indicators for each race/gender group; white men are dropped. 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure 1.3: Decomposing the effects on reservation wages in the promotion experiment



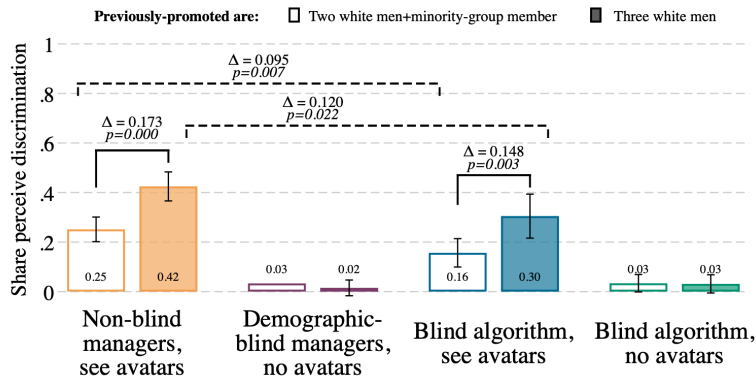
Note: This figure plots the effect of being in the non-blind manager arm relative to the demographic-blind manager arm of the promotion experiment on workers' reservation wages when they will be evaluated by the same manager as in the experiment (top panel) or by an unbiased cutoff rule in baseline quiz scores (bottom panel). Effects are plotted first on the overall measure of reservation wages, then on the measure of reservation wages that accounts for effects on beliefs about the likelihood of promotion, isolating the extent to which the effect is due to anticipated disutility. The sample and controls are the same as in Table 1.2. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure 1.4: Perceived discrimination in the hiring experiment



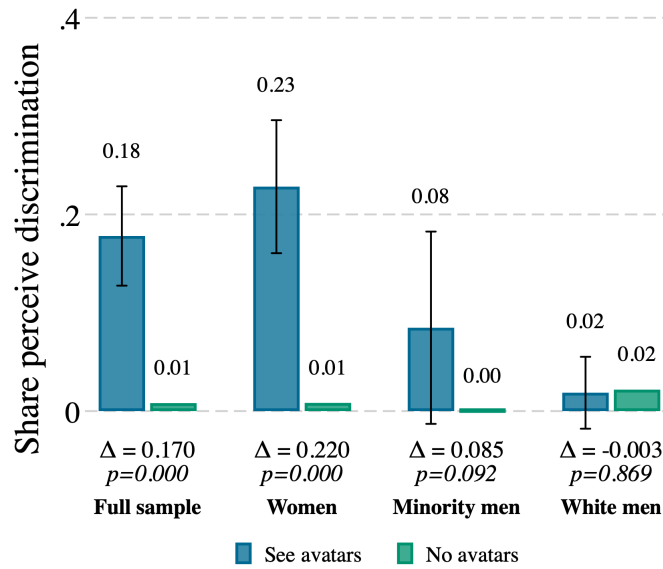
Note: This figure plots the share of workers perceiving discrimination in each treatment arm of the hiring experiment, using the main measure of perceived discrimination. The sample is restricted to workers who would not have been hired under any hiring procedure and workers who see three previous hires who are all white men (a random subsample within each treatment arm). Shares are calculated via regressions with the same controls as Table 1.2. *p*-values and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure 1.5: Effects of seeing one previously-promoted minority-group worker



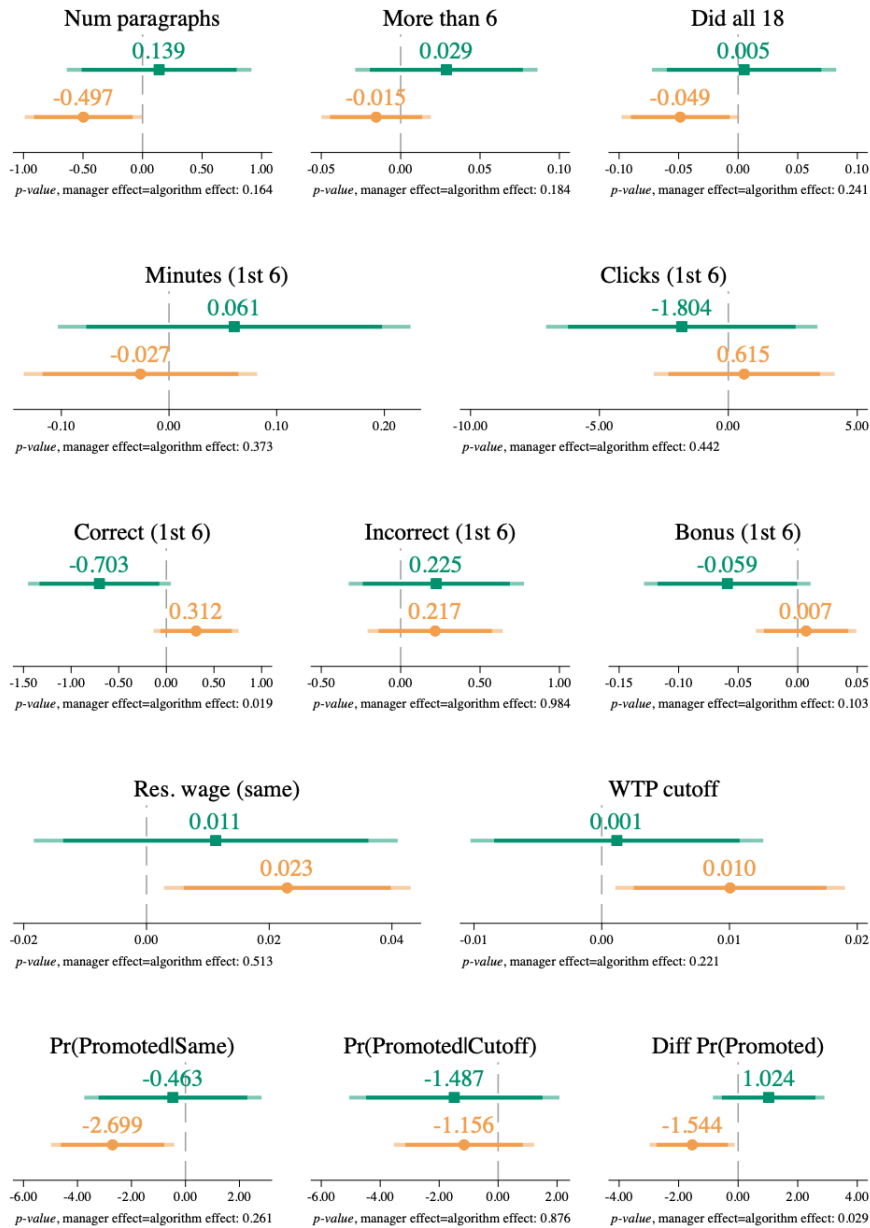
Note: This figure plots the share in each treatment arm of the promotion experiment perceiving discrimination by the main measure, separately by whether workers (would) see three white men previously promoted (40 percent of each arm) or not. Those who do not see three white men see two white men and one member of a minority group. Workers in the demographic-blind manager arm and those in the algorithm arm who do not see worker avatars are split by whether they *would* be shown three white men or not. The sample and controls are the same as Table 1.2. *p*-values and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure 1.6: Perceived discrimination in the algorithm arm of the promotion experiment



Note: This figure is analogous to Figure 1.1 but compares perceived discrimination rates in the two sub-arms of the algorithm treatment arm. The corresponding effects on the secondary measures of perceived discrimination are in Appendix Table A.15.

Figure 1.7: Effects of perceived algorithmic discrimination on worker behavior



- Effect of see avatars in demographic-blind algorithm arm
- Effect of non-blind manager + see avatars

Note: This figure plots the treatment effects of learning that previously-promoted workers were mostly white men in the algorithm arm of the promotion experiment, and compares this effect to the effect of learning that one's manager knew demographic information and previously promoted mostly white men. The outcomes and control variables are the same as in Tables 1.2, 1.3, 1.4, and 1.6. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Table 1.1: Summary statistics and comparison to ACS sample

	Experimental Samples		ACS 2021, ages 18-71	
	Promotion	Hiring	All	Subsample
	(1)	(2)	(3)	(4)
Male	0.29	0.13	0.50	0.31
Race:				
Asian	0.13	0.06	0.06	0.09
Black	0.14	0.12	0.12	0.17
White, Hispanic	0.14	0.10	0.18	0.27
White, Non-Hispanic	0.59	0.71	0.59	0.40
Married	0.54	0.58	0.50	0.48
Kids	0.43	0.46	0.39	0.44
Education:				
Less than high school	0.01	0.01	0.10	0.09
High school graduate	0.13	0.10	0.27	0.24
Some college but no degree	0.27	0.25	0.21	0.21
2 year college degree	0.09	0.10	0.09	0.09
4 year college degree	0.37	0.39	0.21	0.23
Professional or Masters degree	0.11	0.14	0.11	0.13
Doctorate	0.01	0.02	0.01	0.02
Income:				
Less than \$20,000	0.14	0.11	0.12	0.09
\$20,000-\$40,000	0.22	0.18	0.15	0.15
\$40,000-\$60,000	0.18	0.19	0.14	0.14
\$60,000-\$80,000	0.15	0.17	0.12	0.13
\$80,000-\$100,000	0.10	0.13	0.10	0.11
\$100,000-\$120,000	0.07	0.07	0.08	0.09
\$120,000-\$140,000	0.04	0.06	0.06	0.07
\$140,000-\$160,000	0.04	0.04	0.05	0.05
More than \$160,000	0.06	0.07	0.18	0.18
Age:				
18-24	0.17	0.15	0.13	0.13
25-34	0.32	0.33	0.20	0.24
35-44	0.22	0.22	0.19	0.23
45-54	0.14	0.17	0.18	0.20
55-64	0.11	0.10	0.19	0.16
65 or older	0.05	0.04	0.11	0.04
Employment:				
Currently employed outside Prolific	0.67	0.69	–	–
Ever employed outside Prolific	0.96	0.97	–	–
In the labor force	–	–	0.72	1.00
Employed	–	–	0.67	0.93
Employed if in labor force	–	–	0.94	0.93
N	2080	2527	2216940	1005688

Note: The samples in columns 1 and 2 are the analysis samples in the promotion and hiring experiments, respectively, each combining all treatment arms. The sample in column 3 is respondents age 18-71 in the 2021 ACS (the minimum and 99th percentile age in experimental sample, respectively). Column 4 further restricts to women and racial minority men currently in the labor force. All variables are indicators, and means are presented for each variable. ACS estimates weight by the *perwt* variable from IPUMS; experimental sample estimates are unweighted.

Table 1.2: Effects on retention

	Num paragraphs	More than 6	Did all 18
	(1)	(2)	(3)
Non-blind manager, see avatars	-0.497** (0.252)	-0.015 (0.018)	-0.049* (0.025)
N	1387	1387	1387
Control mean	15.916	0.915	0.776

Note: This table reports treatment effects of being in the non-blind manager arm in the promotion experiment on retention (Equation 1.1). The first outcome is the number of paragraphs completed; workers had to do at least six to receive their participation payment and were given the option to quit after each subsequent paragraph. They could proofread up to eighteen paragraphs. The second and third outcomes are indicators for doing more than six paragraphs and finishing all eighteen, respectively. The sample is those evaluated by a manager and the presented coefficient is on the indicator for being in the non-blind manager arm rather than the demographic-blind manager arm, restricting to workers who would not have been promoted under any evaluation procedure in the promotion experiment. All regressions control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table 1.3: Effects on effort and performance in the first six (required) paragraphs

	Minutes	Clicks	Correct	Incorrect	Bonus
	(1)	(2)	(3)	(4)	(5)
Non-blind manager, see avatars	-0.027 (0.055)	0.615 (1.787)	0.312 (0.227)	0.217 (0.217)	0.007 (0.022)
N	1389	1389	1389	1389	1389
Control mean	4.957	36.804	14.817	3.422	0.919

Note: This table reports treatment effects of being in the non-blind manager arm in the promotion experiment on effort and performance. The effort outcomes in columns 1 and 2 are the minutes spent and number of clicks on the first six required paragraphs. The performance outcomes in columns 3, 4, and 5 are the number of correct highlights, number of incorrect highlights, and bonus earned in dollars in the first six paragraphs. Samples and specifications are the same as Table 1.2. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table 1.4: Effects on future labor supply

	Res. wage (same procedure)	WTP unbiased procedure
	(1)	(2)
Non-blind manager, see avatars	0.023** (0.010)	0.010** (0.005)
N	1338	1325
Control mean	0.265	-0.014

Note: This table reports treatment effects of being in the non-blind manager arm in the promotion experiment on measures of future labor supply. The first outcome is the reservation wage (per well-proofread paragraph) to be evaluated again under the same procedure as in the experiment. The second is the reservation wage from column 1 minus the reservation wage when an unbiased cutoff rule is used to decide future promotions. Samples and specifications are the same as Table 1.2. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table 1.5: Effects on sharing with and avoidance of manager

	Cooperative task RW	WTP new manager	Shared zero	Amt bonus shared
	(1)	(2)	(3)	(4)
Non-blind manager, see avatars	0.008 (0.015)	0.041** (0.020)	0.049* (0.029)	-0.255 (0.233)
N	1349	1030	1383	1383
Control mean	0.341	0.223	0.522	2.857

Note: This table reports treatment effects of being in the non-blind manager arm in the promotion experiment on measures of cooperation with and generosity towards managers. The first and second outcomes are workers' reservation wage per paragraph to do a future collaborative task with their manager and their willingness to pay (as a reduction in a \$1 wage per paragraph) to be able to choose their own manager for the collaborative task. The third outcome is an indicator for sharing none of the 20-dollar surprise bonus with their manager and the fourth is the amount of the bonus that they share in dollars. Samples and specifications are the same as in Table 1.2. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table 1.6: Effects on beliefs about future promotion

	Pr(Promoted Same)	Pr(Promoted Cutoff)	Diff Pr(Promoted)
	(1)	(2)	(3)
Non-blind manager, see avatars	-2.699** (1.163)	-1.156 (1.212)	-1.544** (0.728)
N	1385	1385	1385
Control mean	47.219	48.491	-1.272

Note: This table reports treatment effects of being in the non-blind manager arm in the promotion experiment on beliefs about the likelihood of promotion in the future. In the first two columns, the outcomes are the stated probability of being promoted in a future round of the study where evaluation is under the same procedure as in the experiment and using a cutoff rule in baseline quiz scores, respectively. The third outcome is the difference between the two. Samples and specifications are the same as Table 1.2. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Chapter 2

Childhood Confidence, Schooling, and the Labor Market: Evidence from the PSID

With Lucy Page

Abstract

We link over- and under-confidence in math at ages 8-11 to education and employment outcomes 22 years later among the children of PSID households. About twenty percent of children have markedly biased beliefs about their math ability, and beliefs are strongly gendered. Conditional on measured ability, childhood over- and under-confidence predict adolescent test scores, high school and college graduation, majoring or working in STEM, earnings, and unemployment. Across all metrics, higher confidence predicts better outcomes. These biased beliefs persist into adulthood and could continue to affect outcomes as respondents age, since intermediate outcomes do not fully explain these long-run correlations.

We are grateful to David Autor, Esther Dufo, Amy Finkelstein, Kartini Shastry, Rohini Pande, Frank Schilbach, and four anonymous referees and the editor for thoughtful and helpful comments. Both authors are supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1745302. This project is also supported by the George and Obie Shultz Fund at MIT. The majority of data used in this paper is publicly available on the website of the Panel Study of Income Dynamics: <https://psidonline.isr.umich.edu/>. Several control and outcome variables used in the analysis rely on the restricted PSID dataset, which can be obtained via the process described at <https://simba.isr.umich.edu/restricted/ProcessReq.aspx>.

2.1 Introduction

Long-standing research in psychology finds that people have biased beliefs about their abilities in a range of domains.¹ Prior research has focused on “optimism bias,” or over-confidence about one’s performance, belief accuracy, or future outcomes (Moore and Healy, 2008; Sharot et al., 2011; Taylor and Brown, 1988). In contrast, psychologists also document “imposter syndrome,” a form of systematic under-confidence in which people attribute their successes to luck or effort rather than skill (Langford and Clance, 1993; Sakulku, 2011). Recent lab-based work in behavioral economics has sought to microfound this empirical evidence of biased beliefs by documenting that people systematically under-weight or over-weight signals about the truth, especially in ego-relevant domains like intelligence and beauty (see Benjamin (2019) for a review).

Do these confidence gaps matter for economic decision-making in the real world? There are key reasons to expect that they might. For example, if adolescents or young adults perceive ability and educational investment to be complements, under-confident students might exert less effort in school or end their education earlier (Bénabou and Tirole, 2002). Later, under-confident adults may be less likely to complete costly and uncertain job applications, or may select away from jobs with higher returns to performance (Dohmen and Falk, 2011).² Individuals’ beliefs about their own ability could also affect outcomes by shaping how others perceive them. If parents or teachers mistake confidence for aptitude and expect the returns of education to increase with ability, they may invest more in more confident children (Papageorge et al., 2018; Dizon-Ross, 2019). More confident applicants may appear more capable during job interviews, improving their employment prospects (Mobius and Rosenblat, 2006; Schwardmann and van der Weele, 2019).

As yet, there is limited evidence for how confidence affects economic outcomes in realistic settings and over the long term. In addition to the lab-based work on the short-term implications of

¹We refer throughout the paper to *ability* and *beliefs about ability*, but we do not mean to imply that ability or beliefs are innate or fixed. Rather, we are referring to someone’s ability or perceived ability to perform well in a certain domain or task at a particular time.

²Psychological theories of motivation, including Bandura’s (1986) Social Cognitive Theory or Expectancy-Value Theory (see Wigfield and Eccles (2000)) also emphasize that individuals increase effort in domains in which they feel competent.

confidence gaps cited above (Mobius and Rosenblat, 2006; Dohmen and Falk, 2011; Schwarzmann and van der Weele, 2019), a small parallel literature in economics and sociology examines longer-term outcomes and finds that those with higher self-esteem get more education, are more likely to be employed, and earn higher wages (Murnane et al. 2001; Waddell 2006; Drago 2011; de Araujo and Lagos, 2013). However, this literature has struggled to demonstrate that these associations are not driven by omitted variables like unobserved ability. These papers typically control for IQ in an attempt to account for cognitive ability, but it is not feasible to control for subjects’ “ability” across all domains that affect generalized self-esteem.

In this paper, we address the limitations of both prior literatures by examining the real-world and long-term implications of a dimension of confidence in which we can observe and control for demonstrated ability: childhood over- and under-confidence in math.³ We use unique data from the Panel Study of Income Dynamics (PSID) to identify biased beliefs in math in a sample of 2,985 children in core PSID households; we then relate their childhood over- and under-confidence to educational and employment outcomes up to 22 years later, controlling for test scores, general confidence, and other key confounders.

The PSID is an ideal setting in which to examine long-term links with childhood confidence. Our sample is based on child-focused PSID supplements that measure children’s performance on a standardized math test and their own reports of how “good” they are at math. We combine these measures to identify over-confident children as those who scored poorly on the math assessment and yet said they were good at math, and to identify under-confident children as those who scored well but said they were bad at math. The structure of the PSID also allows us to observe much of respondents’ young adulthood: the child supplements and core survey followed our sample from 1997 through 2019, so we observe our oldest respondents from age 12 into their thirties.

Biased beliefs about math ability are prevalent in our sample: 5-20 percent of children are markedly over-confident and 7-16 percent are markedly under-confident (using several definitions

³We report all of the following analysis for parallel measures of reading over- and under-confidence in Appendix Tables B.2-B.5. We discuss our focus on math confidence in Section 2.2.

of biased beliefs, described in more detail below).⁴ Over- and under-confidence in math are highly gendered: girls are 2.3 percentage points (pp) (17 percent) more likely to be under-confident and 2.7 pp (27 percent) less likely to be over-confident in math than boys. In contrast, girls are 30 percent *less* likely to be under-confident in reading than boys. This pattern is consistent with evidence that adults are more likely to be over-confident in stereotypically gender-congruent domains (Coffman, 2014; Coffman et al., 2019; Bordalo et al., 2019; Shastry et al., 2020).

One key concern with our measures of over- and under-confidence is that they may just capture children’s private information about their own ability, driven by measurement error in the cognitive tests. We have several key pieces of evidence against this concern. First, the math assessment that we use has high test-retest reliability (Hicks and Bolen, 1996). Second, over- and under-confidence persist between waves of the child survey among the 60 percent of our sample with multiple measurements, so our measures seem to capture a stable psychological trait. Next, as we’ve noted, our measures show gender variation that is consistent with prior work on gendered patterns in belief updating, and which we would not expect to see in random testing error. Finally, our results largely persist when we use alternate measures of childhood over- and under-confidence that are less vulnerable to measurement error; we calculate these measures based on test scores and self-reported ability averaged over two waves of the PSID child supplement.⁵

Our main analysis is simple: we estimate the associations between biased beliefs about one’s math ability in childhood and later educational and employment outcomes, controlling for childhood math and reading score deciles, working memory, general confidence, and a host of information on respondents’ demographics and family backgrounds.

Children’s biased beliefs in math strongly predict many of their medium- and long-term educational and employment outcomes. First, confidence has large associations with educational achievement: over-confident children score higher than others with comparable prior scores on math assessments five years later, while under-confident children score lower. Biased beliefs in

⁴Using weights that adjust our sample to be nationally representative, these ranges are 6-30 percent and 6-15 percent, respectively.

⁵While these four pieces of evidence strongly suggest that our measures of over- and under-confidence capture more than *random* measurement error on the cognitive test, they do not negate the possibility that children have private information on a form of math ability that the test systematically excludes. We discuss this possibility in detail in Section 2.3.3.

math also predict educational attainment: over-confident children are more likely to graduate from high school and under-confident children are less likely to graduate from college than others with comparable childhood scores. Under-confident children are also less likely to major in STEM during college and attend less selective colleges, though the latter result is imprecisely estimated. Finally, childhood math confidence predicts key employment outcomes at ages 26 and up. Under-confident children are less likely to work in STEM occupations as adults, and we find suggestive evidence that more confident children earn more and are less likely to be unemployed.

While we do not claim that these associations are causal, we do show that they are robust to several key potential confounders. First, children may form inaccurate beliefs about their ability in part because of how their parents or teachers perceive them, and these adult beliefs may themselves affect children’s later success (Papageorge et al., 2018; Jussim and Harber, 2005; Wang et al., 2018). However, our main results are robust to controlling for parent and teacher expectations for children’s later educational attainment, teacher perceptions of children’s competence, and parent-reported measures of investment like often doing homework with their child. Second, children may assess their own ability relative to their school or classroom, while we evaluate their demonstrated ability relative to a national sample. We are limited in our ability to measure school quality, but the measures we do have – proxies for school income, investment, and average achievement – do not correlate with over- and under-confidence, conditional on our other controls. Controlling for these measures of school quality does not change our results. Finally, our results are also robust to controlling for childhood “Big-Five” personality traits, suggesting that over- and under-confidence in math are distinct from these more commonly-studied attributes.⁶

In addition to testing these confounders, we also show that our results hold when we use fourteen different formulations of over- and under-confidence – varying all of the key decision points in constructing our main measures – as our key independent variables.

Two dynamic patterns could underlie the associations we estimate. First, children’s over- and under-confidence could alter early patterns of educational investments by parents, teachers, or chil-

⁶The PSID child assessments do not include standard psychometric scales for the Big Five, so we construct proxies for these traits using parents’ reports of children’s behavior and personality. See Appendices B.2 and B.5 for details.

dren themselves; these early investment patterns could then snowball forward into long-term gaps in education and employment. On the other hand, if children's biased beliefs persist, they may have direct psychological effects on choices and performance at each stage in a young adult's development, conditional on his or her performance up to that point. Our evidence suggests that this latter explanation may play a role in the associations we observe. Over- and under-confidence persist through adolescence and into young adulthood (ages 18-27), so biased beliefs could continue to directly affect young adults' decision-making as they age. Childhood confidence also continues to substantially predict later-life education when we hold fixed all intermediate outcomes.

Our results suggest that over- and under-confidence merit study as psychological traits with key economic implications. While our results are not causally identified, they are consistent with childhood confidence having important effects on later-life outcomes. Our evidence is also consistent with the idea that those with more confidence fare uniformly better: under-confident children have worse outcomes than their peers with comparable test scores, while over-confident children have better.⁷ Our results leave ample room for future work: to experimentally test the impacts of childhood biased beliefs, to clarify the mechanisms underlying the associations we observe, and to design and test interventions that build confidence in childhood and later life.

Our paper contributes to three literatures in economics.⁸ First, we add to a recent literature estimating the returns to psychological or social attributes in the labor market; we provide the first evidence on the returns to over- or under-confidence in the specific academic domain of math. In addition to the work on general self-esteem and long-term outcomes that we cite above, parallel literatures examine the associations between economic outcomes and the Big-Five personality traits (Almlund et al., 2011; Heckman et al., 2019), competitiveness (Buser et al., 2021), and children's time, risk, and social preferences (List et al., 2021). While our data do not measure children's competitiveness or time and risk preferences, our results are robust to controlling for measures of

⁷Since girls are more likely to be under-confident in math and less likely to be over-confident, these associations could help to explain key gender gaps in the labor market. Unfortunately, our results are too imprecise for us to conclude whether controlling for biased beliefs in math reduces the gender gaps in adolescent test scores, majoring or working in STEM, or earnings.

⁸Psychology research on academic confidence studies how these beliefs develop as children age (e.g. Eccles et al. 1984) and depend on social constructs like gender and race (e.g. Herbert and Stipek 2005; Usher and Pajares 2006). This work relies on self-reported psychometric scales and does not compare self-reported ability to a measure of objective ability, as we do in this paper.

the Big-Five traits in childhood. Together with this prior work, our paper suggests that future work should disentangle the economic importance of these various traits.

Second, we extend the literature on asymmetric belief updating in adults by documenting over- and under-confidence in a large sample of children in a real-world setting. This heterogeneity matches the lab-based economics literature, which has found mixed patterns of asymmetric updating (Benjamin, 2019; Zimmermann, 2020). As we've noted, the gender gaps in math confidence that we observe are consistent with lab-based evidence that people over-weight positive ability signals in stereotypically gender-congruent domains (e.g. Coffman et al., 2019).

Finally, the studies most relevant to our own examine the role of beliefs about ability in educational settings. Owen (2020) shows that male college students over-estimate their own ability in STEM and under-estimate the ability of others, while women are more likely to over-estimate others' ability; giving students information about their ability then shrinks gender gaps in beliefs and STEM credits. We find that even children have biased beliefs about their own abilities, with similar gendered patterns. Since children's beliefs may be more malleable than those of college students, our work suggests that interventions like Owen's may be fruitful at younger ages. Owen does not assess whether the de-biasing intervention has effects beyond the same semester, but our results suggest that longer-term effects could be substantial.

While Owen (2020) intervenes specifically to change students' beliefs about their ability, other interventions target self-perceptions more broadly. For example, several studies show that building children's generalized self-efficacy and grit can narrow gender gaps in both confidence and willingness to compete in math (Falco et al., 2010; Alan and Ertac, 2019). Similarly, Carlana et al. (2018) find that a multifaceted career-counseling intervention among high-achieving immigrant students in Italy increases self-efficacy and successfully closes native-immigrant gaps in pursuing a more academic high-school track. In contrast, we study math-specific confidence.⁹ We also show

⁹Contemporaneous work by Anaya et al. (2021) uses the same data from the PSID and its child supplements to examine the relationship between majoring in STEM and early childhood achievement, self-assessed ability, and parent occupation, though they focus on including parent occupation as a novel explanatory variable in this regression. Like theirs, our main specifications include indicators for whether children's parents work in STEM, but adding these controls does not change our results. Anaya et al. also describe similar gender gaps in ability beliefs to those we document, but they do not specifically study over- and under-confidence or their relationships with long-term outcomes. In addition to this difference in our central research questions, we see

that math confidence predicts long-term outcomes even when controlling for general confidence, so interventions to close math confidence gaps may be important complements to interventions that build general self-efficacy or grit.

Finally, [Diamond and Persson \(2017\)](#), the only related paper that considers both biased academic beliefs and long-term outcomes, show that receiving an undeservedly marked-up grade on a test at ages 14-16 leads to higher later test scores, more likely high school and college graduation, and higher earnings. Since marked-up scores in one subject raise later scores across all subjects, the authors argue that these effects arise in part by changing students' beliefs about their own ability. However, they do not actually observe students' beliefs about their own ability, as we do. Together, our papers strongly suggest that students' biased beliefs about ability matter for later educational and employment outcomes.

The paper proceeds as follows. [Section 2.2](#) lays out a conceptual framework for how childhood confidence might affect economic outcomes, and [Section 2.3](#) describes our sample and our measures of biased beliefs. [Section 2.4](#) analyzes the prevalence and predictors of childhood over- and under-confidence in our sample, and [Section 2.5](#) describes our strategy for estimating the links between biased beliefs in math and long-run economic outcomes. [Section 2.6](#) presents results, [Section 2.7](#) describes the stability of our results to potential confounders and alternate definitions of confidence, and [Section 2.8](#) explores the dynamic patterns that these long-term associations might follow.

2.2 How might childhood math confidence affect economic outcomes?

Ability or skill is a primary independent variable in almost every economic model of student and worker decision-making. These include settings where agents are investing in their own futures,

our work as building on theirs in three ways: (1) We use a more comprehensive set of available data from the PSID and its child supplements; (2) We consider a larger set of outcomes observed over a much longer time frame; and (3) We define several new measures of over- and under-confidence to deal with complications with the raw data, an issue that Anaya et al. do not discuss.

like deciding to continue with schooling, choosing a college major or career, or searching for a job (e.g. [Becker, 1964](#); [Roy, 1951](#); [McCall, 1970](#); [Borjas, 1987](#); [Kirkeboen et al., 2016](#)), as well as settings where teachers or parents decide, for example, how to invest in or tailor their pedagogy to a child ([Fryer, 2018](#); [Dizon-Ross, 2019](#)).

Over- and under-confidence would enter any of these models if ability is imperfectly observed: by parents, teachers, and even by the student or worker themselves. Where ability and effort are complements, like college applications, over-confident agents may work harder. Consistent with these cases, psychological theories of motivation, including Bandura's (1986) Social Cognitive Theory or Expectancy-Value Theory (see [Wigfield and Eccles \(2000\)](#)), emphasize that individuals are more likely to attempt and succeed at tasks in which they feel competent. Where ability and effort are substitutes, like some school tests, over-confident agents may reduce their effort. [Bénabou and Tirole \(2002\)](#) model how over-confidence can persist in equilibrium in either setting.

Over- and under-confidence may also affect outcomes in any setting where teachers or parents decide how to invest time and resources into children based on their perceptions of each child's ability. If adults interpret more confident children as more skilled, they may over-invest in over-confident children and under-invest in under-confident children. [Dizon-Ross \(2019\)](#) shows that parents have inaccurate beliefs about their children's academic performance, and that correcting those beliefs causes them to adjust their investments. Similarly, [Papageorge et al. \(2018\)](#) show that having a teacher with higher expectations increases a student's chance of completing college. The same forces could operate in job applications, where potential employers are uncertain about applicants' skill: in lab experiments, [Schwardmann and van der Weele \(2019\)](#) show that interviewers rate more confident job applicants more favorably, and [Mobius and Rosenblat \(2006\)](#) show that employers offer higher wages to more confident workers.

Our focus on confidence gaps in math, not reading

Our data allows us to identify over- and under-confidence in both math and reading, but we focus the remainder of the paper on biased beliefs in math for several reasons. First, performance in math

can be measured more objectively than performance in reading, so children’s beliefs about their math ability may be more precise. Next, past work suggests that math ability during childhood and young adulthood more strongly predicts later achievement than does reading ability (e.g. [Duncan et al., 2007](#); [Castex and Kogan Dechter, 2014](#); [Goodman, 2019](#)). We find similar patterns in our data in Appendix Table B.1, where we regress our main education and employment outcomes on childhood test scores and the set of controls that we will use throughout our main analysis. While both math and reading score percentiles predict later academic achievement and attainment, only math scores predict earnings, unemployment, and majoring in STEM. Thus, children’s perceptions of their own ability in math may also link more strongly with later-life achievements than do their self-perceptions in reading. Finally, the Bureau of Labor Statistics predicts that employment in STEM occupations will continue to grow at faster rates than non-STEM occupations through 2030, so math ability may become an even more important predictor of success in the labor market.

That said, we conduct all of the subsequent analysis for reading confidence (Appendix Tables B.2-B.5). Reading confidence robustly predicts few educational or employment outcomes.

2.3 Measuring confidence and later-life outcomes in the PSID

2.3.1 Sample and survey design

We explore the links between biased childhood beliefs and outcomes in young adulthood using the rich data of the Panel Study of Income Dynamics (PSID). The PSID was first collected in 1968 among 5,000 nationally-representative households from two independent samples: a national sample of low-income families from the Survey of Economic Opportunity (the “SEO sample”) and a national sample drawn by the Survey Research Center (the “SRC sample”). The PSID has since surveyed the descendant households of the original sample annually from 1968 to 1997 and biennially thereafter, adjusting the sample in 1997 to again make it nationally representative.

We combine the core PSID with two supplements that follow respondents from childhood into

young adulthood: the Childhood Development Supplement (CDS) and the Transition into Adulthood Supplement (TAS). The CDS was introduced in 1997, sampling up to two children per PSID household who were then between the ages of 0 and 12 (3,563 children). The CDS collects detailed information from children themselves, from their primary caregivers, and from their elementary school teachers on areas including children's cognitive and emotional development, health, and exposure to parenting practices. The original CDS sample was re-interviewed in 2002-2003, then aged 5-17, and those still below age 18 were included in a third CDS wave in 2007.

In 2005, the PSID introduced the TAS as a bridge between the CDS and the main PSID survey for CDS respondents, the oldest of whom had reached ages 18 to 20 by that year. The TAS has been collected biennially since 2005, with younger CDS respondents aging into the TAS sample at 18. Individuals participate in the TAS until they become economically-independent heads of their own household, at which point they enter the adult PSID sample and are surveyed every two years. The TAS is designed to capture respondents' social and career development as they enter adulthood; we use its modules on education, employment, income, and personality.

The PSID-CDS-TAS data structure is uniquely suited to exploring the links between childhood confidence and long-term educational and employment outcomes. First, the CDS both administers a math test and asks children to evaluate their own math ability; we combine children's test scores and self-assessments to identify over- or under-confidence in math. [Section 2.3.2](#) below details the CDS tests, self-assessments, and our confidence measures. Second, following CDS children into the TAS and then the PSID allows us to observe detailed data on educational and employment outcomes over 22 years, following our oldest respondents into their mid-thirties. Finally, the extensive data on parents' employment and income in the PSID and on parenting practices and other child characteristics in the CDS allows us to control for many covariates that could confound the relationship between biased beliefs and long-run outcomes.

For example, the detailed child module in the CDS allows us to control for other forms of ability and confidence that are distinct from skill and confidence in math, but which may correlate with them. We construct a measure of general confidence as the mean of standardized variables

capturing whether children see themselves as broadly competent (see Appendix B.2 for details); we have no measure of true ability by which to normalize this general confidence scale, so we use it as a control for unobserved abilities and other dimensions of confidence that may correlate with biased beliefs in math and also affect later-life outcomes. We also control for children’s scores on the Digit Span subtest of the Wechsler Intelligence Scale for Children (Revised), a measure of short-term memory. Next, the CDS and core PSID collect detailed household information on total family income, household heads’ education, primary caretakers’ values and mental health, household structure, and financial characteristics like whether the household receives food stamps. Section 2.5 will detail the family and child variables that we control for in our main analysis.

Our final sample consists of the 2,985 CDS respondents with at least one year of math cognitive tests and self-assessments in the CDS, about 84 percent of all CDS respondents.¹⁰ We report summary statistics for this sample in Appendix Table B.6; all variables are observed in the same year in which we first observe childhood over- or under-confidence in math.

Our sample is non-randomly selected from the national population, both because the initial 1968 PSID sample oversampled low-income families and because there is unobserved selection in whether CDS participants report math test scores and self-assessments.¹¹ This selection appears in our sample statistics in notable ways. First, our sample is disproportionately Black: 45.8 percent are White, 41.7 percent are Black, and only 7.5 percent are Hispanic, while the U.S. Census Bureau reports that 69.1 percent of the US residents were White, 12.1 percent Black, and 12.5 percent Hispanic in 2000 (Greico and Cassidy, 2001). While the Census Bureau reports median household income in 1997 of \$55,336, our sample’s median taxable income is slightly lower, at \$52,029 (both in 2016 USD). On the other hand, our study sample performs disproportionately well on the CDS standardized tests: we observe median CDS math and reading score percentiles of 60 and 54,

¹⁰In both the full CDS sample and our final analysis sample, 53 (38) percent of children are descended from the SRC (SEO) sample and 9 percent of children are from the immigrant sample added to the PSID in 1997.

¹¹Most children who are missing test scores or self-assessments lack this data because they skipped the entire section of the CDS administered to the child, while completing the survey portions administered to the primary caregiver. These respondents largely have similar demographics to those for whom we observe confidence measures, but their mothers are less likely to have a high school degree, they have lower total family income, and they are about a year younger. Students who take the math cognitive assessments but do not give self-assessments (about 25 percent of the children who are missing test scores or self-assessments) score much lower on both the math assessment and the Digit Span memory test (Appendix Table B.17).

respectively, relative to national norming samples.

While we do not weight our sample to be nationally representative in our main analysis, we include results that do so in Appendix Tables B.7-B.10. These weights are based on those published by the CDS, which capture the inverse probability of respondents' inclusion in the CDS sample; we then recalibrate these CDS weights via iterative proportional fitting, or raking, to ensure that our sample matches marginal distributions of percentile CDS math scores, race in 2000, and total household income in 1997. Our main results are less precisely estimated when we use weights, though they remain qualitatively similar.¹² We present all descriptive statistics both for the weighted and unweighted samples.

2.3.2 Measuring over- and under- confidence in math

Data on children's self-reported and demonstrated ability in math

The CDS assesses children's math skills using the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R), a test of academic achievement commonly used by school psychologists in the 1990s (Stinnett et al., 1994; Hicks and Bolen, 1996; Duffy and Sastry, 2014). The CDS administers the Applied Problems subtest of the WJ-R, comprising 60 word problems of increasing difficulty that assess math reasoning and knowledge.¹³ Each child completes only a subset of the test, beginning at a "basal" level, where they answer six consecutive questions correctly, and ending at a "ceiling" level, where they get six consecutive questions wrong. The CDS then reports each respondent's percentile rank relative to the nationally-representative WJ-R norming sample for their age group; we use these percentile ranks as our measure of each child's demonstrated ability

¹²Since the error terms in our regressions are unrelated to the sampling criterion, conditional on our extensive controls for family income, race, and other characteristics, weighting may not improve the estimator's consistency and may reduce its precision (Solon et al., 2015). These recalibrated weights put less weight on children with high CDS math scores, in some cases leaving us underpowered to detect the correlations between under-confidence and long-term outcomes. A natural concern is that our unweighted regressions may estimate non-representative partial relationships between confidence and outcomes, if these associations are heterogeneous by race or income. However, weighted least squares estimates are not necessarily closer to the true population average partial relationship than ordinary least squares estimates (Solon et al., 2015). Instead, we directly estimate heterogeneity by characteristics related to the sampling scheme, like family income, race, age, and being in an SRC-sample family. These results are imprecise and show no robust patterns of heterogeneity (results available upon request).

¹³The 1997 CDS wave also included 58 WJ-R questions on calculation skills, and we use this test in the next section to assess the reliability of our over- and under-confidence measures.

in math. Panel A of Figure 2.1 shows the distribution of these scores in our sample.

In addition to collecting this measure of performance in math, the CDS also asks all respondents ages 8 or older to assess their own ability in math, asking them to answer “How good at math are you?” on a scale of 1 (not at all good) to 7 (very good). Children never receive their scores on the WJ-R math test, so these self-reports do not reflect feedback from the CDS. Panel B of Figure 2.1 shows the distribution of these self-assessments. Math self-perceptions are highly skewed towards positive responses, with over 89 percent of respondents ranking themselves as “Okay” or better at math. This skew may be partially explained by the distribution of percentile scores in Panel A, which skews heavily towards higher-performing children. While shifted upwards, children’s self-reports do contain information about objective ability: in Panel C of Figure 2.1, average math test percentiles rise almost linearly with self-reported ability in math.

We measure children’s over- and under- confidence in math in the first wave of the CDS in which they have non-missing cognitive test scores and self-assessments, leaving us with a sample of 2,985 children.¹⁴ We first measure confidence for the median child before age 12, and we observe confidence by age 13 for 83 percent of children. Thus, we will interpret our measures as *childhood* over- and under-confidence in math. Throughout, our analysis will control for both birth year and the age at which we first observe confidence.

Defining binary measures of over- and under-confidence

We first identify over- and under-confidence in math using large mismatches between children’s score percentiles and their self-assessments. In particular, we classify any respondent as under-confident in math if she scored above the 75th percentile nationally and ranked her own ability at 1 to 4, corresponding to the bottom 47 percent of the subjective-ability distribution in our sample, or if she scored above the 50th percentile nationally and ranked herself at 1 to 3, corresponding to the bottom 10 percent of the subjective-ability distribution. We define over-confidence among low-achievers using similar thresholds, but we account for the skewed self-assessment distribution by using stricter cut-offs to identify biased beliefs. In particular, we identify any respondent as

¹⁴We first observe confidence from the 1997 CDS wave for 1,075 children, from the 2002 CDS wave for 1,347 children, and from the 2007 CDS wave for 563 children.

over-confident in math if she scored below the 25th percentile nationally and rated her own ability at 6 or 7, corresponding to the top 39 percent of the subjective-ability distribution in our sample, or if she scored below the 50th percentile and rated herself at 7, corresponding to the top 22 percent of the subjective-ability distribution.

These measures of math over- and under-confidence have several key strengths: they are easy to define and observe, they refrain from putting too much stock in the cardinal value of children's self-assessed ability, and they account for the upward skew in self-assessments, which we consider to be a form of response bias separate from over- or under-confidence.¹⁵

However, our measures also have several limitations. First, we can only identify over-confidence among children scoring below the 50th percentile and under-confidence among those scoring above the 50th percentile; however, this strategy matches the existing literature, which typically documents under-confidence (imposter syndrome) among high-achievers (Sakulku, 2011). Another limitation is that these measures are not directly comparable to measures of over- and under-confidence from the lab-based literature, which can precisely measure respondents' beliefs about their quiz performance or rank relative to a group (e.g. Coffman et al., 2019; Möbius et al., 2014; Eil and Rao, 2011). Our measures of over- and under-confidence, in contrast, identify coarse categories of children with large gaps between their self-assessments and observed scores. Our second measure of biased beliefs, described below, aims to partially address these limitations.

Defining a more continuous measure of biased beliefs

Our second confidence measure identifies biased beliefs as the difference between children's self-reports and their observed performance on the CDS math test. To transform these objects to the same scale, we split the distribution of children's percentile scores uniformly into seven bins, where 1 includes the lowest 14 score percentiles and 7 includes the highest 14 score percentiles relative to the national norming sample. We then assume that students with full information about

¹⁵For example, upward response bias could arise based on children's interpretation of the qualitative labels on the scale ("Not at all good" at 1, "Okay" at 4, and "Very good" at 7) if, for example, they think that nearly everyone is at least "Okay" at math. Upward skew could also arise if self- or social-image concerns make children unwilling to tell a surveyor that they are worse than "Okay" at math. If, on the other hand, this upward skew does reflect true aggregate over-confidence, our estimates for long-term associations with over-confidence would simply reflect links with *particularly* over-confident beliefs in math.

the national distribution of scores and their place in it would have self-reported their math ability as the bin from 1 to 7 in which their score percentile falls; we take the difference between their actual self-report and this bin as our measure of biased beliefs. This measure then takes on integer values from -6 to 6. For ease of interpretation, we standardize this variable to have mean 0 and standard deviation 1 throughout the rest of the paper.

This measure has three strengths relative to our main measure: it allows for more granularity in the extent of biased beliefs, aligns more closely with measurements of biased beliefs in the lab-based literature, and relies on fewer choices by the authors. However, by assuming that we can identify even small biases in beliefs about math ability, it is more likely to conflate actual biased beliefs with children's private information about their math ability (described in more detail in the next section). It may also be confounded by forms of reporting bias other than over-confidence that generate the overall upward skew in self-reports (see footnote 15).

We present results for all outcomes using both the binary and more continuous formulations of biased beliefs, and in general the results are extremely consistent. To ensure that our main results do not arise just from our particular choice of confidence measures, we show that they are robust to a range of alternate definitions of both our indicators for over- and under-confidence and this more continuous measure of biased beliefs. We describe these alternate measures in Section 2.7.

2.3.3 Biased beliefs or measurement error?

One key concern with our measures of biased beliefs in math is that they may conflate over- and under-confidence with children's private information about their math ability, perhaps driven by measurement error in the WJ-R assessment. Four key pieces of evidence support the claim that our measures truly capture biased beliefs in childhood.

First, prior work has shown that the WJ-R assessment is a reliable measure of children's math skills, with test-retest reliability for the applied math problems of about 0.85 in large samples (Hicks and Bolen, 1996).¹⁶ We can also verify WJ-R reliability across math domains in our sample

¹⁶Several studies find test-retest reliability of about 0.75 for certain ages, though these studies use small samples (Shull-Senn et al.,

using the 1997 wave of the CDS, which administered both the Calculation and Applied Problems subtests of the WJ-R. For the 1,450 children who took both tests, the correlation in percentile ranks on the two sections is 0.69. Our binary designations of children as over-confident, under-confident, or neither are also highly consistent whether we measure objective math ability using children's percentile scores on the Calculation or Applied Problem subtest: 81 percent of children with both measures are classified in the same category regardless of which ability measure we use. Another ten (nine) percent switch from under-confident (over-confident) to neither or vice versa.¹⁷

Second, our measures of childhood math confidence persist over time. About 60 percent of the children in our sample appear in two waves of the CDS, allowing us to construct two measures of over- and under-confidence taken five years apart. Children appear in a second CDS wave at ages 13 to 19, so these second-wave measures capture biased beliefs in adolescence. Table 2.1 regresses our *adolescent* measures of biased beliefs on our *childhood* measure of the same variable, controlling for a set of demographics and parent characteristics that we will use throughout our later empirical analysis; we outline these specifications in detail in Section 2.5.¹⁸ These regressions show substantial persistence: respondents who were over-confident in math as children are about 3 times as likely (12pp more likely) to be over-confident in math as adolescents, while under-confident children are about 1.7 times as likely (4pp more likely) to be under-confident as adolescents.¹⁹ Similarly, a one-standard-deviation increase in the degree of biased beliefs in childhood predicts 0.18sd more biased beliefs in adolescence. If our confidence measures just captured random testing variability, we would not expect to see such substantial persistence.

1995)

¹⁷We find similar reliability using our more continuous measure of degrees of confidence, which takes on integer values from -6 to 6. There, 32 percent of children are assigned the same value regardless of which math test we use as the measure of demonstrated ability, 62 percent are within one integer, and 83 percent are within two integers. See Appendix Figure B.3 for the full joint distribution of the more continuous confidence measures based on the two math subtests.

¹⁸The regressions in Table 2.1 add controls for children's *adolescent* test score deciles in math and reading to our main specification. We add these controls to purge any correlations induced by the effects of childhood confidence on adolescent test scores, since childhood over- and under-confidence predict later test scores (see Section 2.6) and higher-scoring (lower-scoring) children are mechanically more likely to be classed as over-confident (under-confident).

¹⁹While our main measure of under-confidence persists only weakly into adolescence, several alternate definitions of under-confidence are strongly persistent (Appendix Figure B.4). Our main measure's limited persistence may relate to the fact that adolescent test scores are much less upward-skewed than childhood test scores, so fewer respondents can be classified as under-confident in adolescence. The more persistent alternate definitions of under-confidence, in contrast, identify under-confidence among respondents with a wider set of test scores and thus are less affected by this distributional shift. Like our main measure, these alternate measures predict substantial gaps in long-run outcomes (Appendix Figures B.5-B.16).

Third, our main results are largely robust to using measures of over- and under-confidence that reduce potential measurement error by combining observations of children’s test scores and self-reported ability across two waves of the CDS. We discuss these measures and results in more detail in Section 2.7. If measurement error is uncorrelated across tests taken 5 years apart, these average confidence measures will be less vulnerable to it than are our main measures.²⁰

Finally, we describe in the next section that we observe substantial gender gaps in math over- and under-confidence, with girls more likely to be under-confident and less likely to be over-confident. This pattern is consistent with gender stereotypes about math ability, which may shape children’s beliefs even at young ages, and mirrors results for adults in the lab (e.g. [Coffman et al., 2019](#)). Our measures of over- and under-confidence could only be entirely explained by measurement error if this error took a similar gendered pattern, beyond its correlation with WJ-R Applied Problems scores and with the many other controls we outline in Section 2.5.²¹ We consider a few possible sources of non-random measurement error that could generate these patterns: skill in some dimension of math that the test does not cover, test-taking anxiety, and test-taking motivation.

First, the CDS data allow us to test for gender gaps in one central dimension of math skill that our main test scores do not directly capture: calculation skills. Using the 1997 CDS sample, when children took both the WJ-R Calculation subtest and the WJ-R Applied Problems subtest, we find no evidence that boys have better calculation skills conditional on the applied problems scores that we use in our main analysis.²²

Next, differential measurement error in the CDS math tests could arise if boys or girls are more

²⁰Despite this benefit, we do not use these averages as our preferred measures of confidence for three reasons: (i) over- and under-confidence at older ages may be more likely to be confounded by unobserved variables; (ii) we are interested in adolescent test scores and confidence measures as outcome variables; and (iii) only 60 percent of our sample has confidence measurements over multiple waves of the CDS.

²¹Differential random error by gender could not fully explain the gendered patterns of over- and under-confidence we observe, since the gender with more variable performance would be more likely to be both over- and under-confident. Nonetheless, comparing boys’ and girls’ performance on the Calculations and Applied Problems subtest in the 1997 CDS sample suggests that neither gender has differentially variable test performance. 81% of both boys and girls receive the same binary confidence designation when calculated using either the Calculations or Applied Problems percentile score as a measure of math skill, and the joint distributions of the more continuous measures are very similar for boys and girls (Appendix Figure B.3).

²²We estimate the following regression: $CALCpctile_i = \beta_0 + \beta_1 APpctile_i + \beta_2 Female_i + \beta_3 APpctile_i \times Female_i + \varepsilon_i$. Coefficient β_3 is not significantly distinguishable from zero, and β_2 is significant and positive. Thus, girls have stronger calculation skills than boys conditional on their Applied Problems scores, which would tend to make girls look more over-confident by our measures, the opposite of what we find.

prone to testing anxiety that impairs performance. While past work finds that boys show higher physiological stress during test-taking (Weekes et al., 2006; Stroud et al., 2002), other research suggests that physiological stress only impairs performance when students psychologically appraise it as an indicator of potential failure (Jamieson et al., 2013; Mattarella-Micke et al., 2011). Girls tend to have higher psychological test anxiety and math anxiety, and most commentary suggests that it is these psychological manifestations of anxiety that pose first-order risks to test performance (Devine et al., 2012; Erturan and Jansen, 2015; Ballen et al., 2017). Thus, we would expect girls' test performance to differentially lag their true skill, producing gender gaps in confidence that would conflict with our empirical results.

Finally, we turn to test-taking motivation. Past work finds that girls are somewhat more motivated than boys to exert effort on low-stakes tests, so boys' CDS math scores may be differentially low relative to their true skill in math (Segal, 2012; DeMars et al., 2013; Gneezy et al., 2019). Then, boys may appear more over-confident by our measures. While it is hard to fully eliminate this possible confounder in our setting, our results are robust to controlling for agreeableness and conscientiousness, two Big-5 personality traits that are positively correlated with unincentivized test effort (DeMars et al., 2013; Segal, 2012). (See Section 2.7 for more details.)

Together, most evidence from our empirical setting and from past work on test-taking strongly suggests that our confidence measures capture a meaningful psychological trait. However, we cannot fully eliminate the risk that these measures capture children's private information on some aspect of math ability that the test systematically excludes. Any such confounder could only explain our results if it is differentially weak among girls and affects outcomes beyond its correlation with demonstrated math ability, general confidence, digit span score, reading ability, and the many other controls we outline in Section 2.5.

2.4 Patterns of over- and under-confidence in the population

This section documents the prevalence and correlates of over- and under-confidence in our sample. Besides documenting biased beliefs in math in a real-world setting, these results are useful both to validate our measures of biased beliefs and to inform our strategy for estimating the links between childhood confidence and long-run outcomes, which we describe in Section 2.5.

2.4.1 Prevalence of biased beliefs

We find substantial over- and under-confidence among children in our sample: using our main binary measures, 8.5 percent of children are over-confident at their first measurement, while 12 percent are under-confident.²³ Since these measures identify large gaps between children’s self-assessed and objective performance, these shares are strikingly high. Turning to our more continuous measure of biased beliefs, 21 percent of children report the same bin as their percentile score would imply, 8.7 percent of children report ability levels that are at least 3 bins lower than that of their score, and 17 percent report ability levels that are at least 3 bins higher, where each bin spans 14 score percentiles. See Appendix Figure B.1 for the full distribution of the continuous confidence measure. It is notable that over- and under-confidence are both prevalent in this large sample, given psychology’s focus on over-confidence (Moore and Healy, 2008) and the mixed evidence from lab experiments on asymmetric belief updating (Benjamin, 2019).

Next, older children have more accurate beliefs. Panel A of Appendix Figure B.2 plots the share of children who are over- or under-confident in math by age; Panel B plots the cumulative density function for the continuous confidence measure for three age groups, pooling respondents’ observations across CDS waves. Both panels show that younger children are more likely to have incorrect beliefs about their math ability, and average belief accuracy increases almost monoton-

²³We find similar results when applying our raked weights to obtain nationally representative estimates: 9.2 percent of children are over-confident and 9.8 percent are under-confident.

ically as children age. We focus on the associations between confidence and later-life outcomes using first-observed confidence, so our confidence observations are drawn from young ages with more biased beliefs. We eliminate bias due to the timing of our confidence measurements by including fixed effects for the age at which confidence was measured in all regressions.

2.4.2 Biased beliefs and other child characteristics

Over- and under-confidence correlate with other child characteristics in largely expected ways (Table 2.2 and Appendix Table B.11). Unsurprisingly, children with higher general confidence are more likely to be over-confident and less likely to be under-confident in math, and children with higher digit span scores are less likely to be under-confident. Math test score deciles strongly predict confidence gaps (though some of this correlation arises mechanically from how our measures are constructed), while reading test score deciles do not (Appendix Table B.11). We will control for children's general confidence, digit span scores, and test score decile fixed effects in math and reading in all regressions of later-life outcomes on childhood biased beliefs.

Conditional on these measures of ability, children who have ever been in a gifted program are 8.7pp less likely to be under-confident in math and 2.6pp more likely to be over-confident. These correlations could reflect that schools and children share private information on children's ability conditional on CDS scores, that being in a gifted program alters children's confidence, or that children's confidence influences their treatment at school conditional on ability. To avoid controlling for mediators of the effects of confidence, our regressions will *not* control for this variable or other signals of ability from schools, like repeating a grade.²⁴

Finally, gender is the strongest demographic predictor of math confidence. Girls are 2.3pp (20 percent) more likely to be under-confident and 2.7pp (27 percent) less likely to be over-confident in math than boys with the same score deciles, and on average, girls' biased beliefs are 0.1 standard deviations (sd) lower than the average boy's. Note that girls do not have more accurate beliefs,

²⁴Math over- and under-confidence also correlate with children's other attitudes towards math and school in reasonable ways (Appendix Table B.18), suggesting that our measures isolate over- and under-confidence in the particular domain of math. See Appendix B.3 for more discussion.

simply more negatively-biased ones.^{25,26} This finding is consistent with prior literature showing that adults are more over-confident in gender-congruent domains (e.g. [Coffman et al., 2019](#)), but it is notable that we find it in children, the majority of whom have not yet entered puberty. These gender differences are present at almost all ages, but due to small sample sizes the patterns are imprecise (available upon request).²⁷

Perhaps surprisingly, we find no significant links between children’s math confidence and their parents’ education or occupation, household income, or race, conditional on all other characteristics. However, noise in these estimates means we cannot reject potentially large correlations.

2.5 Confidence and long-term outcomes: Empirical strategy

Our empirical strategy is simple: we estimate the associations between biased ability beliefs in math and later education and work outcomes, *holding fixed measured childhood ability*. We use the PSID’s rich data on childhood environment and family characteristics to control for extensive pre-determined confounders, but we refrain from interpreting our estimates as the causal effects of confidence. We estimate the following specification:

$$Y_{it} = \alpha + \beta_1 Over_{i0} + \beta_2 Under_{i0} + A'_{i0} \mu + X_{i0}^{Ct} \delta_1 + X_{i0}^{Pt} \delta_2 + \gamma_s + \omega_t + \varepsilon_{it}$$

where Y_{it} is individual i ’s outcome of interest in adolescence or adulthood, measured in wave t of the TAS or PSID, and $Over_{i0}$ and $Under_{i0}$ are indicators for being over- or under-confident in math as a child, respectively. All of our main tables also include regressions in which we replace $Over_{i0}$

²⁵In fact, there is no gender gap in the likelihood of having accurate or almost accurate beliefs (degrees of over- and under-confidence equal to zero, or between -1 and 1, respectively). Results are available upon request.

²⁶Appendix Figure B.17 shows that this gender gap is extremely robust to using alternate definitions of over- and under-confidence and alternate ways of calculating the more continuous degrees of confidence measure. This figure plots the coefficient on the female indicator when we exchange the dependent variables in Table 2.2 with these alternate measures (discussed further in Section 2.7).

²⁷We also test whether childhood gender gaps in math under-confidence explain gender gaps in later education and employment outcomes: adolescent test scores, majoring in STEM, and earnings. Specifically, we estimate the change in the coefficient on gender when we estimate our preferred specification with and without the indicator for under-confidence (following [Buser et al. \(2021\)](#)). The results (available upon request) are quite noisy, so we leave it to future research to determine whether confidence gaps in math can help explain these and other gender gaps.

and $Under_{i0}$ with the single $ZConf_{i0}$ variable, which captures the degree to which a child is over- or under-confident in standard deviations. Due to power limitations, we assume that $ZConf_{i0}$ has a linear relationship with our outcomes of interest.²⁸

Next, all of our regressions include A_{i0} , a vector of controls for childhood ability. In particular, A_{i0} includes linear controls for childhood digit span score and general confidence, as well as fixed effects for test score deciles in both reading and math.²⁹ Our basic specification also includes state fixed effects γ_s , TAS or PSID wave fixed effects ω_t when the outcome is observed multiple times for each individual,³⁰ a set of child controls X_{i0}^C , and a set of parent controls X_{i0}^P . In our first specification, X_{i0}^C and X_{i0}^P include only variables that are certainly unaffected by respondents' childhood math confidence: X_{i0}^C includes fixed effects for race, birth year, quarter of birth, gender, and age at which we observe confidence,³¹ and X_{i0}^P includes family income, its square, and fixed effects for both parents' levels of education. All variables indexed at $t = 0$ are from the first CDS wave in which a child had WJ-R scores and an ability self-assessment. Since about two-thirds of the children in our sample have a sibling in the sample, we cluster standard errors by family. Our coefficients of interest are β_1 and β_2 .

Our second specification takes advantage of the detailed caregiver interviews in the CDS to add additional controls for child and family characteristics that may correlate with both confidence and long-run outcomes. In addition to expanding the set of child controls, X_{i0}^C , with the primary care-

²⁸Appendix Figures B.18, B.19, and B.20 show our main results when we relax this assumption; we plot the coefficients on indicators for each integer value of the variable underlying $ZConf_{i0}$: $Conf_{i0} = -6, Conf_{i0} = -5, \dots, Conf_{i0} = 6$. While these results are noisy, the point estimates suggest that this linearity assumption is reasonable. We also show in Appendix B.4 that we cannot generally reject the null hypothesis that over- and under-confidence predict economic outcomes in similar (opposite-signed) ways, further supporting this linearity assumption.

²⁹One might worry that controlling for general confidence absorbs too much of the variation in math over- and under-confidence if over- and under-confidence in math are dimensions of confidence in general. While the economic impacts of general confidence are certainly of interest, we take the conservative approach of isolating math-specific over- and under-confidence as cleanly as possible by controlling for general confidence. That said, our results are remarkably similar with or without the control for general confidence (available upon request).

³⁰For some outcome variables, like earnings and unemployment, we have multiple years of outcomes across TAS and PSID waves for each respondent. In contrast, we observe our educational outcomes (e.g. whether respondents ever majored in STEM) only once per respondent; we do not include survey wave fixed effects in regressions linking childhood confidence to these outcomes. Note that we do not include respondent fixed effects even in regressions with multiple outcome observations per respondent, since we only measure childhood confidence once. We cluster standard errors by family in all regressions.

³¹Age at which confidence is first observed and birth year are not collinear. For example, children who had their confidence measured when they were 8 years old could have been born in 1989, 1995, or 1999 (and had their confidence measured in the 1997, 2003, or 2007 CDS, respectively).

giver’s assessment of the child’s general health, this specification supplements X_{i0}^P with additional parent and family controls: whether the family receives government transfers; whether the household includes the father or has two adults; parents’ beliefs about gender norms and the qualities that are most important for success; and parent mental health (see Appendix B.2 for details). Finally, we add four indicators for whether the child’s parents work in STEM or another high-education occupation (based on Anaya et al. (2021); see footnote 9). We focus on this specification throughout the text, but results are generally consistent across these two specifications.

2.6 Confidence and long-term outcomes: Results

The following section presents our results, documenting strong associations between childhood under- and over-confidence in math and key later-life outcomes: adolescent test scores, graduation from high school and college, college major, career choice, earnings, and unemployment. We present these results in Tables 2.3, 2.4 and 2.5.

2.6.1 Medium-term educational achievement

We first examine the links between childhood confidence and medium-term educational achievement, measured as adolescent scores on the CDS math assessments. We observe these scores at children’s second CDS observation, about 5 years after we first observe their confidence in math.

Children’s biased beliefs in math significantly predict adolescent math performance (Table 2.3, columns 1 and 2). Using our binary measures (Panel A), children who are over-confident in math score 2.7 percentiles (standard error = 1.5p) higher on the math assessment five years later than others with comparable baseline scores, while under-confident children score 5.9 percentiles (se = 1.5p) lower. Using our more continuous measure (Panel B), a child with 1 standard deviation (sd) higher math confidence in childhood scores 2.8 percentiles (se = 0.57p) higher on the math assessment 5 years later than others with comparable baseline scores. Children marked as over- or

under-confident in our binary metrics differ from others by an average gap of 1.8sd and -1.6sd in continuous degrees of confidence, respectively, so our estimate magnitudes are remarkably consistent across the two panels. In contrast, there is no relationship between childhood math over- or under-confidence and adolescent reading scores using either measure of biased beliefs (Table 2.3, columns 3-4).

These associations are large relative to the links between raw math ability and later scores: increasing one's childhood math score by 10 percentiles is associated with scoring on average 5.3 percentiles higher in adolescence (Column 1 of Appendix Table B.12).³² Thus, being over- (under-) confident in math predicts as large a gap in adolescent test scores as does increasing (decreasing) one's childhood math test score by 5-11 percentiles.

2.6.2 Educational attainment

Biased beliefs in math during childhood also predict important gaps in high school and college graduation. Children who are over-confident in math are 6.2 percentage points (se = 2.6pp) more likely to graduate from high school, and children who are under-confident in math are 5.8pp (se = 2.8pp) less likely to graduate from college (Table 2.3, columns 5-8, Panel A). Since only 30 percent of our sample graduates from college, being under-confident predicts a 20 percent drop in the likelihood of college graduation. We find very similar results using our more continuous measure in Panel B: a child with 1sd higher math confidence in childhood is 1.8pp (se = 1.0pp) more likely to graduate from high school and 3.3pp (se = 1.1pp) more likely to graduate from college, though the first is only marginally significant. Again, the magnitudes of these results are similar regardless of which confidence measure we use.

These gaps are large relative to the associations between childhood math scores and educational attainment in our data: childhood math scores generally do not substantively predict high school

³²While we estimate the relationships between confidence and later outcomes in regressions with test score decile fixed effects, here we run an otherwise identical regression replacing these fixed effects with linear controls for test scores. We benchmark the links between biased beliefs against the coefficients on these linear score controls throughout Section 2.6.

graduation,³³ and increasing test scores by one decile is associated with being 2.9pp more likely to graduate from college on average (Appendix Table B.12, columns 3-4).

2.6.3 College quality, college major choice, and graduate education

Next, we consider later-education outcomes among those who went to college: college quality, college major choice, and whether respondents complete a graduate degree. Since we restrict to college graduates, these regressions use much smaller samples than for our previous outcomes.

First, we find imprecise links between childhood math confidence and the quality of colleges that children later attend. We consider two quality measures: first, an index of general college quality, and second, colleges' 75th-percentile math SAT scores among incoming freshmen – a more specific measure of math quality.³⁴ We focus our discussion on colleges' 75th-percentile math scores (Table 2.4, column 3 and 4), but our results are similar using the more general college quality index (columns 1 and 2). Under-confident children attend schools whose 75th-percentile math SAT scores are 11.3 points (se = 5.9 points) lower than others with the same childhood scores ($p = 0.07$); with 95 percent confidence, we can reject that under-confident children attend schools with math SAT scores that are over 0.3 points higher or 22.9 points lower. Over-confidence is not significantly associated with college quality among childhood low-scorers, but again we observe wide confidence intervals: we cannot reject that over confident children attend colleges that have 20 points lower to 26 points higher SAT scores than their peers. Using our more continuous measure of biased beliefs in Panel B yields consistent, but imprecise, results.

Next, we find that childhood under-confidence in math is starkly associated with major choice among those who go to college (Table 2.4, columns 5 and 6). Among those with a 4-year college

³³Math scores do significantly predict high school graduation, but the coefficient is precisely estimated and very small (increasing test scores by 10 percentiles is associated with being 0.8pp more likely to graduate from high school). The magnitude of this linear coefficient is half of the size of the coefficient on reading scores. Reading skills may be more important for high school graduation than math (e.g. since fewer years of math study are required to graduate).

³⁴Using restricted data from the TAS, we link respondents with college quality data from the National Center for Education Statistics (NCES) for the first college they attended in the first year they attended that college. Following Cohodes and Goodman (2014), we construct an index of college quality as the first component from a principal component analysis of colleges' 75th-percentile math SAT scores among incoming freshmen, graduation rates, and per-pupil instructional expenditures, separately by year. Details on variable construction are available in Appendix B.2. We then standardize this index to have mean 0 and standard deviation 1 in the full sample of four-year colleges in the US by year.

degree, students who were under-confident in math are 16.2pp (se = 3.6pp) less likely to earn a STEM major³⁵ than their peers with comparable childhood scores, an 86 percent drop from the share of STEM majors across all college graduates in our sample. This large gap means that under-confident children who score above the 50th percentile on the CDS math test are only 1.3 times as likely to major in STEM, conditional on going to college, than the average child who *scores below the 50th percentile*; in contrast, other childhood high-scorers are 3.5 times as likely to major in STEM as low-scorers. We obtain very similar results using our more continuous measure of biased beliefs in Panel B: a 1sd increase in confidence is associated with a 7.8pp (se = 2.3pp) increase in the likelihood of majoring in STEM.

Finally, we find no significant relationships between biased beliefs and getting a graduate degree, though again our standard errors are large (Table 2.4, column 7 and 8).

2.6.4 Employment outcomes

Next, we examine the links between childhood over- and under-confidence in math and employment outcomes in young adulthood: occupation type, earnings, and employment status. We follow respondents in the adult PSID when they age out of the TAS, so we observe our oldest CDS respondents through age 36 at the end of our sample period. Since respondents' employment outcomes in their early twenties may not yet be representative of their long-term career trajectories, we restrict the sample to observations in which respondents are older than 25; we observe about 70 percent of our sample above this threshold at least once.³⁶

We first consider job choice. Under-confident children are about 4.9pp (se = 1.6pp) less likely to work in a STEM occupation³⁷ than their peers (Table 2.5, columns 1 and 2), a gap that is approximately equal to the baseline rate at which respondents later work in STEM in our sample.

³⁵We define STEM fields as engineering, math and computer sciences, and natural sciences. We find similar results if we also include health fields.

³⁶Appendix Table B.19 replicates these results using one observation per child, where the dependent variable is calculated as the average outcome observed over ages 28-33. The results are meaningfully the same.

³⁷We define STEM fields to include computer and mathematical occupations, architecture and engineering occupations, and life, physical, and social science occupations. We find similar results if we include healthcare occupations as STEM.

We find a similar result with our measure of the degrees of over- and under-confidence, where a 1sd increase in childhood confidence is associated with a 1.8pp (se = 0.6pp) increase in the likelihood that one works in STEM. These confidence gaps are large relative to the link between childhood math scores and later STEM employment, which is precisely estimated but very close to zero (Appendix Table B.12, column 9).

On the other hand, there are no gaps in the likelihood that over- or under-confident children work in non-STEM high-education occupations³⁸ (Table 2.5, columns 3 and 4). These results are reassuring for our empirical design: the fact that math confidence matters for STEM employment, but not other high-education employment, helps to validate that we properly isolate long-term associations with children's biased beliefs *in math*, rather than picking up correlations with unobserved self esteem or other abilities. Taking these point estimates at face value, about half of the under-confident children who do not pursue STEM careers switch into other high-education occupations, while the rest pursue other work. However, our 95-percent confidence intervals include estimates suggesting that under-confident children are up to 3.9pp less likely or up to 9.0pp more likely to work in other high-education occupations than their peers.

Next, we consider respondents' earnings. Our regression results are imprecisely estimated, but they broadly suggest that higher math confidence is associated with higher earnings later in life (Table 2.5, columns 5 and 6). While our binary measures of over- and under-confidence are not significantly associated with earnings (Panel A), a 1sd increase in the degree of childhood confidence is associated with 5.9 percent (se = 2.9 percent) higher earnings in adulthood. This gap is large relative to the association between childhood math scores and adult earnings: increasing test scores by one decile is associated with 7 percent higher earnings on average (Appendix Table B.12, column 11).

Finally, we consider unemployment. Our regressions suggest that higher confidence may be associated with lower unemployment risk (Table 2.5, columns 7 and 8). Again, our binary indicators

³⁸We define non-STEM high-education occupations as management, business, and financial occupations, legal occupations, education, training, and library occupations, and occupations that focus on writing and communication (a subset of media, arts, and entertainment occupations). We exclude health fields, as they are STEM-adjacent.

for over- and under-confidence are not significantly associated with unemployment (Panel A), but a 1sd increase in childhood confidence is associated with a 2.3pp (se = 0.9pp) lower likelihood of having been unemployed in the previous year. This gap is large relative to the association between childhood math scores and unemployment: increasing test scores by 10 percentiles is associated with 1.6pp lower unemployment risk on average (Appendix Table B.12, column 12).

While most of our results are quite stable – both in magnitude and precision – to the many robustness tests we run in Section 2.7, our results for earnings and unemployment should be interpreted with caution. They are only statistically significant when using our more continuous measure of biased beliefs, which is more vulnerable to measurement error, and we show in Section 2.7 below that they are not robust to using measures of confidence that minimize measurement error by using data from two waves of the CDS. That said, they are suggestive and are consistent with our other findings on the long-term links between childhood confidence and later-life outcomes.

2.7 Robustness

In this section, we show that our main results are robust to controlling for a range of possible confounding variables and to many alternate definitions of our key measures of biased beliefs.

2.7.1 Key confounders: Personality, adult investment, and school quality

First, we show that math over- and under-confidence predict long-run outcomes beyond their correlation with (1) more commonly-studied personality traits, (2) parent and teacher beliefs and investment, and (3) elementary school quality. We do not control for these variables in our main specifications because they are likely jointly determined with math confidence, but they may confound the links we estimate. See Appendix B.5 for more details on data used in this section.

Section 1 of Appendix Table B.13 adds controls for children’s Big-Five personality traits: conscientiousness, agreeableness, neuroticism, openness, and extroversion. The CDS did not administer standard psychometric scales to identify the Big-Five traits among children, so we construct

these measures from caregivers' reports of child behavior (see Appendices B.2 and B.5 for details.) These traits could confound the long-term associations that we observe: other work shows that Big-Five personality traits correlate with contemporaneous educational and employment outcomes (e.g. Almlund et al. 2011; Heckman et al. 2019), and we find some correlations between these common personality traits and measures of over- and under-confidence in our sample (Appendix Table B.14). However, our estimates of the links between over- and under-confidence and long-run outcomes are broadly robust to controlling for them.³⁹

Next, we add controls to our main specification for parent investments, like reading or doing homework with the child, teacher ratings of children's academic, social, and physical competence, and the educational attainment that parents and teachers predict for the child. Note that we only observe teacher perceptions for 20-34 percent of the sample. Teacher and parent beliefs and investments do correlate with children's beliefs in math in our sample (Appendix Table B.15). If these adults' investments affect children's later-life success, they may drive the links between childhood math confidence and later outcomes that we observe (Papageorge et al., 2018; Dizon-Ross, 2019). However, Section 2 of Appendix Table B.13 shows that children's over- and under-confidence continue to predict long-run outcomes in similar ways when we add controls for adult perceptions and investment to our main regressions.

Finally, we show that our results are robust to controlling for the quality of school a child attended when we first observe their biased beliefs in math. If children assess their own ability relative to their peers, not the national distribution, school quality may shape children's self-assessments in math; over-confident children could just be those with low-performing peers, for example. However, these patterns would tend to bias our results *towards zero*, since later-life outcomes may be worse for children from lower-performing schools. We use restricted data from the CDS to match students with data on the percent of students at their school who qualified for free or reduced-price lunch (a proxy for income), the average student-teacher ratio at their school (a proxy for educational inputs), and levels and trends of their school's mean achievement levels in math

³⁹Appendix Table B.20 shows the coefficients on the personality measures in this regression; they correlate with long-run outcomes in expected ways (Almlund et al., 2011).

and reading. Reassuringly, our results do not change meaningfully when we control for school quality (Appendix Table B.13, Section 3).

2.7.2 Alternate definitions of biased beliefs

Next, we show that our results are robust to a range of alternate measures of biased childhood beliefs in math. Appendix B.6 describes each of these alternate measures in more detail. None of these changes affects our main conclusions: that over- and under-confidence strongly and meaningfully predict long-term education and working in STEM.

Redefining over- and under- confidence: We first redefine our binary measures of over- and under-confidence by altering the CDS math score and self-report cutoffs on which they rely, making those designations more or less strict than our main measures. Second, we construct more data-driven measures of over- and under-confidence—what we refer to as the relative confidence measures—that identify over- and under-confident children as those in the tails of the distribution of math scores at each self-reported ability level. Finally, a third class of binary over- and under-confidence measures marks a child as over-confident if the degrees of confidence measure is greater than 2 and under-confident if it is less than -2.

Redefining degrees of confidence: Next, we also test robustness to the key design choice in our more continuous measure of confidence: how we map self-assessed ability and observed scores to the same scale. Our main measure assumes that children with accurate beliefs would report the numbered bin from 1 to 7 in which their CDS score falls when test score percentiles are uniformly distributed across 7 bins (i.e. each bin covers about 14 percentiles). We test robustness to two other transformations: the first assumes that children should have reported the bin from 1-7 in which their test score would fall if they had the CDS' empirical self-assessment distribution in mind, and the second instead differences children's percentiles of self-assessed ability and demonstrated ability. Each of these is converted to standard deviation units to facilitate comparisons.

Measurement error: To reduce the likelihood that our results are driven by measurement error, we also construct alternate confidence measures using testing and self-assessment data from two

waves of the CDS for the 60 percent of children with multiple measures. We take two approaches to redefining our indicators for over- and under-confidence. In the first, we average children's test scores and self-reported ability over two waves and then apply our standard cutoff rules to these average scores and self-reports. In the second, we calculate indicators for being over- or under-confident separately in each of two waves and then average these indicators. We use the same logic in defining multi-wave versions of the more continuous confidence measure.

Results: Appendix Figures B.4-B.16 present specification charts showing results for each of our main outcomes of interest using these alternate measures of biased beliefs.⁴⁰ For simplicity, Appendix Table B.16 presents a subset of these results: we iterate through alternate definitions of biased beliefs for each outcome, always using the control variables from our preferred specification. Panel A shows the results for over- and under-confidence, and Panel B shows the results for our more continuous measure of biased beliefs. Most coefficients that are statistically significant in our main results are remarkably stable, leaving our conclusions unchanged. The only exceptions are our results for earnings and unemployment, which disappear when we use the more continuous measure of biased beliefs based on two waves of the CDS.

2.8 Snowballing investment or persistent over- and under-confidence?

Childhood over- and under-confidence in math are associated with gaps in key educational and employment outcomes down the line, from adolescent math performance to career choices in young adulthood. As we outlined in Section 2.2, these confidence gaps could arise if over- and under-confidence shape children's own investment decisions or those of parents, teachers, or potential employers. In this section, we explore the dynamic patterns through which these confidence gaps open up and persist. On one hand, math confidence could produce investment gaps in childhood that in turn snowball through children's later education and occupational choices. On the other

⁴⁰Besides testing alternate confidence measures, the specification charts also show that our main results are robust to dropping children in the lowest and highest math score deciles from our sample. Across all confidence measures, children at the upper (lower) tail of the score distribution are mechanically most likely to be identified as under-confident (over-confident).

hand, childhood over- and under-confidence in math may persist into adulthood and directly affect choices and performance at each stage of life, conditional on past achievement.

This section explores whether biased beliefs persist into adulthood, and whether gaps in later-life outcomes can be fully accounted for by the links between confidence and intermediate investments that we observe.

2.8.1 The persistence of childhood confidence in math

While Table 2.1, discussed in Section 2.3.2, shows that over- and under-confidence in math persist from childhood into adolescence, we also find that childhood biased beliefs persist even until we last observe respondents in the TAS at ages 18 through 27 (Table 2.6). This persistence is a necessary condition for children's biased beliefs to have direct behavioral effects on their educational and career choices as they age. We use the wealth of questions in the TAS to construct measures of young adults' confidence in math and reading, generalized academic confidence, career confidence, and general confidence.⁴¹ See Appendix B.2 for more detail each of these measures.

First, we calculate an index of adult math confidence as the mean of standardized ratings of how good respondents think they would be in a job requiring math or technology. By this metric, childhood math confidence strongly persists into adulthood. Respondents who were over-confident in math as children score about 0.26sd (se = 0.06sd) higher in math confidence as adults than others with comparable childhood test scores, while under-confident children score about 0.25sd (se = 0.05sd) lower (Table 2.6, columns 1 and 2, Panel A). Likewise, a 1sd increase in our more continuous measure of childhood math confidence predicts 0.17sd (se = 0.02sd) higher math confidence as an adult (Panel B). In contrast, children who were under-confident in math score about 0.17sd (se = 0.05sd) higher in adult reading confidence—measured by standardizing subjects' ratings of how good they would be in a job requiring them to read and write a lot—than others with comparable

⁴¹Unlike our measures of biased beliefs from the CDS, these TAS confidence variables are not paired with measures of demonstrated ability in adulthood. However, the ideal regressions would test the links between childhood over- and under-confidence and *biases* in adult confidence, so as to avoid conflating the persistence of biased beliefs with the links between childhood confidence and adult achievement. We approximate this ideal by controlling for *adolescent* math and reading scores, digit span scores, and general confidence as proxies for adult ability.

childhood test scores (Table 2.6, columns 3 and 4). This pattern may arise because under-confident children are less likely to work in STEM occupations, making them more likely to have a job requiring reading and writing.

Next, childhood over- and under-confidence in math predict gaps in general academic confidence and career confidence in adulthood (Table 2.6, columns 5-8). Generalized academic confidence captures respondents' beliefs in their skill at solving problems, thinking logically, listening, and teaching others, and career confidence captures respondents' belief that they can attain and succeed in their dream job. Children who are over-confident in math score about 0.08sd (se = 0.05sd) higher in adult academic confidence and 0.11sd (se = 0.05sd) higher in adult career confidence than peers with comparable childhood test scores. Similarly, a 1sd increase in childhood math confidence predicts a 0.04sd (se = 0.02) increase in adult academic confidence and a 0.05sd (se = 0.02) increase in adult career confidence. While it is unsurprising that adult math, academic, and career confidence are correlated, it is reassuring that the links between continuous childhood and adult math confidence are 3-4 times as large as those with these other forms of adult confidence.

However, there are no significant relationships between childhood math confidence and a measure of adult general confidence (Table 2.6, columns 9-10), which captures respondents' conviction in their ability to lead and supervise, their independence and decisiveness, and their life's direction. Since these regressions control for childhood and adolescent general confidence, they suggest that while general confidence correlates with math confidence in childhood, childhood math confidence is not significantly linked with the *evolution* of general confidence as respondents age.⁴²

In sum, childhood over- and under-confidence in math persist through childhood and into young adulthood as confidence gaps across academic domains and in one's career. If these biased beliefs directly affect respondents' educational or employment success in adulthood, this persistence may be a key factor in the long-term economic associations that we observe.

⁴²As additional evidence that our results capture links with math confidence, not general self esteem or ability, we consider a set of placebo outcomes: individuals' relationship status, general mental health, social anxiety, alcohol consumption, and dangerous behavior as young adults (all from the TAS). We expect each of these outcomes to be affected by general self-esteem, but not by math over- and under-confidence specifically. Reassuringly, we generally find no relationships between biased beliefs in math and any of these placebo outcomes, except that math over-confidence predicts a lower likelihood of being in a romantic relationship (Appendix Table B.21).

2.8.2 Gaps in intermediate outcomes do not fully explain results

Despite the persistence of childhood confidence, the links we observe between childhood biased beliefs and later-life outcomes could still be fully explained by gaps in intermediate educational investments. In Figure 2.2, we explore the role of past investment by estimating the marginal relationships between childhood biased beliefs and later-life outcomes, conditional on all intermediate, observable outcomes along the chronological chain of education and entry into the labor market. We then compare these results to those from our baseline specification. If childhood biased beliefs continue to predict long-run gaps conditional on intermediate outcomes, these remaining gaps may be related to contemporaneous adult confidence. Of course, this analysis is imperfect, especially since we cannot control for all intermediate investments.

Figure 2.2 reproduces our baseline estimates (Tables 2.3, 2.4, and 2.5, even-numbered columns) for math over- and under-confidence in darker blue, while the lighter blue points present our estimates with controls for all outcomes that precede the outcome of interest. In particular, we re-examine educational outcomes through college holding fixed adolescent math and reading test scores, re-examine having a graduate degree and occupation choice holding fixed all previously-observed educational outcomes, and re-examine log earnings and unemployment history with controls for all educational outcomes and past occupation choices.

Many of the large confidence gaps we've observed in educational and employment outcomes persist when we condition on observable intermediate outcomes. Controlling for adolescent academic achievement does not change the relationship between childhood biased beliefs and any of our educational outcomes, and under-confidence remains half as predictive of working in STEM when we control for all educational outcomes, including whether respondents majored in STEM. Gaps in respondents' earnings fall by up to 60 percent when we condition on intermediate outcomes, though our standard errors remain large. The unemployment coefficients are largely unaffected when we add intermediate outcomes as controls.

Together with the persistence of math confidence into adulthood, these results suggest that over-

and under-confidence may continue to directly affect economic outcomes as respondents age.

2.9 Conclusion

In this paper, we identify over- and under-confidence in math among a large sample of children. In doing so, we are the first to show that even children have markedly biased beliefs about their own math ability. These beliefs are distinct from Big-Five personality traits and general confidence. Girls are less confident in math than boys with the same test scores and general confidence, so gender stereotypes about math may shape ability perceptions even at young ages.

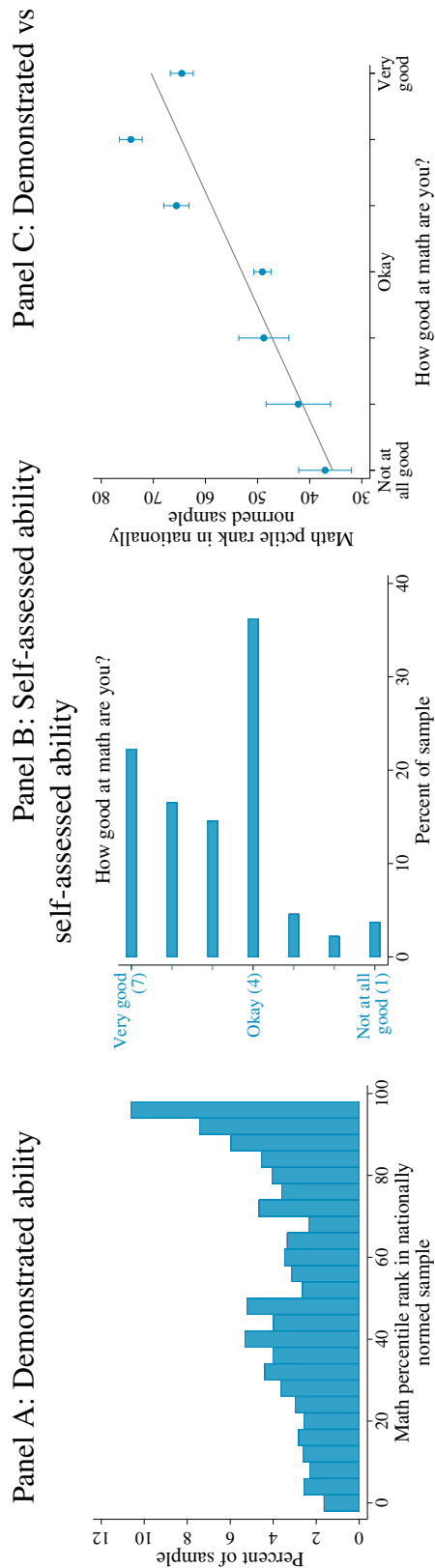
We then estimate striking associations between respondents' childhood over- and under-confidence in math and their educational and employment outcomes up to 22 years later, including comprehensive controls for children's demonstrated ability and family backgrounds. In the near term, under-confident children perform worse on the CDS math tests five years later, while over-confident children score higher. In the longer term, childhood math confidence significantly predicts key aspects of later education and work trajectories: whether respondents graduate from high-school and college, their college major and occupation choices, their earnings, and whether they experience unemployment. We do not observe similar associations with long-run outcomes for childhood confidence in reading, a puzzle that we leave for future work.

Our results suggest that biased beliefs about math ability in childhood may predict later-life outcomes both through accumulated differences in educational investments and by continuing to affect economic outcomes as respondents age. Childhood over- and under-confidence persist into adolescence and adulthood, and childhood confidence continues to broadly predict later-life outcomes, particularly in education, when we control for all observable educational and career investments along the chronological chain of education and labor-market entry.

While our results are not causal, they suggest that confidence in math may crucially shape the education we achieve and jobs we get, with effects possibly taking root as early as childhood. Our results provide key early evidence on the importance of math confidence, but they leave substantial

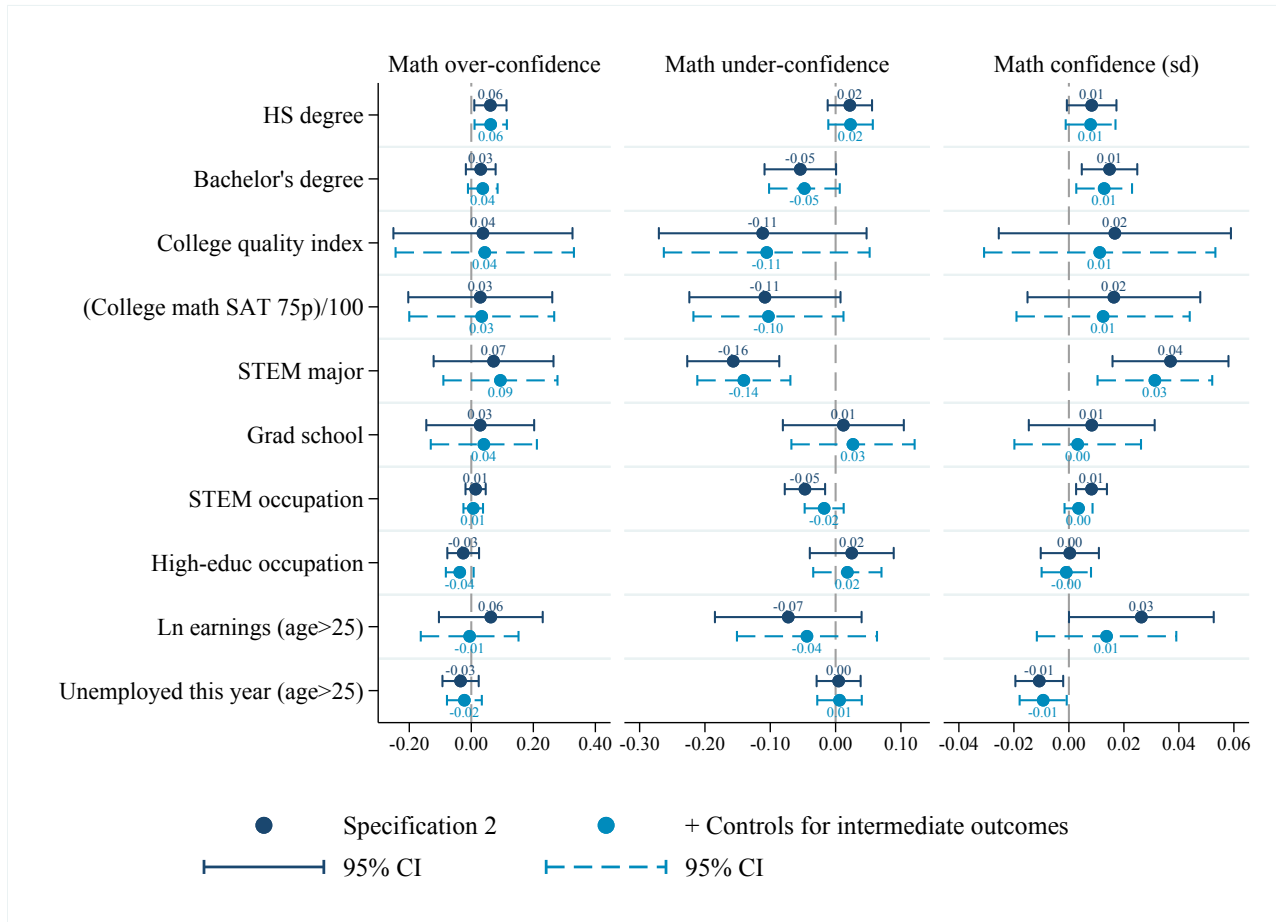
room for future exploration. Besides re-examining the associations we estimate for math over- and under-confidence in an experimental setting, research should explore the mechanisms by which childhood math confidence affects later-life outcomes. For example, do less confident children perform worse later because they get less encouragement from teachers, or do they simply choose to exert less effort at school? Next, we've seen that high-achievers with low confidence are less likely to work in STEM jobs; do they fare worse in job interviews for those positions, or do they simply not apply? Finally, if future research verifies that confidence causally affects later-life outcomes, what interventions can close those gaps?

Figure 2.1: Distributions of self-assessed and demonstrated ability



Note: We plot first-observed math test scores and self-assessments for the 2985 CDS respondents with at least one year of both measurements. We measure respondents' ability and self-beliefs in math at ages ranging from 8 to 19, though we observe the median child at 11 and more than 90% of children by age 13. Panel A plots the distribution of respondents' percentile ranks (calculated relative to a nationally-representative norming sample) on a portion of the Woodcock-Johnson Psycho-Educational Battery Revised (WJ-R) testing math reasoning and knowledge. Panel B plots the distribution of children's responses when asked "How good at math are you?" on a scale from 1 (not at all good) to 7 (very good). Finally, Panel C plots the average math percentile rank within each category from 1 to 7 of children's self-reported ability in math.

Figure 2.2: Controlling for intermediate outcomes



Note: This figure plots the coefficient on over- or under-confidence in our baseline specification (2) and the same coefficient when we add controls for mediating factors. When the outcome is high school or college graduation, majoring in STEM, or college quality, we add controls for adolescent math and reading test scores. When the outcome is earning a graduate degree, we add controls for all previously-observed education outcomes. When the outcome is occupation choice, we add controls for all observed education outcomes: math and reading scores in adolescence, whether the respondent graduated from high school, college, or graduate school, the 75th percentile of the math SAT score distribution of the college he or she attended, and whether he or she majored in STEM. When the outcome is earnings or unemployment, we add controls for all observed educational outcomes and occupational choice.

Table 2.1: The persistence of math over- and under-confidence

	(1)	(2)
Panel A: Math over-confidence	0.118***	0.116***
	(0.031)	(0.032)
N	1747	
Sample mean	0.041	
Panel B: Math under-confidence	0.042*	0.042*
	(0.024)	(0.025)
N	1747	
Sample mean	0.063	
Panel C: Math confidence (SD units)	0.182***	0.185***
	(0.025)	(0.025)
N	1747	
Sample mean	0.000	
Basic controls:	✓	✓
Added background controls:		✓

Notes: This table regresses adolescent confidence outcomes on various definitions of childhood math confidence with various controls. Adolescent confidence is measured five years after the childhood measurement. In each row, the dependent variable is the adolescent measurement of the independent variable described. The measures of over- and under-confidence are our main binary measures. Our secondary measure of degrees of confidence takes on values from -6 to 6 and persistence of that variable is shown in the third row. The fourth row standardizes the degrees of confidence measure to have mean 0 and standard deviation 1, to facilitate ease of interpretation. All controls that are time-variant are observed in the same year as the confidence measures. Basic controls include child gender, race, decile fixed effects for math and reading test percentile scores, digit span test scores, a general confidence index, family taxable income and its square, parent education, quarter-of-birth fixed effects, year-of-birth fixed effects, age at which confidence was measured fixed effects, and state fixed effects. We also include fixed effects for adolescent test score deciles in math and reading. Added background controls are parents' rating of child health, indicators for receiving government transfers, household structure, parenting practices, parent occupation, and parent mental health and confidence measures. All controls are recoded to zero if missing and we include a missing indicator. Standard errors are clustered by family, and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table 2.2: Demographic predictors of over- and under-confidence

	Over-confidence	Under-confidence	Confidence (sd)
<i>Demographic Characteristics</i>			
Female	-0.027*** (0.01)	0.023** (0.01)	-0.097*** (0.03)
Black	0.014 (0.02)	0.015 (0.02)	0.031 (0.04)
Hispanic	-0.038* (0.02)	0.019 (0.03)	-0.037 (0.07)
Asian or Native American	-0.021 (0.02)	0.024 (0.03)	-0.042 (0.06)
Only child	0.005 (0.02)	0.016 (0.02)	0.008 (0.05)
First child	0.015 (0.02)	0.038** (0.02)	-0.012 (0.04)
Second child	0.029* (0.02)	0.004 (0.02)	0.055 (0.04)
Father graduated high school	-0.011 (0.02)	-0.039 (0.02)	0.072 (0.06)
Father has bachelors	0.014 (0.01)	-0.009 (0.02)	0.024 (0.04)
Mother graduated high school	-0.019 (0.02)	-0.010 (0.02)	-0.013 (0.05)
Mother has bachelors	-0.007 (0.01)	-0.025 (0.03)	-0.024 (0.05)
Father works in STEM	0.004 (0.02)	-0.036 (0.03)	0.044 (0.05)
Mother works in STEM	-0.010 (0.01)	-0.006 (0.02)	-0.009 (0.04)
Father works in non-STEM high-educ	0.000 (0.01)	-0.008 (0.03)	0.022 (0.05)
Mother works in non-STEM high-educ	-0.017 (0.01)	0.019 (0.02)	-0.062* (0.04)
Family taxable income (thous 2016 USD)	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)

Table 2.2: Demographic predictors of over- and under-confidence (continued)

	Over-confidence	Under-confidence	Confidence (sd)
<i>Other Child Characteristics</i>			
Child ever in gifted prog	0.026** (0.01)	-0.087*** (0.02)	0.146*** (0.03)
Child ever in special ed prog	0.007 (0.02)	-0.008 (0.02)	0.071 (0.05)
Child has repeated grade	-0.016 (0.02)	-0.012 (0.01)	0.007 (0.05)
Parent's rating of child health	-0.001 (0.01)	-0.013** (0.01)	0.022 (0.01)
<i>School Quality Measures</i>			
Percent FRPL	-0.027 (0.03)	0.058* (0.03)	-0.074 (0.08)
Student-teacher ratio	0.000 (0.00)	-0.000 (0.00)	0.002*** (0.00)
Average math and reading achievement	-0.003 (0.00)	0.003 (0.00)	-0.015 (0.01)
Difference btwn math and reading achievement	0.010 (0.01)	-0.004 (0.02)	0.030 (0.04)
Cohort slope of average achievement	-0.036 (0.06)	0.079 (0.06)	-0.228 (0.15)
Unable to link to NCES id	0.049* (0.03)	0.013 (0.03)	0.109 (0.07)
<i>Other Child Ability Measures</i>			
Digit span score	-0.000 (0.00)	-0.004** (0.00)	0.009** (0.00)
General confidence	0.038*** (0.01)	-0.054*** (0.01)	0.211*** (0.02)
Mean of dependent variable	0.085	0.121	0.000
N	2985	2985	2985
R-squared	0.21	0.21	0.57

Notes: Each column regresses a measure of childhood biased beliefs in math on child characteristics. In columns 1 and 2, the dependent variable is our main indicator for over-confidence or under-confidence, respectively. In column 3, the dependent variable is a linear measure of biased beliefs that ranges from -6 to 6, where negative values represent under-confidence, *which has been standardized to have mean 0 and standard deviation one in our sample*. All variables are taken from the first year in which we observe the child's confidence in math. Additional controls include fixed effects for math and reading test score deciles, birth year, birth quarter, state, and age at which confidence was measured fixed effects. The coefficients on the ability deciles are shown in Appendix Table B.11. All controls are recoded to be zero if missing and the regressions include missing indicators for each variable (not shown). All variables are either continuous or binary indicators, except for child race and birth order. The omitted category for race is non-Hispanic whites, and the omitted category for birth order is any birth order higher than two. Standard errors are clustered by family. *, **, and *** indicate significance at the 10, 5, and 1 percent level, respectively.

Table 2.3: Childhood math confidence and medium-term educational achievement and attainment

Dependent variable:	Adolescent math scores		Adolescent reading scores		High school degree		College degree	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	2.637*	2.666*	-0.362	-0.286	0.057**	0.062**	0.027	0.031
	(1.468)	(1.496)	(1.381)	(1.385)	(0.026)	(0.026)	(0.024)	(0.024)
Under-confidence	-5.705***	-5.860***	0.353	0.162	0.015	0.022	-0.057**	-0.058**
	(1.482)	(1.497)	(1.439)	(1.452)	(0.017)	(0.017)	(0.028)	(0.028)
N	1747	1747	1745	1745	2714	2714	2725	2725
OC = -1*UC? p-value:	0.147	0.138	0.997	0.951	0.022	0.008	0.413	0.457
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	2.806***	2.827***	0.111	0.128	0.019*	0.018*	0.032***	0.033***
	(0.566)	(0.569)	(0.587)	(0.580)	(0.010)	(0.010)	(0.011)	(0.011)
N	1747	1747	1745	1745	2714	2714	2725	2725
Sample mean of dep. var.	50.808		48.231		0.876		0.297	
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses educational achievement and attainment outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table 2.4: Childhood math confidence and college quality, college major choice, and post-college schooling

Dependent variable:	College quality index		College's 75th pctile math SAT score		STEM Major		Graduate degree	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	0.067 (0.145)	0.037 (0.148)	6.244 (12.112)	3.133 (11.829)	0.067 (0.096)	0.076 (0.097)	0.035 (0.083)	0.032 (0.087)
Under-confidence	-0.095 (0.081)	-0.127 (0.082)	-9.312 (5.958)	-11.312* (5.925)	-0.158*** (0.036)	-0.162*** (0.036)	0.014 (0.046)	0.006 (0.048)
N	1107	1107	1117	1117	736	736	810	810
OC = -1*UC? <i>p-value:</i>	0.866	0.601	0.819	0.537	0.365	0.405	0.607	0.704
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	0.044 (0.047)	0.041 (0.046)	4.198 (3.460)	3.631 (3.417)	0.077*** (0.023)	0.078*** (0.023)	0.023 (0.025)	0.022 (0.025)
N	1107	1107	1117	1117	736	736	810	810
Sample mean of dep. var.	0.053		594.172		0.189		0.200	
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses college outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table 2.5: Childhood math confidence and employment outcomes

Dependent variable:	Works in STEM		Non-STEM high-educ occ.		Ln(Earnings)		Unemployed this year	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	0.011 (0.016)	0.014 (0.017)	-0.019 (0.025)	-0.025 (0.026)	0.045 (0.085)	0.064 (0.085)	-0.034 (0.030)	-0.035 (0.030)
Under-confidence	-0.049*** (0.016)	-0.049*** (0.016)	0.029 (0.033)	0.026 (0.033)	-0.067 (0.056)	-0.075 (0.057)	0.006 (0.017)	0.005 (0.017)
N	4592	4592	4592	4592	4423	4423	4975	4975
OC = -1*UC? p-value:	0.096	0.127	0.822	0.987	0.833	0.917	0.437	0.395
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	0.018*** (0.006)	0.018*** (0.006)	-0.001 (0.011)	0.001 (0.012)	0.049* (0.028)	0.059** (0.029)	-0.023** (0.009)	-0.023** (0.009)
N	4592	4592	4592	4592	4423	4423	4975	4975
Sample mean of dep. var.	0.046	0.163	0.163	10.185	0.167	0.167	0.167	0.167
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses employment outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Basic controls also include year fixed effects when the outcome is observed in a panel. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table 2.6: Childhood math confidence and young adult confidence outcomes

Dependent variable:	Math confidence		Reading confidence		Academic confidence		Career confidence		General confidence	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>										
Over-confidence	0.264*** (0.058)	0.260*** (0.058)	0.002 (0.067)	-0.002 (0.067)	0.090* (0.047)	0.086* (0.048)	0.103** (0.050)	0.106** (0.050)	0.056 (0.043)	0.056 (0.043)
Under-confidence	-0.250*** (0.047)	-0.251*** (0.048)	0.159*** (0.053)	0.163*** (0.054)	-0.038 (0.033)	-0.041 (0.033)	-0.058 (0.038)	-0.055 (0.038)	0.002 (0.031)	0.000 (0.031)
N	6632	6632	6634	6634	8096	8096	6265	6265	8050	8050
OC = -1*UC? p-value:	0.850	0.904	0.064	0.062	0.362	0.441	0.487	0.418	0.268	0.289
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>										
Confidence	0.167*** (0.021)	0.168*** (0.021)	-0.046* (0.026)	-0.048* (0.026)	0.037** (0.017)	0.038** (0.017)	0.055*** (0.018)	0.055*** (0.018)	0.020 (0.016)	0.022 (0.016)
N	6632	6632	6634	6634	8096	8096	6265	6265	8050	8050
Sample mean of dep. var.	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.009	0.000	0.000	0.000
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table regresses young adult confidence outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Basic controls also include year fixed effects when the outcome is observed in a panel. In this table, we add controls for adolescent test score deciles in math and reading, as well as adolescent general confidence and digit span scores in all specifications. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Chapter 3

The Narrative of Policy Change: Fiction Builds Political Efficacy and Climate Action

With Lucy Page and James Walsh

Abstract

Can fictional narratives contribute to building political momentum? In an online experiment ($N \approx 6,000$), learning about the Inflation Reduction Act (IRA) strengthens beliefs about government responsiveness to citizen action by only 0.07sd. Watching a short, fictional story about political climate advocacy as a loose backstory to the IRA yields much larger effects on beliefs (0.5sd). While IRA information alone does not affect climate advocacy, the story increases information-gathering about climate marches by 54 percent and donations to lobbying organizations by 19 percent. We show evidence that beliefs and emotions may drive this effect.

We are grateful to Abhijit Banerjee, Esther Duflo, Karla Hoff, Rohini Pande, and Frank Schilbach for their advice and helpful comments. Page and Ruebeck are supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1745302. This project is also supported by the George and Obie Shultz Fund at MIT and the Strengthening American Democracy Program at Beyond Conflict. The pre-registration for this experiment can be found here: <https://www.socialscienceregistry.org/trials/10351>. IRB approval for the project was obtained from the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (Protocol 2208000715).

3.1 Introduction

Concern about climate change is widespread in the US: about two-thirds of Americans report that they are at least somewhat worried about global warming, and over 60% support a range of policies to reduce greenhouse gas emissions. Moreover, 28% of registered voters say they would be willing to contact government officials about climate change. However, few Americans follow through on doing so: only 8% of registered voters say they contacted government officials about global warming in the last year (Leiserowitz et al., 2021).

Longstanding research in psychology and political science suggests that weak political efficacy—the belief that government responds to citizen demands—is a key barrier to political engagement on climate change and other issues.¹ In a survey of 500 young adults fielded on Prolific in June 2022, the most common reason cited for why respondents had not previously pushed for policy change was that it would make no difference (Appendix Figure C.1, Panel A). In this study’s baseline survey, collected in November 2022 through March 2023, only 18% of participants at least somewhat agreed that when groups of citizens push for policy on issues like climate change, the US government responds to their demands (Appendix Figure C.1, Panel B).²

This randomized experiment examines how two interventions aimed at building political efficacy affect subsequent climate action. The first intervention informs participants about the real-world policy progress of the Inflation Reduction Act (IRA), passed in August 2022 as the largest climate bill in US history (Bistline et al., 2023). Our second intervention pairs this information with an explicitly fictional, animated story linking this policy change to citizen advocacy. In this

¹Political scientists distinguish between *external* political efficacy—beliefs about how government responds to citizen demands—and *internal* political efficacy—beliefs about one’s own ability to engage with political processes (e.g. Campbell et al., 1954; Balch, 1974; Niemi et al., 1991; Craig et al., 1990; Scotto et al., 2021). We focus throughout the paper on external political efficacy; for brevity, we refer to it as “political efficacy.” A lengthy literature documents correlations between political efficacy and engagement (e.g. Shaffer, 1981; Abramson and Aldrich, 1982; Finkel, 1985). Political-efficacy beliefs are also related to the social-cognitive concept of collective efficacy: beliefs in a group’s ability to accomplish shared goals (Bandura, 2000).

²While our focus is not the impacts of citizen advocacy on government action, experimental work in subnational contexts finds that both citizen contacts (Bergan, 2009; Bergan and Cole, 2015) and providing information on constituents’ opinions (Butler and Nickerson, 2011) can shift legislators’ votes.

5-minute video, a young woman—devastated by her dog’s death from heatstroke—mobilizes a climate march that attracts national media attention and contributes to policy change.

We conducted our study via three surveys fielded on Prolific, a paid online survey platform, in the six months following the IRA’s passage. From an initial screening survey, we recruited about 6,000 Americans—all of whom believe that climate change is human-caused and were unaware of the IRA’s recent advances—to complete a main survey in which we implemented our treatments and measured political efficacy and costly climate action. Finally, 85% of the sample took an obfuscated follow-up survey with additional outcome measures 1-4 days later, allowing us to estimate treatment effects with little or no experimenter demand and with a moderate delay (Haaland and Roth, 2020, 2023; Settele, 2022).

Learning about the IRA’s real-world policy advance yields small increases in political efficacy (0.07sd) and no effects on climate action. The fictional story, in contrast, has striking effects: it increases political efficacy by an additional 0.5sd, increases donations to climate-lobbying groups by 19%, and makes participants 54% more likely to seek information on nearby climate marches, though it has no detectable effects on efforts to email Congress. The story’s effects persist strongly in the obfuscated follow-up survey.

The story’s impacts on climate action appear to arise both through its effects on efficacy beliefs and its emotional resonance. The story had a range of emotional effects, strongly increasing feelings of hope or strength (0.52sd) and motivation (0.60sd) as well as making participants feel less anxious, sadder, more connected to others, angrier, and less anxious. In suggestive mediation analysis, the story’s treatment effects fall substantially when we control for either efficacy beliefs or motivation-related emotions, with the largest drops when we control for both of these possible mediators. The story does not seem to affect action by changing participants’ beliefs about Americans’ support for or engagement in climate action or by improving recall of the IRA information.

This paper contributes to several literatures in economics, political science, and psychology. First, we add to the large literature on the effect of narratives on social and economic outcomes (Jensen and Oster, 2009; Paluck, 2009; La Ferrara et al., 2012; Kearney and Levine, 2015; Shiller,

2017; Banerjee et al., 2019b,a; Kearney and Levine, 2019; Hoff et al., 2021; Riley, 2022; Walsh et al., 2022). We add to this research in three respects.

Most importantly, we show that a fictional story can increase contributions to a public good and collective action, whereas the existing literature on stories targets behavior with direct private benefits, such as personal health and educational investment. Thus, narratives may be a useful tool to drive efficient mobilization towards common goals. Stories may in fact be particularly useful in promoting behavior with primarily public benefit, like political engagement (Riker and Ordeshook, 1968; Feddersen, 2004; Feddersen and Sandroni, 2006; Fowler, 2006), in which the emotional or self-image returns of doing one’s part are primary drivers of action (Bryan et al., 2011). Second, we show that even low-budget, simple stories can have meaningful effects on political beliefs and behavior. Namely, the effects of fictional narratives embedded in commercial entertainment—“edutainment”—may be explained by other features like celebrities, popular songs, or mass distribution. In contrast, our story is watched in isolation during a survey experiment, is five minutes long, was produced for \$11,000, and was written by this paper’s authors, all of which mitigate these possible confounds. Finally, we find that the climate-action story has large effects both on participants’ causal narratives of policy change and on their emotions, both of which seem to contribute to the story’s effects on climate action. This finding builds on recent theoretical work focusing on how narratives compliment pure information as persuasive tools (Eliaz and Spiegler, 2020; Schwartzstein and Sunderam, 2021; Kendall and Charles, 2022) and the effects of emotions on preferences and decision-making (Elster, 1998; Loewenstein, 2000; Lerner et al., 2015).

Next, we contribute to large political science and environmental psychology literatures on political efficacy. We show for the first time that seeing real-world policy change builds political efficacy, and our short, fictional story about a young climate advocate has effects more than four times as large. These impacts contrast sharply with prior work testing a range of light-touch interventions aiming to build political efficacy around climate change, with limited success (Feldman and Hart, 2016; Hart and Feldman, 2016; Hornsey and Fielding, 2016; Jugert et al., 2016; Xue

et al., 2016; Hamann and Reese, 2020; Angill-Williams and Davis, 2021; Ettinger et al., 2021; Hornsey et al., 2021)

Finally, we add to the growing literature on the drivers of support for climate policy and climate action (e.g. Drews and van den Bergh, 2016; Andre et al., 2022; Dechezlepretre et al., 2022; Bernard et al., 2023). To our knowledge, this project is the first experimental work testing ways to build political climate advocacy. Prior work on climate action focuses on donation outcomes (e.g. Andre et al., 2022) and consumer choices (e.g. Allcott, 2011; Ho and Page, 2023).

The paper proceeds as follows: Section 3.2 describes our experimental design, Section 3.3 presents our results, and Section 3.4 concludes.

3.2 Research design

Appendix Figure C.2 depicts the study procedure. The study unfolds over three surveys: a screening survey (Section 3.2.1), the main survey in which we implement randomized treatments (Section 3.2.2), and an obfuscated follow-up survey (Section 3.2.3).

3.2.1 Sample selection

We recruited a sample of American adults via a 1-minute screening survey on Prolific³ and used two questions to screen participants for the experimental survey. First, participants were only eligible if they answered “No” or “I don’t know” when asked whether, to their knowledge, the US government had made substantial progress on climate change so far during 2022. Second, participants were only eligible if they answered that climate change is mostly human-caused when asked if it is mostly human-caused, caused mostly by natural changes in the environment, not happening, or other. Together, these restrictions allows us to identify participants who likely support the goals

³Participants were recruited to the study in two “waves,” first in November 2022 and again in January 2023. We paused the study due to concerns that proximity to the 2022 midterm elections could affect our results. Our main specifications control for the wave in which a participant completed the survey, and Appendix C.2 disaggregates results by wave.

of climate policy but are unaware of the IRA and its implications.⁴

Of 13,361 participants who completed the screening survey, 8,591 (64%) met these restrictions. We recontacted all qualifying participants, of whom 6,329 participants consented to the main survey and 6,015 completed it. We then exclude 122 participants who failed at least one of two attention checks embedded in the main survey; the first asked participants to select a certain multiple choice answer, while the second asked them to move a 100-point slider to within a 10-point range.

Appendix Table C.1 presents summary statistics for our final sample of 5,879 participants. We stratify recruitment on gender and whether participants are above or below 35 years old; our final sample is 53% female, with an average age of 37. Our sample is predominantly white (74%) and liberal: about 59% identify as Democrats, 28% as Independents, and 9% as Republicans. The sample's baseline political activity broadly matches nationally representative surveys: 25% say that they've contacted elected representatives in the last two years, while 23% of a [Pew Research Center \(2018\)](#) Pew Research sample reported having done so in the last year. Participants are also highly concerned about climate change (Appendix Figure C.3): 85% place themselves at 5 or higher on a 7-point scale of climate worry, and when asked how much they want the federal government to do on climate change, 78% place themselves at 6 or 7 on a scale from 1 (Much less than currently) to 7 (Much more than currently).

3.2.2 Experimental survey

We recontacted all qualifying participants via Prolific to take the main, experimental survey, during which we administered our treatments and measured key outcomes.

⁴Our goal was not to isolate those who had never heard of the IRA, a high-profile bill with extensive media coverage, but rather to identify those who are unaware of the bill's importance in US climate policy. Indeed, 49% of our final sample selected that they had heard of the IRA among a list of four recent bills.

3.2.2.1 IRA information randomization

All participants begin the experimental survey by watching a baseline video (available [here](#)) with visual information on global temperature rise, the Paris Agreement’s goal of limiting warming to 1.5° C, and the speed of global emissions reductions required to meet that goal.

IRA information treatment: Two-thirds of participants are then randomized to watch the IRA information treatment video (available [here](#)). This video highlights the US 2030 Paris commitment and visually plots projected emissions under policies as of February 2022, which would fall only halfway to the 2030 goal. The video then introduces the IRA as a major legislative advance after years of advocacy, explains the magnitude of the bill’s spending, and summarizes its climate provisions. The video plots projected emissions cuts under the IRA—then estimated to achieve 65% of the remaining cuts required to reach the 2030 target (Jenkins et al., 2022)—and ends with the following: “That means that the IRA takes a big step towards US emission commitments, but we still need to make major additional emissions cuts by 2030 to meet our Paris goal and limit catastrophic warming.”

Basic control video: We randomize half of the remaining participants to watch a “basic-control” video (available [here](#)) that exactly reproduces all information and visuals in the IRA treatment *other than information about the IRA itself*. Thus, we control for any effect that essential context on US climate goals and business-as-usual emissions could have on climate action. After presenting projected emissions under February-2022 policies, this video ends with an adaptation of the IRA treatment video’s final sentence: “From this baseline, we would still need to make major emissions cuts by 2030 to meet our Paris goal and limit catastrophic warming.”

Extended control video: While the basic-control video exactly reproduces the beginning of the IRA treatment video, it is 60 seconds shorter. To eliminate concerns that treatment effects arise just from this additional content, we randomize half of the remaining participants to an “extended-control” video (available [here](#)) that adds 60 seconds of filler detail⁵ to the basic-control video. This

⁵This information describes countries’ nationally-determined contributions under the Paris Agreement, the units in which green-

video closes with the same statement as the basic-control video.

3.2.2.2 Fictional climate-advocacy story

Half of those who watched the IRA information video were randomly assigned to subsequently watch a 5-minute fictional, animated story about citizen climate advocacy (available [here](#)). The script was written by the authors of this paper, narrated by professional voice actors, and illustrated, animated, and set to music by a UK-based animation company for a total budget of about \$11,000. See Appendix C.4 for details.

The story centers on a young woman named Annie whose dog, Gilbert, dies in a heatwave. Following Gilbert's death, Annie is angry and hopeless about government progress on climate change. She encounters an elderly man organizing a climate march, and he convinces her that living in a democracy means that citizens can demand change, and that historical movements (e.g. for women's suffrage and civil rights) advanced through collective citizen advocacy. Annie decides to fight for change and begins recruiting people for the march. Thousands show up to march for Gilbert. Annie speaks to a newscaster at the march, and her interview is broadcast across the country. The story ties the climate march to passage of a climate bill, saying that it was part of a movement all over the country that finally forced government action. While the story never explicitly mentions the IRA, it operates as a loose, fictional backstory to policy progress. The story concludes by saying that if we and others around the country don't give up, the government may keep hearing our demands.

Story-duration control: To ensure that the story's effects do not derive just from a longer survey, we cross-randomized half of all participants not assigned to watch the climate-advocacy story to answer filler questions paced by timers to also take five minutes. All results control for whether participants answered these extra questions.⁶

house gases are measured, when the US issued its most recent Paris commitments, example policies that could help achieve US commitments (the same components attributed to the IRA in the treatment video), and a precise numeric statement about how much emissions are expected to fall under February-2022 policies (matching the numeric precision of the IRA treatment).

⁶We used two different sets of filler questions; details are in Appendix C.5. The first wave used open-ended questions that intentionally primed some of the story's themes; in the second, participants took a general science-knowledge quiz. We changed

3.2.3 Experimental fidelity

Attention. In addition to screening the sample with two attention checks (Section 3.2.1), we incentivized attention to our treatments: ahead of each video, participants were told that 10 randomly-selected participants would earn \$5 for correct answers on each of 3 to 7 subsequent comprehension questions (described in Appendix C.6). Overall, participants answered 86% of comprehension questions correctly. Finally, receiving the IRA information substantially increased participants' knowledge of the IRA elicited at the end of the experimental survey (Appendix Table C.2).

Balance. Our sample is largely balanced across treatment conditions (Appendix Table C.1). The exception is that those assigned to receive IRA information, with or without the story, have higher baseline political engagement. Our main specifications control for dummies for each past political behavior, and our results are robust to controlling for a political-engagement index.

Attrition. In total, 95% of those randomized to a treatment status finished the experimental survey and are included in our sample. Those assigned to watch the fictional story are 2pp less likely to finish the experimental survey (Appendix Table C.3), but our main results are robust to Lee (2009) bounding (Appendix Tables C.4 and C.5).

Demand effects. To ensure that our results do not arise from experimenter demand effects, we elicited additional measures of our main outcomes in an “obfuscated” follow-up survey that participants did not know was connected with the previous surveys. Thus, any treatment effects we observe on follow-up outcomes are free of experimenter demand effects (Haaland and Roth 2020, 2023; Settele, 2022). The follow-up survey was advertised under a different researcher's account and described as being about political activity in general rather than climate change, and no participants indicated that they connected the obfuscated follow-up survey with the earlier surveys (see Appendix C.3 for details). 85% of those who finished the experimental survey complete the obfuscated follow-up survey, with no differential completion by treatment conditional on finishing

the design of these questions between the two waves—and updated our pre-analysis plan accordingly—to more cleanly control just for duration. The filler version that cleanly controls for duration has no impacts on our main outcomes of interest (Appendix Tables C.10 and C.11).

the main survey (Appendix Table C.3).

3.2.4 Main outcomes

3.2.4.1 Political efficacy

We elicit both qualitative and quantitative measures of political efficacy in the experimental survey; we detail these and all other outcomes in Appendix C.7. The qualitative measures elicit participants' agreement from 1 (Strongly disagree) to 7 (Strongly agree) with three statements about the role of citizens in climate policy, adapted from Craig et al. (1990). Next, we develop a quantitative measure of political efficacy by asking participants to estimate the probability that a hypothetical climate bill would pass if it was introduced to Congress in the next few months, separately if 2% or 10% of Americans contacted their national representatives to support it. The difference between participants' guesses in each of these cases provides a numeric measure of external collective efficacy: the impact of additional citizen pressure on government action.

In the obfuscated follow-up survey, we measure political efficacy by asking participants to rate their agreement from 0 (Disagree completely) to 7 (Agree extremely strongly) with the statement that "Citizen movements on issues like gun control and climate can make real change." We also ask participants to rate how effective they think (1) marches or rallies and (2) contacting politicians by phone or email are in affecting government policy, from 1 (Not effective at all) to 6 (Extremely effective).⁷

3.2.4.2 Climate action

Donations to climate advocacy organizations. We observe real-stakes donations to climate-advocacy organizations during both the experimental and follow-up surveys. During the experimental survey, we tell participants that we will randomly choose one participant to win an \$80

⁷Note that we added these questions to the follow-up survey shortly after beginning data collection (added as secondary outcomes in an amendment to our pre-analysis plan), so we observe them for only 78% of those who took the obfuscated follow-up survey and 66% of the total sample.

bonus and allow them to earmark any portion of that bonus to one of three policy-oriented climate advocacy organizations in the case that they are chosen. We observe whether and how much participants choose to donate.

In the obfuscated follow-up, participants similarly distribute a \$100 bonus—which one participant will win—between take-home money and donations to advocacy organizations lobbying for environmental policy, abortion access, gun control, and free-market policy. We frame these donation choices as opportunities to advocate for policy change by supporting effective lobbying groups.

Direct citizen advocacy. We also observe participants’ engagement with direct citizen advocacy. During the experimental survey, we offer participants an opportunity to email Congress about climate change via a portal hosted by an NGO. We observe whether participants opt in to the process of writing a letter, whether they compose a custom email to Congress,⁸ and whether they click a link to the portal from which to send the email.

Halfway through data collection, we added an additional outcome to the experimental survey to capture interest in participating in a climate march, since the story centers so heavily on this type of action.⁹ We observe whether participants click a link to a map of upcoming climate marches published by Fridays for Future, a decentralized group that organizes climate marches around the world.

Finally, we observe in the follow-up survey whether participants download “Call the Halls,” a guide to contacting legislators that we suggest they read and share with others.

⁸The portal includes a form letter, so participants do not need to write out a personalized message in order to later send an email. Note that we use these indirect measures of whether participants email Congress, rather than having them send emails directly from our survey, to protect Prolific participants’ anonymity.

⁹We introduced this secondary outcome in an amendment to our pre-registration posted on January 11, 2023 before starting our second round of data collection (see Footnote 3 and Appendix C.2). We elicit this outcome after the other main outcomes to avoid contaminating their interpretation.

3.3 Results

3.3.1 Specifications

We estimate the impacts of the IRA information and fictional story in the following specification:

$$Y_i = \alpha_0 + \beta_1 IRAInfo_i + \beta_2 Story_i + A^T X_i + \varepsilon_i \quad (3.1)$$

where Y_i is our outcome of interest, $IRAInfo_i$ indicates watching the IRA information video (i.e. being in either the T1 or T2 treatment groups in Appendix Figure C.2), and $Story_i$ indicates also watching the fictional story video (i.e. being in the T2 treatment group). X_i is a vector of controls and ε_i is an individual-specific error term. This specification pools the basic and extended control arms, which are rarely statistically or economically distinguishable, as the omitted group. Appendix Figures C.4 through C.15 show that our results are robust to estimating treatment effects relative to either control.

Our main specifications control for demographics, climate worry, desire for additional government climate policy, baseline political efficacy, baseline political engagement, and indicators for whether participants were assigned to the 5-minute filler questions and participated in the first or second wave of data collection. Demographics include sex, 5-year age bins, ethnicity, indicators for having a 4-year college degree interacted with indicators for being over age 25, and political affiliation. Appendix C.7 describes these controls in detail, and Appendix Figures C.4 through C.15 show that our results do not change with any choice of controls.

3.3.2 Political efficacy

While learning about the actual policy progress of the IRA somewhat increases political efficacy, also watching the fictional climate-advocacy story yields much larger effects (Table 3.1). Learning about the IRA increases participants' agreement that the US government responds to citizen de-

mands for policy change by 0.11sd and the index of overall external political efficacy by 0.07sd. In contrast, the story affects all three political efficacy statements by between 0.36 and 0.42sd and increases the overall political efficacy index by 0.51sd. Note that the dependent variables in columns 1 and 2 are agreement with *negative* efficacy statements, which are flipped when added to the index. The story also increases the quantitative measure of political efficacy (column 5): watching the climate story increases participants' beliefs about the effect of an additional 8pp (from 2 to 10%) of Americans calling to support a climate bill on the likelihood that Congress would pass it by 0.9pp, a 10 percent increase over the control mean. Learning about the IRA does not affect this measure.¹⁰

The IRA information's relatively small effects on political efficacy do not persist in the obfuscated follow-up, but the large impacts of the story remain (columns 6-9). The story increases agreement that citizen movements can make real change, beliefs that marches or rallies and contacting Congress are effective in changing government policy, and an index of these measures by 0.23sd, 0.16sd, 0.11sd, and 0.2sd, respectively. In addition to eliminating any demand effects, these results show that the story persistently changes beliefs at least in the short term and that participants substantially extrapolate the story's emphasis on marches to other forms of advocacy – contacting Congress by phone or email – that it did not highlight.

3.3.3 Climate action

Learning about the IRA has no impact on climate action, but the fictional story substantially increases participants' interest in climate marches and climate-advocacy donations in both the main and follow-up surveys (Table 3.2).

Donations to climate advocacy organizations. Learning about the IRA has no effect on climate donations in either the main (columns 1 and 2) or follow-up survey (columns 3 and 4). In contrast, participants who watched the climate-advocacy story are 5pp more likely to donate

¹⁰Appendix Table C.12 separates this result into treatment effects on the likelihood of passing a climate bill if 2% or 10% of Americans contacted Congress to support it, alongside effects on participants' beliefs about the probability that we will meet key national and global climate goals.

to a climate organization in the experimental survey, a 10% increase relative to the control group, and donate \$2.88 more overall, a 19% increase over the average control donation of \$14.94 of a possible \$80. The story had similar effects on donations during the obfuscated follow-up. Those who watch the story are 6pp more likely to donate to climate advocacy, a 13% increase, and donate on average \$1.41 more, a 16% increase over the average control donation of \$8.55 of a possible \$100. Notably, these higher climate donations do not crowd out donations to other causes in the follow-up. The story increases total donations by \$3.02, an effect that is twice as large as that on donations to climate advocacy alone (Appendix Table C.6).¹¹

Citizen advocacy. In contrast, the climate-advocacy story has only narrow effects on interest and engagement in *personal* climate advocacy (Table 3.2, columns 5-9). Neither the IRA information nor the story affect whether participants opt into the letter-writing process, write a custom letter, or click to the portal to send the letter. With 95% certainty, we can rule out that the story made participants more than 2.3pp more or less likely likely to click to the portal, though this range is fairly wide relative to the control mean of 15%. On the other hand, the story does have large effects on participants' revealed interest in climate marches, the form of advocacy it portrays. Participants who watch the story are 4.3pp more likely to click the link to Fridays for Future, a 54% increase relative to the control group.

Neither the IRA information nor the story has a detectable effect on whether participants download the "Call the Halls" guide in the follow-up survey (column 9). With 95% confidence, we can rule out that the story made participants more than 4.9pp more likely or 1.3pp less likely to download the guide (relative to a control mean of 21%).

3.3.4 Mechanisms

While the climate-advocacy story's effects on political efficacy could underlie its impacts on action, other mechanisms could also explain these effects. This section explores additional secondary

¹¹Point estimates suggest that the fictional story comparably increased donations across all of the other causes, though only its impacts on donations to the free-market lobbying group are statistically significant. The story's impacts on donations to the climate organization are twice as large as on donations to any other cause (Appendix Table C.6).

outcomes collected in the main and follow-up surveys to understand the processes through which the story drives action.

3.3.4.1 Emotions

First, the story may drive action through its impacts on emotion. We elicited participants' emotions immediately after the treatment videos in the experimental survey, providing them with three blanks and asking them to list at least one emotion they were currently feeling. Participants then rated how strongly they felt each emotion they listed. Two authors hand-coded these free-response emotions into categories from a treatment-blind list, generating the classification scheme detailed in Appendix C.7. Figure 3.1 plots the impacts of each treatment on standardized measures of how strongly participants felt each emotional category. Note that because we elicited emotions before participants are offered the chance to take action, any impacts on emotion are not due to action itself.

Both learning about the IRA and the climate-advocacy story had sizable effects on participants' emotions, with especially stark effects from the story. Panel A explores the emotional spectrum of motivation versus apathy. While both the IRA information and story substantially increase participants' reports of feelings of hope or strength and reduce expressions of pessimism, the story also increases feelings of motivation (0.6sd) and reduces apathy or fatigue (0.2sd). Turning to other positive and negative emotions in Panels B and C, we find that learning about the IRA increases happiness, peacefulness and connectedness, while reducing sadness, anger, and anxiety. While the story also increases feelings of connectedness, its other emotional effects diverge starkly from those of the IRA: participants feel *less* peaceful, much sadder, and angrier. At the same time, the story sharply reduces feelings of anxiety and doubt.

The disparate emotional effects of the story and IRA information are largely consistent with the story's much larger effects on climate action. Unlike the IRA information, the story pushes participants towards feelings like anger and motivation that have been shown to increase political interest and engagement (Brader, 2005; Valentino et al., 2011), and which are correlated with ac-

tion in our experimental control group (Appendix Table C.7). On the other hand, IRA information pushes participants towards “complacent” emotions, like peacefulness and happiness, which show no or negative associations with action in our control group.

3.3.4.2 Desire for climate policy

The story and IRA treatments’ impact and lack of impact, respectively, on climate action could also arise from their effects on concern about climate change and desire for continuing government action. During the experimental survey, we elicit participants’ worry about climate change from 1 (Not at all worried) to 7 (Extremely worried), how much they want the federal government to do about climate change, from much less (1) to much more (7) than it’s currently doing, and their rankings of how highly Congress should prioritize climate change in a list of policy issues. We elicit a similar measure in the obfuscated follow-up by asking participants how much they want the newly-elected Congress to focus on gun control, climate change, reducing inflation, and reproductive rights, each on a scale from 1 (Not at all) to 6 (Very much so).

Learning about the IRA reduces participants’ desire for government climate action by 0.11sd (Appendix Table C.8).¹² In contrast, the story significantly *increases* all three measures of policy demand: worry about climate change by 0.09sd, desire for more government climate action by 0.16sd, and legislative priority on climate change by 0.07sd. The impacts of the treatments on desire for climate policy are similar in the obfuscated follow-up survey, where the story increases hope that the new Congress will focus on climate change by 0.07sd. The impact of the IRA information treatment on climate priority in the follow-up is statistically insignificant ($p = 0.17$), but the negative point estimate is consistent with results in the experimental survey.

These results are notably consistent with the climate-action patterns we observe, and they suggest that the story could drive action by evoking the urgency of climate change—through Gilbert’s death in the heatwave or depictions of fires and floods—rather than by building political efficacy.

¹²The IRA information should only affect desire for government action by changing participants’ beliefs about current climate policy, not beliefs about the urgency of climate change. The IRA information treatment matches the control videos in stating truthfully that the US is not on track to meet its climate goals, and all three videos end in parallel statements emphasizing the need for continuing emissions cuts.

That said, Section 3.3.4.5 discusses suggestive evidence that the story’s effects on desire for climate policy are not the main drivers of its effects on action.

3.3.4.3 Beliefs about others

Learning about the IRA could signal that many Americans support climate policy or are engaged in the climate movement. Moreover, the story shows a large a climate march and states that “millions of people” across the US could advocate for climate policy. While the story is explicitly fictional, this image and rhetoric could shift participants’ beliefs about other Americans’ climate beliefs or action. Growing research in economics finds that shifting up beliefs about anonymous others’ political participation tends to reduce engagement in collective political action (Cantoni et al., 2019; Hager et al., 2022, 2023). On the other hand, Americans underestimate support for climate policy on average, and correcting these beliefs could increase action if participants conform to the norms of policy support that they perceive (Sparkman et al., 2022).

Learning about the IRA does not change beliefs about support for or engagement in the climate movement (Appendix Table C.9). While the story does not change participants’ belief about the share of Americans who support climate policy, it does increase their beliefs about the share of those Americans who would contact Congress to support a climate bill by 2.5pp (8% of the control mean). Existing work suggests that this increase may *reduce* the story’s impacts on action, rather than driving them.

3.3.4.4 Memory

While recent work suggests that the story could affect action by helping participants encode the IRA information (Graeber et al., 2022), this explanation is unlikely given that information about the IRA has no effect itself on action. Moreover, Appendix Table C.2 shows that the story had no differential effect on whether participants reported having heard of the IRA at the end of the experimental survey.

3.3.4.5 Combining mechanisms

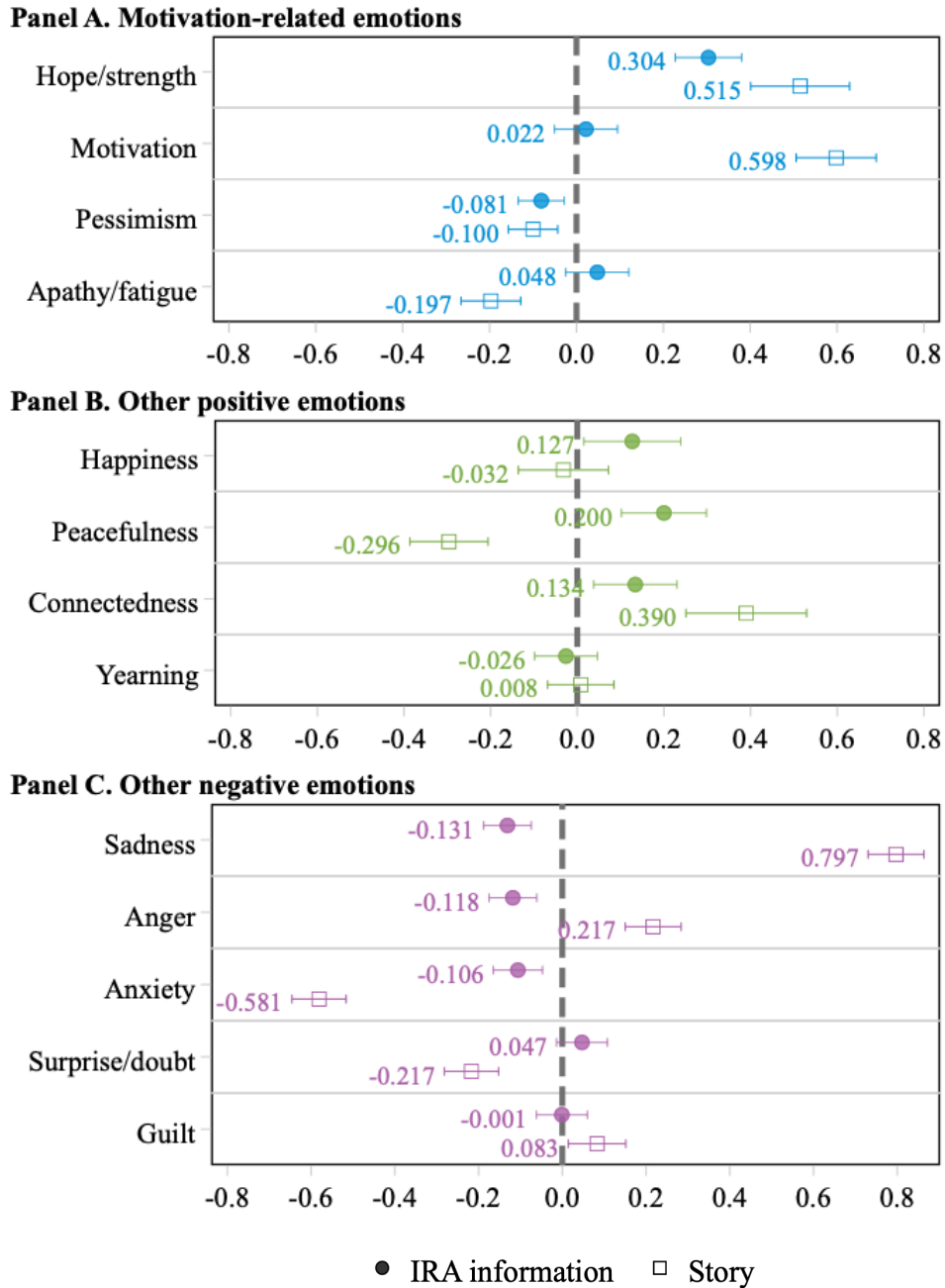
In Figure 3.2, we explore suggestive evidence on the role of each possible mechanism in the story’s effects on action. Here, we plot the story-treatment coefficients in a series of regressions that separately control for each possible mediator—efficacy beliefs, indices of emotion strength, policy desire, and beliefs about others’ political engagement—and then gradually add these controls to a single regression. Across all action outcomes, controlling for political efficacy and the index of motivation-related emotions each substantially reduce the story-treatment coefficient, with the largest drops when controlling for both together. Controlling for policy desire or beliefs about others’ action reduce the story-treatment by a lesser degree and not at all, respectively. These patterns suggest that the story’s effects on action can be explained in large part by its effects on both political-efficacy beliefs and feelings of motivation and strength.¹³

3.4 Conclusion

In a large online experiment, we find that people update their beliefs and behavior substantially more in response to a fictional narrative about citizen climate advocacy than to learning about recent, major legislative progress. These results are all the more striking because of the comparative strength of each treatment: the IRA is the most significant climate legislation ever passed in the United States; the story was produced on a small budget (and written by economists). Suggestive evidence implies that the story’s substantial effects on climate action can be attributed to both its “cold” effects on beliefs about government responsiveness to citizen action and its “hot” effects on emotions of motivation and hope.

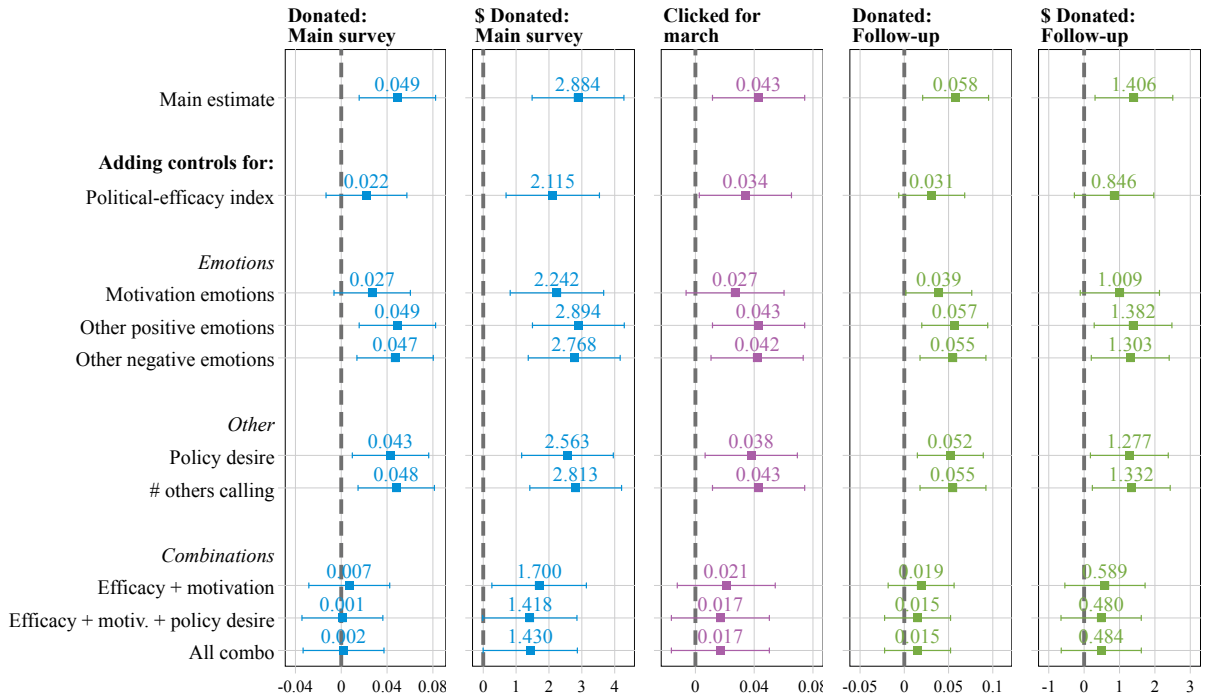
¹³A related, but conceptually distinct, question is what aspect of the story treatment drives its impacts on political efficacy and emotions. For example, these effects could arise from the story’s musical soundtrack, its animated imagery, the fictional storyline itself, informational signals about real-world facts, or, most likely, a combination of these elements. While our treatment variation does not allow us to separate these components, we argue that the story’s informational content is unlikely to play a large role. The only direct quasi-factual statement included in the story is that citizen activism contributed to the success of movements for women’s right to vote, labor laws, and civil rights. While these historical examples could add to the story’s effects on political efficacy, they only take up about 8 seconds near the midpoint of a 5-minute video and are unlikely to play a substantive role relative to the much more salient fictional storyline.

Figure 3.1: Impacts on emotions



Note: This figure plots the impacts of the IRA information treatment and the fictional climate-action story on emotions expressed during the main experimental survey. Panel A presents impacts on motivation-related emotions: Hope or strength, motivation, pessimism, and apathy or fatigue. Panel B presents impacts on other positive emotions: Happiness, peacefulness, connectedness, and yearning. Finally, Panel C presents impacts on other negative emotions: Sadness, anger, anxiety, surprise or doubt, and guilt. We define each emotion outcome as the standardized strength at which participants said they felt that emotion, unprompted. Appendix Section C.7 describes in detail how we constructed these measures of emotions. We estimate treatment impacts by regressing each emotion outcome on an indicator for receiving IRA information and an indicator for additionally watching the climate story. These regressions include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. Points in the figure marked with solid circles and open squares denote coefficients on the IRA information treatment and story treatment, respectively; the error bars denote 95% confidence intervals.

Figure 3.2: Impacts of the story on climate action: Controlling for mediating emotions and beliefs



Note: This figure plots our main estimates for the impacts of the fictional story on key climate-action outcomes and how these estimates change when we control for possibly-mediating beliefs and emotions. In particular, we sequentially add controls for the standardized index of political-efficacy beliefs, for standardized indices of motivation-related emotions, other positive emotions, and other negative emotions, and finally for both the standardized indices of political efficacy and motivation-related emotions. We construct the indices of motivation-related emotions, other positive emotions, and other negative emotions by standardizing the sum of standardized variables for the strength with which each participant reported feeling an emotion in that category, as grouped in Appendix Section C.7. Note that in constructing an index of motivation-related beliefs, we flip the signs of the strength with which participants feel pessimism and apathy or fatigue. The point estimates plotted are the coefficients on the story treatment in regressions of each action outcome on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. In addition to the controls for potentially mediating intermediate outcomes, these regressions including the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. Sample sizes for the regressions involving each outcome are given in the corresponding columns of Table 3.2. The error bars plot 95% confidence intervals.

Table 3.1: Impacts of treatments on political efficacy

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
	Main survey:					Follow-up survey:				
	Agreement that:					Agreement:				
	People like me have no say	Lobbyists have more power	Gov't responds citizens	Index (all +)	$\Delta P(\text{Passbill})$ if 2% to 10% call	Citizen movements make change	How effective on govt policy?	Marches	Contacts	Index
IRA info	-0.029 (0.027)	-0.035 (0.027)	0.105 (0.028)	0.073 (0.024)	-0.431 (0.333)	0.007 (0.035)	-0.003 (0.036)	-0.009 (0.036)	-0.002 (0.034)	
+ Story	-0.363 (0.033)	-0.382 (0.030)	0.416 (0.032)	0.505 (0.029)	0.948 (0.383)	0.231 (0.039)	0.161 (0.040)	0.112 (0.041)	0.198 (0.038)	
N	5879	5879	5879	5879	5879	3899	3899	3899	3899	
Control mean	0.000	0.000	0.000	-0.000	9.029	0.000	-0.000	-0.000	-0.000	

Note: This table estimates the impact of IRA information and the fictional story on political efficacy. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We also control for survey wave, whether participants completed the extra filler questions, demographics (sex, age bins, ethnicity categories, college-by-age groups, and political affiliation), climate attitudes (climate worry and desire for additional government action), political efficacy, and political engagement. Appendix Section C.7.2 defines these control variables in detail. The outcomes presented in columns 1 through 5 are measured during the main experimental survey, while those in columns 6 through 9 are measured during the obfuscated follow-up survey. Columns 1 through 3 present impacts on standardized agreement with three qualitative political-efficacy statements, where negative coefficients in columns 1 and 2 and a positive coefficient in column 3 denote increasing political efficacy. Column 4 presents impacts on a standardized index combining agreement with these qualitative statements, where components are rescaled so that increasing values denote higher political efficacy. Column 5 presents impacts on a numeric measure of political efficacy, defined as participants' estimates for how much more likely Congress would be to pass a climate bill if 10% versus 2% of Americans contacted them to support it. Appendix Table C.12 presents treatment effects on participants' estimates of the likelihood that the bill would pass if 10% or 2% contacted Congress in support. Column 6 presents impacts on standardized agreement that citizen movements can make real change, and columns 7 and 8 present standardized beliefs for how effective marches and contacting Congress are in affecting government policy. Finally, column 9 presents impacts on a standardized index combining the outcomes in columns 6 through 8. Appendix Section C.7 defines all of these outcome variables in detail. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table 3.2: Impacts of treatments on climate donations and citizen advocacy

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Climate donation outcomes:					Direct-action outcomes:				
	<i>Main survey:</i>		<i>Follow-up:</i>		<i>Main survey:</i>				
	Y/N	Amount	Y/N	Amount	Sending letter to interested	Wrote letter	Clicked to send	Clicked for march	<i>Follow-up:</i> Downloaded guide
IRA info	0.006 (0.015)	0.006 (0.623)	-0.015 (0.016)	-0.540 (0.481)	-0.001 (0.015)	0.000 (0.011)	-0.010 (0.011)	-0.008 (0.013)	-0.008 (0.014)
+ Story	0.049 (0.017)	2.884 (0.712)	0.058 (0.019)	1.406 (0.560)	0.016 (0.017)	-0.014 (0.012)	0.009 (0.012)	0.043 (0.016)	0.018 (0.016)
N	5879	5879	5021	5021	5879	5879	5869	2595	5021
Control mean	0.511	14,944	0.438	8,552	0.426	0.126	0.145	0.079	0.210

Note: This table estimates the impact of IRA information and the fictional story on climate action. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. The outcomes presented in columns 1 through 5 are measures of direct citizen action, while those in columns 6 through 9 are measures of donations to climate-lobbying organizations. Columns 1 through 4 estimate impacts on direct-action outcomes measured during the main experimental survey: whether participants said they were interested in emailing Congress (column 1), whether they wrote out text for a custom letter to Congress (column 2), whether they clicked to the portal to send their letter (column 3), and whether they clicked on a link for information about nearby climate marches (column 4). Note that we only observe whether participants click for climate-march information among those in the second survey wave. Column 5 presents whether participants downloaded the guide for contacting legislators offered in the follow-up survey. Columns 6 and 7 present impacts on whether and how much participants donated to a climate organization in the main experimental survey, while columns 8 and 9 present impacts on whether and how much they donated to the climate organization in the follow-up survey. Appendix Table C.6 presents impacts on whether and how much participants donated to non-climate organizations in the follow-up survey. All donation amounts are given in USD. Appendix Section C.7 defines all of these outcome variables in detail. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Bibliography

- Abramson, Paul R. and John H. Aldrich**, “The Decline of Electoral Participation in America,” *The American Political Science Review*, 1982, 76 (3), 502–521.
- Acikgoz, Yalcin, Kristl H. Davison, Maira Compagnone, and Matt Laske**, “Justice Perceptions of Artificial Intelligence in Selection,” *International Journal of Selection and Assessment*, 2020, 28 (4), 399–416.
- Adams-Prassl, Abi, Kristiina Huttunen, Emily Nix, and Ning Zhang**, “Violence Against Women at Work,” *Quarterly Journal of Economics*, *Forthcoming*.
- Agüero, Jorge M., Francisco Galarza, and Gustavo Yamada**, “(Incorrect) Perceived Returns and Strategic Behavior among Talented Low-Income College Graduates,” *AEA Papers and Proceedings*, 2023, 113, 423–426.
- Akerlof, George A.**, “Labor Contracts as Partial Gift Exchange,” *The Quarterly Journal of Economics*, 1982, 97 (4), 543–569.
- Aksoy, Billur, Ian Chadd, and Boon Han Koh**, “Sexual identity, gender, and anticipated discrimination in prosocial behavior,” *European Economic Review*, 2023, 154, 104427.
- Alan, Sule and Seda Ertac**, “Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment,” *Journal of the European Economic Association*, 2019, 17 (4), 1147–1185.
- Allcott, Hunt**, “Social Norms and Energy Conservation,” *Journal of Public Economics*, 2011, 95 (9), 1082–1095.
- Almlund, Mathilde, Angela Lee Duckworth, James J Heckman, and Tim D Kautz**, “Personality psychology and economics,” *NBER Working Paper No. 16822*, February 2011.
- Alston, Mackenzie**, “The (Perceived) Cost of Being Female: An Experimental Investigation of Strategic Responses to Discrimination,” *Working Paper*, 2019.
- Alter, Adam L., Joshua Aronson, John M. Darley, Cordaro Rodriguez, and Diane N. Ruble**, “Rising to the Threat: Reducing Stereotype Threat by Reframing the Threat as a Challenge,” *Journal of Experimental Social Psychology*, January 2010, 46 (1), 166–171.
- Anaya, Lina, Frank Stafford, and Gema Zamarro**, “Gender gaps in math performance, perceived mathematical ability and college STEM education: the role of parental occupation,” *Education Economics*, September 2021, 0 (0), 1–16.
- Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk**, “Misperceived Social Norms and Willingness to Act Against Climate Change,” *Working Paper*, 2022.

- Angeli, Deivis, Ieda Matavelli, and Fernando Secco**, “Expected Discrimination and Job Search,” *Working Paper*, 2023.
- Angill-Williams, Aishlyn and Colin J. Davis**, “Increasing Climate Efficacy is Not a Surefire Means to Promoting Climate Commitment,” *Thinking & Reasoning*, 2021, pp. 1–21.
- Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci**, “Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech,” *Working Paper*, 2023.
- Baker, Andrew, David F. Larcker, Charles McClure, Durgesh Saraph, and Edward M. Watts**, “Diversity Washing,” *Working Paper*, 2023.
- Balch, George I.**, “Multiple Indicators in Survey Research: The Concept “Sense of Political Efficacy”,” *Political Methodology*, 1974, 1 (2), 1–43.
- Ballen, Cissy J., Shima Salehi, and Sehoia Cotner**, “Exams disadvantage women in introductory biology,” *PLoS One*, 2017, 12 (10), e0186419. Publisher: Public Library of Science San Francisco, CA USA.
- Bandura, Albert**, *Social foundations of thought and action: A social cognitive theory.*, Englewood Cliffs, NJ: Prentice Hall., 1986.
- , “Exercise of Human Agency Through Collective Efficacy,” *Current Directions in Psychological Science*, 2000, 9 (3), 75–78.
- Banerjee, Abhijit, Eliana La Ferrara, and Victor Orozco-Olvera**, “The Entertaining Way to Behavioral Change: Fighting HIV with MTV,” *NBER Working Paper 26096*, 2019.
- , —, and —, “Entertainment, Education, and Attitudes Toward Domestic Violence,” *AEA Papers and Proceedings*, 2019, 109, 133–137.
- Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka**, “Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition,” *American Economic Review*, 2016, 106 (6), 1437–1475.
- Becker, Gary S.**, *Human capital: A theoretical and empirical analysis with special reference to education*, The University of Chicago Press, 1964.
- Benjamin, Daniel J.**, “Errors in Probabilistic Reasoning and Judgment Biases,” in “Handbook of Behavioral Economics: Applications and Foundations 1,” Vol. 2, Elsevier, 2019, pp. 69–186.
- Bergan, Daniel E.**, “Does Grassroots Lobbying Work?: A Field Experiment Measuring the Effects of an e-Mail Lobbying Campaign on Legislative Behavior,” *American Politics Research*, 2009, 37 (2), 327–352.
- and **Richard T. Cole**, “Call Your Legislator: A Field Experimental Study of the Impact of a Constituency Mobilization Campaign on Legislative Voting,” *Political Behavior*, 2015, 37 (1), 27–42.
- Bernard, Rene, Panagiota Tzamourani, and Michael Weber**, “Climate Change and Individual Behavior,” *Working Paper*, 2023.
- Bertrand, Marianne and Esther Dufo**, “Field Experiments on Discrimination,” in “Handbook of Economic Field Experiments,” Vol. 1, Elsevier, 2017, pp. 309–393.

- **and Sendhil Mullainathan**, “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination,” *American Economic Review*, 2004, 94 (4), 991–1013.
- Bigman, Yochanan, Kurt Gray, Adam Waytz, Mads Arnestad, and Desman Wilson**, “Algorithmic Discrimination Causes Less Moral Outrage than Human Discrimination,” *Journal of Experimental Psychology: General*, 2023, 152 (1), 4–27.
- Bistline, John, Geoffrey Blanford, Maxwell Brown, Dallas Burtraw, Maya Domeshek, Jamil Farbes, Allen Fawcett, Anne Hamilton, Jesse Jenkins, Ryan Jones, Ben King, Hannah Kolus, John Larsen, Amanda Levin, Megan Mahajan, Cara Marcy, Erin Mayfield, James McFarland, Haewon McJeon, Robbie Orvis, Neha Patankar, Kevin Rennert, Christopher Roney, Nicholas Roy, Greg Schivley, Daniel Steinberg, Nadejda Victor, Shelley Wenzel, John Weyant, Ryan Wisser, Mei Yuan, and Alicia Zhao**, “Emissions and Energy Impacts of the Inflation Reduction Act,” *Science*, 2023, 380 (6652), 1324–1327.
- Blau, Francine D. and Lawrence M. Kahn**, “The Gender Wage Gap: Extent, Trends, and Explanations,” *Journal of Economic Literature*, 2017, 55 (3), 789–865.
- Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg**, “The Dynamics of Discrimination: Theory and Evidence,” *American Economic Review*, 2019, 109 (10), 3395–3436.
- , **Peter Hull, and Alex Imas**, “Systemic Discrimination: Theory and Measurement,” *NBER Working Paper No. 29820*, 2022.
- Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer**, “Beliefs About Gender,” *American Economic Review*, 2019, 109 (3), 739–73.
- , —, —, **Frederik Schwerter, and Andrei Shleifer**, “Memory and Representativeness,” *Psychological Review*, 2021, 128 (1), 71–85.
- Borjas, George J.**, “Immigrants, minorities, and labor market competition,” *ILR Review*, 1987, 40 (3), 382–392.
- Brader, Ted**, “Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions,” *American Journal of Political Science*, 2005, 49 (2), 388–405.
- Brecheisen, Jeremie**, “Research: Where Employees Think Companies’ DEIB Efforts Are Failing,” *Harvard Business Review*, 2023.
- Breza, Emily, Supreet Kaur, and Yogita Shamdasani**, “The Morale Effects of Pay Inequality,” *The Quarterly Journal of Economics*, 2018, 133 (2), 611–663.
- Bryan, Christopher J., Gregory M. Walton, Todd Rogers, and Carol S. Dweck**, “Motivating Voter Turnout by Invoking the Self,” *Proceedings of the National Academy of Sciences*, 2011, 108 (31), 12653–12656.
- Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek**, “Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures,” *NBER Working Paper No. 28916*, 2021.
- Butler, Daniel M. and David W. Nickerson**, “Can Learning Constituency Opinion Affect How Legislators Vote? Results from a Field Experiment,” *Quarterly Journal of Political Science*, 2011, 6 (1), 55–83.

- Bénabou, Roland and Jean Tirole**, “Self-confidence and personal motivation,” *The Quarterly Journal of Economics*, 2002, 117 (3), 871–915.
- Campbell, Angus, Gerald Gurin, and Warren Miller**, *The Voter Decides*, Evanston, Illinois: Row, Peterson and Company, 1954.
- Cantoni, Davide, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang**, “Protests as Strategic Games: Experimental Evidence from Hong Kong’s Antiauthoritarian Movement,” *The Quarterly Journal of Economics*, 2019, 134 (2), 1021–1077.
- Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez**, “Inequality at Work: The Effect of Peer Salaries on Job Satisfaction,” *American Economic Review*, 2012, 102 (6), 2981–3003.
- Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti**, “Goals and gaps: Educational careers of immigrant children,” *CReAM Discussion Paper Series 1812*, 2018, *Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London*.
- Castex, Gonzalo and Evgenia Kogan Dechter**, “The changing roles of education and ability in wage determination,” *Journal of Labor Economics*, 2014, 32 (4), 685–710.
- Chang, Edward H., Katherine L. Milkman, Dena M. Gromet, Robert W. Rebele, Cade Massey, Angela L. Duckworth, and Adam M. Grant**, “The Mixed Effects of Online Diversity Training,” *Proceedings of the National Academy of Sciences*, 2019, 116 (16), 7778–7783.
- Charness, Gary, Ramón Cobo-Reyes, Simone Meraglia, and Ángela Sánchez**, “Anticipated Discrimination, Choices, and Performance: Experimental Evidence,” *European Economic Review*, 2020, 127, 103473.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins**, “Double/debiased Machine Learning for Treatment and Structural Parameters,” *The Econometrics Journal*, 2018, 21 (1), C1–C68.
- Coate, Stephen and Glenn C. Loury**, “Will Affirmative-Action Policies Eliminate Negative Stereotypes?,” *The American Economic Review*, 1993, 83 (5), 1220–1240.
- Coffman, Katherine B., Manuela Collis, and Leena Kulkarni**, “Stereotypes and belief updating,” *Harvard Business School Working Paper No. 19-068*, January 2019.
- Coffman, Katherine Baldiga**, “Evidence on Self-Stereotyping and the Contribution of Ideas,” *The Quarterly Journal of Economics*, 2014, 129 (4), 1625–1660.
- Cohn, Alain, Ernst Fehr, Benedikt Herrmann, and Frédéric Schneider**, “Social Comparison and Effort Provision: Evidence from a Field Experiment,” *Journal of the European Economic Association*, 2014, 12 (4), 877–898.
- Cohodes, Sarah R. and Joshua S. Goodman**, “Merit aid, college quality, and college completion: Massachusetts’ Adams scholarship as an in-kind subsidy,” *American Economic Journal: Applied Economics*, 2014, 6 (4), 251–85.
- Cowgill, Bo**, “Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening,” *Working Paper*, 2020.
- Craig, Stephen C., Richard G. Niemi, and Glenn E. Silver**, “Political Efficacy and Trust: A

- Report on the NES Pilot Study Items,” *Political Behavior*, 1990, 12 (3), 289–314.
- Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler**, “Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence,” *Working Paper*, 2022.
- de Araujo, Pedro and Stephen Lagos**, “Self-esteem, education, and wages revisited,” *Journal of Economic Psychology*, 2013, 34, 120–132.
- de Haan, Thomas, Theo Offerman, and Randolph Sloof**, “Discrimination in the Labour Market: The Curse of Competition Between Workers,” *The Economic Journal*, 2017, 127 (603), 1433–1466.
- de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth**, “Measuring and Bounding Experimenter Demand,” *American Economic Review*, 2018, 108 (11), 3266–3302.
- , **Lise Vesterlund, and Alistair J. Wilson**, “Experimenter Demand Effects,” in “Handbook of Research Methods and Applications in Experimental Economics,” Edward Elgar Publishing, 2019, pp. 384–400.
- Dechezlepretre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva**, “Fighting Climate Change: International Attitudes Toward Climate Policies,” *NBER Working Paper 30265*, 2022.
- DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao**, “Estimating Social Preferences and Gift Exchange at Work,” *American Economic Review*, 2022, 112 (3), 1038–1074.
- DeMars, Christine E., Bozhidar M. Bashkov, and Alan B. Socha**, “The role of gender in test-taking motivation under low-stakes conditions,” *Research & Practice in Assessment*, 2013, 8, 69–82.
- Devine, Amy, Kayleigh Fawcett, Dénes Szucs, and Ann Dowker**, “Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety,” *Behavioral and Brain Functions*, 2012, 8 (1), 1–9.
- Diamond, Rebecca and Petra Persson**, “The long-term consequences of teacher discretion in grading of high-stakes tests,” *NBER Working Paper No. 22207*, 2017.
- Dianat, Ahrash, Federico Echenique, and Leat Yariv**, “Statistical Discrimination and Affirmative Action in the Lab,” *Working Paper*, 2020.
- Dizon-Ross, Rebecca**, “Parents’ beliefs about their children’s academic ability: Implications for educational investments,” *American Economic Review*, August 2019, 109 (8), 2728–2765.
- Doering, Laura, Jan Doering, and András Tilcsik**, “‘Was It Me or Was It Gender Discrimination?’ How Women Respond to Ambiguous Incidents at Work,” *Sociological Science*, 2023, 10, 501–533.
- Dohmen, Thomas and Armin Falk**, “Performance pay and multidimensional sorting: Productivity, preferences, and gender,” *American Economic Review*, 2011, 101 (2), 556–590.
- Drago, Francesco**, “Self-esteem and earnings,” *Journal of Economic Psychology*, June 2011, 32 (3), 480–488.
- Drews, Stefan and Jeroen C.J.M. van den Bergh**, “What Explains Public Support for Climate

- Policies? A Review of Empirical and Experimental Studies,” *Climate Policy*, 2016, 16 (7), 855–876.
- Dube, Arindrajit, Laura Giuliano, and Jonathan Leonard**, “Fairness and Frictions: The Impact of Unequal Raises on Quit Behavior,” *American Economic Review*, 2019, 109 (2), 620–663.
- , **Suresh Naidu, and Adam D. Reich**, “Power and Dignity in the Low-Wage Labor Market: Theory and Evidence from Wal-Mart Workers,” *NBER Working Paper No. 30441*, 2022.
- Duffy, Denise and Narayan Sastry**, “Achievement tests in the Panel Study of Income Dynamics Child Development Supplement,” *PSID Technical Series Paper #14-02*, 2014, p. 29.
- Duncan, Greg J., Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, and Jeanne Brooks-Gunn**, “School readiness and later achievement,” *Developmental psychology*, 2007, 43 (6), 1428.
- Eccles, Jacquelynne, C. Midgley, and T. Adler**, “Grade-related changes in the school environment: Effects on achievement motivation,” in “Advances in Motivation and Achievement: The Development of Achievement Motivation,” Vol. 3 January 1984, pp. 282–331.
- Edmonds, Grant W., Lewis R. Goldberg, Sarah E. Hampson, and Maureen Barckley**, “Personality stability from childhood to midlife: Relating teachers’ assessments in elementary school to observer-and self-ratings 40 years later,” *Journal of Research in Personality*, 2013, 47 (5), 505–513.
- Eil, David and Justin M. Rao**, “The good news-bad news effect: asymmetric processing of objective information about yourself,” *American Economic Journal: Microeconomics*, 2011, 3 (2), 114–38.
- Eliaz, Kfir and Ran Spiegler**, “A Model of Competing Narratives,” *American Economic Review*, 2020, 110 (12), 3786–3816.
- Elster, Jon**, “Emotions and Economic Theory,” *Journal of Economic Literature*, 1998, 36 (1), 47–74. Publisher: American Economic Association.
- Engel, Christoph**, “Dictator Games: A Meta Study,” *Experimental Economics*, November 2011, 14 (4), 583–610.
- Erturan, Selin and Brenda Jansen**, “An investigation of boys’ and girls’ emotional experience of math, their math performance, and the relation between these variables,” *European Journal of Psychology of Education*, 2015, 30 (4), 421–435.
- Ettinger, Joshua, Peter Walton, James Painter, and Thomas DiBlasi**, “Climate of Hope or Doom and Gloom? Testing the Climate Change Hope vs. Fear Communications Debate Through Online Videos,” *Climatic Change*, 2021, 164 (1-2).
- Fahle, Erin M., Belen Chavez, Demitra Kalogrides, Benjamin R. Shear, Sean F. Reardon, and Andrew D. Ho**, “Stanford Education Data Archive: Technical Documentation (Version 4.1).” Technical Report 2021. Retrieved from <http://purl.stanford.edu/db586ns4974>.
- Falco, Lia D., Jessica J. Summers, and Sheri Bauman**, “Encouraging mathematics participation through improved self-efficacy: A school counseling outcomes study,” *Educational Research and Evaluation*, 2010, 16 (6), 529–549.

- Fath, Sean**, “When Blind Hiring Advances DEI — and When It Doesn’t,” *Harvard Business Review*, 2023.
- **and Susan Zhu**, “Preferences for, and Familiarity With, Blinding Among HR Practitioners,” *Working Paper*, 2021.
- Feddersen, Timothy and Alvaro Sandroni**, “A Theory of Participation in Elections,” *The American Economic Review*, 2006, 96 (4), 1271–1282. Publisher: American Economic Association.
- Feddersen, Timothy J.**, “Rational Choice Theory and the Paradox of Not Voting,” *Journal of Economic Perspectives*, February 2004, 18 (1), 99–112.
- Fehr, Ernst and Gary Charness**, “Social Preferences: Fundamental Characteristics and Economic Consequences,” *Working Paper*, 2023.
- , **Lorenz Goette, and Christian Zehnder**, “A Behavioral Account of the Labor Market: The Role of Fairness Concerns,” *Annual Review of Economics*, 2009, 1 (1), 355–384.
- Feldman, Lauren and P. Sol Hart**, “Using Political Efficacy Messages to Increase Climate Activism: The Mediating Role of Emotions,” *Science Communication*, 2016, 38 (1), 99–127.
- Ferrara, Eliana La, Alberto Chong, and Suzanne Duryea**, “Soap Operas and Fertility: Evidence from Brazil,” *American Economic Journal: Applied Economics*, 2012, 4 (4), 1–31.
- Finkel, Steven E.**, “Reciprocal Effects of Participation and Political Efficacy: A Panel Analysis,” *American Journal of Political Science*, 1985, 29 (4), 891–913.
- Fisher, R. A.**, *The Design of Experiments*, Oxford, England: Oliver & Boyd, 1935. Pages: xi, 251.
- Flory, Jeffrey A., Andreas Leibbrandt, Christina Rott, and Olga Stoddard**, “Increasing Workplace Diversity: Evidence from a Recruiting Experiment at a Fortune 500 Company,” *Journal of Human Resources*, 2021, 56 (1), 73–92.
- Folke, Olle and Johanna Rickne**, “Sexual Harassment and Gender Inequality in the Labor Market,” *The Quarterly Journal of Economics*, 2022, 137 (4), 2163–2212.
- Fowler, James H.**, “Altruism and Turnout,” *The Journal of Politics*, August 2006, 68 (3), 674–683.
- Fryer, Roland G.**, “The “pupil” Factory: Specialization and the production of human capital in schools,” *American Economic Review*, March 2018, 108 (3), 616–656.
- , **Jacob K. Goeree, and Charles A. Holt**, “Experience-based Discrimination: Classroom Games,” *The Journal of Economic Education*, 2005, 36 (2), 160–170.
- Gagnon, Nickolas, Kristof Bosmans, and Arno Riedl**, “The Effect of Gender Discrimination on Labor Supply,” *Working Paper*, 2024.
- Gallup Inc.**, “One in Four Black Workers Report Discrimination at Work,” Technical Report 2021.
- , “From Appreciation to Equity: How Recognition Reinforces DEI in the Workplace,” 2023.
- Glover, Dylan, Amanda Pallais, and William Pariente**, “Discrimination as a Self-fulfilling Prophecy: Evidence from French Grocery Stores,” *Quarterly Journal of Economics*, 2017, 132 (3), 1219–1260.
- Gneezy, Uri, John A. List, Jeffrey A. Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu**, “Measuring success in education: the role of effort on the test itself,” *American Economic*

- Review: Insights*, 2019, 1 (3), 291–308.
- Goldin, Claudia and Cecilia Rouse**, “Orchestrating Impartiality: The Impact of “Blind” Auditions on Female Musicians,” *American Economic Review*, 2000, 90 (4), 715–741.
- Goldsmith, Arthur H., Stanley Sedo, William Darity, and Darrick Hamilton**, “The Labor Supply Consequences of Perceptions of Employer Discrimination During Search and On-the-job: Integrating Neoclassical Theory and Cognitive Dissonance,” *Journal of Economic Psychology*, 2004, 25 (1), 15–39.
- Goodman, Joshua**, “The labor of division: Returns to compulsory high school math coursework,” *Journal of Labor Economics*, 2019, 37 (4), 1141–1182.
- Graeber, Thomas, Florian Zimmermann, and Christopher Roth**, “Stories, Statistics, and Memory,” *Working Paper*, 2022.
- Greico, Elizabeth M. and Rachel C. Cassidy**, “Overview of race and Hispanic origin: Census 2000 brief,” Technical Report C2KBR/01-1, US Census Bureau, US Department of Commerce 2001.
- Haaland, Ingar and Christopher Roth**, “Labor Market Concerns and Support for immigration,” *Journal of Public Economics*, 2020, 191.
- and —, “Beliefs about Racial Discrimination and Support for Pro-Black Policies,” *The Review of Economics and Statistics*, 2023, 105 (1), 40–53.
- Hager, Anselm, Lukas Hensel, Johannes Hermle, and Christopher Roth**, “Group Size and Protest Mobilization across Movements and Countermovements,” *American Political Science Review*, 2022, 116 (3), 1051–1066.
- , —, —, and —, “Political Activists as Free Riders: Evidence from a Natural Field Experiment,” *The Economic Journal*, 2023, 133 (653), 2068–2084.
- Hamann, Karen R. S. and Gerhard Reese**, “My Influence on the World (of Others): Goal Efficacy Beliefs and Efficacy Affect Predict Private, Public, and Activist Pro-environmental Behavior,” *Journal of Social Issues*, 2020, 76 (1), 35–53.
- Hampson, Sarah E. and Lewis R. Goldberg**, “A first large cohort study of personality trait stability over the 40 years between elementary school and midlife,” *Journal of Personality and Social Psychology*, 2006, 91 (4), 763.
- Hart, P. Sol and Lauren Feldman**, “The Influence of Climate Change Efficacy Messages and Efficacy Beliefs on Intended Political Participation,” *PLoS ONE*, 2016, 11 (8).
- Heckman, James J., Tomáš Jagelka, and Timothy D. Kautz**, “Some contributions of economics to the study of personality,” *NBER Working Paper No. 26459*, 2019.
- Heidhues, Paul, Botond Köszegi, and Philipp Strack**, “Overconfidence and prejudice,” *Working Paper*, 2019. arXiv: 1909.08497.
- Herbert, J. and D. Stipek**, “The emergence of gender differences in children’s perceptions of their academic competence,” *Journal of Applied Developmental Psychology*, 2005, 26 (3), 276–295.
- Hicks, Peggy and Larry M. Bolen**, “Review of the Woodcock-Johnson Psycho-Educational Battery-Revised,” *Journal of School Psychology*, 1996, 34 (1), 93–102.

- Hoff, Karla, Jyotsna Jalan, and Sattwik Santra**, *Participatory Theater Empowers Women: Evidence from India* Policy Research Working Papers, The World Bank, 2021.
- Hornsey, Matthew J. and Kelly S. Fielding**, “A Cautionary Note about Messages of Hope: Focusing on Progress in Reducing Carbon Emissions Weakens Mitigation Motivation,” *Global Environmental Change*, 2016, 39, 26–34.
- , **Cassandra M. Chapman, and Dexter M. Oelrichs**, “Ripple Effects: Can Information About the Collective Impact of Individual Actions Boost Perceived Efficacy About Climate Change?,” *Journal of Experimental Social Psychology*, 2021, 97.
- Jamieson, Jeremy P., Wendy Berry Mendes, and Matthew K. Nock**, “Improving acute stress responses: The power of reappraisal,” *Current Directions in Psychological Science*, 2013, 22 (1), 51–56.
- Jarrahi, Mohammad Hossein, Gemma Newlands, Min Kyung Lee, Christine T. Wolf, Eliscia Kinder, and Will Sutherland**, “Algorithmic Management in a Work Context,” *Big Data & Society*, 2021, 8 (2).
- Jenkins, Jesse, Erin Mayfield, Jamil Farbes, Ryan Jones, Neha Patankar, Qingyu Xu, and Greg Schivley**, “Preliminary Report: The Climate and Energy Impacts of the Inflation Reduction Act of 2022,” Technical Report, REPEAT Project, Princeton, NJ August 2022.
- Jensen, Robert and Emily Oster**, “The Power of TV: Cable Television and Women’s Status in India,” *The Quarterly Journal of Economics*, 2009, 124 (3), 1057–1094.
- Jones, Kristen P., Chad I. Peddie, Veronica L. Gilrane, Eden B. King, and Alexis L. Gray**, “Not So Subtle: A Meta-Analytic Investigation of the Correlates of Subtle and Overt Discrimination,” *Journal of Management*, 2016, 42 (6), 1588–1613.
- Jugert, Philipp, Katharine H. Greenaway, Markus Barth, Ronja Buchner, Sarah Eisenraut, and Immo Fritsche**, “Collective Efficacy Increases Pro-environmental Intentions Through Increasing Self-Efficacy,” *Journal of Environmental Psychology*, 2016, 48, 12–23.
- Jussim, Lee and Kent D Harber**, “Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies,” *Personality and Social Psychology Review*, 2005, 9 (2), 131–155.
- Kahneman, Daniel and Amos Tversky**, “Subjective Probability: A Judgment of Representativeness,” *Cognitive Psychology*, 1972, 3 (3), 430–454.
- Kaibel, Chris, Irmela Koch-Bayram, Torsten Biemann, and Max Mühlenbock**, “Applicant Perceptions of Hiring Algorithms - Uniqueness and Discrimination Experiences as Moderators,” *Academy of Management Proceedings*, 2019, 2019 (1), 18172.
- Kearney, Melissa S. and Phillip B. Levine**, “Media Influences on Social Outcomes: The Impact of MTV’s “16 and Pregnant” on Teen Childbearing,” *The American Economic Review*, 2015, 105 (12), 3597–3632.
- and —, “Early Childhood Education by Television: Lessons from Sesame Street,” *American Economic Journal: Applied Economics*, 2019, 11 (1), 318–350.
- Kendall, Chad W. and Constantin Charles**, “Causal Narratives,” *NBER Working Paper 30346*, 2022.

- Kessler, Ronald C., Kristin D. Mickelson, and David R. Williams**, “The Prevalence, Distribution, and Mental Health Correlates of Perceived Discrimination in the United States,” *Journal of Health and Social Behavior*, 1999, 40 (3), 208–230.
- Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad**, “Field of study, earnings, and self-selection,” *The Quarterly Journal of Economics*, 2016, 131 (3), 1057–1111.
- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan**, “Algorithmic Fairness,” *AEA Papers and Proceedings*, 2018, 108, 22–27.
- , —, —, and **Cass R. Sunstein**, “Algorithms as Discrimination Detectors,” *Proceedings of the National Academy of Sciences*, 2020, 117 (48), 30096–30100.
- Kline, Patrick M., Evan K. Rose, and Christopher R. Walters**, “Systemic Discrimination Among Large U.S. Employers,” *NBER Working Paper No. 29053*, 2021.
- Langford, Joe and Pauline Rose Clance**, “The imposter phenomenon: Recent research findings regarding dynamics, personality and family patterns and their implications for treatment,” *Psychotherapy: Theory, research, practice, training*, 1993, 30 (3), 495.
- Lee, David S.**, “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” *The Review of Economic Studies*, July 2009, 76 (3), 1071–1102.
- Lee, Min Kyung**, “Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management,” *Big Data & Society*, 2018, 5 (1).
- Leiserowitz, Anthony, Edward Maibach, Seth Rosenthal, and John Kotcher**, “Americans’ Actions to Limit and Prepare for Global Warming, March 2021,” Technical Report, Yale University and George Mason University, New Haven, CT 2021.
- Lepage, Louis-Pierre, Xiaomeng Li, and Basit Zafar**, “Anticipated Gender Discrimination and Grade Disclosure,” *NBER Working Paper No. 30765*, 2022.
- Lerner, Jennifer S., Ye Li, Piercarlo Valdesolo, and Karim S. Kassam**, “Emotion and decision making,” *Annual Review of Psychology*, 2015, 66, 799–823.
- List, John A., Ragan Petrie, and Anya Samek**, “How experiments with children inform economics,” *NBER Working Paper No. 28825*, 2021.
- Liu, Songqi, Pei Liu, Mo Wang, and Baoshan Zhang**, “Effectiveness of Stereotype Threat Interventions: A Meta-analytic Review,” *The Journal of Applied Psychology*, 2021, 106 (6), 921–949.
- Loewenstein, George**, “Emotions in Economic Theory and Economic Behavior,” *American Economic Review*, May 2000, 90 (2), 426–432.
- Lundberg, Shelly J. and Richard Startz**, “Private Discrimination and Social Intervention in Competitive Labor Market,” *The American Economic Review*, 1983, 73 (3), 340–347.
- Mattarella-Micke, Andrew, Jill Mateo, Megan N. Kozak, Katherine Foster, and Sian L. Beilock**, “Choke or thrive? The relation between salivary cortisol and math performance depends on individual differences in working memory and math-anxiety,” *Emotion*, 2011, 11 (4), 1000.
- McCall, John Joseph**, “Economics of information and job search,” *The Quarterly Journal of Economics*, 1970, pp. 113–126.

- Mobius, Markus M. and Tanya S. Rosenblat**, “Why beauty matters,” *American Economic Review*, March 2006, 96 (1), 222–235.
- Moore, Don A. and Paul J. Healy**, “The trouble with overconfidence.,” *Psychological Review*, 2008, 115 (2), 502.
- Mukerjee, Swati**, “Job Satisfaction in the United States: Are Blacks Still More Satisfied?,” *The Review of Black Political Economy*, 2014, 41 (1), 61–81.
- Mummolo, Jonathan and Erik Peterson**, “Demand Effects in Survey Experiments: An Empirical Assessment,” *American Political Science Review*, 2019, 113 (2), 517–529.
- Murnane, Richard J., John B. Willett, M. Jay Braatz, and Yves Duhaldeborde**, “Do different dimensions of male high school students’ skills predict labor market success a decade later? Evidence from the NLSY,” *Economics of Education Review*, 2001, 20 (4), 311–320.
- Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat**, “Managing self-confidence,” *Working paper*, 2014.
- Neumark, David**, “Experimental Research on Labor Market Discrimination,” *Journal of Economic Literature*, 2018, 56 (3), 799–866.
- Newman, David T., Nathanael J. Fast, and Derek J. Harmon**, “When Eliminating Bias isn’t Fair: Algorithmic Reductionism and Procedural Justice in Human Resource Decisions,” *Organizational Behavior and Human Decision Processes*, 2020, 160, 149–167.
- Niederle, Muriel and Lise Vesterlund**, “Do Women Shy Away from Competition? Do Men Compete Too Much?,” *The Quarterly Journal of Economics*, 2007, 122 (3), 1067–1101.
- Niemi, Richard G., Stephen C. Craig, and Franco Mattei**, “Measuring Internal Political Efficacy in the 1988 National Election Study,” *The American Political Science Review*, 1991, 85 (4), 1407–1413.
- Noble, Sean M., Lori L. Foster, and S. Bartholomew Craig**, “The Procedural and Interpersonal Justice of Automated Application and Resume Screening,” *International Journal of Selection and Assessment*, 2021, pp. 1–15.
- Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan**, “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations,” *Science*, 2019, 366, 447–453.
- Owen, Stephanie**, “College field specialization and beliefs about relative performance: An experimental intervention to understand gender gaps in STEM,” *Working Paper*, 2020.
- Pager, Devah and David S. Pedulla**, “Race, Self-Selection, and the Job Search Process,” *American Journal of Sociology*, 2015, 120 (4), 1005–1054.
- **and Hana Shepherd**, “The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets,” *Annual Review of Sociology*, 2008, 34, 181–209.
- Paluck, Elizabeth Levy**, “Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda,” *Journal of Personality and Social Psychology*, 2009, 96 (3), 574–587.
- Papageorge, Nicholas W., Seth Gershenson, and Kyung Min Kang**, “Teacher expectations mat-

- ter,” *NBER Working Paper No. 25255*, 2018.
- Pew Research**, “Gender Discrimination Comes in Many Forms for Today’s Working Women,” Technical Report 2017.
- Pew Research Center**, “The Public, the Political System and American Democracy,” Technical Report 2018.
- Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy**, “Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices,” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 469–481.
- Ridley, Matthew**, “Mental Illness Discrimination,” *Working Paper*, 2022.
- Riker, William H. and Peter C. Ordeshook**, “A Theory of the Calculus of Voting,” *The American Political Science Review*, 1968, 62 (1), 25–42. Publisher: [American Political Science Association, Cambridge University Press].
- Riley, Emma**, “Role Models in Movies: The Impact of *Queen of Katwe* on Students’ Educational Attainment,” *The Review of Economics and Statistics*, 2022, pp. 1–48.
- Romano, Joseph P. and Michael Wolf**, “Stepwise Multiple Testing as Formalized Data Snooping,” *Econometrica*, 2005, 73 (4), 1237–1282.
- and —, “Efficient Computation of Adjusted p-values for Resampling-based Stepdown Multiple Testing,” *Statistics & Probability Letters*, 2016, 113, 38–40.
- Rosenbaum, Paul R.**, “Randomized Experiments,” in Paul R. Rosenbaum, ed., *Observational Studies*, Springer Series in Statistics, New York, NY: Springer, 2002, pp. 19–66.
- Roy, Andrew Donald**, “Some thoughts on the distribution of earnings,” *Oxford Economic Papers*, 1951, 3 (2), 135–146.
- Sakulku, Jaruwat**, “The impostor phenomenon,” *The Journal of Behavioral Science*, 2011, 6 (1), 75–97.
- Schwardmann, Peter and Joel van der Weele**, “Deception and self-deception,” *Nature Human Behaviour*, 2019, 3 (10), 1055–1061.
- Schwartzstein, Joshua and Adi Sunderam**, “Using Models to Persuade,” *American Economic Review*, 2021, 111 (1), 276–323.
- Scotto, Thomas J., Carla Xena, and Jason Reifler**, “Alternative Measures of Political Efficacy: The Quest for Cross-Cultural Invariance With Ordinally Scaled Survey Items,” *Frontiers in Political Science*, 2021, 3, 76.
- Segal, Carmit**, “Working when no one is watching: Motivation, test scores, and economic success,” *Management Science*, 2012, 58 (8), 1438–1457.
- Settele, Sonja**, “How Do Beliefs about the Gender Wage Gap Affect the Demand for Public Policy?,” *American Economic Journal: Economic Policy*, 2022, 14 (2), 475–508.
- Shaffer, Stephen D.**, “A Multivariate Explanation of Decreasing Turnout in Presidential Elections, 1960–1976,” *American Journal of Political Science*, 1981, 25 (1), 68–95.
- Sharot, Tali, Christoph W. Korn, and Raymond J. Dolan**, “How unrealistic optimism is main-

- tained in the face of reality,” *Nature Neuroscience*, 2011, *14* (11), 1475–1479.
- Shastri, Gauri Kartini, Olga Shurchkov, and Lingjun Lotus Xia**, “Luck or skill: How women and men react to noisy feedback,” *Journal of Behavioral and Experimental Economics*, October 2020, *88*, 101592.
- Shiller, Robert J.**, “Narrative Economics,” *The American Economic Review*, 2017, *107* (4), 967–1004.
- Shull-Senn, Shannon, Michael Weatherly, Sandra Kanouse Morgan, and Sharon Bradley-Johnson**, “Stability reliability for elementary-age students on the Woodcock-Johnson Psychoeducational Battery—Revised (Achievement section) and the Kaufman Test of Educational Achievement,” *Psychology in the Schools*, 1995, *32* (2), 86–92.
- Small, Mario L. and Devah Pager**, “Sociological Perspectives on Racial Discrimination,” *Journal of Economic Perspectives*, 2020, *34* (2), 49–67.
- Sockin, Jason**, “Show Me the Amenity: Are Higher-Paying Firms Better All Around?,” *Working Paper*, 2021.
- Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge**, “What are we weighting for?,” *Journal of Human Resources*, March 2015, *50* (2), 301–316.
- Soto, Christopher J.**, “The Little Six personality dimensions from early childhood to early adulthood: Mean-level age and gender differences in parents’ reports,” *Journal of Personality*, August 2016, *84* (4), 409–422.
- Sparkman, Gregg, Nathan Geiger, and Elke U. Weber**, “Americans Experience a False Social Reality by Underestimating Popular Climate Policy Support by Nearly Half,” *Nature Communications*, 2022, *13* (1), 4779.
- Spencer, Steven J., Christine Logel, and Paul G. Davies**, “Stereotype threat,” *Annual Review of Psychology*, 2016, *67* (1), 415–437.
- Stinnett, Terry A., J. Michael Havey, and Judy Oehler-Stinnett**, “Current test usage by practicing school psychologists: A national survey,” *Journal of Psychoeducational Assessment*, 1994, *12* (4), 331–350.
- Stroud, Laura R., Peter Salovey, and Elissa S. Epel**, “Sex differences in stress responses: social rejection versus achievement stress,” *Biological psychiatry*, 2002, *52* (4), 318–327.
- Taylor, Shelley E and Jonathon D Brown**, “Illusion and well-being: A social psychological perspective on mental health,” *Psychological Bulletin*, 1988, *103* (2), 193–210.
- Usher, Ellen L. and Frank Pajares**, “Sources of academic and self-regulatory efficacy beliefs of entering middle school students,” *Contemporary Educational Psychology*, April 2006, *31* (2), 125–141.
- Valentino, Nicholas A., Ted Brader, Eric W. Groenendyk, Krysha Gregorowicz, and Vincent L. Hutchings**, “Election Night’s Alright for Fighting: The Role of Emotions in Political Participation,” *The Journal of Politics*, 2011, *73* (1), 156–170.
- van den Akker, Alithe, Maja Deković, J.J. Asscher, and Peter Prinzie**, “Mean-level personality development across childhood and adolescence: A temporary defiance of the maturity principle

- and bidirectional associations with parenting,” *Journal of Personality and Social Psychology*, August 2014, 107.
- Waddell, Glen R.**, “Labor-market consequences of poor attitude and low self-esteem in youth,” *Economic Inquiry*, 2006, 44 (1), 69–97.
- Walsh, James, Naomi Vaida, Alin Coman, and Susan T. Fiske**, “Stories in Action,” *Psychological Science in the Public Interest*, 2022, 23 (3), 99–141.
- Wang, Shengnan, Christine M. Rubie-Davies, and Kane Meissel**, “A systematic review of the teacher expectation literature over the past 30 years,” *Educational Research and Evaluation*, 2018, 24 (3-5), 124–179.
- Weekes, Nicole, Richard Lewis, Falgooni Patel, Jared Garrison-Jakel, Dale E. Berger, and Sonia J. Lupien**, “Examination stress as an ecological inducer of cortisol and psychological responses to stress in undergraduate students,” *Stress*, 2006, 9 (4), 199–206.
- Wigfield, Allan and Jacquelynne S. Eccles**, “Expectancy–value theory of achievement motivation,” *Contemporary Educational Psychology*, January 2000, 25 (1), 68–81.
- Xue, Wen, Donald W. Hine, Anthony D. Marks, G, Wendy J. Phillips, Patrick Nunn, and Shouying Zhao**, “Combining Threat and Efficacy Messaging to Increase Public Engagement with Climate Change in Beijing, China,” *Climatic Change*, 2016, 137 (1-2), 43–55.
- Zhang, Lixuan and Christopher Yencha**, “Examining Perceptions Towards Hiring Algorithms,” *Technology in Society*, 2022, 68, 101848.
- Zimmermann, Florian**, “The dynamics of motivated beliefs,” *American Economic Review*, 2020, 110 (2), 337–361.

Appendix A

Appendix to *Perceived Discrimination at Work*

Appendix [A.1](#) contains supplementary tables and figures. Appendix [A.2](#) describes manager recruitment and the manager task. Appendix [A.3](#) describes the algorithms and how they were trained. Appendix [A.4](#) shows the elicitation of all survey-based measures and includes detailed definitions of all variables used in the analysis. Appendix [A.5](#) outlines the process used to code the free-text response-based variables. Appendix [A.6](#) reproduces the main results for the effects of perceived discrimination using an IV strategy, instrumenting for the main measure of perceived discrimination with treatment assignment in the promotion experiment. Appendix [A.7](#) contains the heterogeneity analysis that was pre-registered but not included in the main text. Appendix [A.8](#) describes the replication of the effects of perceived discrimination on future labor supply in the hiring experiment. Appendix [A.9](#) shows the robustness of the results to standard and specific concerns.

A.1 Supplementary figures and tables

Figure A.1: Information workers received about the job(s)

Panel A: Promotion experiment

Thank you for returning to complete the proofreading tasks. **You are working on a team that proofreads scientific articles.** A manager assigned tasks to you and other workers, and there are two kinds of tasks:

1. The lower-paying, easier task: If assigned to this task, you'll proofread short paragraphs from science articles written for elementary or middle school students. You will earn \$0.25 for each paragraph that you do a good job proofreading.

2. The higher-paying, but harder task: Proofread excerpts from articles published in leading scientific journals, and earn \$0.50 for each paragraph you do a good job proofreading. Because these paragraphs are so much more complicated, we also need you to summarize each paragraph in one sentence. You can earn \$0.25 for each clear, concise, and accurate summary.

In both types of tasks, about half of the paragraphs are proofread or summarized well. So people assigned to the harder task earn 2-3 times as much per paragraph.

Keep in mind that if you complete this survey, you will be eligible to be evaluated again and offered one of these two types of tasks again in the future, even if you are assigned to the easier task today.

Panel B: Hiring experiment

Thank you for returning to complete this follow-up survey.

Our previous survey included spelling, science, and grammar quizzes. We collected information on these skills because in this study, we are hiring workers to proofread scientific articles. A manager decided who to hire for the job.

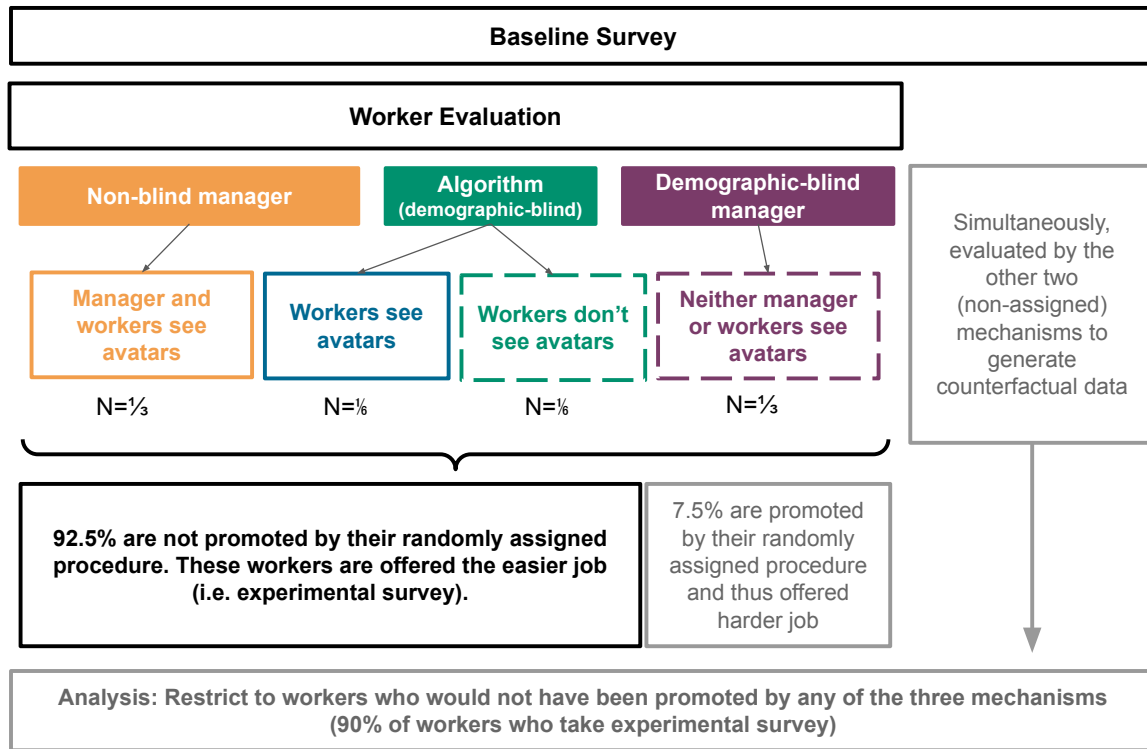
The job is to **proofread excerpts from articles published in leading scientific journals**, and earn \$0.50 for each paragraph you do a good job proofreading. Because these paragraphs are so complicated, you would also summarize each paragraph in one sentence. The job pays \$0.25 for each clear, concise, and accurate summary.

About half of the paragraphs are proofread or summarized well, so the average worker would earn about \$0.38 per paragraph that they proofread and summarize.

Keep in mind that if you complete this survey, you will be eligible to be evaluated again and potentially hired again in the future, even if you are not hired today.

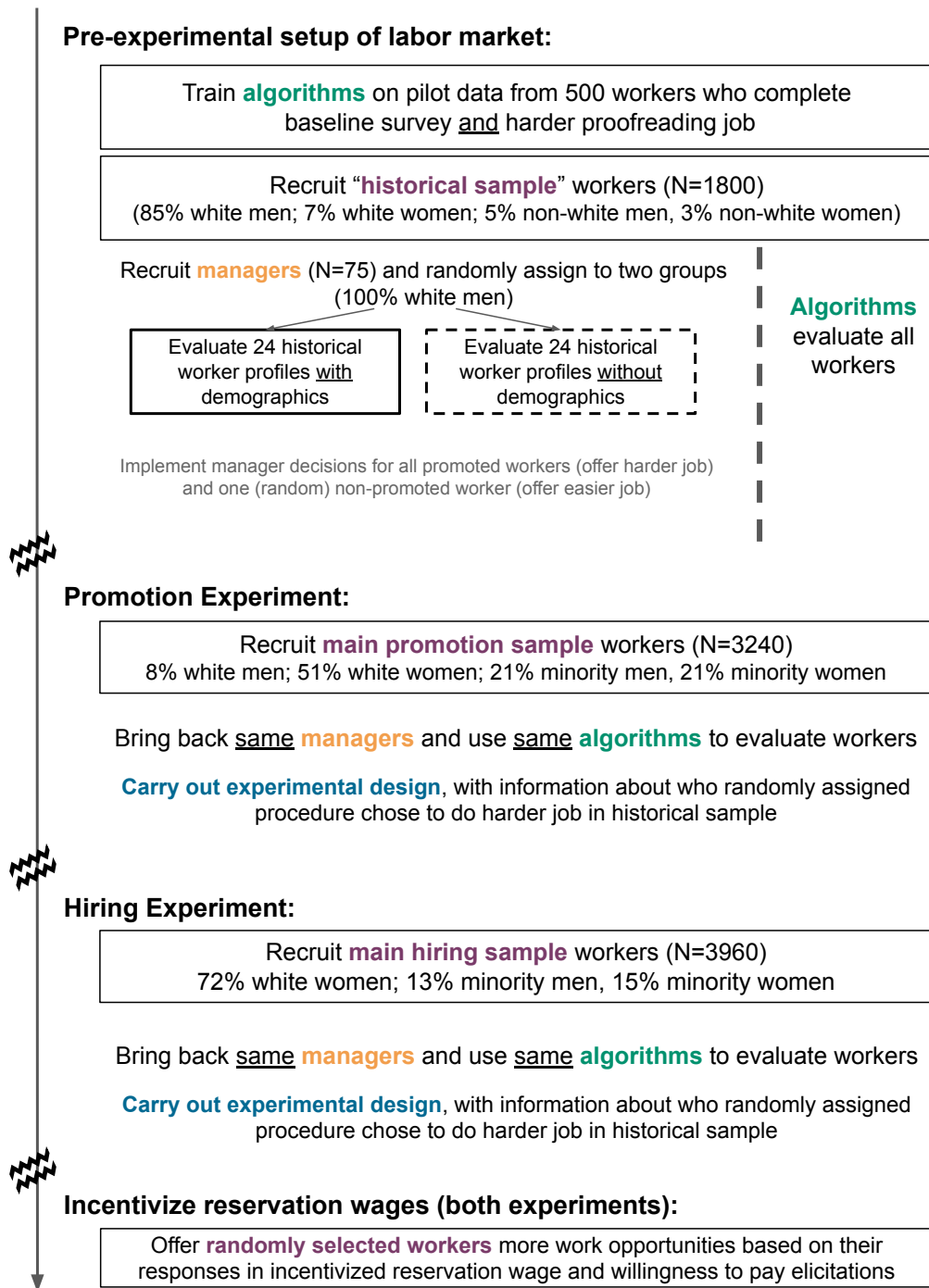
Note: This figure shows the information that workers received about the easier, lower-paying (“non-promotion”) job and the harder, higher-paying (“promotion”) job in the promotion experiment (Panel A) or the job that they were not hired to do in the hiring experiment (Panel B).

Figure A.2: Experimental design of the promotion experiment



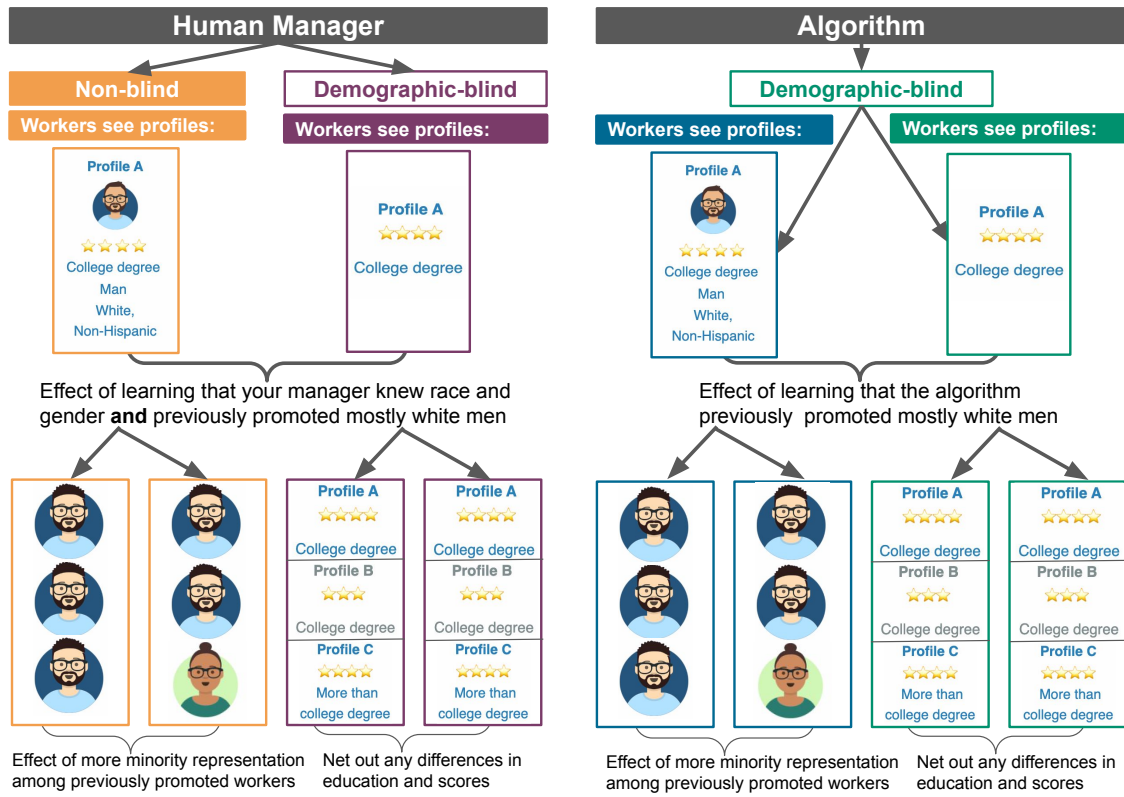
Note: This figure describes the design of the promotion experiment. The baseline survey includes grammar, science, and spelling quizzes and confidence in each, demographics, work history, and past experiences of and beliefs about the prevalence of discrimination. The set-up for workers in the historical sample was similar, except that only one worker in the non-promoted group was offered the easier job and did not complete the survey afterwards to reduce costs. The design of the hiring experiment was the same except that workers were assigned to be evaluated by one of four procedures (see Figure A.9). In addition, 97.5% of workers were not hired under their randomly assigned procedure.

Figure A.3: Timeline of study



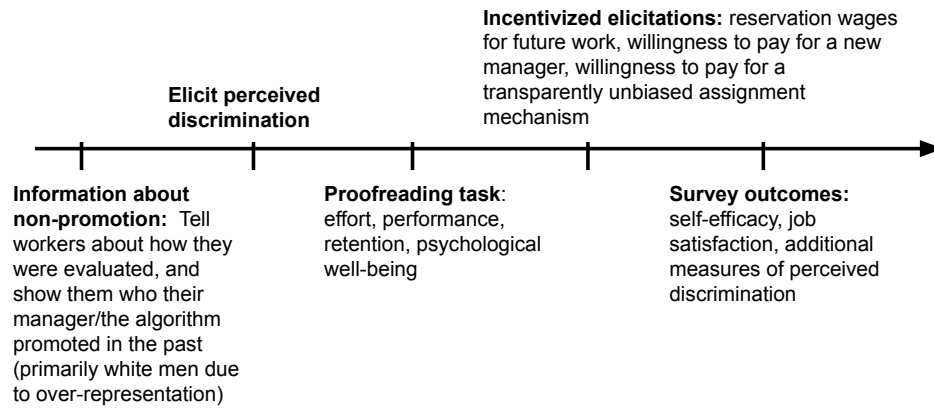
Note: This figure describes the dynamics of the labor market created on Prolific to generate repeated interactions with workers and managers. The pre-experimental setup describes the recruitment of the historical sample and managers. The same managers were brought back to evaluate the workers in the experimental samples. The main difference between the historical and experimental samples was the demographic composition, in order to approximate a scenario in which under-represented workers saw that their manager had mostly promoted workers from the majority group in the past. After the conclusion of the experiment, randomly-selected workers are offered additional proofreading jobs that implement the scenarios in the reservation wage and willingness to pay elicitation.

Figure A.4: Design of promotion experiment



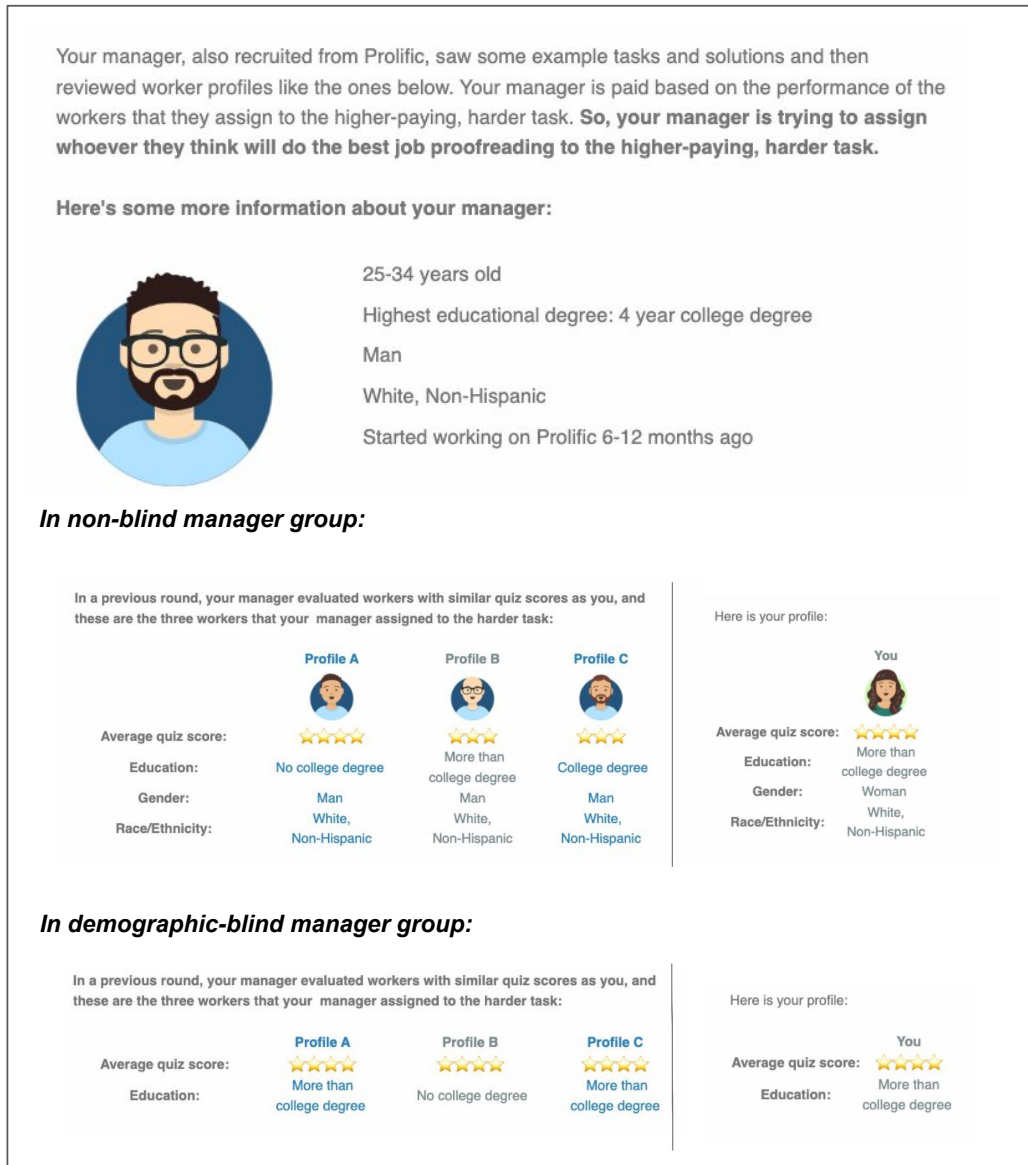
Note: This figure illustrates the treatment arms in the promotion experiment and how they are used to answer various research questions. Randomization is indicated by gray arrows. In the manager sample, workers are randomly assigned to managers. Some previously promoted three white men, some promoted two white men and someone else. This is uncorrelated with other manager characteristics (Appendix Table A.14). In the algorithm sample, workers are randomly assigned to groups of workers jointly evaluated by the algorithm (analogously to how the managers jointly evaluated groups of workers). Similarly, in some groups, three white men were previously promoted, and in others, it was two white men and someone else.

Figure A.5: Timeline of experimental survey (promotion experiment)



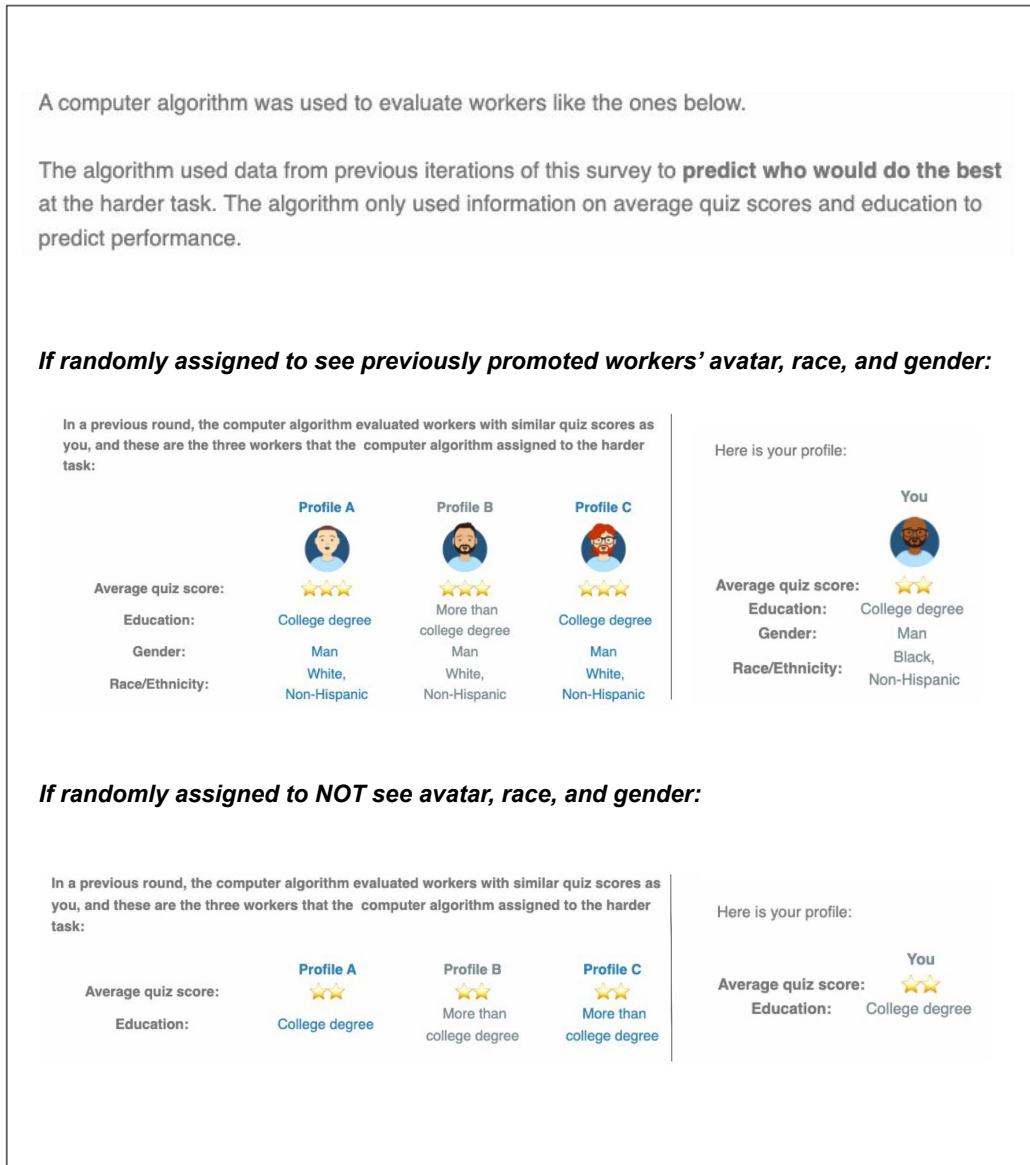
Note: This figure describes the timeline of the experimental survey in the promotion experiment. The experimental survey in the hiring experiment was the same except that workers did not complete the easier proofreading job so I do not observe the corresponding outcome variables, and the survey outcomes additionally included comprehension questions about the hiring procedures.

Figure A.6: Treatment variation in manager groups



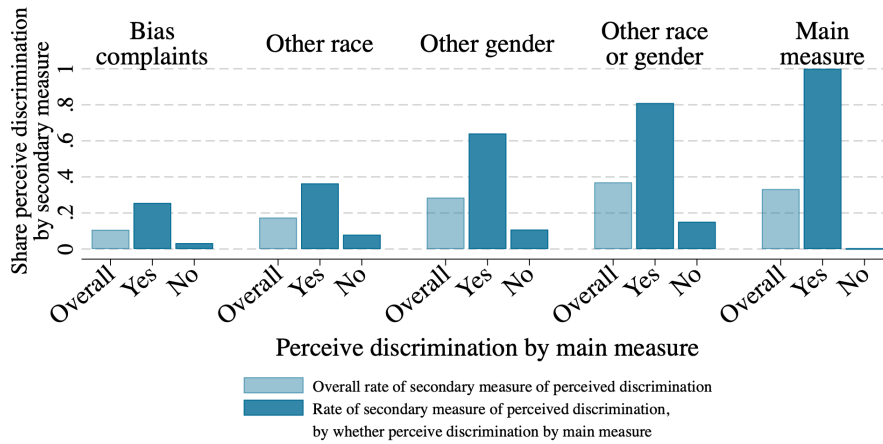
Note: This figure shows what workers saw about their manager, regardless of whether their manager saw avatars, race, and gender or not, and the information they saw about the workers their manager previously promoted in the non-blind manager group and demographic-blind manager group.

Figure A.7: Treatment variation in algorithm groups (promotion experiment)



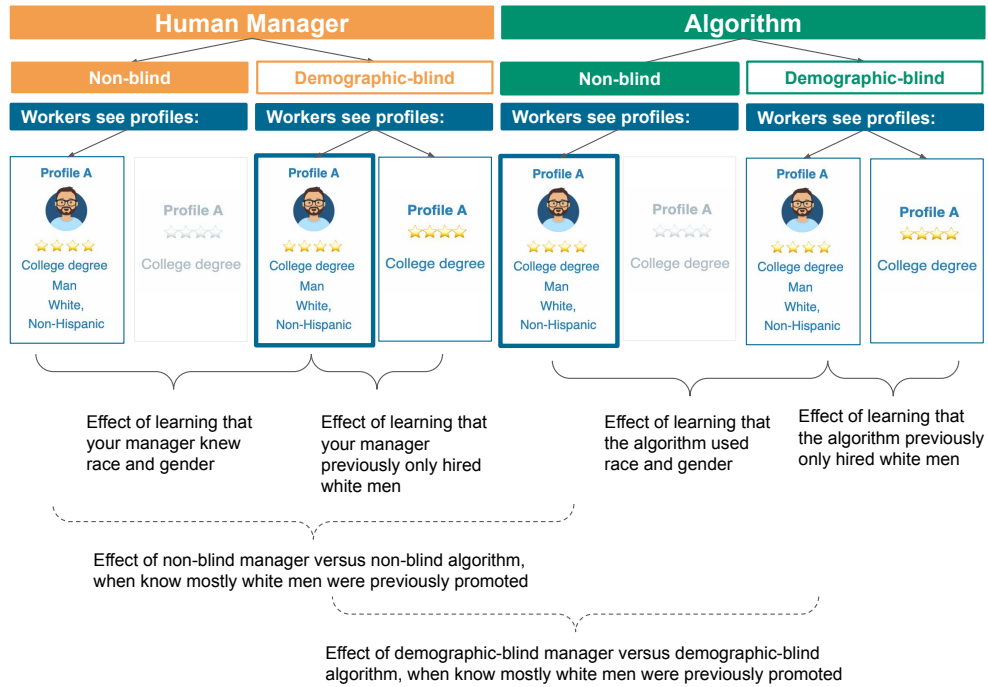
Note: This figure shows what workers saw about the algorithm and the information they saw about the workers their manager previously promoted depending on whether they were randomly assigned to see those workers' avatars, race, and gender or not.

Figure A.8: Correlation of perceived discrimination measures



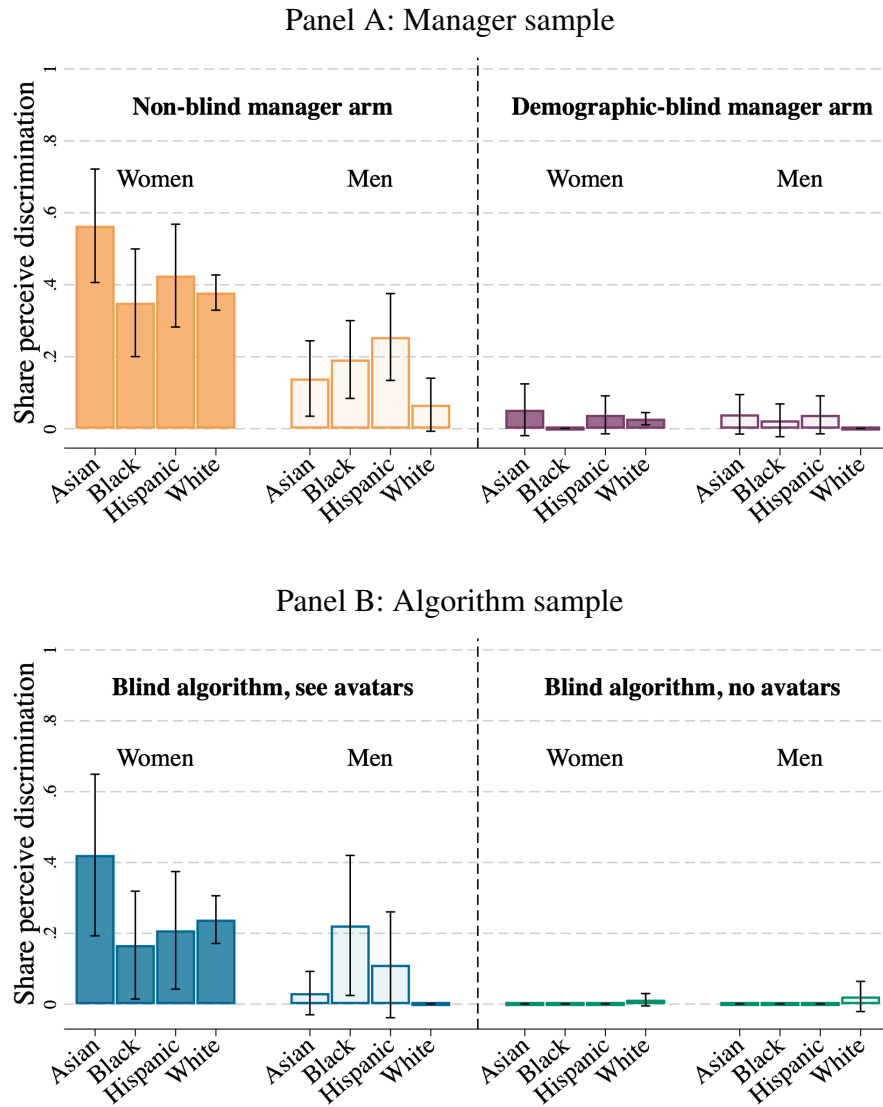
Note: This figure shows the correlation between the primary and secondary measures of perceived discrimination and the overall frequency of each in the non-blind manager arm of the promotion experiment. The primary measure is an indicator for whether a worker mentions demographics in their response to the open-ended question, “*what needed to be different about your profile in order to be assigned to the harder task?*” The secondary measures are an indicator for mentioning discrimination or bias in a open-ended response describing their complaint about the promotion procedure if they said they had one, an indicator for saying they thought they would have been promoted if they had a different race, the same for gender, and an indicator that is the maximum of the two. The light blue bars plot the share of workers who perceive discrimination by each measure in the full sample, and then the dark blue bars split the sample into workers who perceive discrimination by the main measure and those who don’t.

Figure A.9: Design of hiring experiment



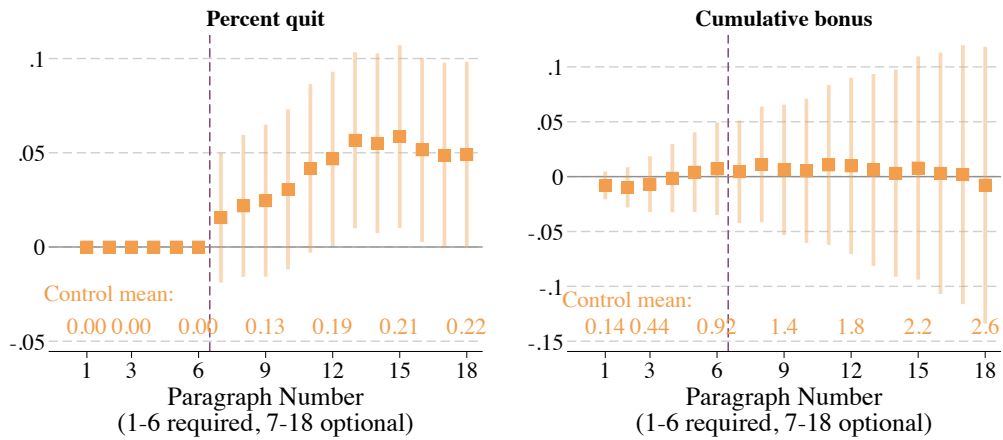
Note: This figure illustrates the treatment arms in the hiring experiment, which replicates the four original treatment arms and adds two new arms, bolded. The grayed-out boxes indicate arms that would be included in the full 2x2x2 factorial design, but are not truthfully or realistically implementable in a way parallel to the other arms. Workers are randomized equally across the six arms, as indicated by the gray arrows. The figure shows how comparisons between the arms are used to answer various research questions.

Figure A.10: Perceived discrimination by race \times gender



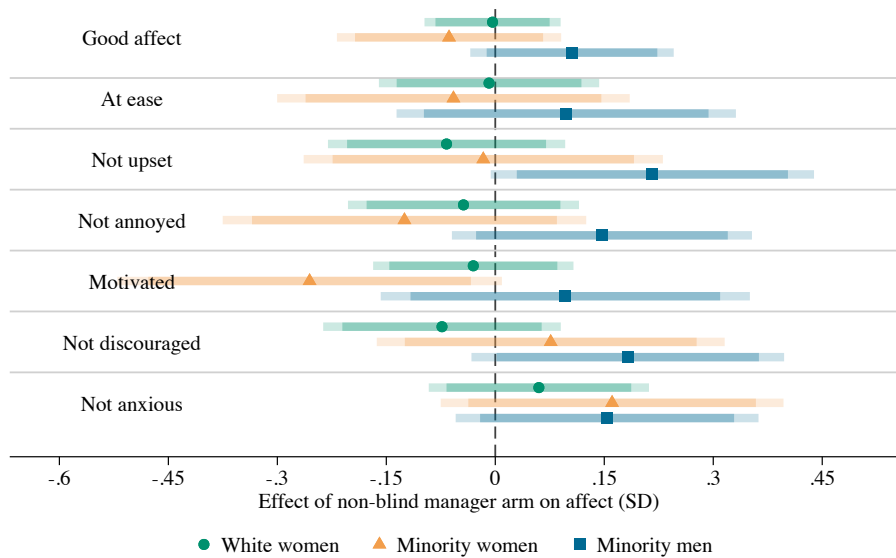
Note: This figure plots the share of workers in the experimental sample who perceived discrimination using the main measure (mentioning demographics in their response to the open-ended question, “*what needed to be different about your profile to be assigned to the harder, higher-paying task?*” separately by race \times gender in the promotion experiment, separately by treatment arm. 95 percent confidence intervals are indicated by the black bars.

Figure A.11: Effects on retention and overall earnings in the promotion experiment



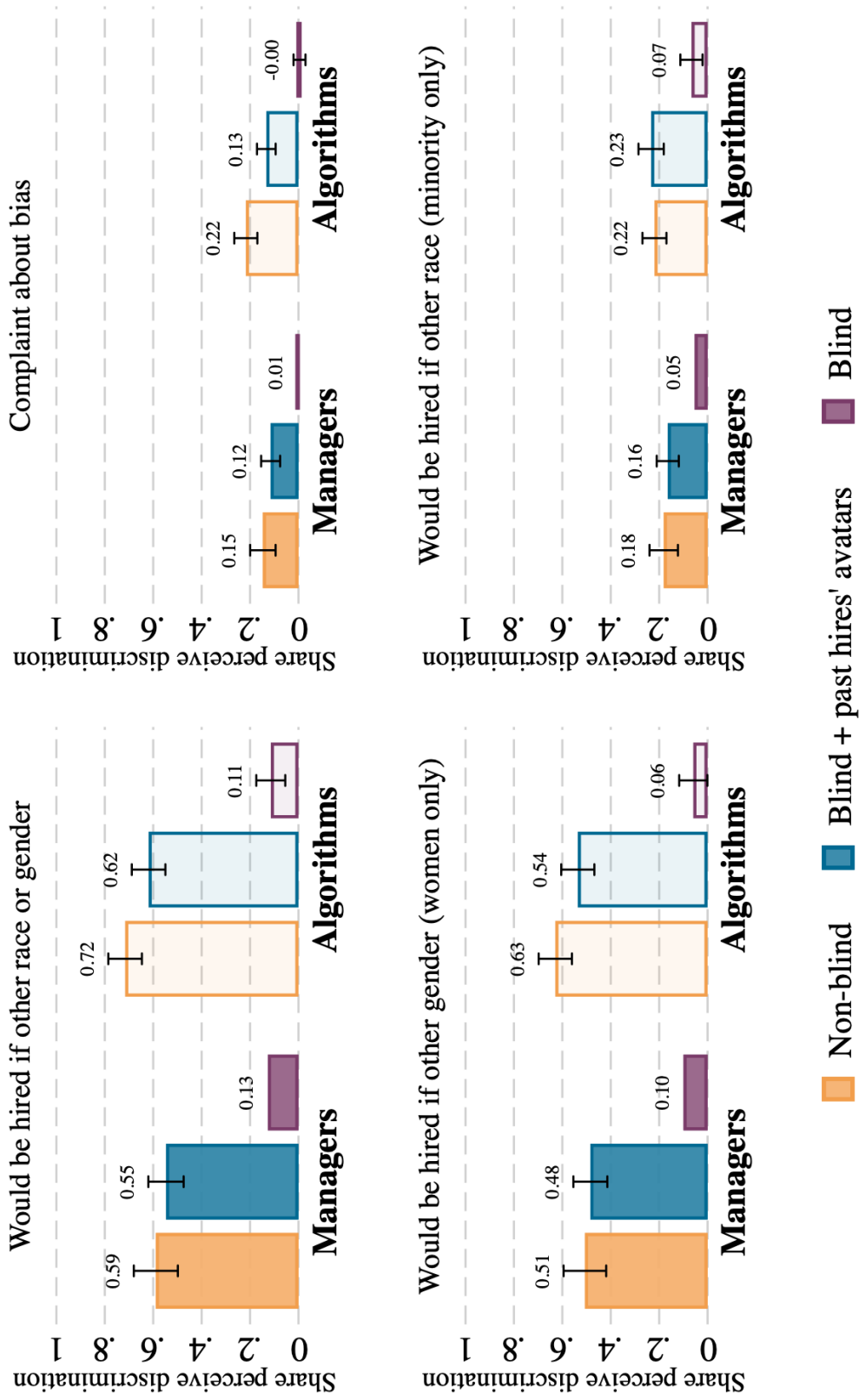
Note: This figure plots the effect of being in the non-blind manager arm relative to the demographic-blind manager arm (Panel A) and the effect of seeing the avatars, race, and gender of previously-promoted workers in the algorithm arm (Panel B) of the promotion experiment on whether workers have quit and how much they have earned, cumulatively by the given paragraph on the x-axis. Workers were required to proofread the first six paragraphs, after which they could quit after any paragraph, so treatment effects absent selection are available in the first six paragraphs. The regressions restrict to workers who would not have been promoted under any promotion procedure. Regressions control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.12: Effects on psychological well-being, gender and race heterogeneity



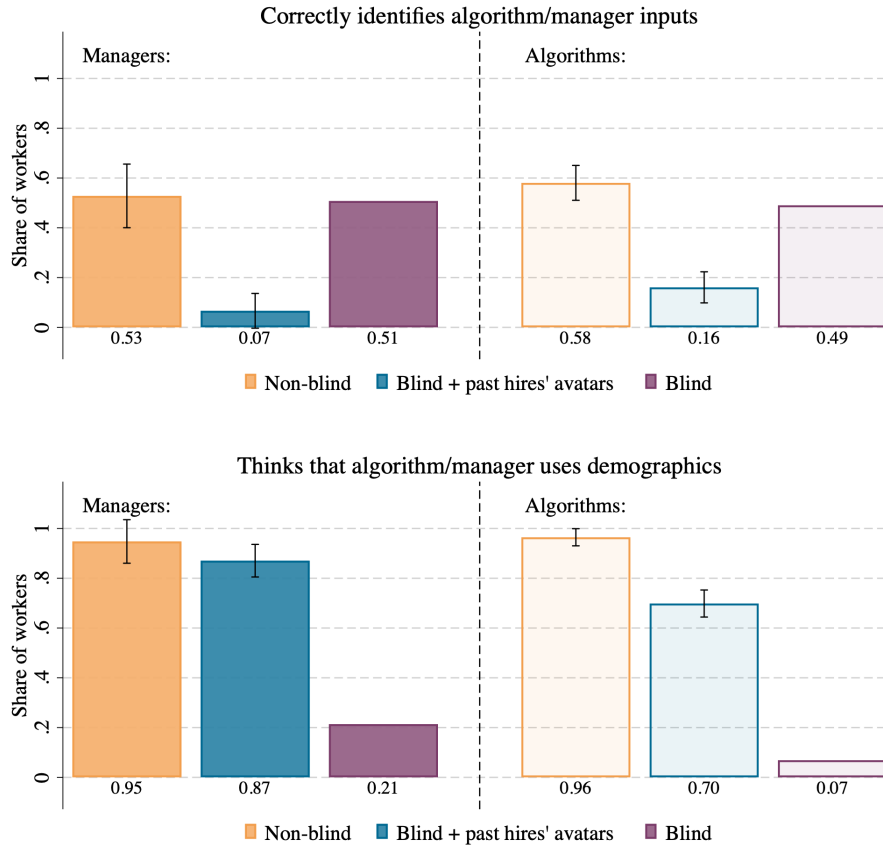
Note: This figure plots the effect of being in the non-blind manager arm relative to the demographic-blind manager arm in the promotion experiment on emotional states, measured with an affective well-being scale and reported in standard deviations, separately by race and gender. The regressions restrict to workers who would not have been promoted under any promotion procedure. Regressions control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.13: Perceived discrimination in the hiring sample, secondary measures of perceived discrimination



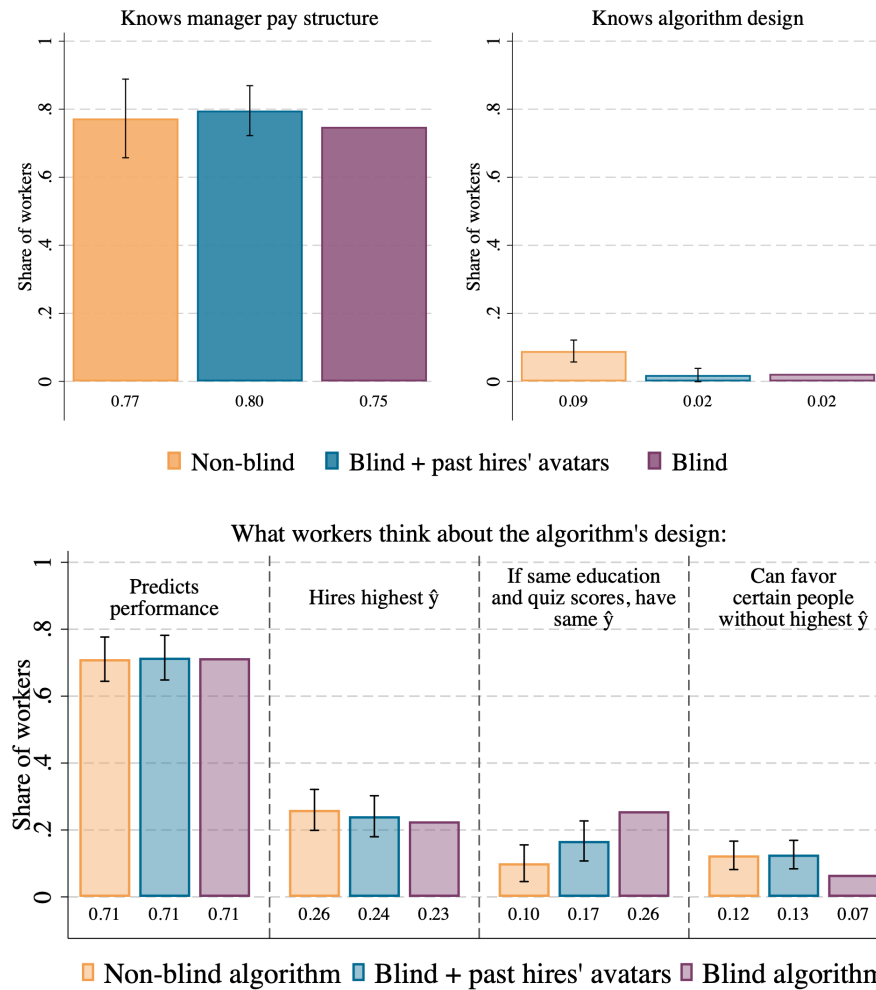
Note: This figure replicates Figure 1.4 using the four secondary measures of perceived discrimination. In the top graphs, the sample is restricted to workers who would not have been hired under any procedure and those who saw that three white men previously promoted. In the bottom graphs, the sample is further restricted to women only when perceived gender discrimination is the outcome and racial minorities only when perceived racial discrimination is the outcome. The secondary measures of perceived discrimination are the same as in Appendix Figure A.8. 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

Figure A.14: Comprehension of hiring procedure inputs



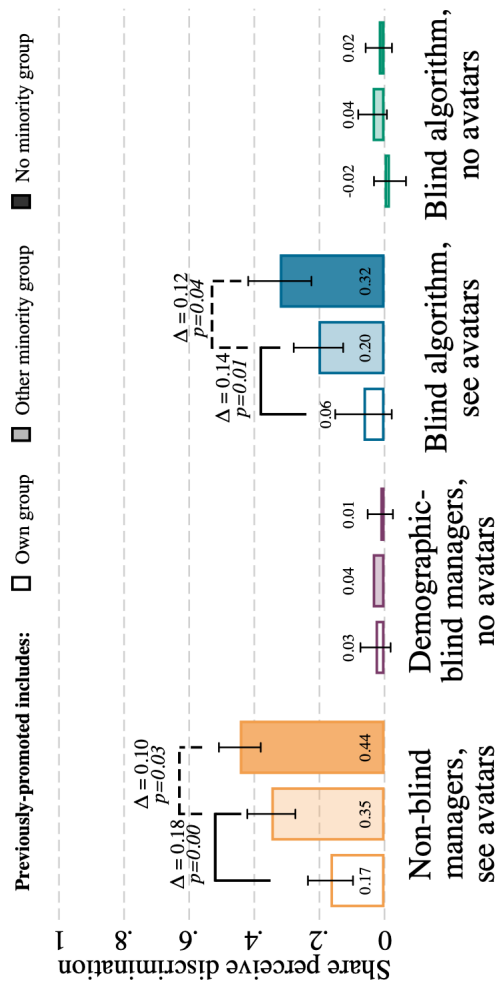
Note: This figure shows worker knowledge of the decision-making processes by treatment arm in the hiring experiment. In the top panel, the y-axis is the share of workers correctly identifying the inputs (what decision-makers know about workers) to hiring decision-makers. When asked what information their manager or the algorithm had about workers when deciding who to hire, this requires checking off “An avatar,” “Gender,” “Race/Ethnicity,” “1-5 stars representing the average quiz scores,” and “Education” (in the non-blind group), and just quiz scores and education in the demographic-blind groups, and **not** checking off “Age,” “All three numeric quiz scores from the baseline survey (science, spelling, and grammar),” “The average quiz score (for example, 85%),” “Work history,” and “Time on Prolific.” In the bottom panel, it is the share of workers who think that the decision-maker knew either their race and gender (i.e., checked either of those boxes on the same question just described). The sample is restricted to workers who would not have been promoted under any hiring procedure, and workers who saw that all three previous hires were white men (a random subsample within each treatment arm—results are the same for the full sample). Shares are calculated via regressions that control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

Figure A.15: Comprehension of hiring procedure incentives (manager arms) and design (algorithm arms)



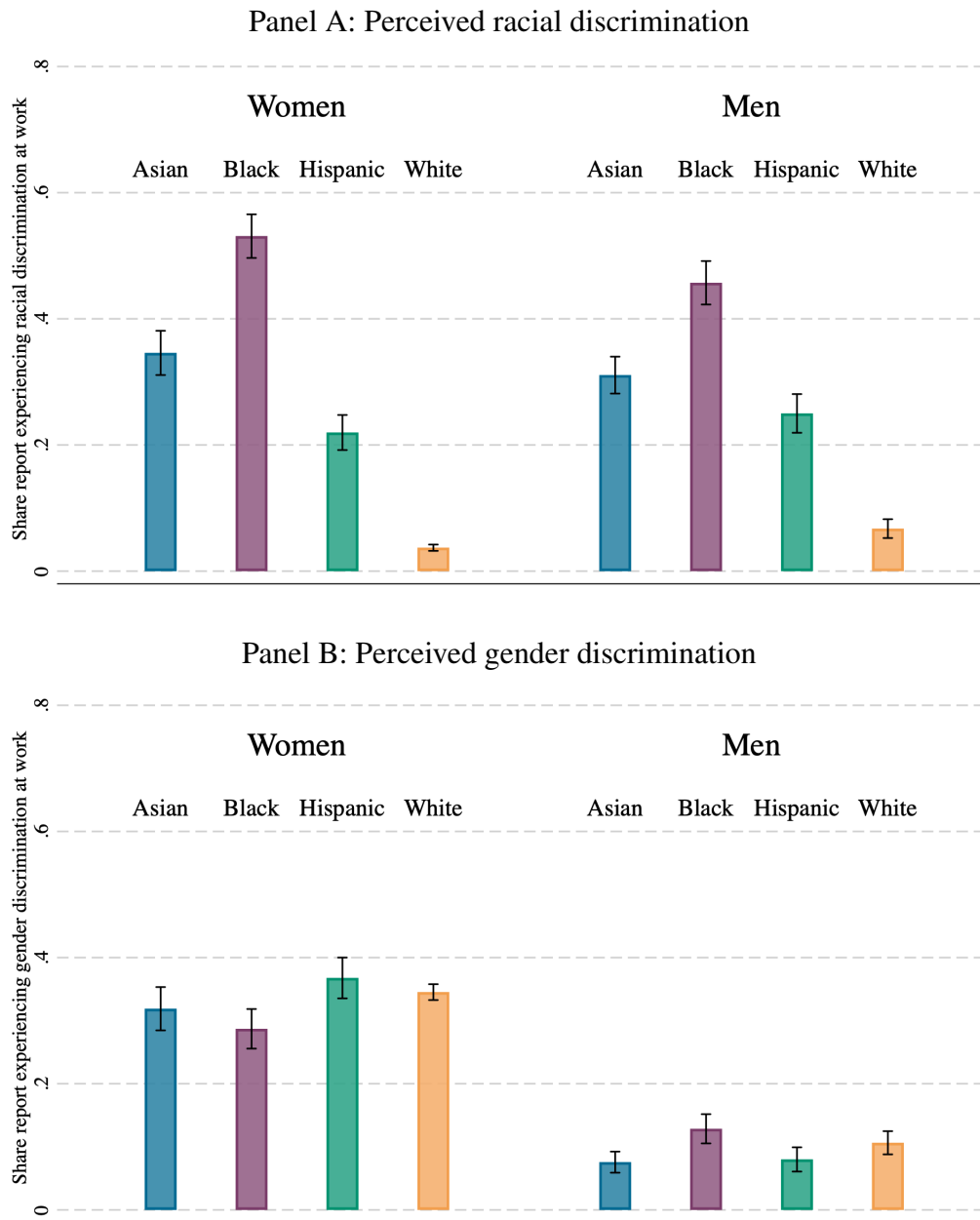
Note: This figure shows worker knowledge of managers' incentives and algorithm design. In the top left, it plots the share of workers in the manager arms who know the manager's pay structure. When asked, "what was the basis of your manager's bonus payment?," this requires checking off "The performance in the proofreading task of the workers they hired," and "How many workers they hired showed up to do the proofreading task," and **not** checking off, "How many workers they hired," "Whether they chose workers with the highest screening quiz scores," or "Whether they chose diverse workers." On the right, it plots the share of workers in the algorithm arms who correctly identify the algorithm's design. When asked "what do you know about how the algorithm was designed?," this requires checking off "It was predicting worker performance at the proofreading task," "It hired whoever had the highest predicted [performance]," and, in the demographic-blind decision-maker groups only, "It would provide the same predicted [performance] for any worker with the same education and quiz score," and **not** checking off "It was predicting whether a manager would have hired a worker," or "It could favor certain people and hire them even if they didn't have the highest predicted [performance]." The lower panel plots the share of workers checking off each of these options. Approximately zero workers thought it was predicting managers' previous decisions. The sample is restricted to workers who would not have been promoted under any hiring procedure, and workers who saw that all three previous hires were white men (a random subsample within each treatment arm—results are the same for the full sample). Shares are calculated via regressions that control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

Figure A.16: Effects of seeing one previously-promoted minority-group worker of a same or different demographic group



Note: This figure plots the share in each treatment arm of the promotion experiment perceiving discrimination by the main measure, separately by whether workers (would) see three white men previously promoted or not, analogous to Figure 1.5, but now splitting the sample who see two white men and one member of a minority group by whether the minority-group representative is of the worker's own demographic group or a different one. The sample restricts to workers who would not have been promoted under any procedure and are not white men. Shares are calculated via regressions that control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. p -values and 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

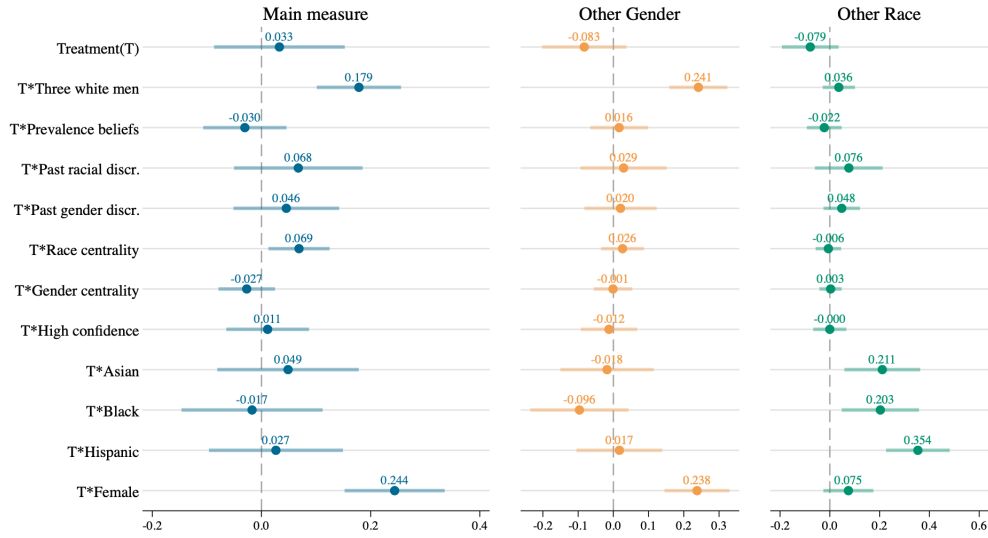
Figure A.17: Reported past experiences of discrimination at work in the experimental sample



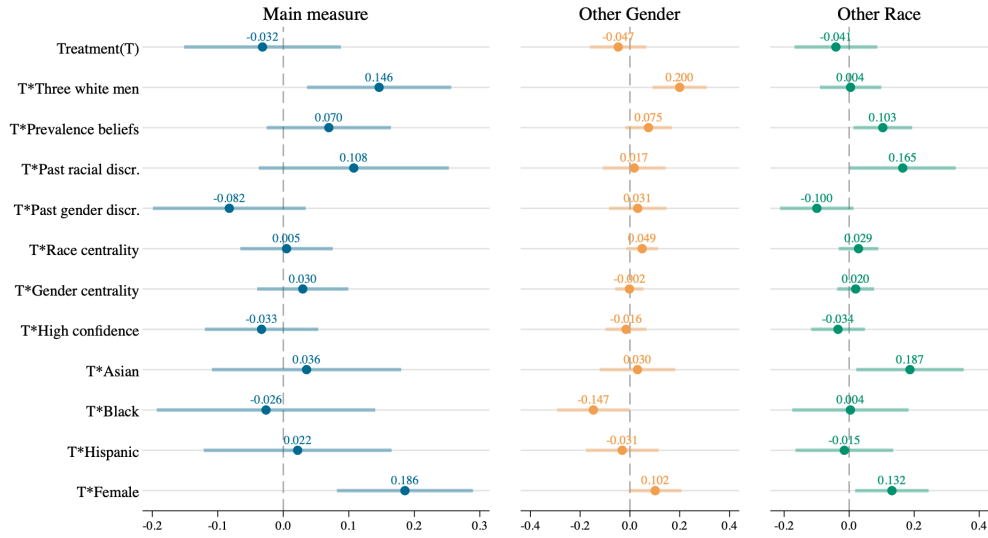
Note: The sample is 3,240 workers recruited for the promotion experiment on Prolific. The share of each group who reported having experienced discrimination at work in the past is plotted with standard errors in this sample indicated by black bars. Perceived discrimination includes during job search, promotion, termination, and daily work activities. “Asian” refers to those who identify as Asian (no participants identified as Asian and Hispanic), “Black” is those who identify as Black or African American, regardless of whether they identify as Hispanic, “Hispanic” is those who identify as white and Hispanic, and “White” refers to those who identify as white and non-Hispanic. 9 non-binary participants are included with “Women” (those who are more likely to experience and perceive discrimination).

Figure A.18: All predictors of perceived discrimination in the promotion experiment

Panel A: Manager sample



Panel B: Algorithm sample



Note: This figure plots coefficients from a regression of the primary and secondary measures of perceived discrimination on an indicator for treatment, that indicator interacted with each of the worker characteristics listed on the left, the worker characteristics, and the control variables in the main specification. The sample is workers who would not have been promoted under any procedure in the promotion experiment. In Panel A, *Treatment* is assignment to the non-blind manager arm among those evaluated by a manager. In Panel B, it is assignment to see previously-promoted workers' avatars, gender, and race among those evaluated by the algorithm. 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.19: Avatar-making procedure

Step 1:
In the next few questions, you will **build an avatar** that looks most like you.

Which is the closest to your skin tone? Which is the closest to your **hair color**?



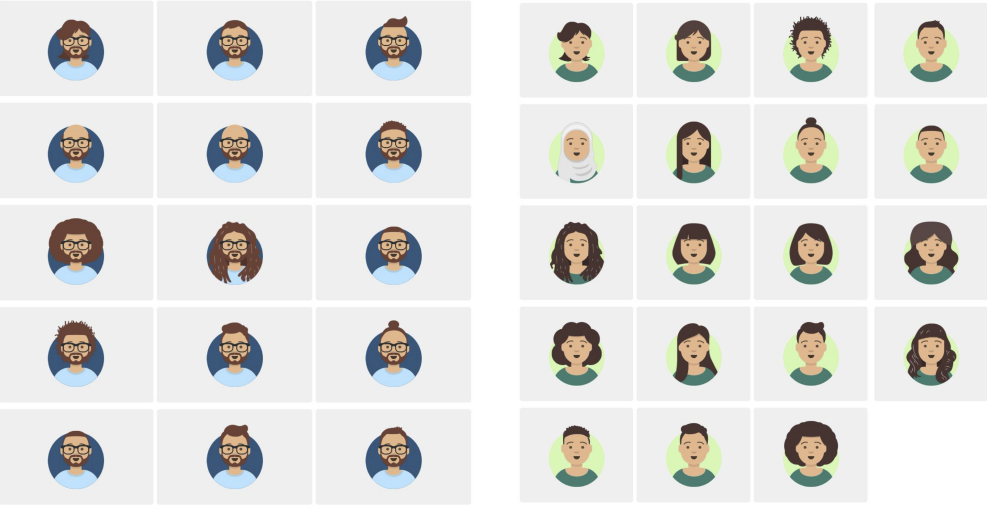
Step 2:
Workers see the following, with the skin tone and hair color they have chosen:

Do you wear glasses, or not? Do you have facial hair, or not?
(self-identified men only)



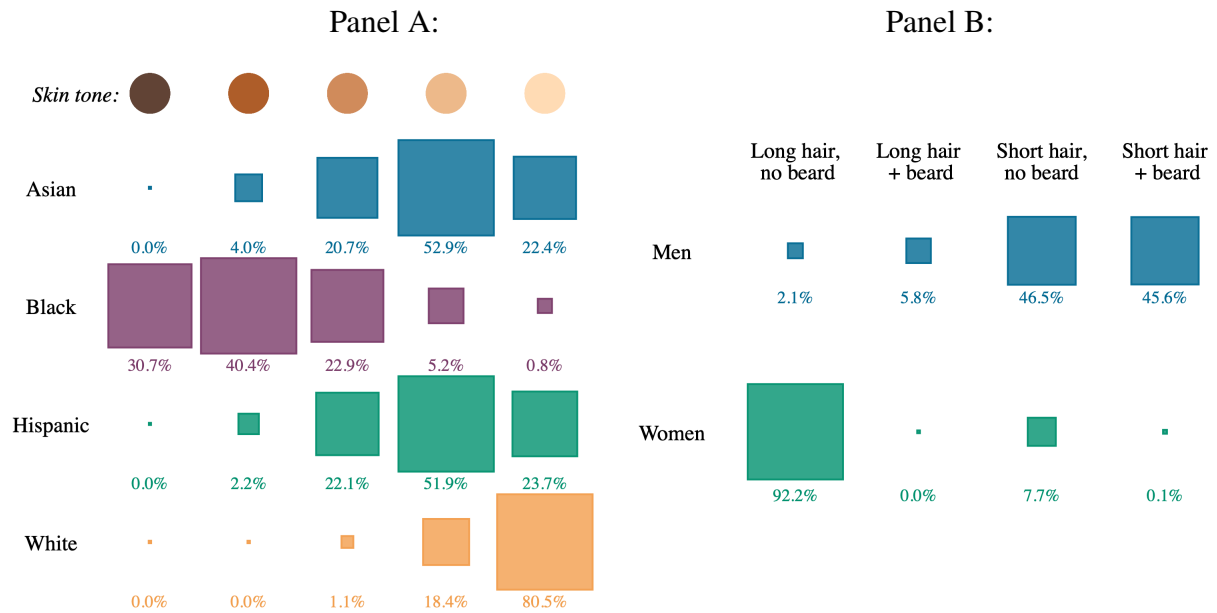
Workers see the set of possible hair styles, with all of their other choices implemented. If a man had chosen that they have a beard and glasses, they would see the options on the left. If a woman had chosen that they did not have glasses, they would see the options on the right. Option order was randomized.

Step 3:
We've put it all together and now you just need to choose a hairstyle.
Please choose the avatar that looks **the most like you**:

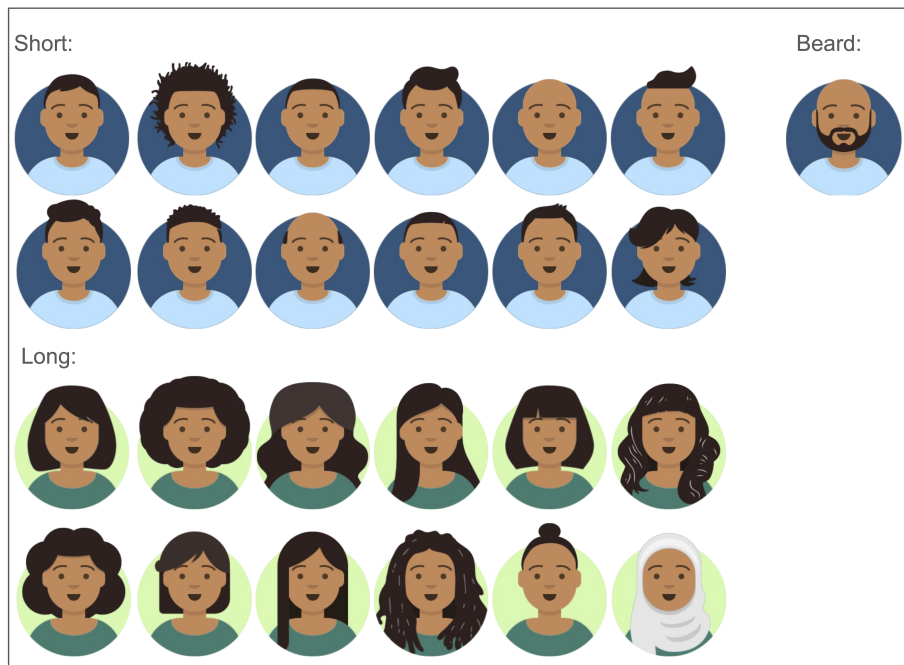


Note: This figure depicts the procedure by which workers created their avatars in the baseline survey.

Figure A.20: Avatar characteristics by self-reported race and gender

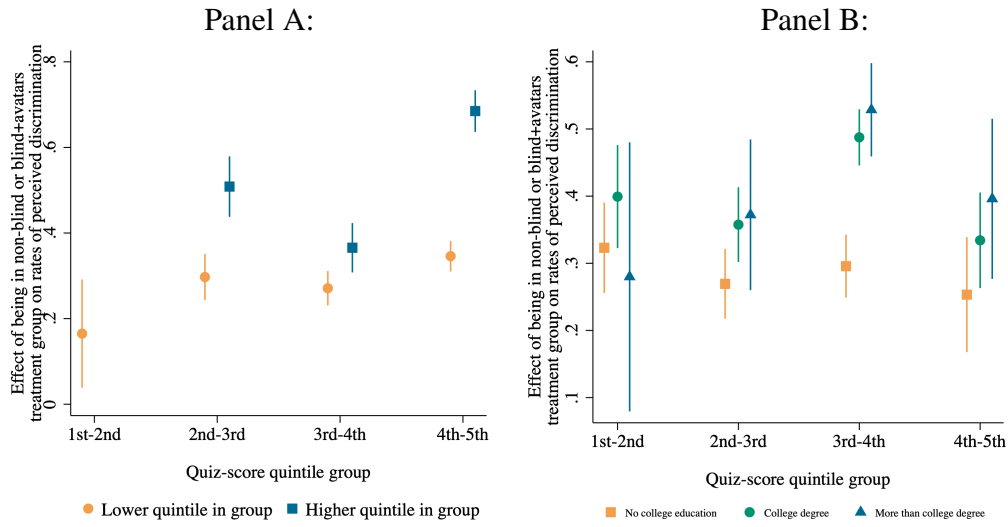


Panel C:



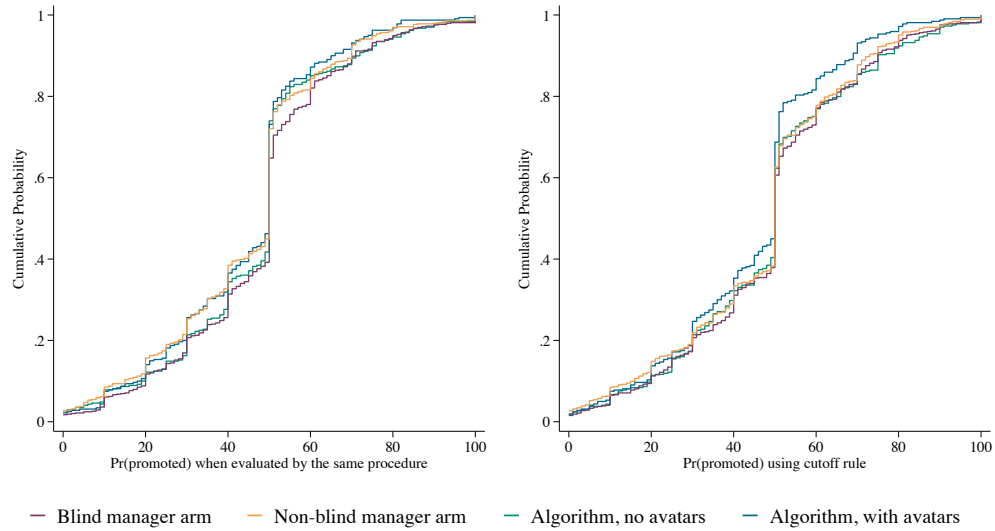
Note: The sample in panels A and B is all workers in the hiring and promotion experiments who took the baseline survey (N=6720). Each row plots the share of participants in that racial or gender group who choose each skin tone (Panel A) or hair style (Panel B). Panel C shows the hair styles by whether they are “long” or “short” and the representation of a beard; men could choose to add a beard to any hairstyle. Men could choose long hair with blue shirts and women could choose short hair with green shirts; these are not shown for the sake of space.

Figure A.21: Perceived discrimination by quiz-score quintile group and quintile



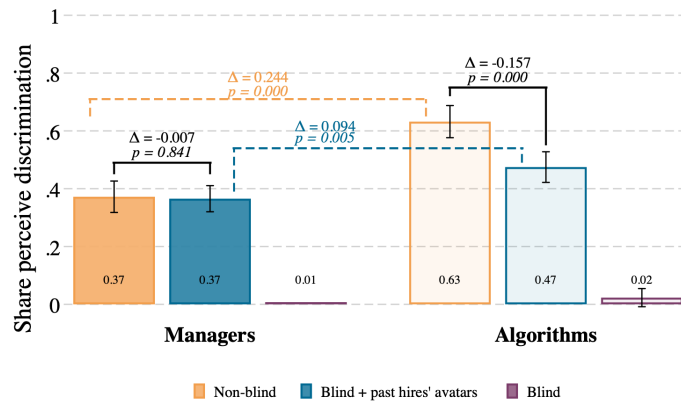
Note: The sample is all workers in the hiring and promotion experiments who would not have been hired/promoted by any of the relevant mechanisms, and the “treatment effect” pools the effect of being in any of the arms with positive perceived discrimination in Figures 1.1 and 1.4 relative to the arms with no perceived discrimination (the arms with demographic-blind decision-makers and no demographics of previously-selected workers shown); the experiments and treatment arms are pooled for power and ease of exposition but the result is very similar when looking at the non-blind manager arm in the promotion experiment separately (the other arms are too small to consider separately). In each panel, the treatment effects by quiz-score quintile group and whether the worker is in the higher quintile in their group or the lower quintile in their group (panel A) or whether they have no degree, a college degree, or more than a college degree (panel B) are jointly estimated in one regression with all of the main analysis control variables. 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.22: CDFs of workers' beliefs about the likelihood of future promotion



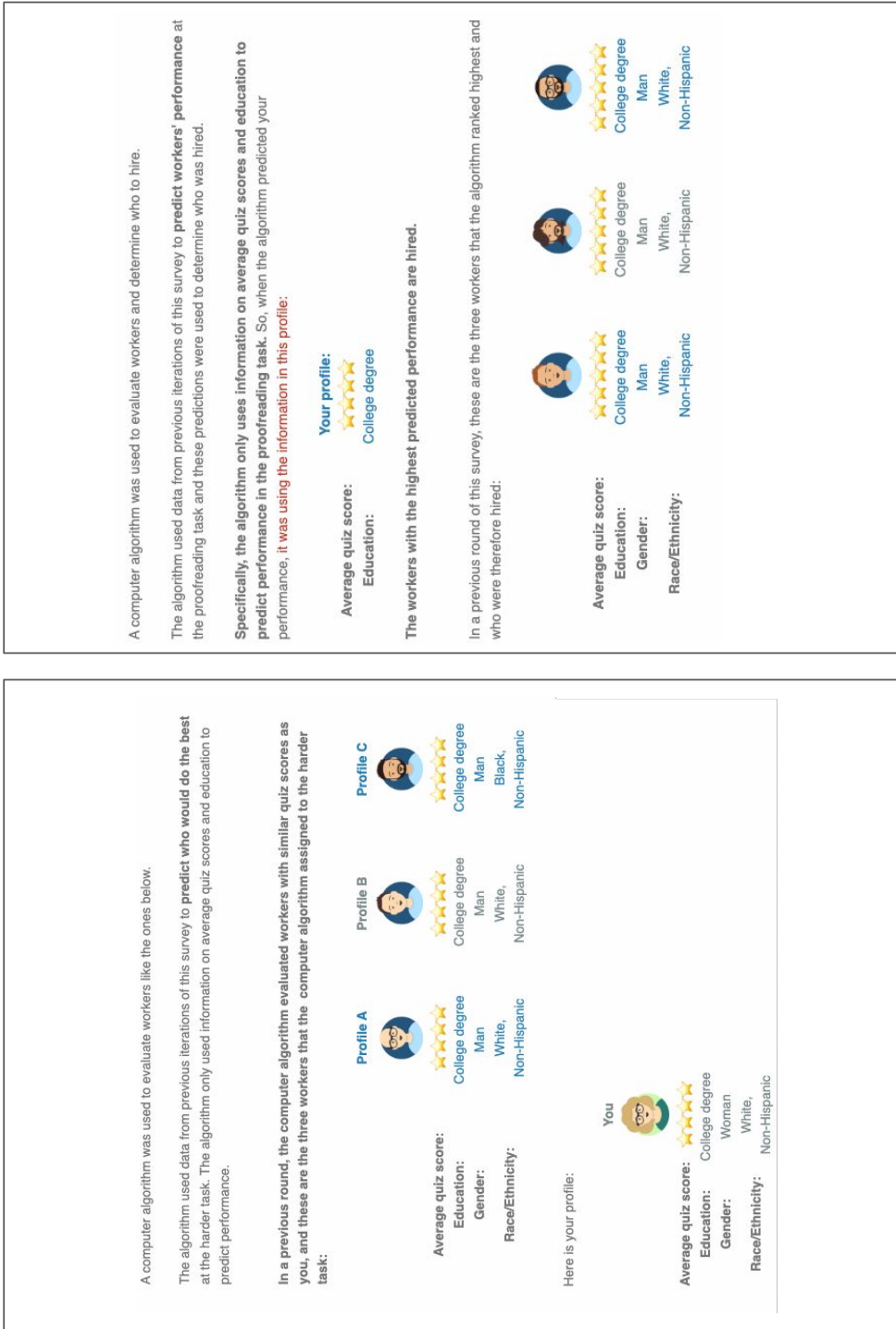
Note: This figure plots the raw CDF of workers' reported beliefs about the likelihood that they will be selected to do the harder task in the future, separately by treatment arm in each experiment. The sample is workers who would not have been selected under any of the relevant promotion or hiring procedures.

Figure A.23: Perceived discrimination in the hiring experiment (full sample)



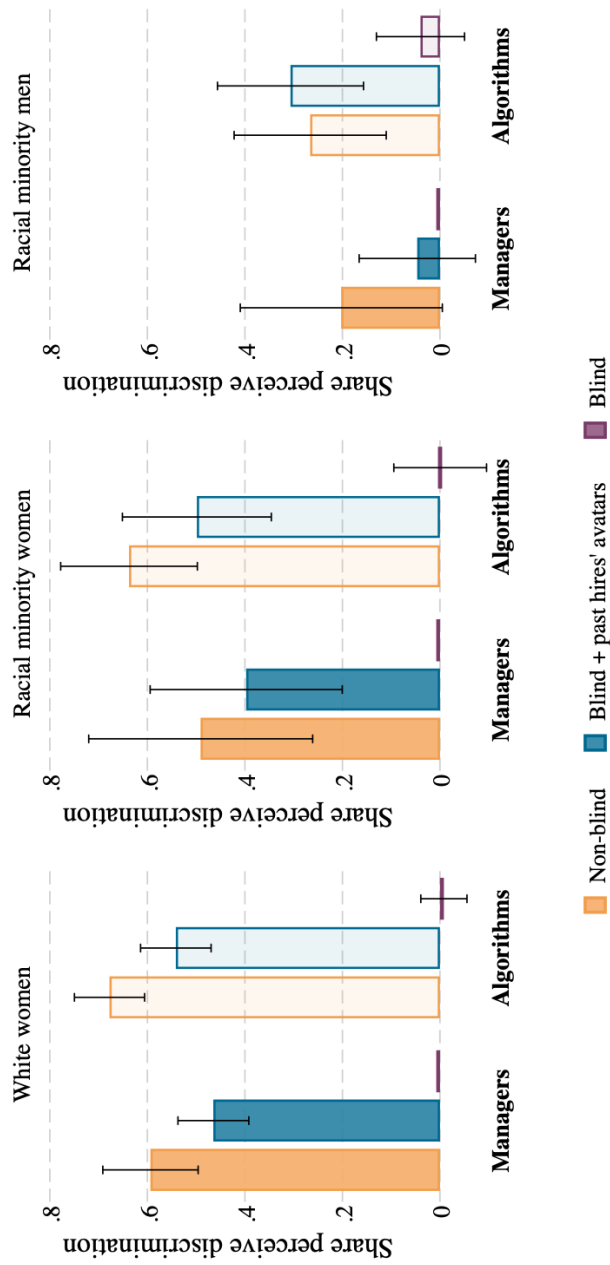
Note: This figure replicates Figure 1.4 but for all workers regardless of the demographic makeup of the previously-hired workers they saw, without accounting for differences across hiring procedures in the rate at which workers saw that all three previous hires were three white men versus seeing two white men and some-one else. It plots the share of workers perceiving discrimination in each treatment arm of the hiring experiment, using the main measure of perceived discrimination. The sample is restricted to workers who would not have been promoted under any hiring procedure. p -values and 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

Figure A.24: Making information about decision-making inputs visually salient in the hiring experiment



Note: This figure shows the only difference in implementation of the replicated treatment arms in the hiring experiment relative to the promotion experiment. In the promotion experiment, workers evaluated by a non-blind algorithm but who saw previously-promoted workers' avatars, race, and gender may have missed the sentence explaining what information the algorithm used (on the left). In the hiring experiment (on the right), this was made equivalently salient by communicating it in words and visually, like the information about the previously-promoted workers, using workers' own profile. This means that workers were also not shown their own demographics in their profile. For consistency, the change in placement of the workers' own profile from the bottom to the top was also changed in the other treatment arms, but in those cases, there was no change in what information was provided.

Figure A.25: Perceived discrimination in the hiring experiment, heterogeneity by race \times gender



Note: This figure replicates Figure 1.4 separately by race \times gender. Regressions are estimated separately in each sample. The sample is 1806 white women, 389 racial minority women, and 332 racial minority men. 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

Table A.1: Balance table, promotion experiment

	Manager arms			Algorithm sub-arms		
	Blind	Non-blind	<i>p-value</i>	No avatars	Avatars	<i>p-value</i>
	(1)	(2)	(1)=(2)	(3)	(4)	(3)=(4)
Male	0.28	0.27	<i>0.84</i>	0.35	0.31	<i>0.25</i>
Race:						
Asian	0.13	0.12	<i>0.62</i>	0.14	0.16	<i>0.42</i>
Black	0.13	0.13	<i>0.93</i>	0.17	0.13	<i>0.19</i>
White, Hispanic	0.15	0.14	<i>0.64</i>	0.14	0.13	<i>0.66</i>
White, Non-Hispanic	0.59	0.61	<i>0.46</i>	0.55	0.58	<i>0.50</i>
Married	0.54	0.52	<i>0.37</i>	0.56	0.54	<i>0.60</i>
Kids	0.43	0.45	<i>0.46</i>	0.42	0.42	<i>0.85</i>
Education:						
Less than high school	0.01	0.01	<i>0.40</i>	0.01	0.01	<i>0.91</i>
High school graduate	0.14	0.15	<i>0.65</i>	0.11	0.13	<i>0.35</i>
Some college but no degree	0.29	0.27	<i>0.36</i>	0.25	0.27	<i>0.54</i>
2 year college degree	0.10	0.10	<i>0.78</i>	0.10	0.07	<i>0.23</i>
4 year college degree	0.36	0.34	<i>0.39</i>	0.39	0.41	<i>0.62</i>
Professional or Masters degree	0.09	0.13	<i>0.03</i>	0.13	0.10	<i>0.15</i>
Doctorate	0.01	0.02	<i>0.13</i>	0.01	0.01	<i>0.85</i>
Income:						
Less than \$20,000	0.13	0.14	<i>0.44</i>	0.15	0.14	<i>0.78</i>
\$20,000-\$40,000	0.24	0.21	<i>0.14</i>	0.20	0.23	<i>0.40</i>
\$40,000-\$60,000	0.17	0.20	<i>0.19</i>	0.16	0.20	<i>0.17</i>
\$60,000-\$80,000	0.15	0.15	<i>0.83</i>	0.14	0.16	<i>0.49</i>
\$80,000-\$100,000	0.12	0.11	<i>0.54</i>	0.10	0.08	<i>0.49</i>
\$100,000-\$120,000	0.07	0.07	<i>0.91</i>	0.07	0.05	<i>0.29</i>
\$120,000-\$140,000	0.04	0.02	<i>0.05</i>	0.06	0.04	<i>0.23</i>
\$140,000-\$160,000	0.03	0.04	<i>0.16</i>	0.04	0.03	<i>0.41</i>
More than \$160,000	0.05	0.05	<i>0.72</i>	0.08	0.07	<i>0.55</i>
Age:						
18-24	0.17	0.16	<i>0.71</i>	0.17	0.18	<i>0.76</i>
25-34	0.31	0.32	<i>0.70</i>	0.32	0.33	<i>0.78</i>
35-44	0.22	0.22	<i>0.78</i>	0.20	0.23	<i>0.36</i>
45-54	0.15	0.13	<i>0.43</i>	0.15	0.12	<i>0.25</i>
55-64	0.11	0.11	<i>0.87</i>	0.12	0.08	<i>0.14</i>
65 or older	0.04	0.06	<i>0.18</i>	0.04	0.06	<i>0.34</i>
Employment:						
Currently employed outside Prolific	0.66	0.65	<i>0.67</i>	0.67	0.72	<i>0.23</i>
Ever employed outside Prolific	0.96	0.95	<i>0.36</i>	0.95	0.97	<i>0.14</i>
Satisfied with current employer	0.78	0.79	<i>0.66</i>	0.77	0.75	<i>0.67</i>
Satisfied with most recent employer	0.66	0.62	<i>0.38</i>	0.64	0.65	<i>0.93</i>
Past experienced discrimination:						
Job search	0.26	0.24	<i>0.23</i>	0.28	0.24	<i>0.27</i>
Promotion	0.19	0.18	<i>0.40</i>	0.18	0.14	<i>0.15</i>
Termination	0.12	0.09	<i>0.09</i>	0.09	0.09	<i>0.85</i>
Daily work activities	0.34	0.32	<i>0.30</i>	0.36	0.28	<i>0.04</i>
Any discrimination	0.52	0.50	<i>0.62</i>	0.53	0.46	<i>0.05</i>
Beliefs about prevalence (own group)	0.47	0.44	<i>0.05</i>	0.46	0.45	<i>0.74</i>
N	694	695		371	320	

Note: This table shows means of worker demographic characteristics, all measured in the baseline survey, by treatment group in the promotion experiment and tests for balance across the manager arms and the algorithm sub-arms. All variables except for the final one are indicators; the final variable is a continuous, self-reported belief about the share of workers in one's own race \times gender group who report experiencing discrimination in the past in the baseline survey. *p*-values are calculated with standard errors robust to heteroskedasticity.

Table A.2: Balance table, hiring experiment

	Manager arms			Algorithm arms			<i>joint</i> <i>p-value</i>
	Blind	Blind+	Non-Blind	Blind	Blind+	Non-Blind	
	(1)	(2)	(3)	(4)	(5)	(6)	
Male	0.12	0.16	0.17	0.15	0.13	0.11	0.55
Race:							
Asian	0.06	0.06	0.09	0.07	0.06	0.05	0.61
Black	0.13	0.13	0.10	0.12	0.13	0.12	0.69
White, Hispanic	0.09	0.10	0.10	0.12	0.11	0.12	0.44
White, Non-Hispanic	0.73	0.71	0.71	0.69	0.70	0.71	0.66
Married	0.60	0.61	0.53	0.55	0.54	0.61	0.23
Kids	0.45	0.46	0.39	0.44	0.45	0.50	0.52
Education:							
Less than high school	0.01	0.00	0.01	0.00	0.01	0.00	0.99
High school graduate	0.10	0.09	0.08	0.08	0.12	0.09	0.20
Some college but no degree	0.29	0.27	0.26	0.25	0.24	0.23	0.53
2 year college degree	0.10	0.12	0.13	0.12	0.10	0.08	0.52
4 year college degree	0.36	0.37	0.34	0.38	0.37	0.44	0.14
Professional or Masters degree	0.13	0.13	0.16	0.15	0.14	0.14	0.76
Doctorate	0.02	0.01	0.02	0.02	0.03	0.01	0.14
Income:							
Less than \$20,000	0.11	0.10	0.10	0.11	0.12	0.10	0.85
\$20,000-\$40,000	0.18	0.17	0.21	0.16	0.18	0.15	0.59
\$40,000-\$60,000	0.19	0.25	0.18	0.19	0.16	0.19	0.28
\$60,000-\$80,000	0.17	0.14	0.15	0.16	0.17	0.18	0.92
\$80,000-\$100,000	0.15	0.09	0.07	0.11	0.13	0.14	0.19
\$100,000-\$120,000	0.07	0.07	0.11	0.08	0.07	0.07	0.26
\$120,000-\$140,000	0.06	0.07	0.08	0.07	0.05	0.06	0.96
\$140,000-\$160,000	0.03	0.02	0.04	0.04	0.05	0.04	0.55
More than \$160,000	0.05	0.08	0.07	0.07	0.07	0.07	0.44
Age:							
18-24	0.17	0.13	0.17	0.14	0.15	0.13	0.61
25-34	0.31	0.33	0.28	0.34	0.33	0.35	0.61
35-44	0.20	0.25	0.22	0.22	0.21	0.22	0.86
45-54	0.19	0.17	0.19	0.17	0.14	0.16	0.33
55-64	0.09	0.10	0.09	0.10	0.10	0.12	0.89
65 or older	0.03	0.02	0.05	0.04	0.06	0.02	0.16
Employment:							
Currently employed outside Prolific	0.66	0.72	0.70	0.70	0.67	0.72	0.45
Ever employed outside Prolific	0.97	0.98	0.97	0.97	0.97	0.97	0.99
N	411	449	389	420	443	415	

Note: This table shows means of worker demographic characteristics, all measured in the baseline survey, by treatment group in the hiring experiment. *p*-values test the null of equality across all groups, and are calculated with standard errors robust to heteroskedasticity.

Table A.3: Who is selected by each promotion/hiring procedure

Selected by:	Promotion experiment			Hiring experiment			
	Blind manager (1)	Non-blind manager (2)	Blind algorithm (3)	Blind manager (4)	Non-blind manager (5)	Blind algorithm (6)	Non-blind algorithm (7)
Male	0.338	0.325	0.274	0.111	0.067	0.156	0.367
Race:							
Asian	0.154	0.201	0.175	0.056	0.089	0.122	0.222
Black	0.107	0.145	0.137	0.178	0.167	0.078	0.178
White, Hispanic	0.098	0.141	0.111	0.033	0.056	0.100	0.356
White, Non-Hispanic	0.641	0.513	0.577	0.733	0.689	0.700	0.244
Quiz score quintile:							
1 star	0.004	0.030	0.000	0.000	0.000	0.000	0.000
2 stars	0.085	0.047	0.077	0.000	0.000	0.000	0.000
3 stars	0.141	0.167	0.154	0.222	0.200	0.200	0.200
4 stars	0.363	0.346	0.346	0.356	0.311	0.300	0.300
5 stars	0.406	0.410	0.423	0.422	0.489	0.500	0.500
Age:							
18-24	0.068	0.103	0.150	0.067	0.089	0.222	0.267
25-34	0.359	0.329	0.380	0.322	0.267	0.300	0.389
35-44	0.269	0.278	0.209	0.278	0.289	0.189	0.156
45-54	0.128	0.158	0.115	0.156	0.222	0.167	0.133
55-64	0.090	0.077	0.064	0.111	0.122	0.100	0.044
65 or older	0.085	0.056	0.081	0.067	0.011	0.022	0.011
N	234	234	234	90	90	90	90

Note: This table shows means of worker characteristics for the workers who are selected by each procedure in each experiment, where the worker characteristics shown are the ones that managers and the algorithm could have used (though only the non-blind procedures could use the information on race and gender).

Table A.4: Balance table, never versus ever promoted/hired

	Promotion experiment			Hiring experiment		
	Main	Ever	<i>p-value</i>	Main	Ever	<i>p-value</i>
	sample	promoted		sample	promoted	
	(1)	(2)	(1)=(2)	(3)	(4)	(3)=(4)
Male	0.29	0.32	<i>0.34</i>	0.13	0.16	<i>0.30</i>
Race:						
Asian	0.13	0.19	<i>0.03</i>	0.06	0.13	<i>0.01</i>
Black	0.14	0.11	<i>0.12</i>	0.12	0.07	<i>0.02</i>
White, Hispanic	0.14	0.14	<i>0.74</i>	0.10	0.17	<i>0.03</i>
White, Non-Hispanic	0.59	0.57	<i>0.60</i>	0.71	0.63	<i>0.03</i>
Married	0.54	0.53	<i>0.80</i>	0.58	0.61	<i>0.39</i>
Kids	0.43	0.35	<i>0.02</i>	0.46	0.42	<i>0.37</i>
Education:						
Less than high school	0.01	-0.00	<i>0.00</i>	0.01	0.00	<i>0.00</i>
High school graduate	0.13	0.02	<i>0.00</i>	0.10	0.01	<i>0.00</i>
Some college but no degree	0.27	0.05	<i>0.00</i>	0.25	0.02	<i>0.00</i>
2 year college degree	0.09	0.09	<i>0.77</i>	0.10	0.07	<i>0.26</i>
4 year college degree	0.37	0.49	<i>0.00</i>	0.39	0.45	<i>0.12</i>
Professional or Masters degree	0.11	0.29	<i>0.00</i>	0.14	0.39	<i>0.00</i>
Doctorate	0.01	0.05	<i>0.00</i>	0.02	0.05	<i>0.06</i>
Income:						
Less than \$20,000	0.14	0.09	<i>0.02</i>	0.11	0.07	<i>0.04</i>
\$20,000-\$40,000	0.22	0.17	<i>0.04</i>	0.18	0.10	<i>0.01</i>
\$40,000-\$60,000	0.18	0.15	<i>0.26</i>	0.19	0.24	<i>0.17</i>
\$60,000-\$80,000	0.15	0.12	<i>0.26</i>	0.17	0.15	<i>0.60</i>
\$80,000-\$100,000	0.10	0.16	<i>0.03</i>	0.13	0.14	<i>0.70</i>
\$100,000-\$120,000	0.07	0.10	<i>0.19</i>	0.07	0.11	<i>0.13</i>
\$120,000-\$140,000	0.04	0.05	<i>0.38</i>	0.06	0.03	<i>0.02</i>
\$140,000-\$160,000	0.04	0.08	<i>0.03</i>	0.04	0.07	<i>0.12</i>
More than \$160,000	0.06	0.08	<i>0.20</i>	0.07	0.10	<i>0.20</i>
Age:						
18-24	0.17	0.11	<i>0.01</i>	0.15	0.16	<i>0.82</i>
25-34	0.32	0.38	<i>0.07</i>	0.33	0.29	<i>0.31</i>
35-44	0.22	0.24	<i>0.38</i>	0.22	0.24	<i>0.57</i>
45-54	0.14	0.12	<i>0.37</i>	0.17	0.17	<i>0.91</i>
55-64	0.11	0.08	<i>0.17</i>	0.10	0.09	<i>0.65</i>
65 or older	0.05	0.07	<i>0.28</i>	0.04	0.06	<i>0.33</i>
Employment:						
Currently employed outside Prolific	0.67	0.76	<i>0.00</i>	0.69	0.81	<i>0.00</i>
Ever employed outside Prolific	0.96	0.98	<i>0.05</i>	0.97	0.99	<i>0.01</i>
Past experienced discrimination:						
Job search	0.25	0.29	<i>0.22</i>	–	–	–
Promotion	0.18	0.17	<i>0.78</i>	–	–	–
Termination	0.10	0.07	<i>0.14</i>	–	–	–
Daily work activities	0.33	0.38	<i>0.12</i>	–	–	–
Any discrimination	0.50	0.56	<i>0.14</i>	–	–	–
Beliefs about prevalence (own group)	0.46	0.44	<i>0.54</i>	–	–	–
N	2080	237		2527	153	

Note: This table shows means of worker demographic characteristics, all measured in the baseline survey, for workers who are included in the analysis sample (because they would not have been selected by any of the relevant promotion or hiring procedures) and those who are not, separately for the promotion and hiring experiments. The *p*-values test the null that ever-selected and never-selected workers are the same and are calculated with standard errors robust to heteroskedasticity.

Table A.5: Effects on secondary measures of perceptions of discrimination in the manager sample

	All Workers	Women	Minority Men	White Men
<i>Panel A: Free response, primary reason is discrimination</i>				
Non-blind manager, see avatars	0.182*** (0.016)	0.195*** (0.021)	0.168*** (0.033)	0.081* (0.042)
Control mean	0.010	0.014	0.000	0.000
<i>Panel B: Yes, would be promoted if different race or gender</i>				
Non-blind manager, see avatars	0.251*** (0.024)	0.288*** (0.029)	0.166*** (0.056)	0.084 (0.052)
Control mean	0.127	0.130	0.159	0.000
<i>Panel C: Complaint about job assignment mentions discrimination</i>				
Non-blind manager, see avatars	0.099*** (0.013)	0.116*** (0.016)	0.050** (0.022)	0.059 (0.049)
Control mean	0.003	0.004	0.000	0.000
<i>Panel D: Members of my race * gender group underrepresented among promoted</i>				
Non-blind manager, see avatars	0.162*** (0.027)	0.208*** (0.032)	-0.009 (0.064)	0.015 (0.047)
Control mean	0.386	0.378	0.538	0.020
<i>Panel E: White men are overrepresented among promoted</i>				
Non-blind manager, see avatars	0.157*** (0.026)	0.196*** (0.030)	0.024 (0.062)	0.327** (0.141)
Control mean	0.585	0.580	0.614	0.551
N	1389	1004	291	94

Note: This table shows the effects on secondary measures of perceived discrimination of being in the non-blind manager arm, relative to the demographic-blind manager arm in the promotion experiment. The secondary measures are the following: In Panel A, an indicator for the worker's *primary* reason listed for what needed to be different about their profile being demographics (as opposed to the main measure, which indicates mentioning demographics at all); in Panel B, an indicator for saying they think they would have been promoted if their race or gender was different; and in Panel C, an indicator for mentioning bias or discrimination in their open-ended description after saying they had a complaint about the promotion procedure. These are alternative measures of discrimination against the person themselves. Panels D and E measure perceptions of discrimination in general: an indicator for thinking that people in their race \times gender group are under-represented among the promoted workers, and an indicator for thinking that white men are over-represented among the promoted workers. Specifications are otherwise the same as in Figure 1.1. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.6: Effects on retention, gender and racial heterogeneity

	Num paragraphs (1)	More than 6 (2)	Did all 18 (3)
<i>Panel A: Race and gender heterogeneity</i>			
Non-blind manager arm*White women	-0.645** (0.316)	-0.039* (0.021)	-0.046 (0.032)
Non-blind manager arm*Minority women	-1.001* (0.557)	-0.029 (0.038)	-0.120** (0.057)
Non-blind manager arm*Minority men	0.808 (0.560)	0.065 (0.042)	0.065 (0.054)
<i>p-values:</i>			
White=minority, women	0.566	0.809	0.247
Men=women, minority	0.020	0.094	0.016
N	1293	1293	1293
<i>Panel B: Gender heterogeneity</i>			
Non-blind manager arm*Women	-0.741*** (0.285)	-0.036* (0.019)	-0.066** (0.029)
Non-blind manager arm*Minority men	0.808 (0.560)	0.065 (0.042)	0.065 (0.054)
<i>p-values:</i>			
Men=women	0.011	0.026	0.027
N	1293	1293	1293
Control Mean, white women	16.402	0.944	0.818
Control Mean, minority women	15.829	0.921	0.764
Control Mean, minority men	14.538	0.821	0.662

Note: This table estimates the effect of being in the non-blind manager arm on retention relative to the demographic-blind manager arm in the promotion experiment, separately for white women, racial minority women, and racial minority men. The outcomes and sample are the same as in Panel A of Table 1.2, but white men are additionally dropped from the sample. The specification is also the same, but the indicator for treatment is replaced with three interactions between treatment and indicators for being a white woman, an Asian, Black, or Hispanic woman, and an Asian, Black, or Hispanic man. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.7: Effects on effort and performance in the manager sample of the promotion experiment, gender and racial heterogeneity

	Minutes (1)	Clicks (2)	Correct (3)	Incorrect (4)	Bonus (5)
<i>Panel A: Race and gender heterogeneity</i>					
Non-blind manager arm*White women	-0.060 (0.073)	1.107 (2.390)	-0.055 (0.296)	0.070 (0.257)	-0.017 (0.028)
Non-blind manager arm*Minority women	-0.012 (0.116)	4.397 (4.401)	0.803 (0.522)	0.798** (0.406)	0.043 (0.049)
Non-blind manager arm*Minority men	0.099 (0.119)	-4.937* (2.622)	1.320*** (0.469)	-0.109 (0.487)	0.080* (0.045)
<i>p-values:</i>					
White=minority, women	0.721	0.513	0.139	0.109	0.268
Men=women, minority	0.496	0.061	0.454	0.154	0.566
N	1295	1295	1295	1295	1295
<i>Panel B: Gender heterogeneity</i>					
Non-blind manager arm*Women	-0.048 (0.063)	1.989 (2.098)	0.175 (0.267)	0.265 (0.229)	-0.001 (0.025)
Non-blind manager arm*Minority men	0.099 (0.119)	-4.933* (2.619)	1.321*** (0.469)	-0.108 (0.487)	0.080* (0.045)
<i>p-values:</i>					
Men=women	0.259	0.033	0.032	0.485	0.106
N	1295	1295	1295	1295	1295
Control Mean, white women	4.968	36.903	15.581	3.475	0.976
Control Mean, minority women	4.932	36.179	14.686	2.707	0.905
Control Mean, minority men	4.904	33.503	12.552	3.938	0.752

Note: This table estimates the effect of being in the non-blind manager arm on effort and performance relative to the demographic-blind manager arm in the promotion experiment, separately for white women, racial minority women, and racial minority men. The outcomes and sample are the same as in Panel A of Table 1.3, but white men are additionally dropped from the sample. The specification is the same as Appendix Table A.6. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.8: Effects on self-reported interest in future work

	Work for employer (1)	Tasks w/ assignment (2)	Easy paragraphs (3)	Hard paragraphs (4)	Prefer hard job (5)
Non-blind manager, see avatars	-0.079*** (0.028)	-0.053* (0.029)	-0.047* (0.029)	-0.063*** (0.029)	-0.026 (0.027)
N	1380	1380	1380	1380	1389
Control mean	0.657	0.561	0.584	0.532	0.689

Note: This table estimates the effect of being in the non-blind manager arm on self-reported interest in future work, relative to the demographic-blind manager arm (in panel A) and the effect of seeing that previously-promoted workers were mostly white men in the algorithm arm (in panel B) in the promotion experiment. In columns 1-4, the outcomes are indicators for strongly agreeing with statements that they would like to do more tasks (on Prolific) for this employer, more tasks with this type of job assignment, more tasks of the same difficulty level, and tasks that are the harder job that they didn't do in the experiment. The outcome in column 5 is an indicator for probably or definitely preferring the harder job to the easier job. Specifications are the same as in Table 1.2. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.9: Balance table by whether see two or three white men previously promoted

	Non-blind manager		Blind manager		Algorithm + avatars		Algorithm, no avatars		
	<3/3	p	<3/3	p	<3/3	p	<3/3	p	
Male	0.27	0.29	0.31	0.23	0.30	0.32	0.36	0.33	0.580
Race:									
Asian	0.13	0.11	0.15	0.09	0.18	0.12	0.14	0.14	0.941
Black	0.15	0.12	0.14	0.13	0.12	0.16	0.17	0.16	0.727
White, Hispanic	0.14	0.14	0.14	0.16	0.14	0.11	0.15	0.14	0.863
White, Non-Hispanic	0.59	0.64	0.57	0.62	0.56	0.62	0.55	0.56	0.741
Married	0.51	0.52	0.55	0.53	0.53	0.57	0.60	0.50	0.066
Kids	0.49	0.40	0.43	0.42	0.40	0.48	0.44	0.38	0.277
Education:									
Less than high school	0.01	0.01	0.01	0.01	0.02	-0.00	0.01	0.01	0.974
High school graduate	0.16	0.13	0.14	0.14	0.12	0.15	0.11	0.11	0.924
Some college but no degree	0.26	0.27	0.28	0.29	0.25	0.32	0.23	0.27	0.386
2 year college degree	0.09	0.10	0.10	0.10	0.07	0.07	0.10	0.09	0.557
4 year college degree	0.34	0.34	0.36	0.38	0.42	0.38	0.41	0.36	0.279
Professional or Masters degree	0.12	0.13	0.10	0.08	0.11	0.08	0.12	0.15	0.341
Doctorate	0.02	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.704
Income:									
Less than \$20,000	0.13	0.16	0.13	0.13	0.14	0.14	0.11	0.21	0.010
\$20,000-\$40,000	0.21	0.20	0.21	0.29	0.25	0.18	0.23	0.17	0.121
\$40,000-\$60,000	0.20	0.19	0.17	0.16	0.20	0.18	0.17	0.14	0.450
\$60,000-\$80,000	0.14	0.16	0.15	0.14	0.15	0.17	0.15	0.11	0.250
\$80,000-\$100,000	0.09	0.12	0.14	0.08	0.07	0.11	0.09	0.11	0.495
\$100,000-\$120,000	0.09	0.05	0.07	0.08	0.05	0.07	0.07	0.08	0.682
\$120,000-\$140,000	0.03	0.02	0.04	0.04	0.04	0.03	0.05	0.07	0.508
\$140,000-\$160,000	0.04	0.05	0.03	0.04	0.03	0.04	0.05	0.03	0.192
More than \$160,000	0.06	0.05	0.05	0.04	0.06	0.08	0.07	0.09	0.489
Age:									
18-24	0.14	0.18	0.17	0.17	0.20	0.14	0.19	0.15	0.395
25-34	0.30	0.35	0.32	0.29	0.29	0.40	0.30	0.34	0.501
35-44	0.24	0.18	0.22	0.22	0.26	0.16	0.19	0.22	0.516
45-54	0.14	0.13	0.15	0.14	0.13	0.10	0.14	0.16	0.632
55-64	0.11	0.11	0.09	0.14	0.06	0.14	0.14	0.09	0.202
65 or older	0.06	0.05	0.04	0.04	0.06	0.05	0.04	0.04	0.955
Employment:									
Currently employed outside Prolific	0.64	0.65	0.66	0.65	0.69	0.77	0.71	0.62	0.081
Ever employed outside Prolific	0.93	0.97	0.96	0.96	0.97	0.99	0.95	0.95	0.968
Satisfied with current employer	0.78	0.80	0.78	0.77	0.72	0.80	0.78	0.76	0.714
Satisfied with most recent employer	0.59	0.65	0.64	0.68	0.63	0.70	0.67	0.62	0.623
N	376	319	422	272	216	104	220	151	

Note: This table shows means of worker demographic characteristics, all measured in the baseline survey, by treatment group in the promotion experiment, separately for workers who were randomly paired with managers who previously promoted three white men or not (in the manager arms) or randomly paired with a group of twenty-four historical-sample workers jointly evaluated by the algorithm among whom three white men were previously promoted (in the algorithm arm). p -values test the null of equality between workers who do or do not see three white men in each treatment group, and are calculated with standard errors robust to heteroskedasticity.

Table A.10: Effects on other reasons for non-promotion in manager sample of the promotion experiment

	All Workers	Women	Minority Men	White Men
<i>Panel A: Free response, first answer is more school</i>				
Non-blind manager, see avatars	-0.109*** (0.024)	-0.116*** (0.028)	-0.146** (0.057)	0.099 (0.104)
Control mean	0.414	0.410	0.428	0.408
<i>Panel B: Free response, first answer is higher quiz score</i>				
Non-blind manager, see avatars	-0.056** (0.027)	-0.079** (0.031)	0.005 (0.066)	-0.041 (0.110)
Control mean	0.484	0.488	0.476	0.469
<i>Panel C: Free response, first answer is it was random</i>				
Non-blind manager, see avatars	0.000 (0.011)	0.007 (0.013)	0.001 (0.021)	-0.081 (0.064)
Control mean	0.045	0.046	0.028	0.082
N	1389	1004	291	94

Note: This table shows the effects of being in the non-blind manager arm on the other reasons that workers gave for what needed to be different about their profile to be assigned to the harder job (the question underlying the main measure of perceived discrimination), relative to the demographic-blind manager arm in the promotion experiment. In Panel A the outcome is an indicator for thinking they needed more education to be promoted, in Panel B it is an indicator for thinking they needed higher quiz scores, and in Panel C it is an indicator for thinking it was random and there was nothing they could do. Specifications are otherwise the same as Figure 1.1. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.11: Effects on psychological outcomes in the promotion experiment

	Total good affect (SD)	Self-efficacy index (SD)	Very satisfied with work
	(1)	(2)	(3)
Non-blind manager, see avatars	0.006 (0.035)	-0.013 (0.040)	-0.038 (0.029)
N	1389	1382	1389
Control mean	0.011	-0.013	0.549

Note: This table estimates the effect of being in the non-blind manager arm on psychological well-being, relative to the demographic-blind manager arm (in panel A) and the effect of seeing that previously-promoted workers were mostly white men in the algorithm arm (in panel B) in the promotion experiment. The first outcome is an index of overall affect (emotional state) and the second is the average of two self-efficacy indices, one for skills related to the proofreading job and one over the proofreading jobs themselves. Both are measured in standard deviation units. The third is an indicator for reporting being very satisfied with the job in the experimental survey. The specifications and sample are the same as in Table 1.2. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.12: Effects on psychological outcomes, gender and racial heterogeneity

	Total good affect (SD)	Self-efficacy index (SD)	Very satisfied with work
	(1)	(2)	(3)
<i>Panel A: Race and gender heterogeneity</i>			
Non-blind manager arm*White women	-0.004 (0.048)	0.006 (0.052)	-0.015 (0.039)
Non-blind manager arm*Minority women	-0.064 (0.079)	-0.055 (0.094)	-0.056 (0.062)
Non-blind manager arm*Minority men	0.106 (0.071)	-0.026 (0.085)	-0.062 (0.059)
<i>p-values:</i>			
White=minority, women	0.506	0.563	0.562
Men=women, minority	0.110	0.818	0.937
N	1295	1288	1295
<i>Panel B: Gender heterogeneity</i>			
Non-blind manager arm*Women	-0.020 (0.042)	-0.010 (0.047)	-0.026 (0.034)
Non-blind manager arm*Minority men	0.106 (0.071)	-0.026 (0.085)	-0.062 (0.059)
<i>p-values:</i>			
Men=women	0.124	0.870	0.580
N	1295	1288	1295
Control Mean, white women	0.038	-0.043	0.572
Control Mean, minority women	-0.063	-0.079	0.471
Control Mean, minority men	0.026	0.111	0.566

Note: This table estimates the effect of being in the non-blind manager arm on future labor supply relative to the demographic-blind manager arm in the promotion experiment, separately for white women, racial minority women, and racial minority men. The outcomes and sample are the same as in Panel A of Appendix Table A.11, but white men are additionally dropped from the sample. The specification is the same as Appendix Table A.6. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.13: Effects on future labor supply, gender and racial heterogeneity

	RW (Same) (1)	WTP cutoff (2)	Coop. RW (3)	WTP new mgr (4)	Shared zero (5)	Amt shared (6)
<i>Panel A: Race and gender heterogeneity</i>						
Non-blind manager arm*White women	0.033** (0.014)	0.012** (0.006)	0.025 (0.019)	0.051* (0.027)	0.038 (0.039)	-0.234 (0.317)
Non-blind manager arm*Minority women	-0.007 (0.021)	0.011 (0.011)	-0.022 (0.029)	0.115*** (0.044)	0.009 (0.062)	-0.585 (0.442)
Non-blind manager arm*Minority men	0.025 (0.023)	0.013 (0.012)	-0.014 (0.032)	-0.015 (0.046)	0.103* (0.058)	-0.001 (0.496)
<i>p-values:</i>						
White=minority, women	0.108	0.912	0.182	0.201	0.680	0.500
Men=women, minority	0.298	0.865	0.853	0.043	0.261	0.373
N	1247	1235	1256	957	1289	1289
<i>Panel B: Gender heterogeneity</i>						
Non-blind manager arm*Women	0.022* (0.012)	0.012** (0.005)	0.012 (0.017)	0.067*** (0.023)	0.030 (0.034)	-0.329 (0.269)
Non-blind manager arm*Minority men	0.025 (0.023)	0.013 (0.012)	-0.014 (0.032)	-0.015 (0.046)	0.103* (0.058)	-0.002 (0.496)
<i>p-values:</i>						
Men=women	0.902	0.890	0.461	0.109	0.268	0.549
N	1247	1235	1256	957	1289	1289
Control Mean, white women	0.253	-0.009	0.325	0.205	0.529	2.815
Control Mean, minority women	0.280	-0.019	0.349	0.218	0.550	2.693
Control Mean, minority men	0.280	-0.026	0.366	0.264	0.517	2.924

Note: This table estimates the effect of being in the non-blind manager arm on future labor supply relative to the demographic-blind manager arm in the promotion experiment, separately for white women, racial minority women, and racial minority men. The outcomes and sample are the same as in Panel A of Table 1.4, but white men are additionally dropped from the sample. The specification is the same as Appendix Table A.6. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.14: Correlations of manager characteristics with previously selected workers' characteristics

	3 white men (1)	3 men (2)	3 white (3)	3 max stars (4)	2 max stars (5)	1 max stars (6)	3 grad sch (7)	3 college (8)	Any grad sch (9)	Any coll (10)	Any no coll (11)
Age: 25-34	-0.026 (0.184)	-0.081 (0.195)	0.123 (0.175)	0.086 (0.159)	-0.080 (0.150)	-0.005 (0.064)	-0.049 (0.130)	-0.169 (0.124)	0.198 (0.152)	-0.062 (0.161)	0.131 (0.163)
Age: 35-44	0.415** (0.194)	0.201 (0.205)	0.394** (0.185)	0.083 (0.168)	-0.119 (0.158)	0.035 (0.067)	0.183 (0.137)	0.007 (0.131)	0.005 (0.160)	-0.122 (0.169)	-0.088 (0.172)
Age: 45-54	-0.018 (0.230)	-0.022 (0.243)	0.142 (0.218)	0.083 (0.198)	-0.052 (0.187)	-0.031 (0.080)	-0.043 (0.162)	-0.031 (0.155)	0.096 (0.189)	0.036 (0.200)	-0.001 (0.204)
Age: 55-64	0.237 (0.291)	-0.073 (0.307)	0.044 (0.276)	-0.074 (0.251)	0.065 (0.237)	0.009 (0.101)	0.507** (0.205)	-0.188 (0.196)	0.237 (0.239)	-0.419 (0.253)	-0.202 (0.258)
Age: 65 or older	-0.212 (0.523)	0.512 (0.496)	-0.419 (0.496)	0.340 (0.451)	-0.336 (0.426)	-0.004 (0.181)	-0.085 (0.369)	-0.148 (0.352)	-0.796* (0.430)	0.201 (0.455)	0.790* (0.463)
Educ: 4 year college degree	-0.029 (0.252)	-0.041 (0.266)	0.148 (0.239)	-0.299 (0.217)	0.299 (0.205)	0.000 (0.087)	-0.004 (0.178)	-0.159 (0.169)	0.094 (0.207)	-0.095 (0.219)	0.170 (0.223)
Educ: Doctorate	-0.479 (0.561)	0.545 (0.592)	-0.315 (0.533)	0.115 (0.484)	-0.102 (0.457)	-0.013 (0.194)	-0.597 (0.396)	-0.119 (0.378)	0.061 (0.462)	0.525 (0.489)	0.161 (0.497)
Educ: High school graduate	0.179 (0.284)	-0.086 (0.300)	0.622** (0.270)	-0.110 (0.245)	0.120 (0.231)	-0.010 (0.098)	0.304 (0.200)	-0.105 (0.191)	-0.013 (0.234)	-0.277 (0.247)	0.167 (0.252)
Educ: Professional or Masters degree	0.006 (0.286)	0.051 (0.302)	0.321 (0.271)	-0.128 (0.246)	0.165 (0.233)	-0.037 (0.099)	-0.146 (0.202)	-0.070 (0.192)	-0.054 (0.235)	0.176 (0.249)	0.026 (0.253)
Educ: Some college but no degree	-0.069 (0.270)	-0.106 (0.285)	0.347 (0.257)	-0.139 (0.233)	0.155 (0.220)	-0.015 (0.094)	-0.184 (0.191)	-0.154 (0.182)	0.165 (0.222)	0.020 (0.235)	0.360 (0.240)
On Prolific: 3-6 months ago	-0.153 (0.209)	-0.302 (0.221)	0.157 (0.199)	0.125 (0.180)	-0.212 (0.170)	0.088 (0.073)	0.020 (0.148)	0.338** (0.141)	-0.501*** (0.172)	0.224 (0.182)	-0.061 (0.186)
On Prolific: 6-12 months ago	0.515* (0.286)	0.210 (0.302)	0.419 (0.271)	-0.314 (0.246)	0.057 (0.233)	0.257** (0.099)	0.243 (0.202)	0.298 (0.192)	-0.185 (0.235)	-0.062 (0.249)	-0.057 (0.253)
On Prolific: more than 2 years ago	-0.177 (0.175)	-0.274 (0.185)	0.076 (0.166)	-0.026 (0.151)	0.026 (0.143)	-0.000 (0.061)	0.053 (0.124)	0.109 (0.118)	-0.056 (0.144)	0.002 (0.153)	-0.090 (0.155)
On Prolific: within the last 3 months	-0.124 (0.185)	-0.119 (0.195)	0.056 (0.176)	-0.086 (0.160)	0.085 (0.151)	0.001 (0.064)	0.226* (0.131)	0.057 (0.125)	-0.013 (0.152)	-0.059 (0.161)	-0.220 (0.164)
N	75	75	75	75	75	75	75	75	75	75	75
F-statistic p-value	0.141	0.674	0.151	0.867	0.892	0.592	0.021	0.738	0.205	0.475	0.433
R-squared	0.259	0.156	0.255	0.120	0.114	0.169	0.333	0.145	0.240	0.188	0.195

Note: This table shows estimated coefficients from regressions of the characteristics of workers each manager promoted in the historical sample on the manager characteristics that were observable to workers (other than race and gender, which were constant). The dependent variables are indicators for whether the three workers are all white men, all men, and all white in the first three columns, and the rest are the previously promoted worker educational and performance measures that are included as controls in the main specification. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.15: Effects on secondary measures of perceptions of discrimination in the algorithm sample of the promotion experiment

	All Workers	Women	Minority Men	White Men
<i>Panel A: Free response, primary reason is discrimination</i>				
See avatars	0.086*** (0.019)	0.107*** (0.025)	0.044 (0.043)	–
Control mean	0.003	0.004	0.000	0.000
<i>Panel B: Yes, would be promoted if different race or gender</i>				
See avatars	0.166*** (0.030)	0.204*** (0.036)	0.111 (0.079)	0.003 (0.066)
Control mean	0.054	0.045	0.096	0.022
<i>Panel C: Complaint about job assignment mentions discrimination</i>				
See avatars	0.039*** (0.012)	0.042*** (0.015)	0.030 (0.032)	–
Control mean	0.000	0.000	0.000	0.000
<i>Panel D: Members of my race * gender group underrepresented among promoted</i>				
See avatars	0.107*** (0.039)	0.115** (0.049)	0.114 (0.107)	0.041 (0.110)
Control mean	0.294	0.285	0.446	0.065
<i>Panel E: White men are overrepresented among promoted</i>				
See avatars	0.127*** (0.043)	0.134** (0.053)	0.258** (0.114)	-0.273 (0.202)
Control mean	0.472	0.479	0.494	0.391
N	691	464	151	76

Note: This table shows the effects of seeing that previously-promoted workers were mostly white men in the algorithm sample of the promotion experiment on secondary measures of perceived discrimination. The secondary measures are the same as in Appendix Table A.5. Specifications are otherwise the same as in Figure 1.6. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.16: Effects on other reasons for non-promotion in algorithm sample of the promotion experiment

	All Workers	Women	Minority Men	White Men
<i>Panel A: Free response, first answer is more school</i>				
See avatars	-0.069** (0.034)	-0.037 (0.040)	-0.058 (0.082)	-0.073 (0.156)
Control mean	0.340	0.310	0.349	0.478
<i>Panel B: Free response, first answer is higher quiz score</i>				
See avatars	0.000 (0.040)	-0.053 (0.048)	0.163 (0.102)	-0.119 (0.165)
Control mean	0.526	0.562	0.482	0.413
<i>Panel C: Free response, first answer is it was random</i>				
See avatars	-0.014 (0.019)	0.005 (0.024)	-0.062* (0.034)	0.012 (0.102)
Control mean	0.059	0.058	0.048	0.087
N	691	464	151	76

Note: This table shows the effects of learning that mostly white men were previously promoted on the other reasons that workers gave for what needed to be different about their profile to be assigned to the harder job (the question underlying the main measure of perceived discrimination), relative to other workers evaluated by the algorithm. The outcomes are the same as in Appendix Table A.10. Specifications are otherwise the same as in Figure 1.6. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

A.2 Manager recruitment and manager task

Managers were recruited from Prolific. They were required to be white men and employed outside of Prolific. Having all managers be white men serves three purposes: (i) to best proxy for cases in which a woman or racial minority man might be the most likely to feel discriminated against, (ii) to minimize noise in the experiment caused by having large variation in manager characteristics, since only 54 managers participated in the promotion experiment and only 20 in the hiring experiment, and (iii) to preserve the sample of possible women and racial minority men to participate in the experiment itself. As illustrated in Figure A.3, initially 75 managers were recruited to evaluate the 1800 workers in the historical sample, at which point they were randomly assigned to be a demographic-blind or non-blind manager. They were independently randomly assigned to evaluate workers who scored in the bottom two quintiles of average baseline quiz scores, the second and third quintiles, third and fourth quintiles, or fourth and fifth quintiles. Demographic-blind managers saw workers' education level (no college, college degree, or more than college) and 1-5 stars indicating the workers' approximate quintile of average baseline quiz scores, which are predictive of the workers' proofreading skill. Non-blind managers also saw workers' self-identified race/ethnicity and gender as well as an avatar that workers built to look like themselves in the baseline survey.

Managers were shown examples of the paragraphs that workers would be asked to proofread in the harder job and then chose three out of 24 historical workers to promote, viewing worker profiles that had either just quiz scores and education, or also avatars, race, and gender, depending on the manager's random assignment. The promoted workers' performance in the harder job determined the managers' bonus payment, which represented the majority of their total payment, and thus managers were incentivized to choose the workers they thought would do the best at the harder job. Specifically, managers were paid \$1 to complete the survey, which took on average 7 minutes, and on average earned \$4 in bonuses based on the performance of the workers in the historical sample that they had promoted.

The same managers were brought back to evaluate the workers in the experimental samples. They were randomly assigned to a group of 120 workers in the promotion (hiring) experiment who fell into the managers' (previously randomly assigned) quiz-score quintile group but had otherwise been randomly grouped together, with one demographic-blind and one non-blind manager assigned to evaluate the workers in each group. Managers saw workers in the same quiz-score quintile groups in both the historical and main samples and the same information about workers in the historical and main samples.

After a reminder about the harder proofreading job and that they had done this worker-selection task before, each manager evaluated workers in groups of forty. In the promotion experiment, they evaluated three sets of forty workers independently, promoting three in each group, and in the hiring experiment, they evaluated nine sets of forty workers independently, hiring one in each group. Managers knew that their decisions for **one** of the three or nine groups would be randomly chosen to be implemented, and the workers they chose in that group would determine their bonus payment. This ensured that every worker was evaluated by both a demographic-blind and non-blind manager, though only the decisions of one—or (one of) the algorithm(s)—were implemented. This generated counterfactual data on how workers would have been assigned to jobs under all three (four) procedures, which were used in the analysis (but not in the implementation of the experiment). In the promotion experiment, managers were paid \$2 to complete the survey, which took on average 10 minutes, and on average earned \$4 in bonuses based on their chosen workers' performance. In the hiring experiment, managers were paid \$2 to complete the survey, which took on average 15 minutes, and on average earned \$3.25 in bonuses based on their chosen workers' performance.

Of course, there was attrition between the original manager recruitment and their return to evaluate the main sample workers. Initially 75 managers were recruited in order to account for up to 20 percent attrition. Then, after the promotion experiment baseline survey was complete, one demographic-blind and one non-

blind manager were each matched with a random group of 120 workers in their same (randomly assigned to managers) quiz-score quintile group. After several days and several reminders, when a manager failed to return to evaluate the new workers, they were replaced with a randomly chosen manager who matched their random assignment (demographic-blind or not \times quiz-score quintile group) from the group of managers that had not been assigned to evaluate workers initially. This continued until all workers had been evaluated. 54 managers participated in the promotion experiment.

The same managers were invited back for the hiring experiment, now eight months following their initial recruitment. Again, one demographic-blind manager and one non-blind manager was each matched with a random with a group of 360 workers in their same (randomly assigned to managers) quiz-score quintile group. The number of groups of workers per manager was larger so that fewer managers would be needed, assuming there would be major attrition given the amount of time since the initial survey. This time, managers were matched with groups randomly but those who had returned to evaluate workers in the promotion experiment (a signal that they would return again) and had previously promoted mostly white men (in order to try to minimize variation in that dimension of treatment) had a higher chance of being initially paired with workers and offered the worker-evaluation task. The same replacement procedure was followed when managers did not return after several days and reminders. 20 managers participated.

In one group of 120 workers in the promotion experiment, I ran out of managers who could be paired with the group because none of the eligible managers returned. I invited two managers who had already evaluated a group of 120 workers in the same quiz-score quintile group to return and evaluate another set. However, only one returned. In the hiring experiment, in one group of 360, one of the initially-assigned managers returned but none of the other eligible managers returned. In both experiments, I implemented the decisions of the one manager who returned and paid them accordingly, but the workers are dropped from the experimental sample. Results are fully robust to their inclusion in the promotion experiment (in the hiring experiment, only the worker who was hired, not the ones who were not, were offered the corresponding survey due to budget constraints and the fact that they could not be included in the main sample). These results are available upon request.

A.3 Training the algorithm

The demographic-blind algorithm is a regression-based random forest model that uses workers' average quiz-score quintile fixed effects and education fixed effects (no college, college degree, more than college) to predict which workers will do the best if they are promoted. The algorithm used these variables to estimate a model predicting good performance in the harder job. As described in Section 1.2.2.1, the harder job involved proofreading 12 paragraphs from articles published in leading scientific journals.

The algorithm predicted the average number of mistakes correctly highlighted minus the number of non-mistakes incorrectly highlighted, each standardized to have mean zero and standard deviation one. Since workers could choose to stop proofreading at any time in the harder job, using the total performance across all 12 paragraphs also includes a component of retention. For the purposes of the experiment, the accuracy and predictiveness of the algorithm were irrelevant – what mattered was its inputs and that it predicted performance, as this is what was communicated to workers.

The non-blind algorithm, used only in the hiring experiment, simply adds interactions of each of the above predictors with race and gender fixed effects. Both algorithms are unbiased using calibration: conditional on predicted performance, race and gender are not independently predictive of performance in the training sample or the small number of promoted/hired workers for whom I observe both predicted performance and performance. If anything, the non-blind algorithm promotes *more* racial minorities in the experimental sample.

The algorithms were trained on a sample of 500 workers recruited from MTurk using Cloud Research's pre-approved participant pool. The sample for the entire study was originally going to be recruited from MTurk, since these types of tasks are more common there. However, in recruiting this training sample, it became evident that it would be impossible, or at least prohibitively expensive, to recruit enough racial minority participants from CloudResearch's pre-approved pool of MTurk workers. Like the experimental sample, women and racial minorities are also over-represented in the training sample: 33 percent are white women, 18 percent are non-white women, 18 percent are non-white men, and 30 percent are white men. There are no significant race or gender differences in performance in the training sample. These workers completed (in one survey) both the baseline survey *and* paragraphs the proofreading job.

The random forest model was used to predict worker performance in the harder job after completing only the baseline survey in the historical and two experimental samples. Like the managers, the algorithm selected three workers in each group of 24 (in the historical sample) or 40 (in the promotion experiment sample) to do the harder job. In the hiring experiment, again like the managers, the algorithm selected one worker in each group of 40. In each group, the algorithm ranked workers by their predicted performance, and assigned the top three or one worker(s) to the harder job. Ties were resolved randomly (e.g. if four workers tied for the second-highest predicted performance, two were randomly assigned to the higher-status job along with the worker with the highest predicted performance, etc.). Similarly to how the managers actually evaluated all worker groups in order to generate counterfactual data to use in the analysis, the algorithm's "decisions" were determined for all workers, not just those in groups randomly assigned to have the algorithms' decisions implemented.

A.4 Variable definitions

Unless otherwise specified, definitions apply in both the promotion experiment and hiring experiment. The measures of perceived discrimination are described in full detail in the text and so are not included here.

A.4.1 Outcome variables

Retention. (Promotion experiment only). After the sixth paragraph, workers had the option to skip to the end of the survey after proofreading each paragraph. The measures of retention are the number they proofread, whether they do more than six (i.e. don't quit right away), and whether they do all eighteen. The question about skipping to the end of the survey was the following: "Would you like to continue to the next paragraph? You've done the required six paragraphs to receive your participation payment. Remember, you can earn \$0.25 per paragraph that you do a good job proofreading. There are X more paragraphs to proofread. There is no penalty for stopping now, but you can only be paid for paragraphs that you complete."

- "I'd like to proofread another paragraph"
- "I do not want to proofread any more paragraphs. (By selecting this option, you will continue to the end of the survey)"

Effort and performance. (Promotion experiment only). The effort measures come from data collected passively while workers do the proofreading job. This includes the number of times they click on the page per paragraph, the amount of time they spend on each paragraph, how many words they correctly highlight that are mistakes, how many words they incorrectly highlight that are not mistakes, and the bonus they earn.

Future labor supply. All reservation wages are elicited using a multiple price list. Workers are told that in each case, one of the wages/wage schedules in the rows below would be randomly selected and their answer in that row would determine whether they would be evaluated in the future and offered the corresponding wage/job depending on their evaluation, and that workers who said they were interested at a given wage may be randomly selected to be offered the job if interest exceeds the number of workers we needed to hire. *Note that there are many long blocks of text below; workers were always shown text in shorter blocks and with key words bolded for emphasis which has been removed here.*

Promotion experiment

- *Same procedure:* "We will offer some workers a chance to be evaluated again and assigned to more proofreading tasks like this in the future. In one future round, the mechanism used to assign people to the easier, lower-paying task versus the harder, higher-paying task will be the same as today:"

In the demographic-blind manager arm: "a manager will review your average quiz score and education and assign you to proofread science articles for kids or articles published in leading scientific journals"

In the non-blind manager arm: “a manager will review your applicant profile (avatar, average quiz score, etc.) and assign you to proofread science articles for kids or articles published in leading scientific journals”

In the algorithm arm: “an algorithm will use your average quiz score and education to determine which task you would be assigned to”

[All arms] “You can earn twice as much per hard paragraph than easy paragraph. At what wages would you be interested in being evaluated again under this task assignment mechanism?”

- *Cutoff rule:* “In another future round, we will use screening quiz scores to determine who is assigned to the easier, lower-paying task vs the harder, higher-paying task: Among workers interested in the job, the top scorers on the screening quizzes will be offered the harder job. You can earn twice as much per hard paragraph than easy paragraph. At what wages would you be interested in being evaluated again under this task assignment mechanism?”

Multiple price list in both cases:

	I would not want this job (at these wages)	I would want this job (at these wages)
Earn \$0.05 for each high-quality easy paragraph; \$0.10 for each high-quality hard paragraph	<input type="checkbox"/>	<input type="checkbox"/>
Earn \$0.10 for each high-quality easy paragraph; \$0.20 for each high-quality hard paragraph	<input type="checkbox"/>	<input type="checkbox"/>
...
Earn \$0.50 for each high-quality easy paragraph; \$1.00 for each high-quality hard paragraph	<input type="checkbox"/>	<input type="checkbox"/>

The calculation of the outcome variables from the multiple price lists are described in detail in the paper; the overall reservation wage is a function of the point at which they switch from not wanting the job to wanting the job and the decomposed reservation wage imagines that each worker saw an individualized multiple price list with their expected wages, given their beliefs about the probability that they would be assigned to the harder, higher-paying task.

Hiring experiment:

- *Same procedure:* “We will offer some workers a chance to be evaluated again and potentially hired for the proofreading task in the future. In one future round, workers will be evaluated and hired the same way as today:”

In the demographic-blind manager arms, regardless of whether they see previous hires’ avatars, race and gender: “a manager will review your and others’ average quiz scores and education again and

decide who to hire for the proofreading task. The outcome may be different from today (you might be hired this time) because they would be reviewing a different set of workers.”

In the non-blind manager arm: “a manager will review your and others’ applicant profiles (avatar, average quiz score, etc.) and decide who to hire for the proofreading task. The outcome may be different from today (you might be hired this time) because they would be reviewing a different set of workers.”

At what wages would you be interested in being evaluated again under this task assignment mechanism?”

- *Cutoff rule:* “In another future round, we will use screening quiz scores to determine who is hired to do the proofreading task: Among workers interested in the job, the top scorers on the screening quizzes will be offered the harder job. At what wages would you be interested in being evaluated again under this task assignment mechanism?”

Multiple price list in both cases:

	I would not want this job (at these wages)	I would want this job (at these wages)
Earn \$0.10 for each high-quality paragraph	<input type="checkbox"/>	<input type="checkbox"/>
Earn \$0.20 for each high-quality paragraph	<input type="checkbox"/>	<input type="checkbox"/>
...
Earn \$1.00 for each high-quality paragraph	<input type="checkbox"/>	<input type="checkbox"/>

The reservation wage is the midpoint between the two wages at which the worker switches from not wanting the job to wanting the job.

Cooperation and sharing with managers. (Promotion experiment only, manager sample only). This section of the survey started with, “The next three questions are going to ask you about whether you would want to work together with your manager on a task in the future and how you would want to share a thank-you bonus with your manager. So, we’ll give you a reminder about what you know about your manager.” Workers then saw the same manager profile and three previously-promoted workers from the beginning of the survey (Appendix Figure A.6).

- **Cooperative task reservation wage.** Workers were told the following: “In other future jobs, we will also be asking workers to work together with their manager from today’s task or a similar manager to produce high-quality summaries of the more complicated scientific texts. Instead of using managers to assign workers to the harder summarizing task: (1) All workers will summarize complicated scientific paragraphs (summarizing round) (2) Managers will review worker summaries, leave comments

and choose a bonus payment for the worker (3) Workers will have another chance to revise their work (editing round). The manager will have some discretion over how much workers are paid, but there will be a base payment per paragraph. Below, we would like to know if you would be interested in the job at each of the given base payments per paragraph, where this base payment applies to both to first summarizing round and the second editing round. We will randomly choose one of the base payments below, and among the workers who said they would be interested in the job at that wage, we will randomly choose 20 workers to be offered the job.” The reservation wage is the midpoint between the two wages at which the worker switches from not wanting the job to wanting the job.

	I would not want this job (with this base payment)	I would want this job (with this base payment)
Earn \$0.05 for each high-quality summary and edit	<input type="checkbox"/>	<input type="checkbox"/>
Earn \$0.10 for each high-quality summary and edit	<input type="checkbox"/>	<input type="checkbox"/>
...
Earn \$1.00 for each high-quality summary and edit	<input type="checkbox"/>	<input type="checkbox"/>

- Willingness to pay to choose own manager.** Next, workers are told the following: “Now, imagine that we will pay \$1.00 per high-quality completed summary and you are interested in doing this future task at this wage. The default in this future task is that you will be assigned to the same or a similar manager as in the survey today. This manager will review your summaries, leave comments, and suggest your bonus payment. But, you can give up part of your wage in order to be able to choose who you want to review your work from a list of 5 managers, including the one who assigned you today. Below, we would like to know if you would want to keep your same manager or pay to choose your manager, if it costs the amount of money in the leftmost column to choose your own manager. We will be randomly choosing 20 additional workers who are interested in this job if we pay \$1.00 per high-quality summary in the first round and edit in the second round, and offer them this job. We will randomly choose one of the prices below and implement their choices at that price. For example, if we choose the price \$0.10, and you said that you would want to choose your own manager at that price, you would get to choose your own manager, and be paid \$0.90 per high-quality summary. Otherwise, you will work with the same manager as today and be paid \$1.00 per high-quality summary. In either case, your manager would determine your bonus payment.” The willingness to pay is the midpoint between the two prices at which the worker switches from wanting to switch to not wanting to switch.

	I would keep the manager from today (at this price)	I would want to choose a new manager (at this price)
Pay \$0.05 per summary/edit to choose a manager	<input type="checkbox"/>	<input type="checkbox"/>
Pay \$0.10 per summary/edit to choose a manager	<input type="checkbox"/>	<input type="checkbox"/>
...
Pay \$0.95 per summary/edit to choose a manager	<input type="checkbox"/>	<input type="checkbox"/>
Pay \$1.00 per summary/edit to choose a manager	<input type="checkbox"/>	<input type="checkbox"/>

- **Generosity.** Finally workers play a standard dictator game with their manager. They are told, “Finally, as a thank-you for your participation in the study, 20 workers will be randomly selected to receive an extra \$20 bonus. If you are chosen, you have the option of allocating some of your thank-you bonus to your manager, whose job was to assign you and other workers to different tasks. If you are selected as to receive the extra \$20 bonus, how much, if any, of your extra \$20 bonus would you share with your manager? If you are selected to receive the extra bonus, your choice here will be implemented. (Your manager will not know who did or did not share thank-you bonuses with them.)”

Workers chose how much they would like to keep on a sliding scale, with labels “You keep nothing, Manager gets \$20” at 0, “You keep \$10, Manager gets \$10” at 10, and “You keep \$20, Manager gets nothing” at 20. The outcome is the chosen amount they share.

Beliefs about promotion. After the reservation wage elicitation for the case where they would be evaluated by the same procedure as in the experiment, they were asked, “if you are selected to participate in this round, what do you think is the probability that [your manager/the algorithm] ... will assign you to the higher-paying, harder task? (in the promotion experiment) ... would choose you to do the proofreading task? (in the hiring experiment).”

After the reservation wage elicitation for the case where they would be evaluated using the cutoff rule in baseline quiz scores, the question was instead “If you are selected to participate in this round, what do you think is the probability that you would be among the highest scorers on the quizzes and therefore ... assigned to the higher-paying, harder task? (in the promotion experiment) ... hired to do the proofreading task (in the hiring experiment).”

In both cases, workers answered by dragging a sliding scale from 0 to 100, labeled as “I will definitely be assigned to the lower-paying, easier task” at 0, “It’s the same as a coin flip” at 50, and “I will definitely be assigned to the higher-paying, harder task” at 100. The outcome is workers’ reported probability.

Comprehension of evaluation procedures. At the end of the hiring experiment, workers learned that 100

workers would be randomly selected to earn a small bonus for each of two comprehension questions they answered correctly.

- **Decision-maker inputs.** In both the manager and algorithm samples, workers were asked to select all that apply to the question “What information did [your manager/the algorithm] have about workers when deciding who to hire for the proofreading job? The list below was in a random order.

Age; Gender; Race/Ethnicity; An avatar; All three numeric quiz scores from the baseline survey (science, spelling, and grammar); The average quiz score (for example, 85%); 1-5 stars representing average quiz scores; Work history; Time on Prolific; Education

Workers were correct if they selected only the italicized and italicized+underlined options above in the non-blind decision-maker arms, and if they selected only the italicized (without an underline) options in the demographic-blind decision-maker arms.

- **Manager incentives (payment structure).** In the manager sample, workers were asked to select all that apply to the question, “What as the basis of your manager’s bonus payment?” The list below was in a random order.

The performance in the proofreading task of the workers they hired; How many workers they hired; Whether they choose workers with the highest screening quiz scores; Whether they chose diverse workers; How many workers they hired showed up to do the proofreading task

Workers were correct if they only selected the italicized options in any manager treatment arm

- **Algorithm design.** In the algorithm sample, workers were also to select all that apply to the question, “What do you know about how the algorithm was designed?”

It was predicting worker performance at the proofreading task; It was predicting whether a manager would have hired a worker; Call the thing the algorithm was predicting Y. The algorithm hired whoever had the highest predicted Y; Call the thing the algorithm was predicting Y. The algorithm would provide the same predicted Y for any worker with the same education and quiz score; Call the thing the algorithm was predicting Y. The algorithm could favor certain people and hire them even if they didn’t have the highest predicted Y.

Workers were correct if the selected only the italicized and italicized+underlined options above in the demographic-blind decision-maker arms, and if they selected only the italicized (without an underline) options in the non-blind decision-maker arms.

Psychological mechanisms. Note: before these questions were asked, workers were assured that their answers to them were confidential and would not affect their evaluation or chances of future work.

- **Affect.** Workers are asked to indicate to what extent they feel each of the following emotions right now, on a scale from 1 (not at all) to 6 (very much): happy, at ease, anxious, annoyed, motivated, calm, tired, bored, gloomy, active. This is the standalone short- form 10-item Daniels five-factor measure of affective well-being (D-FAW; Russell and Daniels, 2018). Mixed in with the standard items, they are also asked how discouraged and upset they are in order to validate responses to how

motivated and at ease they are, key emotions of interest. I standardize their rating of each emotion to have mean zero and standard deviation one in the demographic-blind manager group and the overall index of psychological well-being is the mean of the twelve standardized variables, with the sign of negative emotions (anxious, annoyed, tired, bored, gloomy, discouraged, upset) flipped.

- **Self-efficacy.** Work self-efficacy is one's confidence in their ability to do the tasks required of them in a particular job. To assess participants' work self-efficacy about the tasks at hand, workers are asked how much they agree or disagree with the following statements on a Likert scale from 1 (Strongly disagree) to 5 (Strongly agree):

(1) I am capable of doing the harder proofreading job well; (2) I would have liked a chance to do the harder proofreading task; (3) I am confident in my ability to work under pressure; (4) I am capable of doing the easier proofreading job well; (5) I did a good job on the proofreading task today; (6) I was able to improve as I proofread more paragraphs

And to understand their self-efficacy related to the underlying skills they possess, they are asked to indicate their skill level in the following areas on a Likert scale from 1 (Not at all skilled) to 5 (Completely skilled):

(1) Written communication; (2) Oral communication; (3) Problem solving; (4) Numeracy; (5) Motivation; (6) Learning new material

For both measures, I standardize workers' responses to each component to have mean zero and standard deviation one in the control group and take the average to form an index of task-specific and skills-based self-efficacy. In the promotion experiment, I average both together as a total index in Appendix Table A.11, but the (lack of) effects on the two indices separately are similar. In the hiring experiment, I only observe the index of underlying skills.

- **Self-reported interest in future work.** Workers were asked to indicate how much they agreed or disagreed with the following on a Likert scale from 1 (Strongly disagree) to 5 (Strongly agree):

(1) I want to complete more surveys for this employer; (2) I want to complete more surveys with tasks that are assigned in this way; (3) I want to complete more proofreading tasks of this level of difficulty; (4) I want to complete more difficult proofreading tasks.

The outcomes in columns 1-4 of Appendix Table A.8 are indicators for whether workers strongly agreed with each of the above statements. The outcome in column 5 comes from a question in which they were briefly reminded of their evaluation and non-promotion and asked, "Would you have preferred to be assigned to the harder, higher-paying task?" The outcome of interest is an indicator for answering "Definitely" or "I think so," rather than "I don't think so," "Definitely not," or "I don't care." These questions were only asked in the promotion experiment.

- **Job satisfaction.** Workers were asked to indicate how satisfied they were with the following on a Likert scale from 1 (Very dissatisfied) to 4 (Very satisfied):

(1) How satisfied are you on the whole with the work that you were offered in this survey?; (2) How satisfied are you on the whole with the work that is available on Prolific?; (3) How satisfied are you on the whole with the work that you do outside of Prolific (if applicable)?

The measure of job satisfaction is an indicator for whether they said that they were very satisfied with the work they were offered in this survey. These questions were only asked in the promotion experiment.

A.4.2 Control variables

All control variables are the same in the promotion and hiring experiments.

Demographic controls.

- Education: *What is the highest level of education you have attained?* I control for seven mutually-exclusive groups {Less than high school, high school graduate, some college but no degree, 2 year college degree, 4 year college degree, Professional or Masters degree, Doctorate}
- Income: *What is your annual household income (pre-tax)?* I control for nine mutually-exclusive groups {Less than \$20,000, \$20,000-40,000, \$40,000-60,000, \$60,000-80,000, \$80,000-\$100,000, \$100,000-120,000, \$120,000-140,000, \$140,000-160,000, More than \$160,000}
- Married: *Are you married or in a long-term partnership?* I control for an indicator for answering “Yes” rather than “No”
- Kids: *Do you have children?* I control for an indicator for answering “Yes” rather than “No”
- Age: *How old are you?* I control for six mutually-exclusive groups {18-24, 25-34, 35-44, 45-54, 55-64, over 65}
- Race: *With what racial group do you most strongly identify? and Do you identify as Hispanic or Latino?* I code workers as one of Asian, Black, white and Hispanic, or white and non-Hispanic and control for these four mutually-exclusive categories. Workers who identified as multiracial, other race, or Native American were grouped with Asian, Black, or Hispanic workers based on the race recorded in Prolific’s administrative data, which matched one of those groups. Black, Hispanic workers were grouped with Black, Non-Hispanic workers.
- Gender: *What is your gender?* I control for an indicator for answering “Man” rather than “Woman,” “Non-binary,” or “Other”

Baseline quiz scores. I control for the number of questions workers got correct on each of the following three quizzes:

- Spelling quiz: Workers had 12 seconds to play an audio recording of the word and type it into a text box. They earned a bonus of \$0.025 for each word they spelled correctly. The words were: hypothesis, paleontology, equilibrium, herbivore, aerobic, enzyme, homeostasis, chlorophyll, sedimentary, and vertebrae. They were reminded to have their audio on and volume up before beginning, and had two practice words to make sure the audio was working.
- Grammar quiz: Workers had 4 minutes to answer as many of 15 questions as they could. They earned a bonus of \$0.04 for each question they answered correctly.

- Which of these is not a word or phrase? [a lot, *alot*, allot]
 - You’ve probably heard the phrase “i before e except after c” ...but which of these words defies this rule? [species, science, policies, *all of them*]
 - A plural subject needs [a singular verb, *a plural verb*]
 - When two singular subjects are connected by ‘or,’ use [*a singular verb*, a plural verb]
 - Which phrase is incorrect? [should’ve, should have, *should of*]
 - If an opinion-adjective and a fact-adjective are used before a noun, which comes first? [a fact-adjective, *an opinion-adjective*, it doesn’t matter]
 - Fill in the blank: “Bad weather can _____ people’s ability to work” [*affect*, effect]
 - Every sentence must have a subject and [an object, *a verb*, an adjective, an adverb, a phrase]
 - If someone says “I’m sorry” you can _____ their apology [except, *accept*]
 - The order of a basic positive sentence is [*subject-verb-object*, verb-object-subject, object-verb-subject, subject-object-verb]
 - Which is correct? [*they’re looking good*, their looking good, there looking good]
 - Fill in the blank. “_____ so hot outside!” [*it’s*, its]
 - Which is correct? [the boy’s dog bark loudly, the boy’s dog loudly barks, *the boy’s dog barks loudly*, they boys dog barks loudly]
 - Select all of the sentences that are written in the active voice [grass is eaten by cows, the books were written by that author I like, *she drove over the bridge*, *horses eat hay*, *they clued us in*]
 - Select all that apply: The contraction “she’s” can mean... [*she is*, *she has*, she was]
- Science quiz: Workers had 4 minutes to answer as many of 11 questions as they could. They earned a bonus of \$0.05 for each question they answered correctly.

When large areas of forest are removed so land can be converted for other uses, such as farming, which of the following occurs?

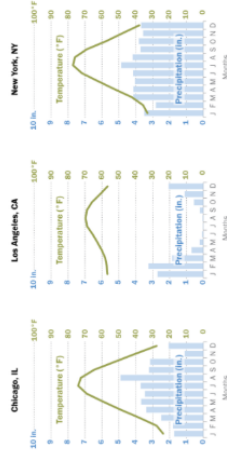
Decreased carbon dioxide

Colder temperature

Increased erosion

Greater oxygen production

These graphs show the monthly precipitation and average temperature for three cities in the United States over the course of one year. Based on the graphs, which city has the greatest annual range of temperatures?



Chicago

Los Angeles

New York

The all have the same annual temperature range

Which of these is a major concern about the overuse of antibiotics?

It can lead to antibiotic-resistant bacteria

There will be an antibiotic shortage

Antibiotics can cause secondary infections

Antibiotics will get into the water system

Which of the following is an example of genetic engineering?

Growing a whole plant from a single cell

Finding the sequences of bases in plant DNA

Attaching the root of one type of plant to the stem of another type of plant

Inserting a gene into plants that makes the plant resistant to insects

Many diseases have an incubation period. Which of the following best describes what an incubation period is?

The recovery period after being sick

The period during which someone has an infection, but is not showing symptoms

The period during which someone builds up immunity to a disease

The effect of a disease on babies

Oil, natural gas, and coal are examples of:

Geothermal resources

Biofuels

Fossil fuels

Renewable resources

An antacid relieves an overly acidic stomach because the main components of antacids are ...

Acids

Neutral

Isotopes

Bases

A scientist is conducting a study to determine how well a new medication treats ear infections. The scientist tells the participants to put 10 drops in their infected ear each day. After two weeks, all participants' ear infections had healed.

Which of the following changes to the design of this study would most improve the ability to test if the new medication effectively treats ear infections?

Create a second group of participants with ear infections who use 15 drops a day

Have participants put ear drops in both their infected ear and healthy ear

Have participants use ear drops for only one week

Create a second group of participants with ear infections who do not use any ear drops

The time a computer takes to start has increased dramatically. One possible explanation for this is that the computer is running out of memory. This explanation is a scientific...

Experiment

Hypothesis

Observation

Conclusion

What is the main cause of seasons on the Earth?

The tilt of the Earth's axis in relation to the sun

The distance between the Earth and the sun

Changes in the amount of energy coming from the sun

The speed that the Earth rotates around the sun

A car travels at a constant speed of 40 miles per hour. How far does the car travel in 45 minutes?

25 miles

30 miles

35 miles

40 miles

Previously-selected worker education and baseline performance. In all treatment groups, workers saw the education and baseline performance (as 1-5 stars representing quintiles of average quiz scores) of three workers who were previously promoted or hired by their manager or the algorithm that evaluated them. Recall that workers were in “quiz-score groups” that only included workers within one quiz-score quintile of each other, and I include controls for the quiz-score group in all regressions. I also include controls that measure how a worker compares to the previously-promoted workers, conditional on their quiz-score group:

- Performance: I control for four mutually-exclusive indicators: Whether they see three previously-selected workers with the maximum number of stars, two previously-selected workers with the maximum number of stars (and thus one with the minimum), one worker with the maximum number of stars (and thus two with the minimum), or three previously-selected workers with the minimum number of stars.
- Education: I control for whether they see three previously-selected workers with more education than themselves, two with more education than themselves, one with more education than themselves, three with the same education as themselves, one with less education than themselves, two with less education as themselves, or three with less education than themselves

A.4.3 Variables used in heterogeneity

Past experiences of discrimination. Workers answered a module on their prior work history at the end of the baseline survey. They are told that the remaining questions are about if they’ve ever felt treated unfairly at work because of their age, race, gender, etc., and asked:

- Do you think you have ever been discriminated against while applying for a job? That is, do you think you were or weren’t hired unfairly because of some characteristic like your race, age, gender, disability status, religion, etc. rather than because of your qualifications?
- Do you think you have ever been discriminated against by your employer in a promotion decision? That is, do you think you were or weren’t promoted unfairly because of some characteristic like your race, age, gender, disability status, religion, etc. rather than because of your qualifications?
- Do you think you have ever been discriminated against by your employer in a termination decision? That is, do you think you were or weren’t fired unfairly because of some characteristic like your race, age, gender, disability status, religion, etc. rather than because of your qualifications?
- During day-to-day activities at work or while working, do you feel that you have been treated differently than others based only on your gender, race, age, or some other demographic characteristic?

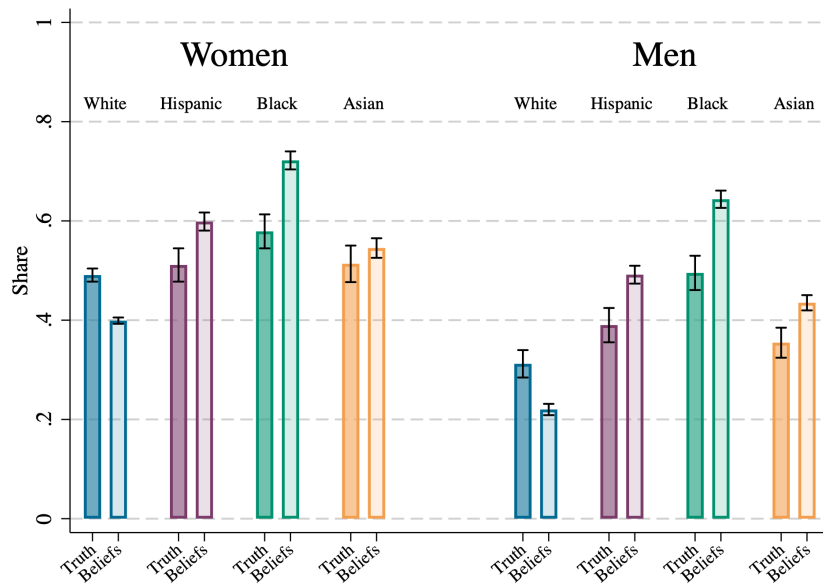
After each question, they could answer “Yes,” “No,” or “It’s impossible to know.” If they answered yes, they were asked if they thought they were discriminated against based on gender, race, age or something else, and to select all that apply, including an option to write in “other.” Finally, they were asked if they experienced positive or negative discrimination: e.g. “**I was fired** because of some demographic characteristic and **I shouldn’t have been,**” or “**I was not fired** because of some demographic characteristic and **I should have been fired.**”

In the heterogeneity analysis, I code someone as having experienced discrimination in the past if they say yes to any of the four questions and say that the discrimination was on the basis of race and gender and say that it was negative discrimination (almost none of the workers report experiencing positive discrimination, likely because they did not think of that as discrimination in the first place and so were not asked whether it was positive or negative).

Confidence. After each of the quizzes, workers were asked how many questions they thought they got right from a multiple choice list from 0-10, 11, or 15 in the case of spelling, science, and grammar, respectively. They earned a \$0.20 bonus if they were correct about how many questions they got right for each quiz. The difference from their actual score is a measure of domain-specific confidence. As a general measure, I standardize the difference between their belief and their score to have mean zero and standard deviation one in the demographic-blind manager group, average the standardized variables across the three quizzes, and use an indicator for having this statistic be above-median to test whether confidence matters in heterogeneity analysis.

Beliefs about the prevalence of discrimination. Directly after the questions about past experiences of discrimination, I ask workers, “What fraction of each of the following demographic groups they think would answer “yes” to any of the previous questions? That is, what percent of each demographic group do you think reports that they have been discriminated against while searching for a job, in promotion or termination decisions, or in day-to-day work?” I tell them that the sample is other workers recruited on Prolific. They answer on a slider for the eight {Asian, Black, Hispanic, White} × {men, women} groups from 0 to 100, labeled at 0 and 100 “No one in this group reports experiencing discrimination” and “Everyone in this group reports experiencing discrimination,” respectively.

Their answer is incentivized; I randomly select one group and workers are paid a \$0.50 bonus if their answer is within 5pp of the truth. The measure used in heterogeneity analysis is an indicator for whether their beliefs about the prevalence of discrimination for their *own group* is above- or below-median. The figure below shows the average belief compared to the truth separately for workers in each group.



Note: This figure plots the fraction of workers in each race \times gender group who report ever experiencing discrimination at work in the past and the average belief about workers in that group about the same statistic. Standard errors are shown in black bars.

Racial- and gender-identity centrality. Workers completed two standard scales that measure racial- and gender-identity centrality, or the degree to which race and gender are, respectively, important to one’s sense of self. In both cases, workers were asked to rate statements on a Likert scale from 1 (Strongly disagree) to 7 (Strongly agree). Indices are formed by standardizing each response to have mean zero and standard deviation one and taking the average, flipping the items marked with a (-) to have the opposite sign.

- Racial identity centrality:
 - Overall, my racial group identity has very little to do with how I feel about myself (-)
 - In general, identifying with my racial group is an important part of my self-image (+)
 - My destiny is tied to the destiny of other people in my racial group (+)
 - The racial group I identify with is unimportant to my sense of what kind of person I am (-)
 - I have a strong sense of belonging to my racial group (+)
 - I have a strong attachment to other people in my racial group (+)
 - The racial group I identify with is an important reflection of who I am (+)
 - Belonging to the racial group I identify with is not a major factor in my social relationships (-)

- Gender identity centrality:
 - I often think about the fact that I am a [man/woman] (+)

- Overall, being a [man/woman] has very little to do with how I feel about myself (-)
- The fact that I am a [man/woman] rarely enters my mind (-)
- I am not usually conscious of the fact that I am a [man/woman] (-)
- In general, being a [man/woman] is an important part of my self-image (+)
- Being a [man/woman] is an important reflection of who I am (+)
- In my everyday life, I often think about what it means to be a [man/woman] (+)

A.5 Coding the text data

As the first step for coding all three variables that came from open-ended responses, I read every response and coded them myself. These measures are not used in the paper, but correlate extremely highly with the final measures. To ensure that the process was entirely blind to the study purpose and treatment assignment, external coders generated equivalent measures and these are what are used in the paper. In the promotion experiment, the measures used in the paper were coded by two MIT undergraduate economics majors and their disagreements were resolved by a professional hired on Upwork (all three blind to study purpose and treatment assignment). The process with these coders was as follows:

1. Given my initial reading, I generated coding schemes (described in each subsection below) for each variable and provided them in an Excel spreadsheet with all of the open-ended responses, stripped of all identifiers and other variables.
2. The two undergraduate coders each received the same spreadsheet and separately coded each variable according to the coding scheme using a drop-down menu in Excel. Each variable was a different sheet in the spreadsheet. When they had questions, the questions and my response were shared with both of them, but they did not otherwise communicate about the exercise. They could attach up to three codes to each response answer.
3. Any observations with a discrepancy between the two coders (details on frequency follow) were reviewed by a PhD psychologist who professionally codes qualitative data hired on Upwork. Discrepancies included having the same codes but in a different order or having a different set of codes from each other. She could either indicate agreeing with one of the two original coders or she could re-code the response (which generally meant combining/re-ordering the original coders' tags).

In the hiring experiment, due to budget and time constraints, I only hired the professional coder on Upwork to code the data. Now, she was carrying out the same process as the undergraduate coders did for the data from the promotion experiment. Again, her measures correlate extremely highly with those that I coded myself. The codebooks for each variable were the same in both experiments and are described next.

In the promotion experiment, the external coders generated the indicators used in the paper for the main measure of perceived discrimination, whether workers complained about bias, and whether they knew the study purpose. In the hiring experiment, due to budget and time constraints, the external coder generated the indicators used in the paper for the main measure of perceived discrimination and whether workers complained about bias, but I generated the indicators for whether workers knew the study purpose myself.

A.5.1 The main measure of perceived discrimination

Reasons workers gave for their perception of what needed to be different about their profile to be assigned to the harder job could be classified into one of the following five categories, with the descriptions below provided to each coder. They ranked up to three given reasons.

1. **More school:** Needing more school or a higher degree
2. **Higher quiz score:** Needing a higher score, higher quiz score, or “more stars”

3. **Demographics:** They would have been promoted if they had (a) different demographic(s) characteristic(s) or reference bias or discrimination
4. **Random:** It was random, or they don't know because they have the same qualifications as the others
5. **Other**

The undergraduate coders disagreed for 317 of 2,397 observations and these were resolved by the coder on Upwork. Of these, she agreed with coder 1 for 33 percent of observations, coder 2 for 58 percent of observations, and provided new codes for 8 percent of observations.

The main measure of perceived discrimination is an indicator for whether the final code was that the worker mentioned demographics as a reason they weren't promoted, regardless of its rank in their reasons. Of the workers who mentioned demographics at all, 58 percent mentioned it as their primary reason.

The control group means and treatment effects of evaluation by a non-blind manager on whether the *primary* reason was demographics, more school, higher quiz scores, or random are in Appendix Tables A.5 and A.15 (Panel A) and Appendix Tables A.10 and A.16 (Panels A, B, and C), respectively.

A.5.2 Secondary measure of perceived discrimination (complaints about bias)

Workers were asked near the end of the survey if they had any complaints about how they were assigned to jobs. If they said yes (N=368, 15 percent of workers), they were asked to describe their complaint. Workers' complaints could be classified into one of the following five categories, with the descriptions below provided to each coder. They ranked up to three reasons the worker had a complaint.

1. **Discrimination:** On the basis of any demographics
2. **Unfair:** Any unfairness other than discrimination on the base of demographics
3. **College degree:** College degree is irrelevant and shouldn't have been used to make promotion decisions, or a worker "was discriminated against because they didn't have a degree."
4. **Confusion:** They seem confused about the job or job-assignment mechanism, or don't remember taking the screening quiz
5. **Other**

The undergraduate coders disagreed for 128 of 368 observations and these were resolved by the coder on Upwork. Of these, she agreed with coder 1 for 35 percent of observations, coder 2 for 61 percent of observations, and provided new codes for 4 percent of observations. The complaints-based measure of perceived discrimination is an indicator for whether the final code was that the worker complained about discrimination, regardless of its rank. Of those who complained about discrimination, it was the primary complaint for 93 percent (only 33 workers gave more than one reason for their complaint). Of those who had a complaint, 30 percent referenced discrimination in their description (23 percent referenced other types of unfairness, 18 percent referenced college degrees, and 10 percent were confused; 29 percent had some other complaint).

A.5.3 Study topic

Workers were asked about their beliefs about the study topic at two times in the survey. In the proofreading experiment, half of the sample was randomly chosen to be asked for the first time between the proofreading section and the reservation wage and willingness to pay elicitation, the other half was asked after the reservation wage elicitation, and all workers were asked a second time at the very end of the survey. In the hiring experiment, all were asked after the reservation wage elicitation and at the very end of the survey. Their beliefs about the study topic could be broadly categorized into six categories, with the following descriptions given to each coder. The coders ranked up to three codes per response (no response referenced more than three topics).

1. **Proofreading:** The study is about proofreading, including gender, race, or education differences in proofreading ability
2. **Motivation/persistence:** The study is about persevering through a long and difficult or tedious job
3. **Perceptions of job assignment:** The study is about perceptions of job assignment but don't mention discrimination or bias specifically. Can include fairness.
4. **Discrimination/bias:** The study is about (perceptions of) discrimination or bias in job assignment
5. **I don't know**
6. **Other**

Study topic 1: The undergraduate coders disagreed for 933 of 2,397 observations and these were resolved by the coder on Upwork. Of these, she agreed with coder 1 for 38 percent of observations, coder 2 for 52 percent of observations, and provided new codes for 10 percent of observations. Study topic 2: The undergraduate coders disagreed for 885 of 2,397 observations and these were resolved by the coder on Upwork. Of these, she agreed with coder 1 for 55 percent of observations, coder 2 for 33 percent of observations, and provided new codes for 12 percent of observations.

Appendix Table A.30 shows treatment effects on whether workers mentioned discrimination/bias. Of those who mentioned discrimination/bias when asked what they thought the study topic was the first time (N=102), 67 percent mentioned discrimination primarily. Of those who mentioned discrimination/bias at the end of the survey (N=1013), 79 percent mentioned discrimination primarily.

In the promotion experiment, the primary study topic participants hypothesized the first time and second time they were asked followed this distribution:

	1st time asked	2nd time asked
Proofreading	28 percent	11 percent
Motivation	16 percent	5 percent
Perceptions of assignment	12 percent	13 percent
Discrimination/bias	3 percent	33 percent
I don't know	31 percent	30 percent
Other	10 percent	8 percent

The share of workers who thought the study was about discrimination in the hiring experiment is plotted in Appendix Figure A.48.

A.6 Instrumental variables (IV) specification results

As described in Section 1.3.1, an IV strategy estimating the effects of perceived discrimination has interpretation challenges due to the amorphous nature of perceived discrimination. The general exclusion restriction—that treatment only affects outcomes via “perceived discrimination”—is valid, because the experimental design carefully eliminates any other channels. That said, IV estimates effectively scale the ITT estimates by the “first stage,” in this case, the effect of treatment assignment on perceived discrimination.

I measure many (highly correlated) measures of perceived discrimination, but they vary substantially in the size of the first stage. Thus, IV estimates vary greatly depending on which measure of perceived discrimination is used. For example, the first stage for the main measure of perceived discrimination is 31pp, but the first stage for the measure based on whether workers make a complaint about bias is only 10pp. One can think of the frequency of a variable as the inverse of its seriousness or severity, so it is expected that these two measures are different in this way.

Interpreting an IV estimate in this case requires taking a stand on the “type” of perceived discrimination through which the treatment has effects, and assuming away treatment heterogeneity depending on the severity of the perceived discrimination. For example, if only those who make a complaint about bias are affected, the implied effect of perceived discrimination is 3.3 times as large as the case where anyone who thinks demographics played a role in their evaluation is affected.

Thus, the paper presents the ITT estimates. These are a lower bound on the average effects of perceived discrimination in the case that everyone perceives some discrimination. Here, I present the results from an IV specification that instruments the main measure of perceived discrimination with treatment assignment. The main measure of perceived discrimination has the largest first stage of all of the measures of perceived discrimination, making these the most conservative IV estimates using any observed measure.

Table A.17: The effect of perceived discrimination (IV estimates) on retention

	Num paragraphs (1)	More than 6 (2)	Did all 18 (3)
<i>Panel A: Manager sample, promotion experiment</i>			
Perceived manager bias	-1.596** (0.806)	-0.050 (0.056)	-0.156* (0.081)
N	1387	1387	1387
Control mean	15.916	0.915	0.776
<i>Panel B: Algorithm sample, promotion experiment</i>			
Perceived algorithmic bias	0.820 (2.266)	0.169 (0.170)	0.029 (0.226)
N	691	691	691
Control mean	15.456	0.881	0.730

Note: This table shows instrumental variables estimates of the effect of the main measure of perceived discrimination on retention, instrumented with random assignment to treatment. This effectively transforms the ITT effects in Table 1.2 by dividing them by the first-stage ITT effects on the main measure of perceived discrimination (in column 1 of Figure 1.1). Outcomes and control variables are otherwise the same as Table 1.2. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.18: The effect of perceived discrimination (IV estimates) on effort and performance

	Minutes	Clicks	Correct	Incorrect	Bonus
	(1)	(2)	(3)	(4)	(5)
<i>Panel A: Manager sample, promotion experiment</i>					
Perceived manager bias	-0.085 (0.175)	1.975 (5.652)	1.003 (0.719)	0.698 (0.687)	0.023 (0.068)
N	1389	1389	1389	1389	1389
Control mean	4.957	36.804	14.817	3.422	0.919
<i>Panel B: Algorithm sample, promotion experiment</i>					
Perceived algorithmic bias	0.356 (0.481)	-10.615 (15.439)	-4.137* (2.238)	1.321 (1.612)	-0.348* (0.206)
N	691	691	691	691	691
Control mean	4.817	37.941	15.016	3.679	0.935

Note: This table shows instrumental variables estimates of the effect of the main measure of perceived discrimination on effort and performance in the first six required paragraphs, instrumented with random assignment to treatment (analogous to Appendix Table A.17 for the outcomes in Table 1.3). Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.19: The effect of perceived discrimination (IV estimates) on future labor supply

	Res. wage (same)	WTP cutoff
	(1)	(2)
<i>Panel A: Manager sample, promotion experiment</i>		
Perceived manager bias	0.072** (0.032)	0.032** (0.014)
N	1338	1325
Control mean	0.265	-0.014
<i>Panel B: Algorithm sample, promotion experiment</i>		
Perceived algorithmic bias	0.066 (0.086)	0.007 (0.033)
N	672	660
Control mean	0.282	-0.008

Note: This table shows instrumental variables estimates of the effect of the main measure of perceived discrimination on reservation wages, instrumented with random assignment to treatment (analogous to Appendix Table A.17 for the outcomes in Table 1.4). Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.20: The effect of perceived discrimination (IV estimates) on beliefs about promotion

	Pr(Promoted Same) (1)	Pr(Promoted Cutoff) (2)	Diff Pr(Promoted) (3)
<i>Panel A: Manager sample, promotion experiment</i>			
Perceived manager bias	-8.668** (3.675)	-3.711 (3.831)	-4.957** (2.298)
N	1385	1385	1385
Control mean	47.219	48.491	-1.272
<i>Panel B: Algorithm sample, promotion experiment</i>			
Perceived algorithmic bias	-2.722 (9.552)	-8.743 (10.455)	6.021 (5.582)
N	689	689	689
Control mean	45.908	48.201	-2.293

Note: This table shows instrumental variables estimates of the effect of the main measure of perceived discrimination on beliefs about promotion, instrumented with random assignment to treatment (analogous to Appendix Table A.17 for the outcomes in Table 1.6). Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.21: The effect of perceived discrimination (IV estimates) on sharing with and avoidance of manager

	Cooperative task RW (1)	WTP new manager (2)	Shared zero (3)	Amt bonus shared (4)
Perceived manager bias	0.025 (0.046)	0.124** (0.060)	0.158* (0.091)	-0.815 (0.733)
N	1349	1030	1383	1383
Control mean	0.341	0.223	0.522	2.857

Note: This table shows instrumental variables estimates of the effect of the main measure of perceived discrimination on sharing and willingness to work with managers, instrumented with random assignment to treatment (analogous to Appendix Table A.17 for the outcomes in Table 1.5). Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

A.7 Other heterogeneity

In the main analysis, I focus on heterogeneity by race and gender as it yields the most consistent patterns of the pre-registered heterogeneity analysis. Here, I test for heterogeneity along other dimensions.

All pre-registered dimensions of heterogeneity. Appendix Tables A.22, A.23, and A.24 show results for these outcomes from a single multivariate regression that simultaneously tests gender and all other pre-registered dimensions of heterogeneity (an indicator for whether an individual reported experiencing discrimination at work in the past, has above-median beliefs about the prevalence of discrimination against their group, or has above-median confidence in related domains). The effects on retention are most negative for workers with below-median confidence. The pattern of gender heterogeneity for the retention and effort/performance outcomes persists strongly. The effect on willingness to pay to be able to choose one's own manager is largest for workers who reported experiencing discrimination at work in the past. There are no other significant dimensions of heterogeneity.

Heterogeneity by outside options. This dimension of heterogeneity was not pre-registered, but I also test for treatment heterogeneity by whether workers are currently employed outside of Prolific, a proxy for their outside options. Perhaps surprisingly, there is no differential effect on retention, but the negative effect on performance is driven by workers who are currently employed outside Prolific (Appendix Tables A.25 and A.26), so perhaps workers with higher outside options are more likely to mentally disengage when they perceive discrimination. There is no differential effect on any measures of future labor supply by whether workers are currently employed outside Prolific or not, though the effects are driven by the workers who are currently employed outside Prolific (Appendix Table A.27).

Table A.22: Effects on retention, all measures of heterogeneity

	Num paragraphs	More than 6	Did all 18
	(1)	(2)	(3)
Non-blind mgr	-0.916*	-0.034	-0.107**
	(0.480)	(0.034)	(0.048)
Non-blind mgr*Minority Men	1.429**	0.100**	0.112*
	(0.614)	(0.046)	(0.060)
Non-blind mgr*Exp Past Discr	-0.545	0.004	-0.066
	(0.488)	(0.034)	(0.049)
Non-blind mgr*High Discr Beliefs	0.258	-0.011	0.038
	(0.483)	(0.033)	(0.049)
Non-blind mgr*High Confidence	0.610	0.003	0.108**
	(0.481)	(0.033)	(0.048)
N	1293	1293	1293
Control Mean	15.916	0.915	0.776

Note: This table estimates the effect of being in the non-blind manager arm on retention relative to the demographic-blind manager arm in the promotion experiment, allowing the effect to vary by all pre-registered dimensions of experiment-specific heterogeneity. The outcomes and sample are the same as in Table 1.2, but white men are additionally dropped from the sample. The specification is also the same, but the treatment indicator is also interacted with being a racial minority man (omitted group: women), reporting experiencing discrimination in the past (omitted group: not), having above-median beliefs about the prevalence of past experienced discrimination in this sample (omitted group: below-median), and having above-median confidence in the spelling, science, and grammar quizzes in the baseline survey (omitted group: below-median). Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.23: Effects on effort and performance, all measures of heterogeneity

	Minutes	Clicks	Correct	Incorrect	Bonus
	(1)	(2)	(3)	(4)	(5)
Non-blind mgr	-0.056 (0.104)	2.656 (3.119)	0.070 (0.441)	0.583* (0.335)	-0.015 (0.040)
Non-blind mgr*Minority Men	0.155 (0.131)	-6.437* (3.295)	1.136** (0.539)	-0.284 (0.534)	0.079 (0.051)
Non-blind mgr*Exp Past Discr	0.111 (0.109)	1.589 (3.606)	0.436 (0.448)	0.189 (0.394)	0.012 (0.043)
Non-blind mgr*High Discr Beliefs	-0.087 (0.109)	-1.102 (3.447)	-0.396 (0.455)	-0.450 (0.418)	0.014 (0.043)
Non-blind mgr*High Confidence	-0.006 (0.106)	-1.691 (3.479)	0.167 (0.442)	-0.377 (0.384)	-0.002 (0.042)
N	1295	1295	1295	1295	1295
Control Mean	4.957	36.804	14.817	3.422	0.919

Note: This table estimates the effect of being in the non-blind manager arm on effort and performance relative to the demographic-blind manager arm in the promotion experiment, allowing the effect to vary by all pre-registered dimensions of experiment-specific heterogeneity. The specification is the same as in Appendix Table A.22 and the outcomes are the same as in Table 1.3. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.24: Effects on future labor supply, all measures of heterogeneity

	RW (Same)	WTP cutoff	Coop. RW	WTP new mgr	Shared 0	Amt shared
	(1)	(2)	(3)	(4)	(5)	(6)
Non-blind mgr	0.017 (0.019)	0.015 (0.009)	0.005 (0.028)	0.064 (0.041)	-0.042 (0.055)	0.491 (0.421)
Non-blind mgr*Minority Men	0.006 (0.025)	0.002 (0.013)	-0.027 (0.035)	-0.074 (0.051)	0.070 (0.066)	0.394 (0.551)
Non-blind mgr*Exp Past Discr	0.018 (0.020)	0.003 (0.009)	0.004 (0.028)	0.064* (0.039)	0.079 (0.057)	-0.757* (0.456)
Non-blind mgr*High Discr Beliefs	0.009 (0.020)	-0.003 (0.010)	0.002 (0.028)	-0.052 (0.038)	0.044 (0.057)	-0.297 (0.460)
Non-blind mgr*High Confidence	-0.019 (0.020)	-0.007 (0.009)	0.007 (0.028)	0.007 (0.039)	0.019 (0.056)	-0.584 (0.454)
N	1247	1235	1256	957	1289	1289
Control Mean	0.265	-0.014	0.341	0.223	0.522	2.857

Note: This table estimates the effect of being in the non-blind manager arm on future labor supply relative to the demographic-blind manager arm in the promotion experiment, allowing the effect to vary by all pre-registered dimensions of experiment-specific heterogeneity. The specification is the same as in Appendix Table A.22 and the outcomes are the same as in Table 1.4. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.25: Effects on retention, heterogeneity by employment outside Prolific

	Num paragraphs (1)	More than 6 (2)	Did all 18 (3)
Non-blind mgr*Employed	-0.574* (0.296)	-0.015 (0.021)	-0.060** (0.030)
Non-blind mgr*Not employed	-0.355 (0.410)	-0.017 (0.028)	-0.028 (0.041)
<i>p-values:</i>			
Employed = not	0.649	0.952	0.507
N	1387	1387	1387
Control Mean, employed	15.927	0.914	0.780
Control Mean, not employed	15.895	0.916	0.768

Note: This table estimates the effect of being in the non-blind manager arm on retention relative to the demographic-blind manager arm in the promotion experiment, separately for workers who are currently employed outside prolific and those who are not. The outcomes and sample are the same as in Table 1.2. The specification is the same, but the indicator for treatment is replaced with two interactions between treatment and indicators for being employed or not being employed. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.26: Effects on effort and performance, heterogeneity by employment outside Prolific

	Minutes	Clicks	Correct	Incorrect	Bonus
	(1)	(2)	(3)	(4)	(5)
Non-blind mgr*Employed	-0.001 (0.067)	0.809 (2.221)	0.523* (0.277)	0.149 (0.262)	0.014 (0.026)
Non-blind mgr*Not employed	-0.076 (0.087)	0.312 (3.074)	-0.063 (0.372)	0.370 (0.319)	-0.005 (0.035)
<i>p-values:</i>					
<i>Employed = not</i>	0.475	0.896	0.196	0.564	0.661
N	1389	1389	1389	1389	1389
Control Mean, employed	4.909	37.302	14.807	3.635	0.925
Control Mean, not employed	5.050	35.844	14.835	3.013	0.909

Note: This table estimates the effect of being in the non-blind manager arm on future labor supply relative to the demographic-blind manager arm in the promotion experiment, separately for workers who are currently employed outside prolific and those who are not. The specification is the same as Appendix Table A.25 and the outcomes are the same as in Table 1.3. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.27: Effects on future labor supply, heterogeneity by employment outside Prolific

	RW (Same) (1)	WTP cutoff (2)	Coop. RW (3)	WTP new mgr (4)	Shared 0 (5)	Amt shared (6)
Non-blind mgr*Employed	0.026** (0.012)	0.012** (0.006)	0.017 (0.017)	0.046* (0.025)	0.072** (0.035)	-0.352 (0.288)
Non-blind mgr*Not employed	0.018 (0.016)	0.007 (0.008)	-0.009 (0.024)	0.032 (0.031)	0.009 (0.046)	-0.086 (0.364)
<i>p-values:</i>						
<i>Employed = not</i>	0.711	0.601	0.351	0.727	0.264	0.552
N	1338	1325	1349	1030	1383	1383
Control Mean, employed	0.269	-0.016	0.337	0.228	0.533	2.762
Control Mean, not employed	0.257	-0.010	0.347	0.214	0.502	3.038

Note: This table estimates the effect of being in the non-blind manager arm on future labor supply relative to the demographic-blind manager arm in the promotion experiment, separately for workers who are currently employed outside prolific and those who are not. The specification is the same as Appendix Table A.25 and the outcomes are the same as in Table 1.4. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

A.8 Effects on future labor supply in the hiring experiment

Unlike in the promotion experiment, both perceived manager and algorithmic discrimination increase beliefs about the likelihood of future discrimination (thus reducing beliefs about the likelihood of future job offers) when a worker will be evaluated by the same procedure in the future. There are two possible explanations for this difference between experiments. The difference may be due to differences in the timing of the reservation wages and belief elicitation relative to when workers learned about how they were evaluated, or due to the size of the “first stage,” both of which differ between the two experiments and make it more likely for there to be a larger effect on anticipated future discrimination in the hiring experiment. In Panel A of Table A.28, I pool the two manager arms with positive perceived discrimination and compare to the demographic-blind manager arm where workers do not see previous hires’ avatars, race, and gender; in Panel B, I do the same in the algorithm arms. I pool the arms to improve power (as pre-registered), but results are similar in each arm separately (Appendix Figure A.26). Column 1 shows the effect of treatment on perceived discrimination in the pooled treatments (36pp and 53pp when evaluated by a manager or algorithm, respectively, $se=2pp$ for both). Being in the manager arms with perceived discrimination lowers beliefs about the probability of being hired by 2.5pp (5 percent, $se=1.2pp$) in the manager arms or 7.8pp (15 percent, $se=1.1pp$) in the algorithm arms (Table A.28, column 2). 90 percent and 51 percent of these effects, respectively, remain after subtracting workers’ beliefs about the likelihood of promotion under the cutoff rule, and is therefore due to increased anticipated discrimination (column 3).

I also measure reservation wages for future work in the hiring experiment. Effects on these reservation wages are purely indicative of psychological costs from anticipated discrimination: workers’ choices depend only on one wage, not an expected wage, so movement of their switching point along the multiple price list cannot be caused by changes in beliefs about the likelihood of being hired.

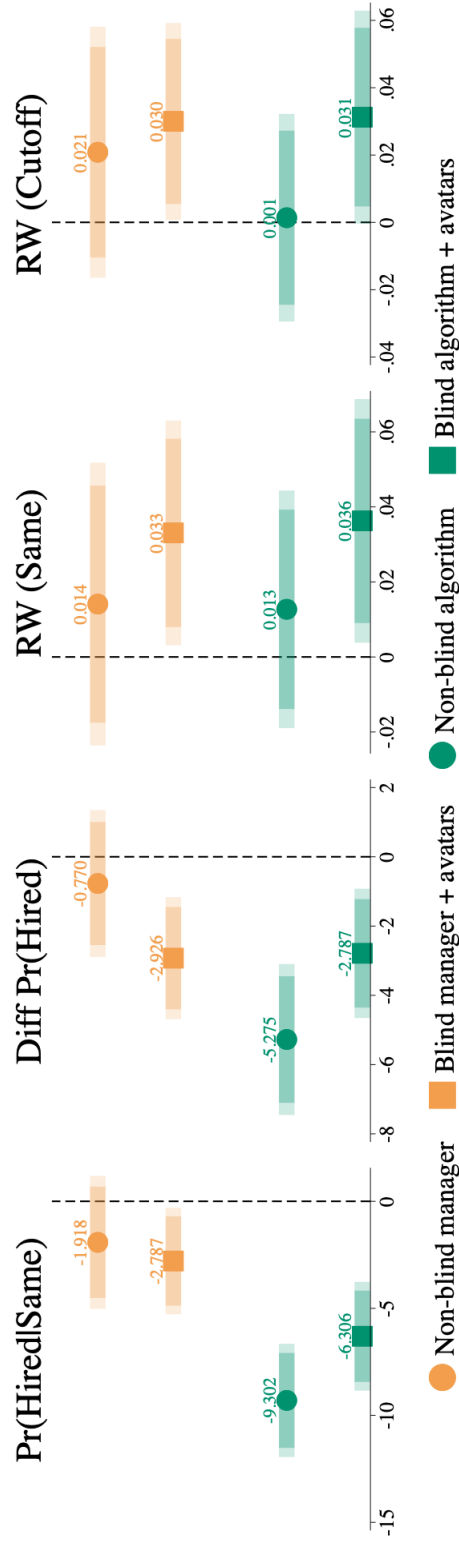
Workers do have higher reservation wages when they perceive (and anticipate) discrimination, *even though they will not interact with any manager or employer in the future if they are **not** hired*. That is, they only want to be offered the job at higher wages, implying a direct utility cost from anticipated discrimination or future interactions, just like in the promotion experiment. Unlike in the promotion experiment, this seems to be largely due to avoidance of the employer who hires biased managers or uses biased algorithms, because reservation wages increase similarly when workers will be evaluated by the same procedure as in the experiment (Table A.28, column 4) as when they will be evaluated by a cutoff rule in baseline performance (column 5). As with the promotion experiment, these results imply that workers will sort in or out of jobs based on anticipated discrimination and require higher wages in situations where they may face discrimination. Consistently, I find suggestive evidence that the treatments worsen workers’ psychological well-being.

Table A.28: Effects on future labor supply and beliefs in the hiring experiment

	Perceived Discrimination (1)	Pr(Hired) (Same mech) (2)	Diff Pr(Hired) (Same-Cutoff) (3)	RW (Same) (4)	RW (Cutoff) (5)
<i>Panel A: Manager sample, hiring experiment</i>					
Non-blind manager	0.361*** (0.020)	-2.516** (1.166)	-2.252*** (0.827)	0.027* (0.014)	0.027** (0.014)
or blind mgr + avatars	1246	1246	1246	1231	1228
N	0.007	51.572	-2.418	0.388	0.383
Control mean					
<i>Panel B: Algorithm sample, hiring experiment</i>					
Non-blind algorithm	0.529*** (0.018)	-7.753*** (1.119)	-3.988*** (0.869)	0.025* (0.014)	0.017 (0.014)
or blind alg + avatars	1278	1278	1278	1254	1254
N	0.002	51.545	-2.762	0.414	0.415
Control mean					

This table reports treatment effects on perceived discrimination, future labor supply, and beliefs about the likelihood of being hired in future. The outcome in the first column is the main measure of perceived discrimination. The second and third outcomes concern beliefs about being hired in the future: the probability that workers think they will be hired in the future if they are evaluated by the same procedure as in the experiment (column 2) and the same probability minus the probability they assign if they will be evaluated by a cutoff rule in baseline quiz scores instead (column 3). In columns 4 and 5, the outcome is their reservation wage per paragraph to be evaluated again and potentially hired if they are evaluated by the same procedure as in the experiment (column 4) or the cutoff rule (column 5). In panel A, the sample is those evaluated by a manager and the presented coefficient is on an indicator for being in *either* the non-blind manager arm or the demographic-blind manager arm where workers see previous hires' avatars, race, and gender, rather than the demographic-blind manager arm that does not see previous hires' avatars, race and gender. Panel B is the analogous specification in the three algorithm arms. Appendix Figure A.26 shows results that do not pool the arms in this way. Both samples restrict to workers who would not have been hired under any procedure. The effect on perceived discrimination is different from Figure 1.4 because the sample does *not* restrict to workers who saw three white men previously hired. All regressions control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously promoted-workers each worker saw. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Figure A.26: Effects on beliefs, reservation wages, and affect in the hiring experiment, separately by treatment arm



Note: This figure plots the estimated treatment effect of being in each arm with positive rates of perceived discrimination in the hiring experiment. The outcomes, specifications, and samples are the same as in Table A.28—the only difference is that the effect of the two manager arms with positive perceived discrimination are estimated separately to the demographic-blind manager arm that does not see previous hires’ avatars, race, and gender, and analogously in the algorithm arms. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

A.9 Robustness analyses

A.9.1 Alternative inference methods

Randomization inference. Randomization inference is a standard method used to determine the statistical significance of an estimate without making assumptions about its asymptotic distribution (Fisher, 1935; Rosenbaum, 2002). Specifically, I re-assign treatment status using the same procedure as in the experiments 500 times, indexed by $k \in (1, 500)$, and for each iteration k , estimate the parameter of interest $\hat{\beta}^k$ using the re-assigned “treatment” with the same specification as the main analysis. The distribution of $\hat{\beta}^k$'s provides the empirical distribution of the treatment effect parameter under the null hypothesis that $\beta = 0$. The two-sided randomization inference p -value is the fraction of absolute values of $\hat{\beta}^k$'s that are greater than the absolute value of the observed $\hat{\beta}^{true}$ estimated using the real treatment assignment.

All of the main results and conclusions of the paper are unaffected by using randomization inference. Appendix Figure A.27 plots the p -values of the effects of assignment by a non-blind manager in the promotion experiment on all of the outcomes of interest. Of the outcomes for which the asymptotic standard errors suggest a statistically significant effect at a less than 5 percent level, all of the randomization-inference p -values remain below 0.06. The p -value on whether workers share with their manager (originally significant at the 10 percent level) is only marginally significant with a p -value of 0.16 when using randomization inference. Appendix Figure A.28 plots randomization-inference p -values for the *differences* in rates of perceived discrimination between the arms of the hiring experiment. The effects of blinding decision-makers to demographics and the effect of using an algorithm instead of a manager when the decision-maker is non-blind remain significant at the 5 percent level; the effect of using an algorithm instead of a manager when the decision-maker is demographic-blind remains significant at the 10 percent level. The effects of increasing minority-group representation retain p -values less than 0.02 (Appendix Figure A.29). The effects of learning that the demographic-blind algorithm previously promoted mostly white men on performance retain significance at the 10 percent level (Appendix Figure A.30).

Multiple hypothesis testing. All of the figures reference above also plot the corresponding asymptotic p -values when they are corrected for multiple hypothesis testing; all results except the effects of perceived algorithmic discrimination are unaffected. Specifically, I control the family-wise error rate using the Romano-Wolf (2005; 2016) correction with 500 bootstrap iterations. In Appendix Figure A.27 the four “families” are the groups in bold: retention, effort and performance in the first six paragraphs, future labor supply (including cooperation and sharing with managers), and beliefs about future promotion. The effects of being in the non-blind manager arm relative to the demographic-blind manager arm are generally robust; p -values change the most for the future labor supply outcomes, which in some cases are now only marginally significant (p -values around 0.16). Again, the effects of blinding decision-makers to demographics and the effect of using an algorithm instead of a manager when the decision-maker is non-blind remain significant at the 5 percent level; the effect of using an algorithm instead of a manager when the decision-maker is demographic-blind remains significant at the 10 percent level (Appendix Figure A.28). Appendix Figure A.29 does not include MHT-corrected p -values as there are only two estimates in the “family.” A simple Bonferroni-style correction that multiplied the asymptotic p -values by two would imply p -values that are still less than 0.02. The results that are not robust to using MHT-corrected inference are the effects of learning that the demographic-blind algorithm previously promoted mostly white men on performance; these p -values increase to more than 0.2 when they are adjusted.

A.9.2 Attrition and attention

Differential attrition and attention also cannot explain any of the results Appendix Table A.29, columns (1) and (2) shows that there is no differential attrition in the promotion experiment between the demographic-blind and non-blind manager arms or between workers in the algorithm sample that do or do not see demographics (Panels A and B). Panels C and D show that there is also no differential attrition between workers who saw three white men or two white men and someone else in the non-blind manager arm or in the demographic-blind algorithm arm when workers saw previously-promoted workers' demographics, respectively. Appendix Figure A.47 shows that there is similarly no differential attrition in the hiring experiment.

The same tables and figures show that there is no difference in the rate at which workers failed the attention check in the final part of the experimental survey, except between workers who saw three white men versus two white men and someone else in the algorithm + avatars arm of the promotion experiment (Appendix Table A.29, Panel D, column 3). Note that workers who failed the attention check are included in the main sample, and the results are robust to dropping them, as discussed next. Section 1.2.4 shows that worker attention to and comprehension of the experiment was high overall.

A.9.3 Specification choices

I test the sensitivity of the main estimates to the choice of control variables and sample definition by estimating the main results while iterating over all possible decision-points. Appendix Figures A.31-A.38 show the resulting estimates for the effects of being in the non-blind manager arm relative to the blind manager arm in the promotion experiment on rates of perceived discrimination and all of the main worker-behavior outcomes. Appendix Figures A.39 and A.40 test the robustness of the effects of anti-bias hiring policies on rates of perceived discrimination, Appendix Figure A.41 tests the robustness of the effects of minority-group representation, and A.42-A.46 test the robustness of the effects of perceived algorithmic discrimination.

A.9.3.1 Sample definitions

The main results restrict the sample to workers who were not selected by any of the procedures, as pre-registered. This eliminates concerns that different procedures may have selected different types of workers, generating differences between the treatment arms in the non-promoted or non-hired sample other than those caused by different perceptions of the reason they were not selected. The specification charts iterate over other sample definitions: further restricting to workers who passed the “attention check” in the final survey section or expanding to the full sample of workers who completed the experimental survey. The effects of perceived manager discrimination on behavior, effects of anti-bias hiring procedures on perceived discrimination, and effects of minority-group representation on perceived discrimination are generally unchanged by either of these changes. The effects of perceived algorithmic discrimination are again more sensitive.

Specifically, the effects of learning that the demographic-blind algorithm previously promoted mostly white men on earnings and performance are less negative and lose significance when restricting to those that passed the attention check (Appendix Figure A.45). Being in the non-blind manager arm also seems to negatively affect time spent per paragraph when restricting to these workers (Appendix Figure A.33).

A.9.3.2 Included controls

The main specification controls for demographic information (age, education, income, race, gender, family status), baseline quiz scores, quiz-score group fixed effects, and the educational experience and quiz-

score quintiles of the previously-promoted workers each worker saw.

A second specification (in the manager arms only) replaces the quiz-score group fixed effects with evaluation group fixed effects (groups within which workers were jointly assigned to be evaluated by the same counterfactual managers) but leaves the rest of the main specification controls unchanged. I pre-registered using evaluation-group fixed effects rather than quiz-score group fixed effects. I primarily present results using quiz-score group fixed effects because this allows for the same exact specification to be used for the manager and algorithm samples (the non-blind/blind manager randomization was within evaluation groups, whereas the with/without avatars algorithm randomization was across evaluation groups). As seen here, this does not matter for the results.

An alternate set of controls was selected by a double-post-lasso procedure, following [Chernozhukov et al. \(2018\)](#) as pre-registered. Specifically, the procedure was the following: I run a linear lasso model separately for each outcome and treatment-group indicators including all possible control variables as independent variables (about 350, each standardized to have mean zero and standard deviation one).¹ These models are run separately for those evaluated by a manager and those evaluated by an algorithm and separately for each sample definition in the previous subsection. For each outcome \times sample, the lasso-selected controls are the union of controls selected by lasso when predicting the outcome in that sample and when predicting treatment assignment in that sample. The baseline survey was shortened for the hiring experiment to minimize costs so I observe fewer covariates, and only carry out this exercise for the promotion experiment.

A fourth set of controls removes the controls for the education and performance of the previously-promoted workers, since these were not pre-registered. As seen here, this does not matter for the results. They are included in the main specification because there are significant differences in these variables across treatment arms. Finally, the figures also show estimates from regressions that include no control variables.

The estimated treatment effects of evaluation by a non-blind manager are very stable regardless of the choice of control variables, consistent with effective randomization and the lack of observable differences between the treatment arms. This is true for all four sets of results (the effects of perceived manager discrimination or perceived algorithmic discrimination on worker behavior or the effects of anti-bias hiring policies and minority-group representation on rates of perceived discrimination. Unsurprisingly, adding evaluation-group fixed effects tends to improve the precision of the estimated coefficient in the manager arms of the promotion experiment, and removing all controls does the opposite for all results.

A.9.4 Demand effects

A common critique of survey experiments is the potential for differential experimenter demand effects: if participants perceive that they are in an experiment, figure out the treatments, and guess the researcher's hypothesis, differences between treatment arms could be induced by a desire to help (or harm) the researcher's "agenda" ([de Quidt et al., 2019](#)). That said, recent evidence suggests that quantitative social scientists running survey experiments (and their readers) may be unnecessarily or overly concerned with bias from experimenter demand effects, relative to an older literature of lab experiments in psychology in

¹All controls in the main specification, plus: indicators for reporting having experienced gender or racial discrimination in past job search, promotion, termination, or daily work activities, second-order beliefs about those probabilities for each of eight race \times gender groups, indicators for being currently or ever employed outside of Prolific, job satisfaction with their current or most recent employer, whether they've ever been unemployed and looking for work, whether they've experienced an unemployment spell longer than six weeks, and scales of racial and gender identity centrality. The set of possible controls also includes 250 indicators for and interactions of the elements of the profiles that non-blind managers had access to when making decisions about who to assign to higher-status job: 11 self-identified race and ethnicity categories, gender, having a college degree or more than a college degree, quiz-score quintiles, five skin tones, eight hair colors, and 24 hairstyles, and interactions of race/ethnicity \times gender \times quintiles \times degree status as shown on the profiles.

which the phenomena was originally documented.²

Unsurprisingly, Appendix Table A.30 (promotion experiment) and Appendix Figure A.48 (hiring experiment) show that regardless of when they were asked, workers in the arms that made the possibility of discrimination salient were much more likely to say that they thought the study was about discrimination.³ In the promotion experiment 0-1 percent and 3-4 percent of the control (no salient demographics) and treatment (salient demographics) groups, respectively, know that the study is about perceived discrimination when asked after the proofreading task, and 0-4 and 4-13 percent say the same when asked after the reservation wage, willingness to pay, and dictator game elicitation. A randomly-selected half of the sample was asked at each of these times. All workers are asked again at the end of the study. Now, 35-40 percent of the control groups think the study was about discrimination, and being in a treatment arm where the possibility of discrimination is salient increased this by 5pp (se=4pp) and 10pp (se=2pp) when evaluated by the algorithm or a manager, respectively. This is likely due to the final questions in the survey that ask about perceptions of discrimination most explicitly. Among those evaluated by a non-blind manager, workers were 5-9pp (depending on when they were asked more likely to report that they thought the study was about discrimination when they saw three white men than when they saw two white men and someone else were previously promoted; workers in the demographic-blind algorithm arm that learned previously-promoted workers' demographics were 2-9pp more likely to do so.

In the hiring experiment, workers are asked what they thought was the topic of the study first after the reservation wage elicitation and again at the end of the survey. At both times, workers are more likely to think that the study was about discrimination in the arms in which there were positive rates of perceived discrimination (Appendix Figure A.48). At the first elicitation, 0-3 percent of workers in the control groups and 16-24 percent of workers in the groups where the possibility of discrimination was salient think the study was about discrimination. At the second elicitation, this was 8-18 percent and 30-40 percent, respectively.

Even when workers differentially think that the study was about discrimination in the treated groups, there still may not be differential demand effects if workers do not alter their behavior based on this knowledge. I estimate conservative bounds on all of the main results that account for the possibility of experimenter demand effects by replacing the value of each outcome variable with its value plus or minus 0.2 times the standard deviation of that variable in the demographic-blind manager arm if the worker guesses that the topic of the study was discrimination (for the outcomes measuring worker behavior) and, more conservatively, I replace the perceived discrimination outcome with 0 or 1 if the worker guesses that the topic of the study was discrimination.⁴ Appendix Figures A.49 and A.50 present the results for the effect of being in the non-blind manager arm relative to the demographic-blind manager arm on worker behavior with these bounds. By subtracting (adding) 0.2sd to the outcome, I obtain an estimate for the effect of treatment if participants who know the study topic have a higher (lower) value of the outcome than they would have if they did not know the topic. There are three upper and three lower bounds; these use the first, second,

²Mummolo and Peterson (2019) show that giving participants information about experimenter hypotheses does not affect participant behavior or estimated treatment effects, even when participants are financially incentivized to conform to those hypotheses and participants are more knowledgeable about the study purpose than participants in the “no information” condition. de Quidt et al. (2018) show that telling participants they will “do the researcher a favor” by choosing a higher or lower action than they otherwise would generates substantial bias in estimated treatment effects, and saying that the researcher “expects participants shown these instructions to [choose a higher or lower action] than they normally would” generates some, but much less, bias.

³These indicators were generated via coding workers' answers to the open-ended question, “*What do you think this study is about? It's fine if you're uncertain, please still write something down.*” The independent variable in these regressions is an indicator for workers' answers being about discrimination (Appendix A.5). In the promotion experiment, these were coded by external coders. Due to budget and time constraints, in the hiring experiment, I generated the coded variables from the text responses.

⁴de Quidt et al. (2018) show that a strong message telling workers to work harder or less hard changes outcomes by about 0.1sd in the incentivized effort task they test, though they are not powered to detect this small of an effect. They do not test effects on multiple price list elicitation like the ones I use to measure reservation wages, which are less cleanly incentivized. So, as a conservative bound, I use 0.2 as my multiplier on the standard deviation. I use an even more conservative bound on the perceived discrimination outcome which is not incentivized at all.

and third elicitation of beliefs about the study purpose to determine whose outcomes to replace (recall, the first and second elicitations were each asked of half of the sample). The treatment effect point-estimates are remarkably similar and my conclusions are unchanged, though some of the lower bounds are no longer statistically significant. Appendix Figures A.51 and A.52 shows that the effects of perceived algorithmic discrimination on performance are also robust to this adjustment.

Appendix Figures A.53 show the analogous bounds for the effects of minority-group representation on perceived discrimination rates, now replacing the perceived discrimination measure with 0 or 1 if a worker believed the study was about discrimination. The lower bound is about half the size of the main point estimate (but still statistically significant) and the upper bound is similar to the main point estimate. Thus, a more conservative interpretation would be that increasing minority-group representation from 0 to 1 of 3 cuts rates of discrimination by about 25 percent rather than about 50 percent.

In the hiring experiment, I estimate similar bounds on the differences between rates of perceived discrimination in each of the treatment arms, replacing the perceived discrimination indicator with a 0 or 1 if a worker thinks the study was about discrimination. I use knowledge of the study topic measured after the reservation wage elicitation (which was asked of all workers). The effect of blinding managers and algorithms to demographics are unchanged. The lower bound on the effects of using an algorithm rather than a manager fall to about half the magnitude and are no longer statistically significant (Appendix Figure A.54).

A.9.5 Multiple price list elicitations

The four measures of future labor supply were elicited using multiple price lists (MPL).⁵ These were reservation wages for future work when (1) one would be evaluated by the same manager as in the experiment and promoted or not, and (2) when baseline quiz scores would alone determine promotion, reservation wages to do a collaborative task with their manager, and willingness to pay to be able to choose a different manager to work with in the collaborative task. Appendix A.4 shows the exact implementation of each MPL.

Following the literature, these were incentivized by telling workers that I would randomly choose one row of the MPL and implement the worker's choice in that row for randomly-selected workers. That said, workers could report low reservation wages and not take the job later. This was uncommon and does not seem to matter for the results: only 10-15 percent of workers answered "I would want the job at this wage" in all rows for each of the three reservation wages, there is no treatment effect on whether workers answer in this way, and dropping these workers does not change the estimated treatment effects on reservation wages.

Some workers may not have understood or paid attention to the instructions for the MPL questions. In the main results, I follow the literature and drop any worker who does not respond to the MPL questions in a sensible way: these are workers who display multiple switching points throughout the list or those who switch in the wrong direction (i.e, report downward-sloping labor supply or upward-sloping demand for choosing their own manager). Appendix Table A.31 compares the workers who are dropped for this reason with other workers. These workers are less attentive and perform worse on the proofreading task than those who answer the MPL questions sensibly, but have similar education (Panel A). Consistently, the treatment effects on all other outcome variables are driven by the workers who *do* answer the MPL questions sensibly (Panel B), with the exception of whether and how much workers share with their manager.

A.9.6 Deviations from the Pre-Analysis Plan:

Exploiting random variation in the fraction of previously-promoted workers who are white men.

⁵Budget constraints prohibited me offering all workers a second round of work to measure effects on *observed* future labor supply.

I could not predict which workers managers or the algorithm would select out of twenty-four randomly-grouped historical-sample workers. I did not anticipate much variation in what fraction of previously-promoted workers were white men. However, there was substantial variation, with 40 percent of workers in the experimental sample seeing three white men being previously promoted and the rest seeing two white men and someone else. Workers were randomly paired with managers or groups of workers jointly evaluated by the algorithm, and indeed seeing that all previously-promoted workers are white men is not correlated with any observable worker characteristics (Appendix Table A.9). The documented differences in rates of perceived discrimination between workers who saw three white men previously promoted and those who saw two white men and someone else have important implications and thus are included in the paper despite not being pre-registered. There were also differences across arms in what fraction of workers who saw three white men; Appendix Figure 1.5 controls for this difference and allows for comparisons across evaluation procedures conditional on whether workers saw two or three white men previously promoted.

Given these results, a second related deviation from the pre-analysis plan was to focus the analysis in the hiring experiment on the subsample of workers who see only three white men previously promoted. All of the main exhibits are duplicated with the full sample in the appendix and all secondary analysis is available upon request in the full sample.

Change in sample composition. The demographic make-up of the sample also deviates from the pre-analysis plan. I managed to over-sample racial minorities relative to the population but not to the extent originally planned. I also exhausted a larger fraction of the white men interested in my study in recruiting the historical sample than anticipated, and so I recruited fewer white men to the experimental sample than intended. Thus, a much larger fraction of the sample is white women than planned. That said, I am still able to test for racial and gender heterogeneity and obtain moderately precise results.

Change from pre-registered sample size. First, I recruited 120 fewer workers to the screening survey for the promotion experiment than pre-registered. This was mainly because of difficulty recruiting the 75 managers, since they were required to be white men (the pool of which had been largely exhausted in recruiting the historical sample). Second, in order to recruit as many racial minority participants as I did, I had to raise the base wage paid for the screening survey for the last part of the sample. Stopping recruitment to the screening survey early helped make up this unexpected expense in the budget. All of this occurred before any workers were randomly assigned to treatment groups or took the experimental survey.

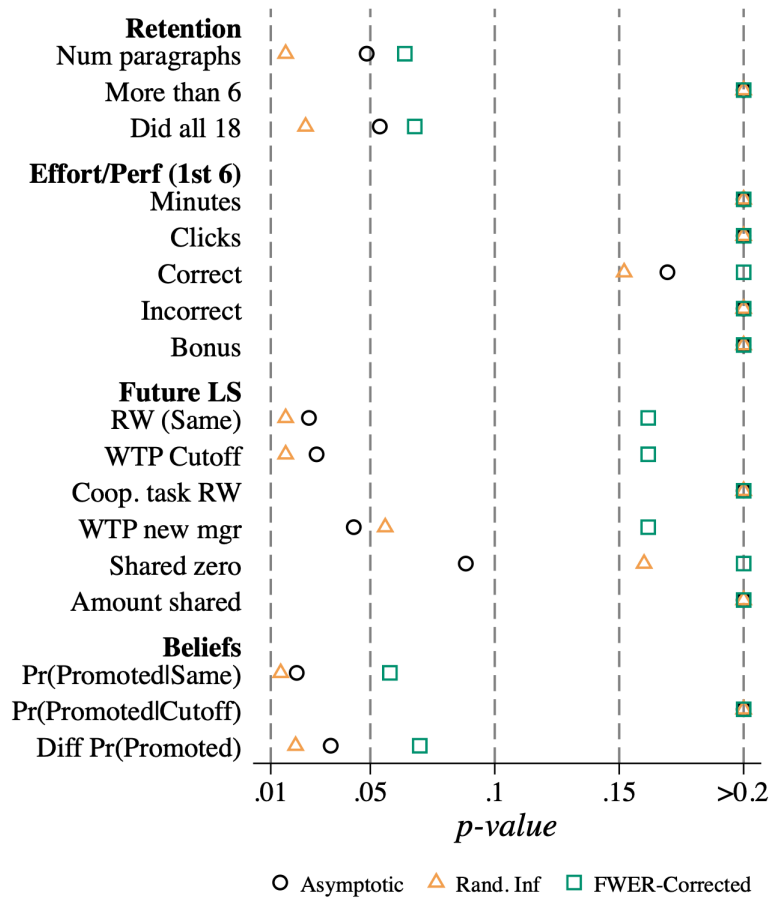
Second, in both the promotion and hiring experiments, one group of workers (one evaluation group of 120 (360) in the promotion (hiring) experiment) was not offered the experimental survey after completing the screening survey because I exhausted all possible managers that could be brought back to evaluate them and none did. Since they had no data on counterfactual decision-making by that manager, they would not have been able to be in the main analysis sample. Since treatment was assigned within evaluation groups, this had no implications for the analysis. Again, this decision was made before any workers were assigned a treatment status. It also helped to make up the above-mentioned slight budgetary shortfall.

Added controls for previously selected workers' education and baseline performance. In the control groups, workers saw the education and baseline performance of the workers that their manager or the algorithm had selected in the historical sample. Comparing the treated groups, which also saw previously selected workers' race, gender, and avatar, to the control group nets out the effect of knowing that one was not selected and that previous workers with similar quiz scores as you were previously selected. However, the different evaluation procedures selected workers in the historical sample with systematically different education and baseline performance. To account for these differences—for example, to make sure that the effects are driven by perceived racial and gender discrimination rather than feeling mis-evaluated because one did not have a college degree—I added controls for the education and baseline performance of the three previously-selected workers to the baseline specification. As discussed in Section A.9.3 however, this does not affect the main conclusions. Estimates without these controls are in Appendix Figures A.39-A.46.

De-emphasis of IV specification. Section 1.3.1 and Appendix A.6 discusses the challenges associated with interpreting the IV specification, despite it being pre-registered, and the appendix presents all of the results for the pre-registered specifications.

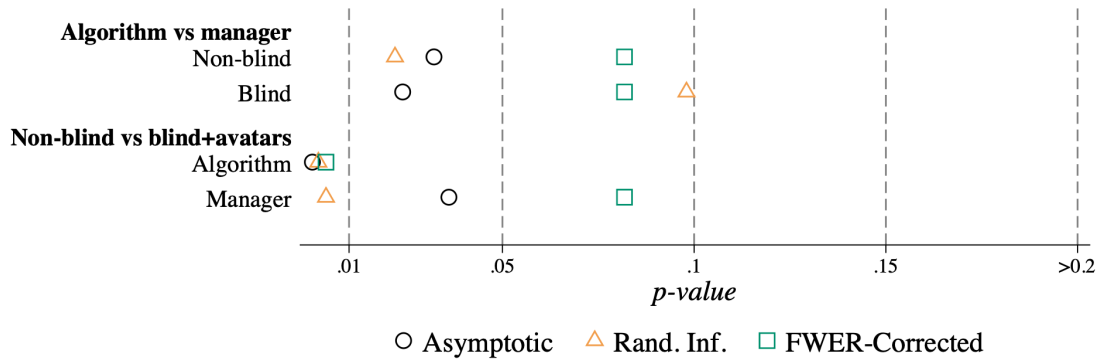
Implicit measures of perceived discrimination. The paper focuses on one primary measure of perceived discrimination and shows results for the other pre-registered measures of explicit perceived discrimination, but does not show results for the pre-registered measures of *implicit* perceived discrimination. These are differences in the number of stars (quiz-score quintiles) that workers report thinking they would have needed to be promoted compared to members of different demographic groups. The signs of the treatment effects on these variables are consistent with the explicit measures, but generally imprecise. Due to space constraints, these are available upon request.

Figure A.27: Randomization inference and multiple hypothesis testing (effects of perceived discrimination)



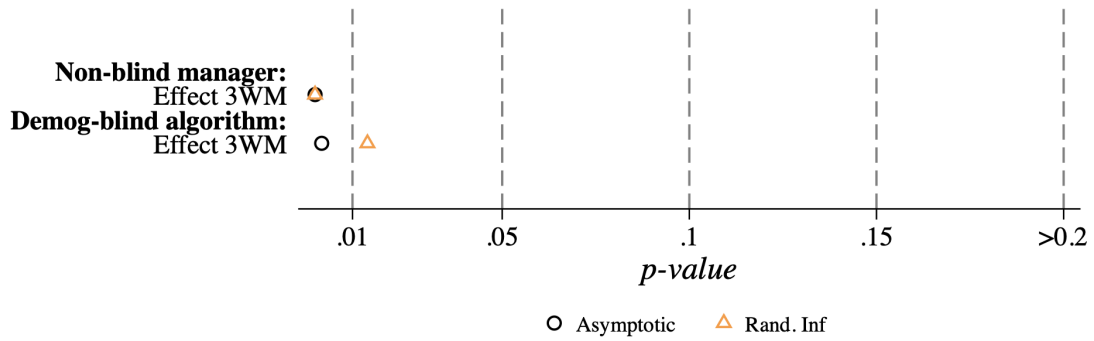
Note: This figure plots the p -values for the effects of being in the non-blind manager group relative to the demographic-blind manager group in the promotion experiment on the key outcomes of interest for the specifications in Tables 1.2, 1.3, 1.4, 1.5, and 1.6, with the p -values calculated asymptotically (as in the main analysis, they are robust to heteroskedasticity), with randomization inference, and corrected for multiple hypothesis testing using the Romano-Wolf correction.

Figure A.28: Randomization inference and multiple hypothesis testing (hiring experiment)



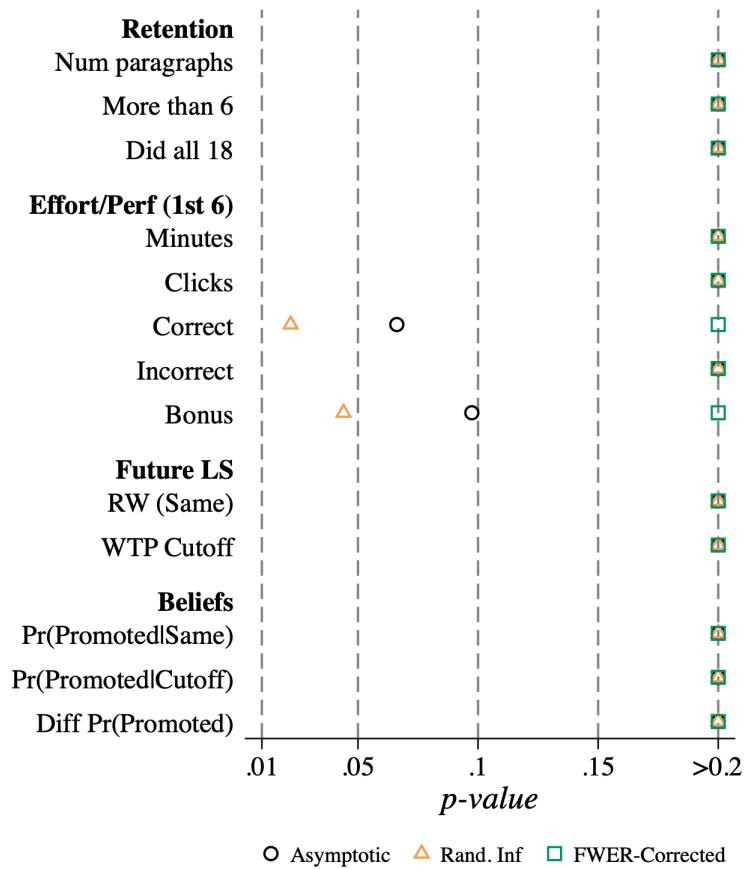
Note: This figure plots the p -values for the differences in perceived discrimination between treatment arms in Figure 1.4, with the p -values calculated asymptotically (as in the main analysis, they are robust to heteroskedasticity), with randomization inference, and corrected for multiple hypothesis testing using the Romano-Wolf correction.

Figure A.29: Randomization inference and multiple hypothesis testing (historical representation)



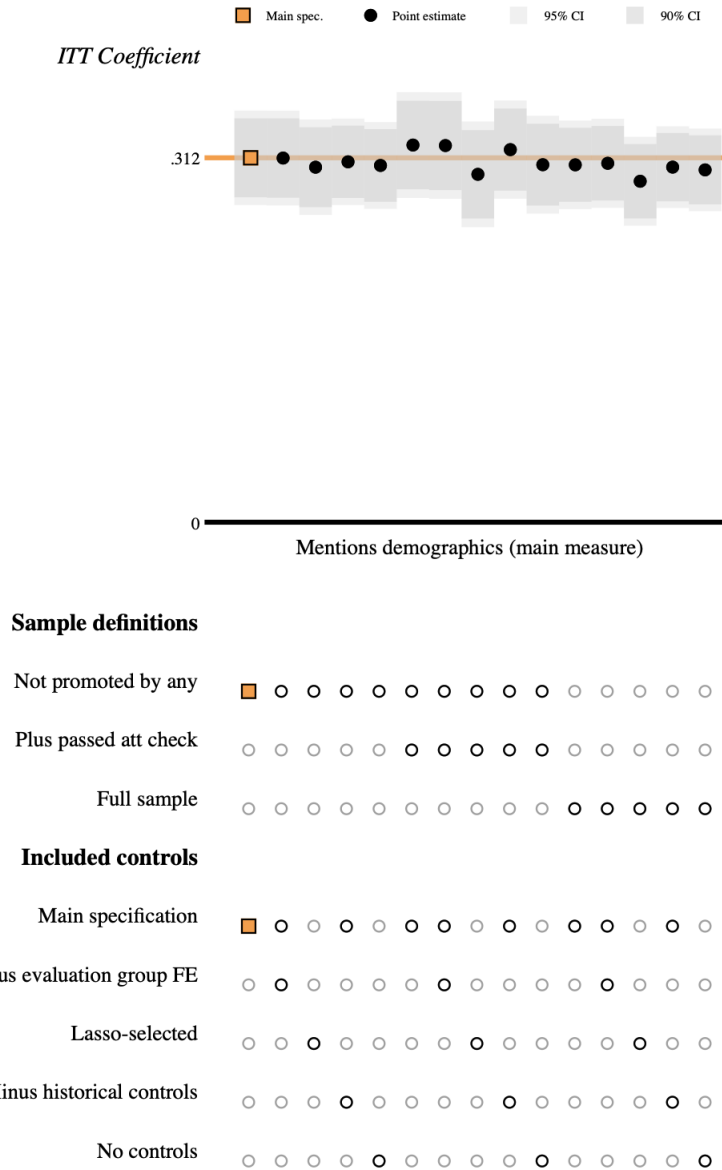
Note: This figure plots the p -values for the differences in perceived discrimination between treatment arms in Figure 1.4, with the p -values calculated asymptotically (as in the main analysis, they are robust to heteroskedasticity) and with randomization inference.

Figure A.30: Randomization inference and multiple hypothesis testing (perceived alg. discrimination)



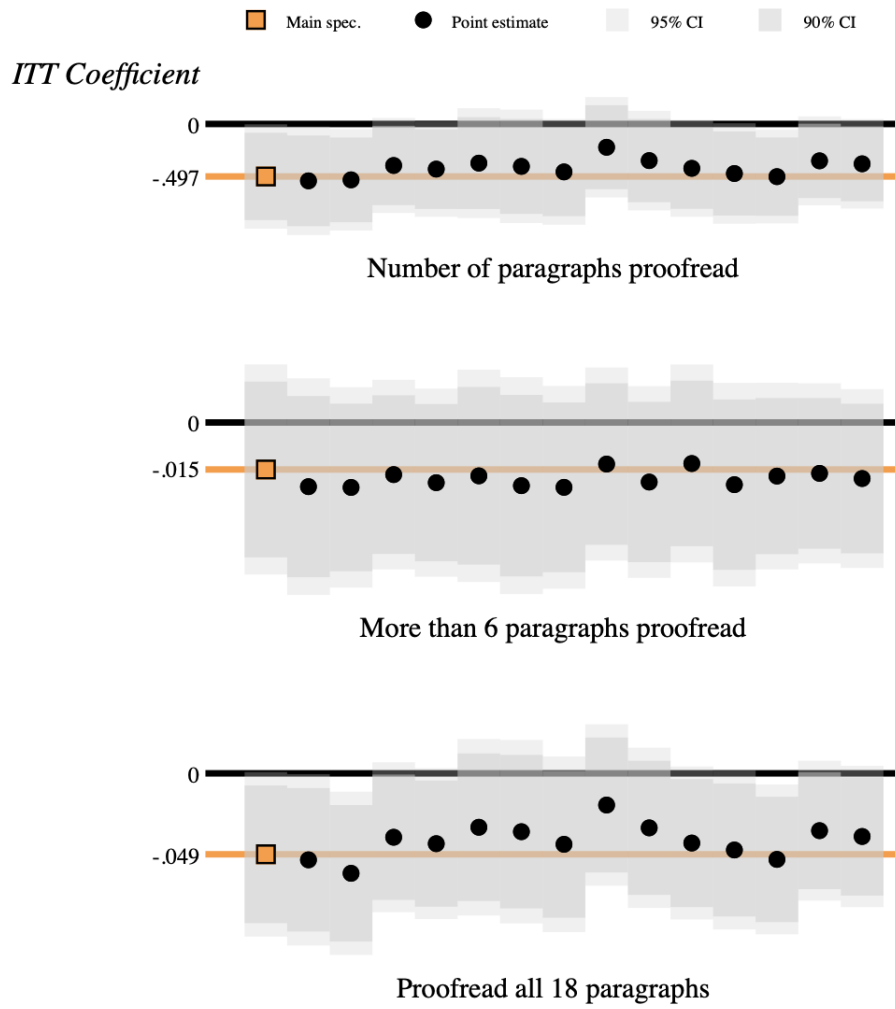
Note: This figure plots the p -values for the effects of learning that the demographic-blind algorithm previously promoted mostly white men in the promotion experiment on the key outcomes of interest for the specifications in Figure 1.7, with the p -values calculated asymptotically (as in the main analysis, they are robust to heteroskedasticity), with randomization inference, and corrected for multiple hypothesis testing using the Romano-Wolf correction.

Figure A.31: Alternative specifications: Effect of being in non-blind manager arm on perceived discrimination



Note: This figure plots the treatment effect of being in the non-blind manager group on the main measure of perceived discrimination in the promotion experiment, relative to the demographic-blind manager group. The figure iterates over sample restrictions and choices of control variables, starting with the main specification. Starting with the main sample definition (workers who would not have been promoted under any procedure), it then additionally restricts to workers who passed an attention check near the end of the survey, then expands to the full sample. The controls begin with the main specification, then add evaluation group fixed effects (the groups of 120 who were counterfactually evaluated by the same managers). Then, I remove the controls for the composition of previous hires' education and performance, as they were not pre-registered. Next, I include double-post lasso-selected controls, and then drop all control variables. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

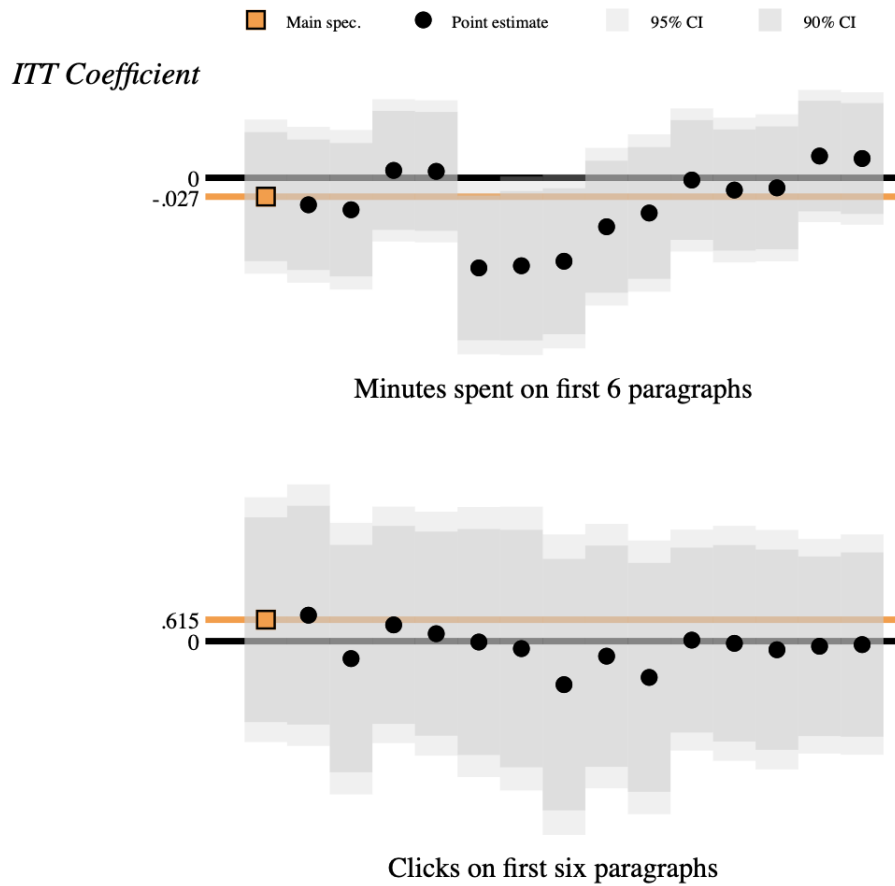
Figure A.32: Alternative specifications: Effect of being in non-blind manager arm on retention



Sample definitions	
Not promoted by any	■ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Plus passed att check	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Full sample	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Included controls	
Main specification	■ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Plus evaluation group FE	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Lasso-selected	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Minus historical controls	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
No controls	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Note: This figure plots the treatment effect of being in the non-blind manager group on retention in the promotion experiment, relative to the demographic-blind manager group. The outcomes and main specification are those in Table 1.2. Otherwise, the figure is analogous to Appendix Figure A.31.

Figure A.33: Alternative specifications: Effect of being in non-blind manager arm on effort



Sample definitions

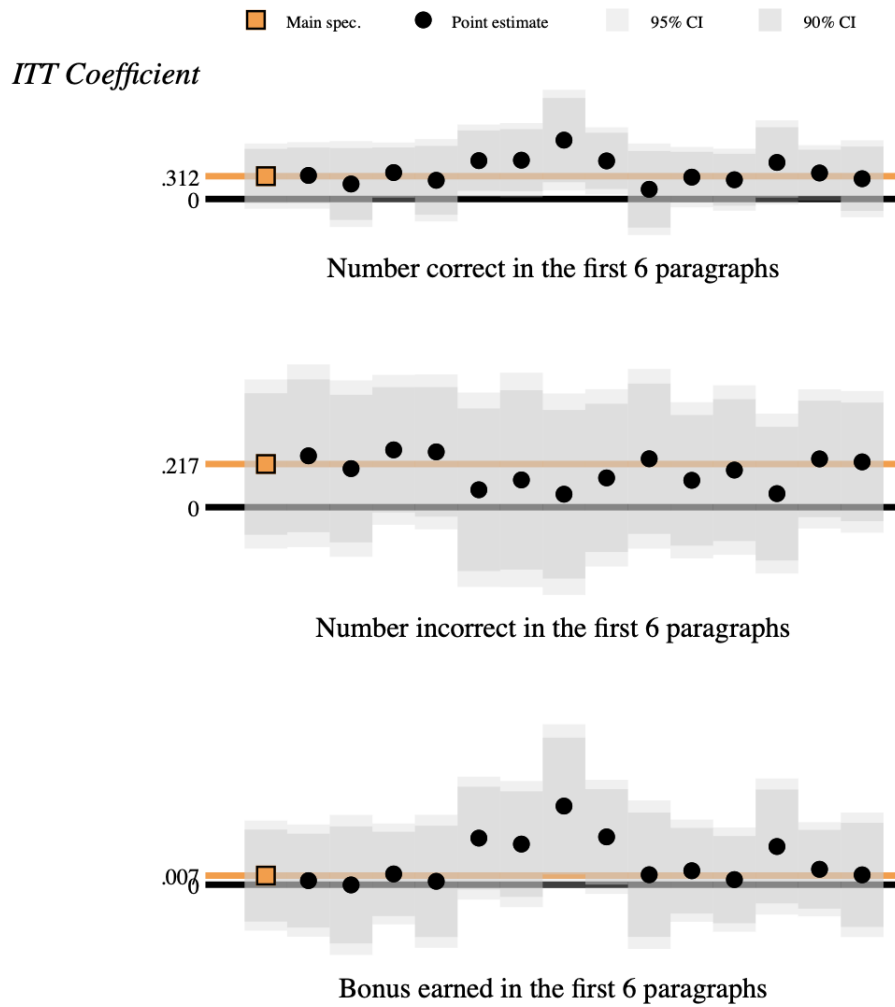
Not promoted by any	■	○	○	○	○	○	○	○	○	○	○	○	○	○
Plus passed att check	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Full sample	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Included controls

Main specification	■	○	○	○	○	○	○	○	○	○	○	○	○	○
Plus evaluation group FE	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Lasso-selected	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Minus historical controls	○	○	○	○	○	○	○	○	○	○	○	○	○	○
No controls	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Note: This figure plots the treatment effect of being in the non-blind manager group on effort in the promotion experiment, relative to the demographic-blind manager group. The outcomes and main specification are those in columns 1 and 2 of Table 1.3. Otherwise, the figure is analogous to Appendix Figure A.31.

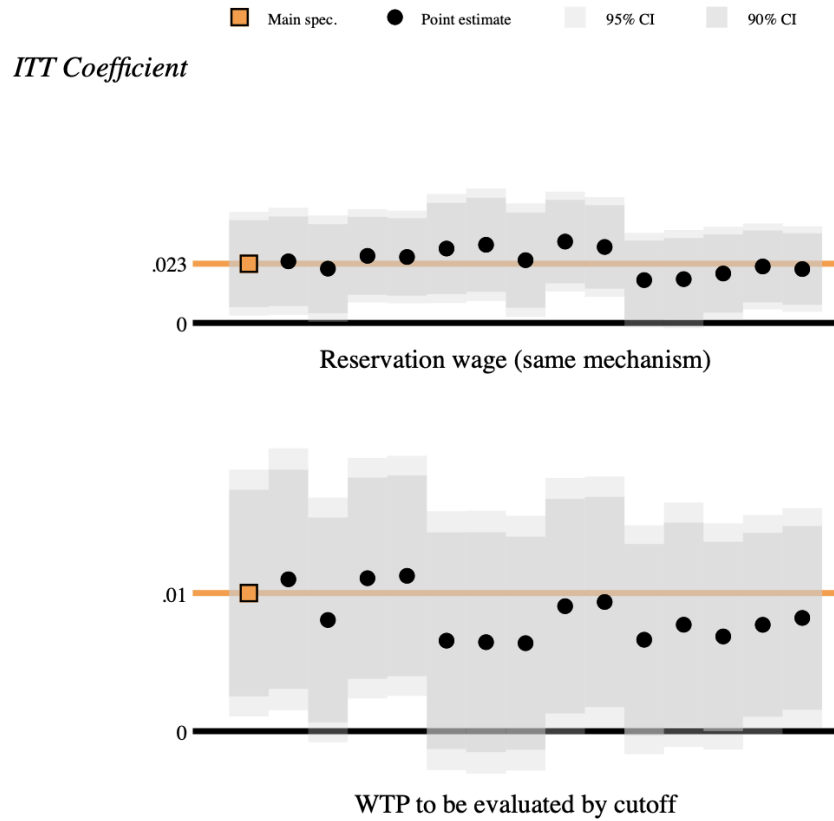
Figure A.34: Alternative specifications: Effect of being in non-blind manager arm on performance



Sample definitions	
Not promoted by any	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Plus passed att check	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Full sample	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Included controls	
Main specification	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Plus evaluation group FE	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Lasso-selected	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Minus historical controls	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
No controls	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Note: This figure plots the treatment effect of being in the non-blind manager group on performance in the promotion experiment, relative to the demographic-blind manager group. The outcomes and main specification are those in columns 3-5 of Table 1.3. Otherwise, the figure is analogous to Appendix Figure A.31.

Figure A.35: Alternative specifications: Effect of being in non-blind manager arm on future labor supply for the same job



Sample definitions

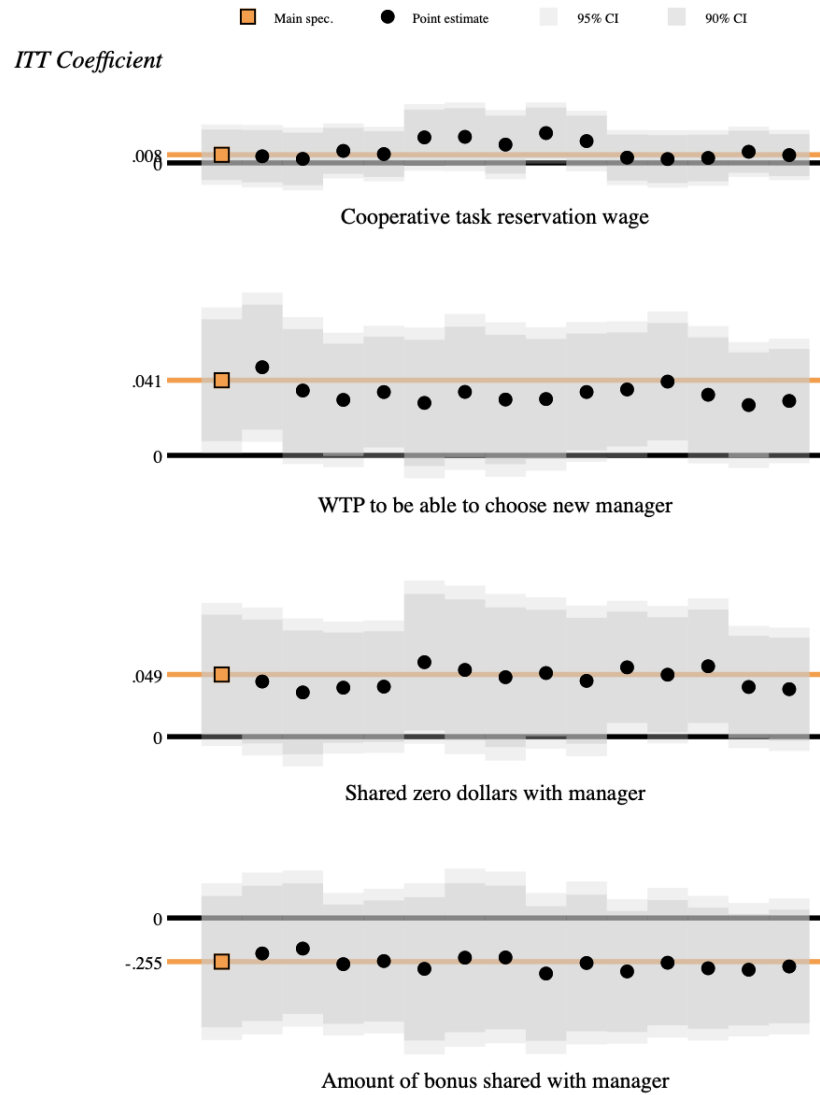
Not promoted by any	■	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Plus passed att check	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Full sample	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Included controls

Main specification	■	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Plus evaluation group FE	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Lasso-selected	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Minus historical controls	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
No controls	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Note: This figure plots the treatment effect of being in the non-blind manager group on reservation wages in the promotion experiment, relative to the demographic-blind manager group. The outcomes and main specification are those in Table 1.4. Otherwise, the figure is analogous to Appendix Figure A.31.

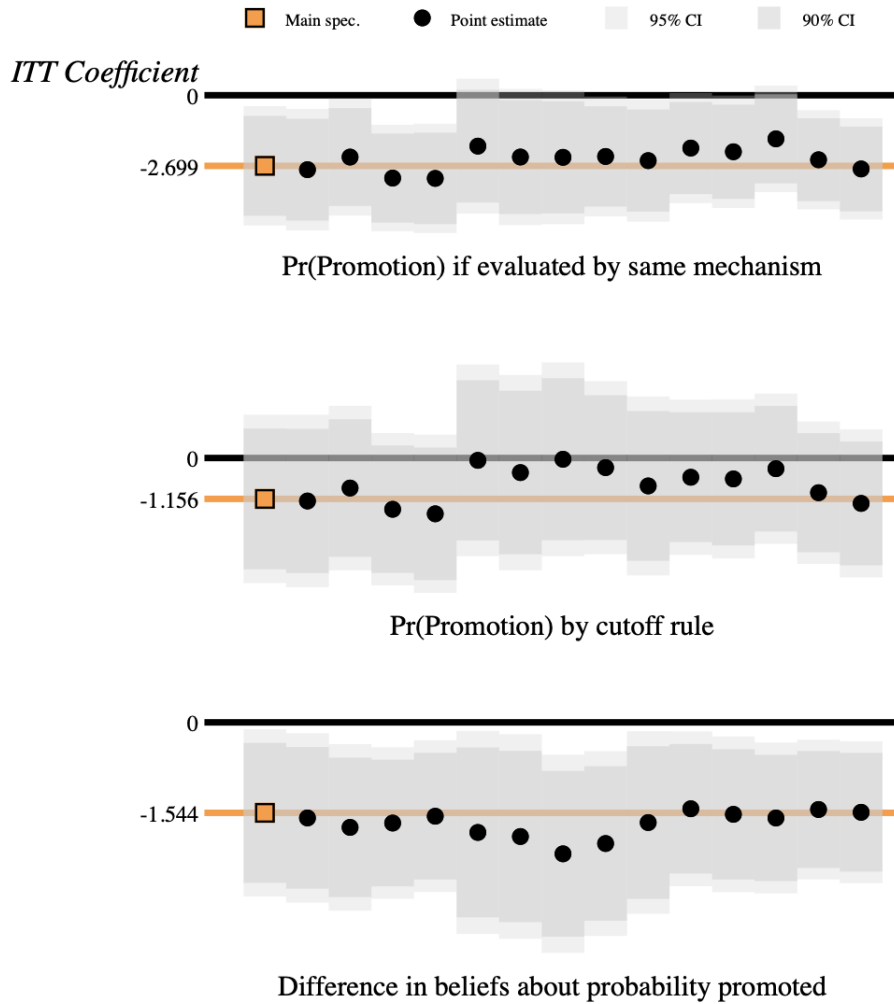
Figure A.36: Alternative specifications: Effect of being in non-blind manager arm on future labor supply (cooperative task) and generosity



Sample definitions														
Not promoted by any	■	○	○	○	○	○	○	○	○	○	○	○	○	○
Plus passed att check	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Full sample	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Included controls														
Main specification	■	○	○	○	○	○	○	○	○	○	○	○	○	○
Plus evaluation group FE	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Lasso-selected	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Minus historical controls	○	○	○	○	○	○	○	○	○	○	○	○	○	○
No controls	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Note: This figure plots the treatment effect of being in the non-blind manager group on willingness to interact and share with managers in the promotion experiment, relative to the demographic-blind manager group. The outcomes and main specification are those in Table 1.5. Otherwise, the figure is analogous to Appendix Figure A.31.

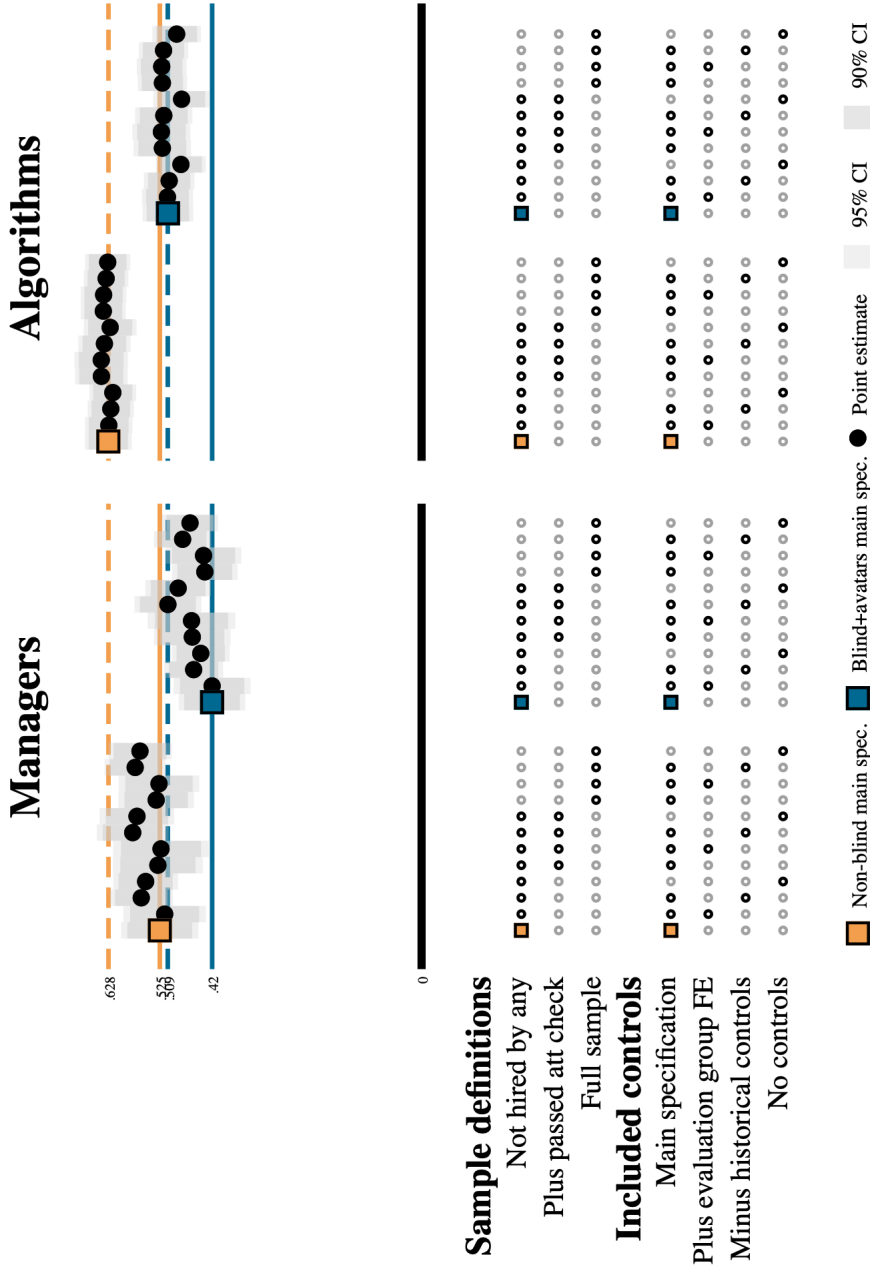
Figure A.37: Alternative specifications: Effect of being in non-blind manager arm on beliefs about promotion



Sample definitions	
Not promoted by any	■ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Plus passed att check	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Full sample	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Included controls	
Main specification	■ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Plus evaluation group FE	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Lasso-selected	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Minus historical controls	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
No controls	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

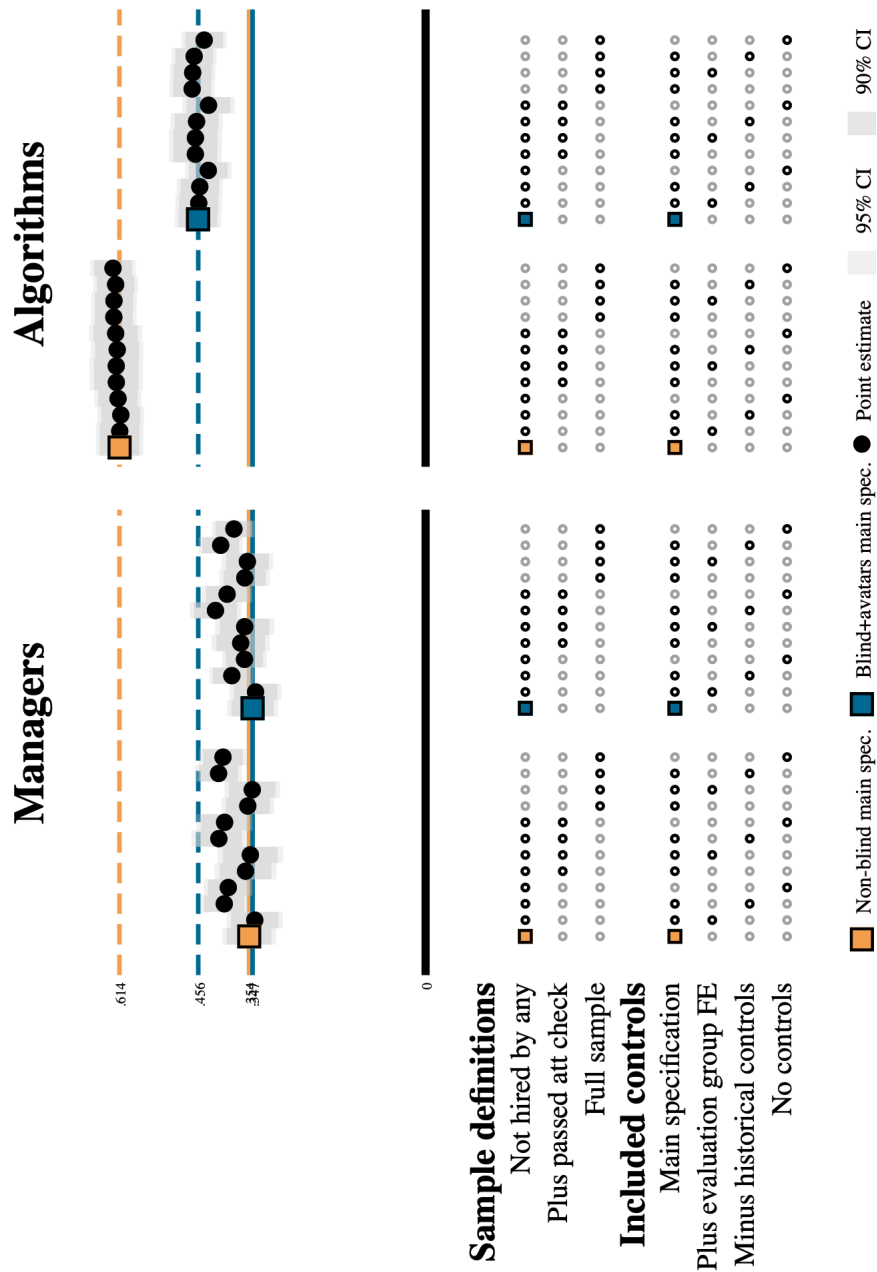
Note: This figure plots the treatment effect of being in the non-blind manager group on beliefs about promotion in the promotion experiment, relative to the demographic-blind manager group. The outcomes and main specification are those in Table 1.6. Otherwise, the figure is analogous to Appendix Figure A.31.

Figure A.39: Alternative specifications: Hiring experiment (if previous hires are all white men)



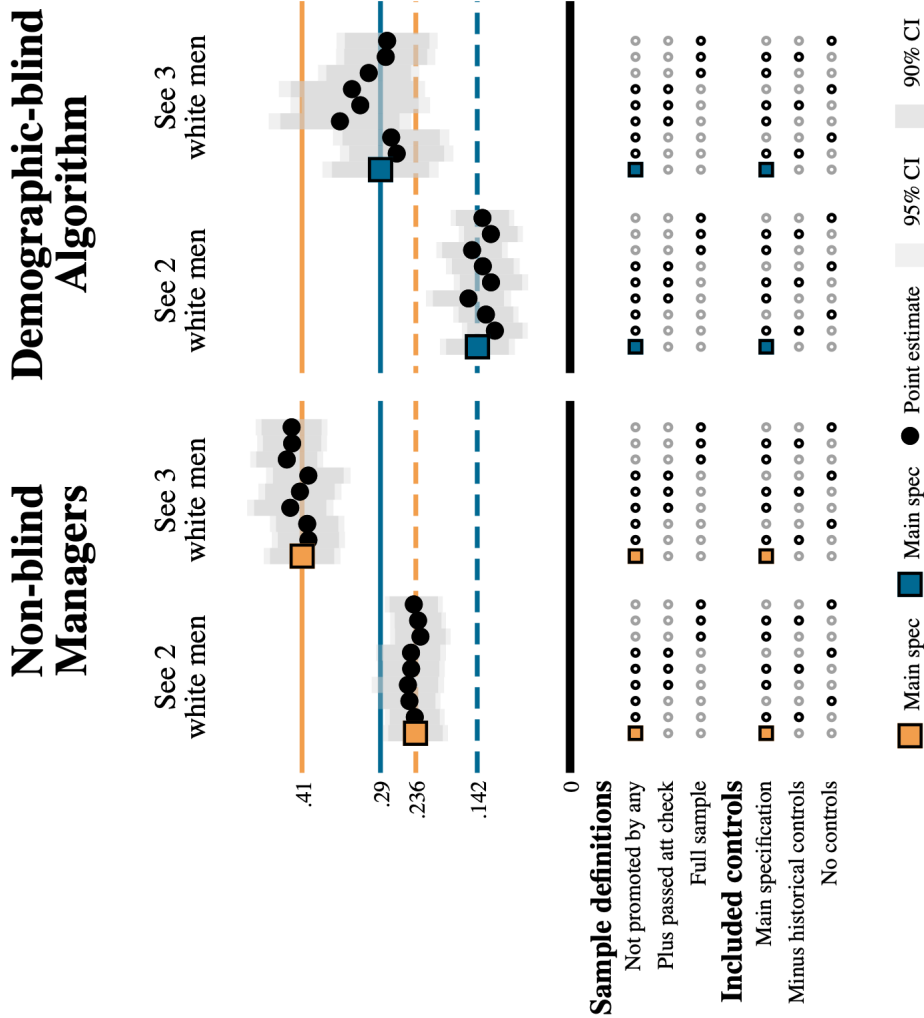
Note: This figure plots the estimated share of workers perceiving discrimination by the main measure in the hiring experiment. It parallels Figure 1.4, with the sample restricted to workers who saw only white men being previously hired. The figure iterates over other sample restrictions and choices of control variables, starting with the main specification. Starting with the main sample definition (workers who would not have been hired under any procedure), it then additionally restricts to workers who passed an attention check near the end of the survey, then expands to the full sample. The controls begin with the main specification, then add evaluation group fixed effects (the groups of 120 who were counterfactually evaluated by the same managers). Then, I remove the controls for the composition of previous hires' education and performance, as they were not pre-registered. Next, I drop all control variables. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.40: Alternative specifications: Hiring experiment (regardless of demographic makeup of previous hires)



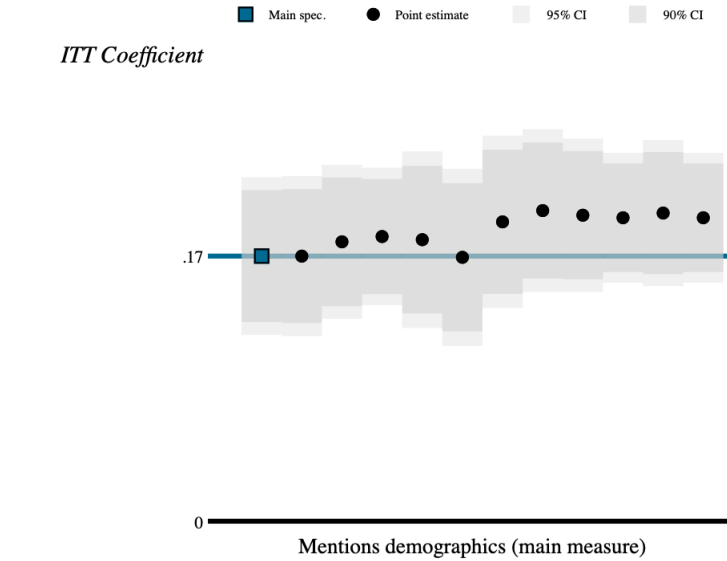
Note: This figure plots the estimated share of workers perceiving discrimination by the main measure in the hiring experiment, paralleling Appendix Figure A.23 in using the full sample (not accounting for differences in the share of previous hires who are white men). Otherwise, it is analogous to Appendix Figure A.39.

Figure A.41: Alternative specifications: Effects of minority-group representation



Note: This figure plots the estimated share of workers perceiving discrimination by the main measure in the promotion experiment. It parallels Figure 1.5, plotting only the share perceiving discrimination in the non-blind manager arm and the demographic-blind algorithm arm in which workers saw previously-promoted workers' demographics (though these rates are estimated using the same specification as Figure 1.5). The figure iterates over other sample restrictions and choices of control variables, starting with the main specification. Starting with the main sample definition (workers who would not have been hired under any procedure), it then additionally restricts to workers who passed an attention check near the end of the survey, then expands to the full sample. The controls begin with the main specification, then, I remove the controls for the composition of previous hires' education and performance, as they were not pre-registered. Next, I drop all control variables. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.42: Alternative specifications: Effects of seeing avatars on perceived algorithmic discrimination (promotion experiment)



Sample definitions

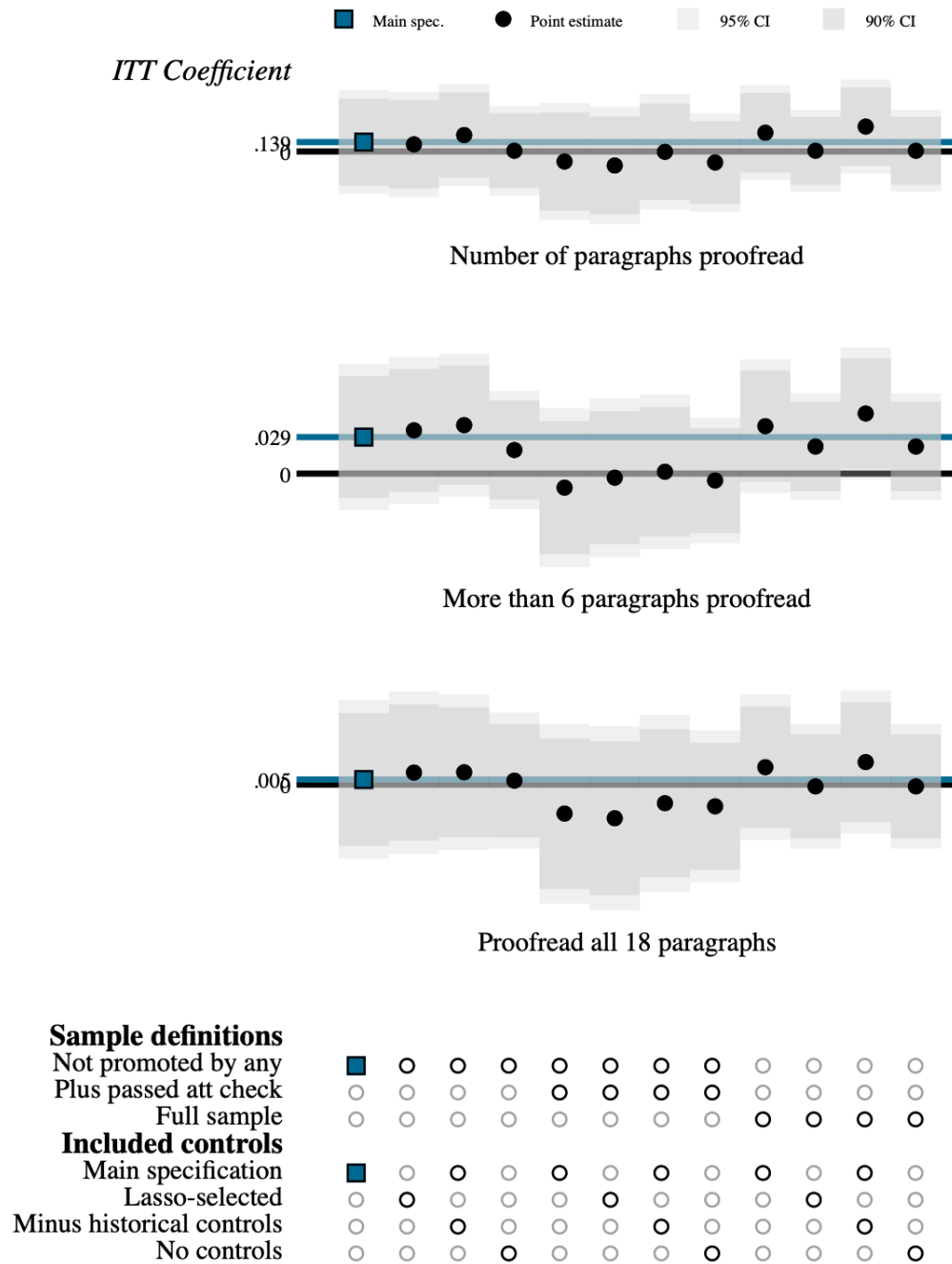
Not promoted by any	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Plus passed att check	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Full sample	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Included controls

Main specification	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Lasso-selected	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Minus historical controls	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
No controls	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

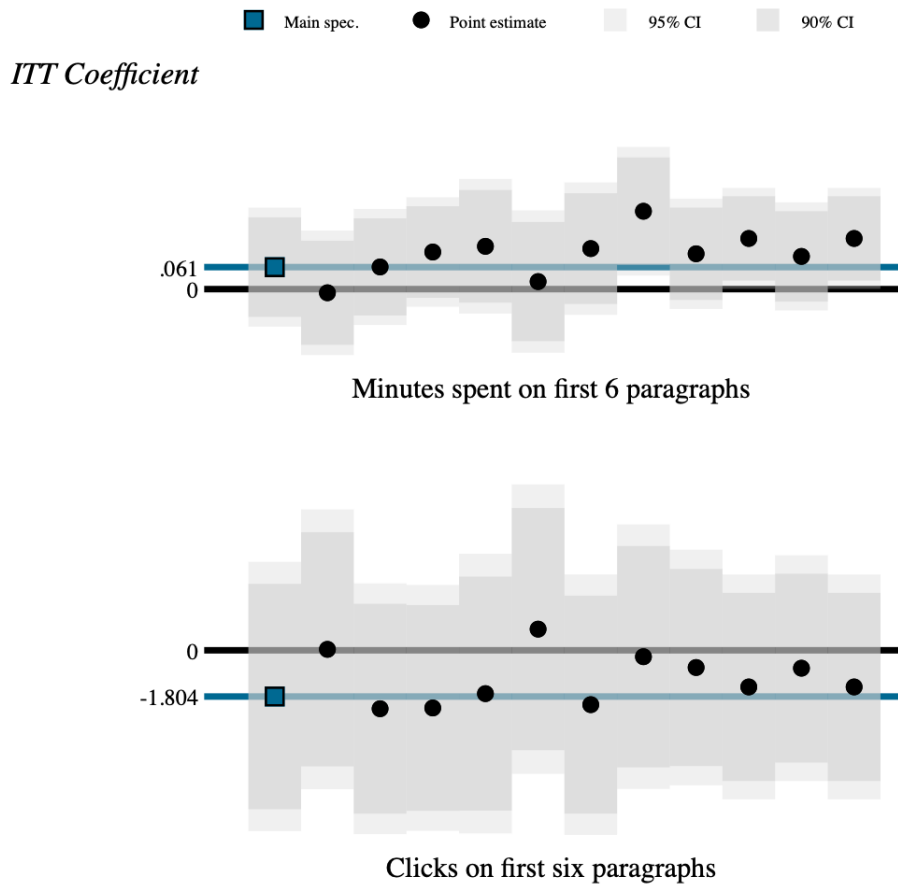
Note: This figure plots the treatment effect of seeing that the demographic-blind algorithm previously promoted mostly white men on perceived discrimination in the promotion experiment, relative to the other workers in the demographic-blind algorithm group who do not see the avatars, race, and gender of previously-promoted workers. Because the randomization within the algorithm arm was across rather than within evaluation groups, this chart does not include a specification that adds evaluation group fixed effects. Otherwise, the figure is analogous to Appendix Figure A.31.

Figure A.43: Alternative specifications: Effects of seeing avatars in the algorithm arm on retention



Note: This figure plots the treatment effect of seeing previously-promoted workers' avatars on retention in the avatar arm of the promotion experiment. The outcomes and main specification are those in Table 1.2. Otherwise, the figure is analogous to Appendix Figure A.42.

Figure A.44: Alternative specifications: Effects of seeing avatars in the algorithm arm on effort



Sample definitions

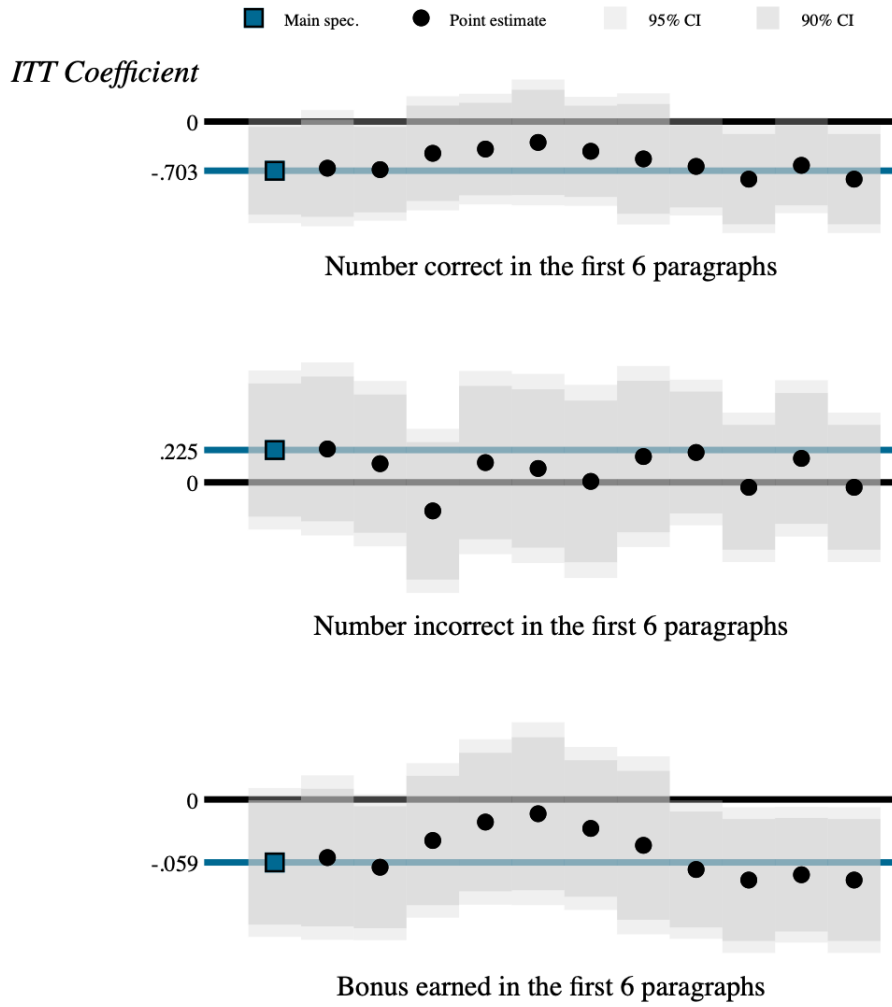
Not promoted by any	■	○	○	○	○	○	○	○	○	○	○
Plus passed att check	○	○	○	○	○	○	○	○	○	○	○
Full sample	○	○	○	○	○	○	○	○	○	○	○

Included controls

Main specification	■	○	○	○	○	○	○	○	○	○	○
Lasso-selected	○	○	○	○	○	○	○	○	○	○	○
Minus historical controls	○	○	○	○	○	○	○	○	○	○	○
No controls	○	○	○	○	○	○	○	○	○	○	○

Note: This figure plots the treatment effect of seeing previously-promoted workers' avatars on effort in the avatar arm of the promotion experiment. The outcomes and main specification are those in columns 1 and 2 of Table 1.3. Otherwise, the figure is analogous to Appendix Figure A.42.

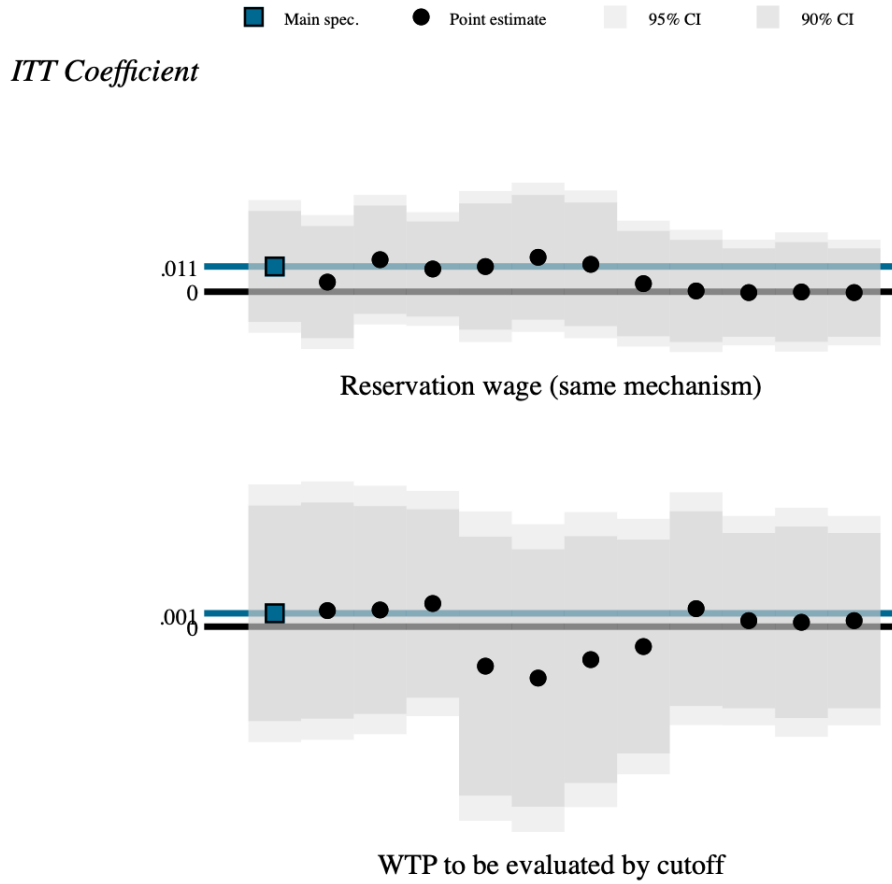
Figure A.45: Alternative specifications: Effects of seeing avatars in the algorithm arm on performance



Sample definitions											
Not promoted by any	■	○	○	○	○	○	○	○	○	○	○
Plus passed att check	○	○	○	○	○	○	○	○	○	○	○
Full sample	○	○	○	○	○	○	○	○	○	○	○
Included controls											
Main specification	■	○	○	○	○	○	○	○	○	○	○
Lasso-selected	○	○	○	○	○	○	○	○	○	○	○
Minus historical controls	○	○	○	○	○	○	○	○	○	○	○
No controls	○	○	○	○	○	○	○	○	○	○	○

Note: This figure plots the treatment effect of seeing previously-promoted workers' avatars on performance in the avatar arm of the promotion experiment. The outcomes and main specification are those in columns 3-5 of Table 1.3. Otherwise, the figure is analogous to Appendix Figure A.42.

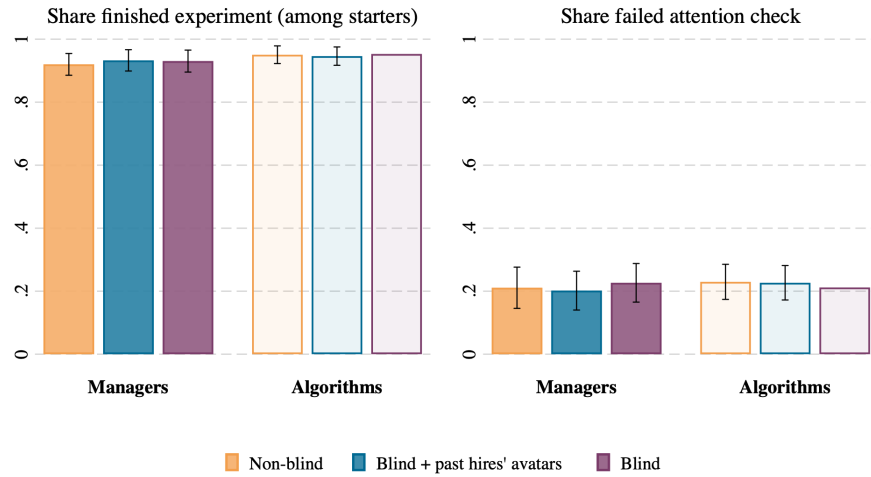
Figure A.46: Alternative specifications: Effects of seeing avatars in the algorithm arm on future labor supply



Sample definitions											
Not promoted by any	■	○	○	○	○	○	○	○	○	○	○
Plus passed att check	○	○	○	○	○	○	○	○	○	○	○
Full sample	○	○	○	○	○	○	○	○	○	○	○
Included controls											
Main specification	■	○	○	○	○	○	○	○	○	○	○
Lasso-selected	○	○	○	○	○	○	○	○	○	○	○
Minus historical controls	○	○	○	○	○	○	○	○	○	○	○
No controls	○	○	○	○	○	○	○	○	○	○	○

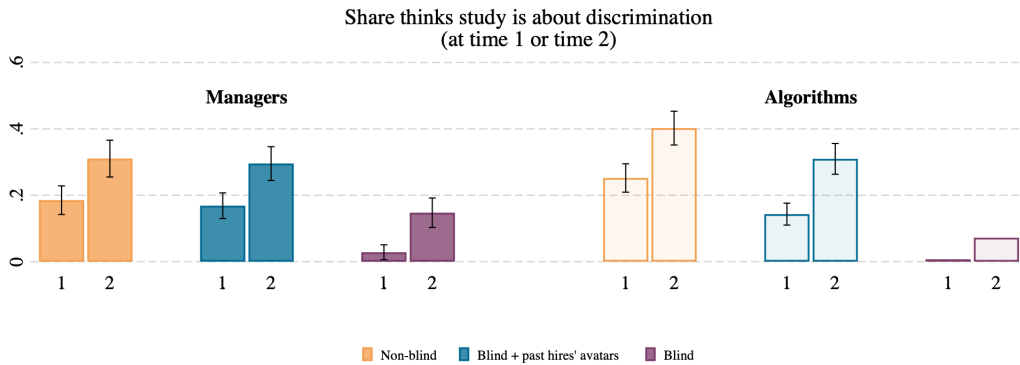
Note: This figure plots the treatment effect of seeing previously-promoted workers' avatars on reservation wages in the avatar arm of the promotion experiment. The outcomes and main specification are those in Table 1.4. Otherwise, the figure is analogous to Appendix Figure A.42.

Figure A.47: Attrition and attention in the hiring experiment



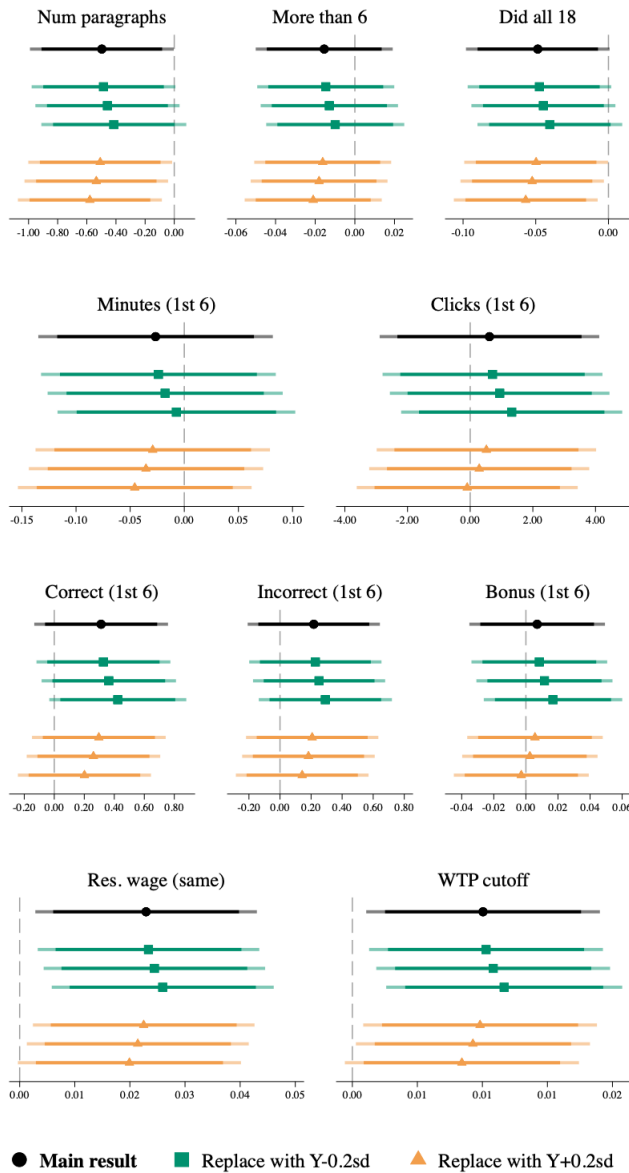
Note: This figure shows the fraction of workers in each arm of the hiring experiment who finished the experiment, among those who started it (on the left) and the share of the main sample that failed the attention check (on the right). The sample is restricted to workers who would not have been promoted under any hiring procedure, and workers who saw that all three previous hires were white men (a random subsample within each treatment arm—results are the same for the full sample). Shares are calculated via regressions that control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

Figure A.48: Knowledge of study topic in the hiring experiment



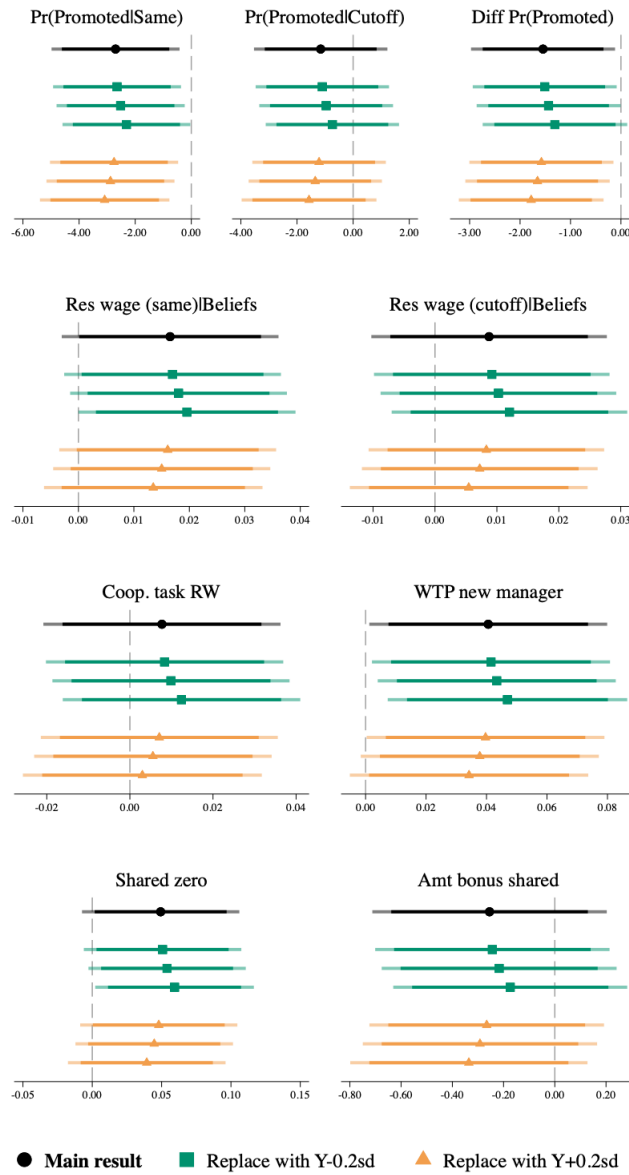
Note: This figure shows the fraction of workers in each arm of the hiring experiment who thought that the study was about discrimination the first time they were asked (after learning that they weren't hired and the reservation wage elicitation, before any survey questions) and the second (at the end of the survey). The sample is restricted to workers who would not have been promoted under any hiring procedure, and workers who saw that all three previous hires were white men (a random subsample within each treatment arm—results are the same for the full sample). Shares are calculated via regressions that control for quiz scores, education, income, age, marital and parental status, race, gender, quiz-score group fixed effects, and the educational and previous-performance composition of the previously-promoted workers each worker saw. 95 percent confidence intervals (in black bars) are calculated with standard errors robust to heteroskedasticity.

Figure A.49: Accounting for possible demand effects in the manager arms of the promotion experiment: effort and future labor supply



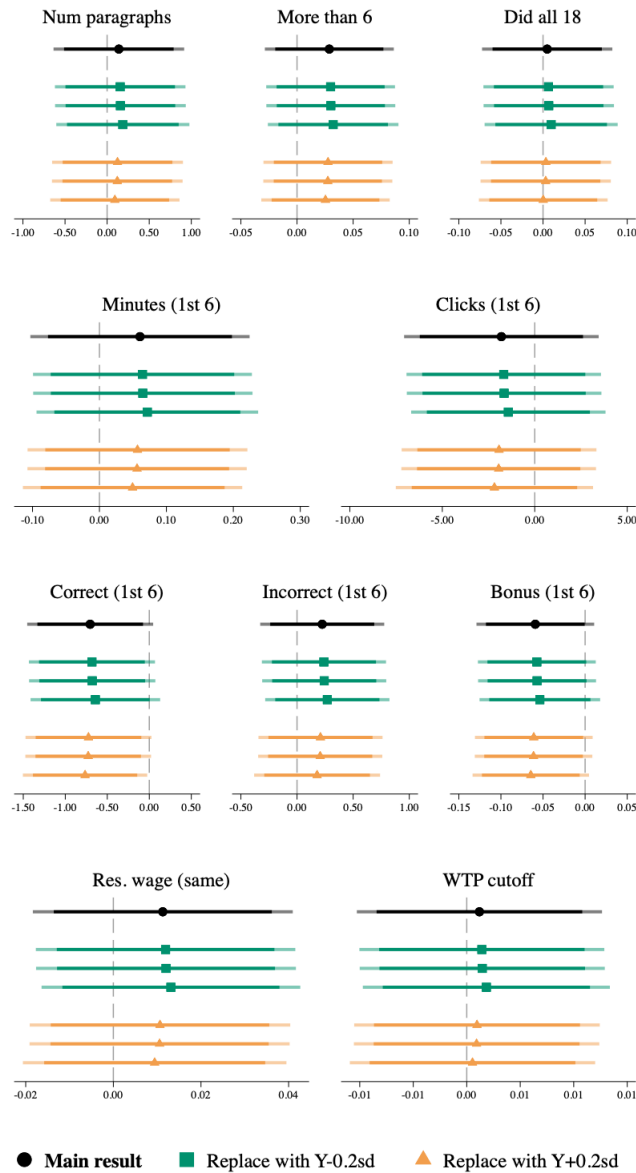
Note: This figure plots the treatment effects of being in the non-blind manager arm in the promotion experiment, relative to the demographic-blind manager arm, for the outcomes in Tables 1.2, 1.3, and 1.4, with adjustments to account for possible experimenter demand effects. These bounds replace the outcome variable with its value plus or minus 0.2 standard deviations of that variable in the demographic-blind manager arm if workers thought that the study was about discrimination when they were asked after the proofreading task, willingness to pay elicitation, and at the end of the study (in the first, second, and third estimate for each outcome). The treatment effects estimated using these alternate outcome variables are plotted below the main estimate. This accounts for workers who knew the study topic potentially changing their answer to the question about perceived discrimination based on what they thought the researcher “wanted” to hear, differentially by treatment group. The specifications are otherwise identical to the main tables. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.50: Accounting for possible demand effects in the manager arms of the promotion experiment: secondary outcomes



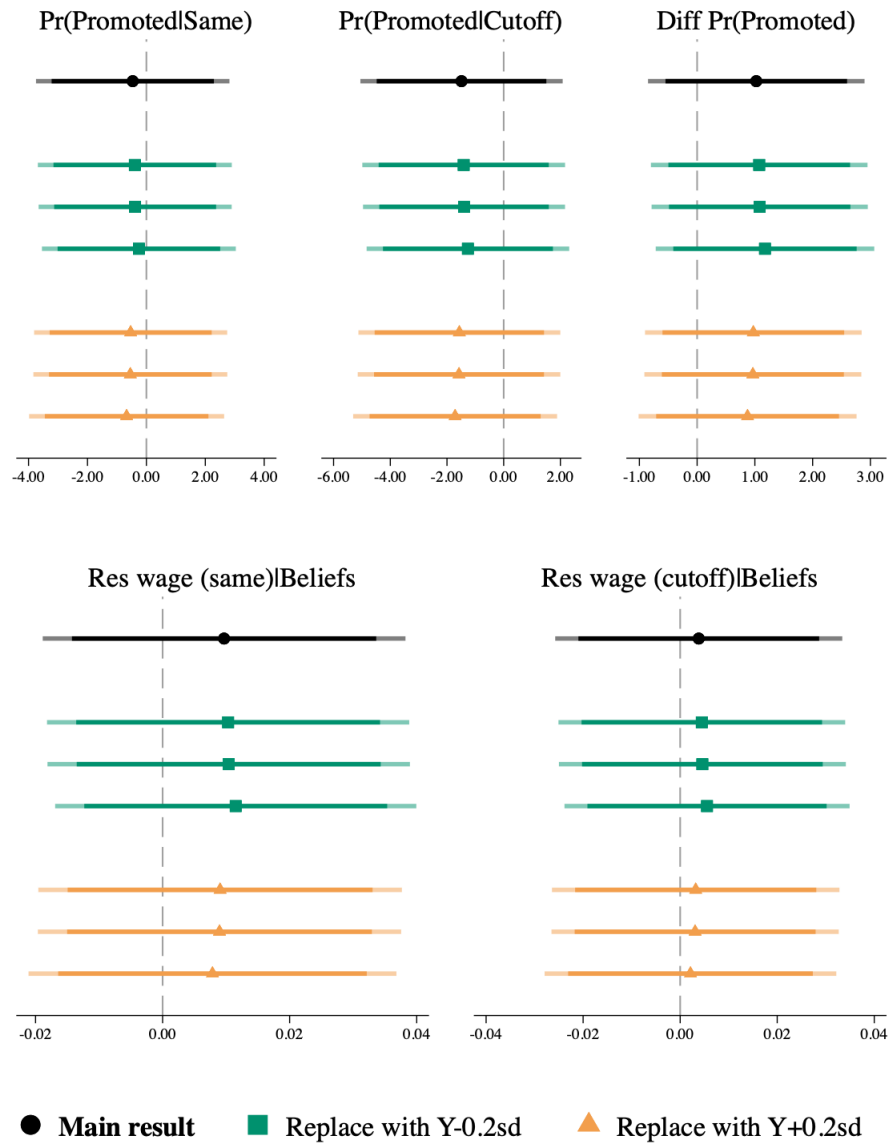
Note: This figure plots the treatment effects of being in the non-blind manager arm in the promotion experiment, relative to the demographic-blind manager arm, for the outcomes in Tables 1.6 and 1.5, and Figure 1.3, with adjustments to account for possible experimenter demand effects. The figure is otherwise analogous to Appendix Figure A.49.

Figure A.51: Accounting for possible demand effects in the algorithm arm of the promotion experiment: effort and future labor supply



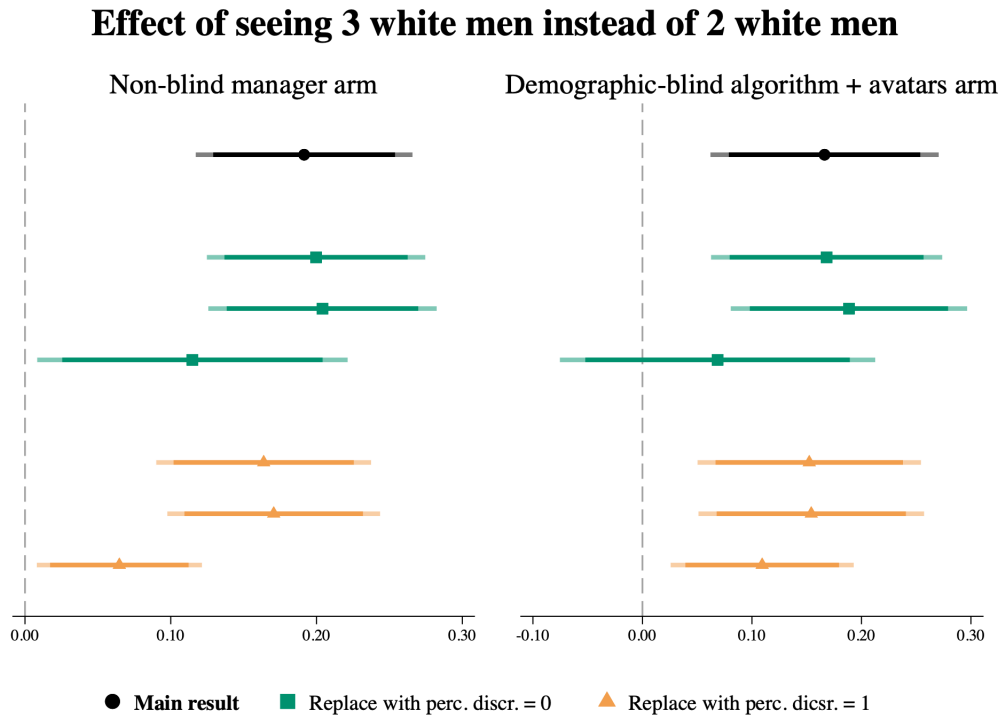
Note: This figure plots the treatment effects of learning the demographics of previously-promoted workers in the promotion experiment among workers evaluated by the demographic-blind algorithm with adjustments to account for possible experimenter demand effects. These bounds replace the outcome variable with its value plus or minus 0.2 standard deviations of that variable (for workers in the algorithm arm who do not learn the demographics of previously-promoted workers) if workers thought that the study was about discrimination when they were asked after the proofreading task, willingness to pay elicitation, and at the end of the study (plotted first, second, and third, respectively). The treatment effects estimated using these alternate outcome variables are plotted below the main estimate. This accounts for workers who knew the study topic potentially changing their answer to the question about perceived discrimination based on what they thought the researcher “wanted” to hear, differentially by treatment group. The specifications are otherwise identical to the main tables. The three estimates for each replacement outcome variable use workers’ ideas of the study topic at its first, second, and third elicitation, respectively. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.52: Accounting for possible demand effects in the algorithm arm of the promotion experiment: secondary outcomes



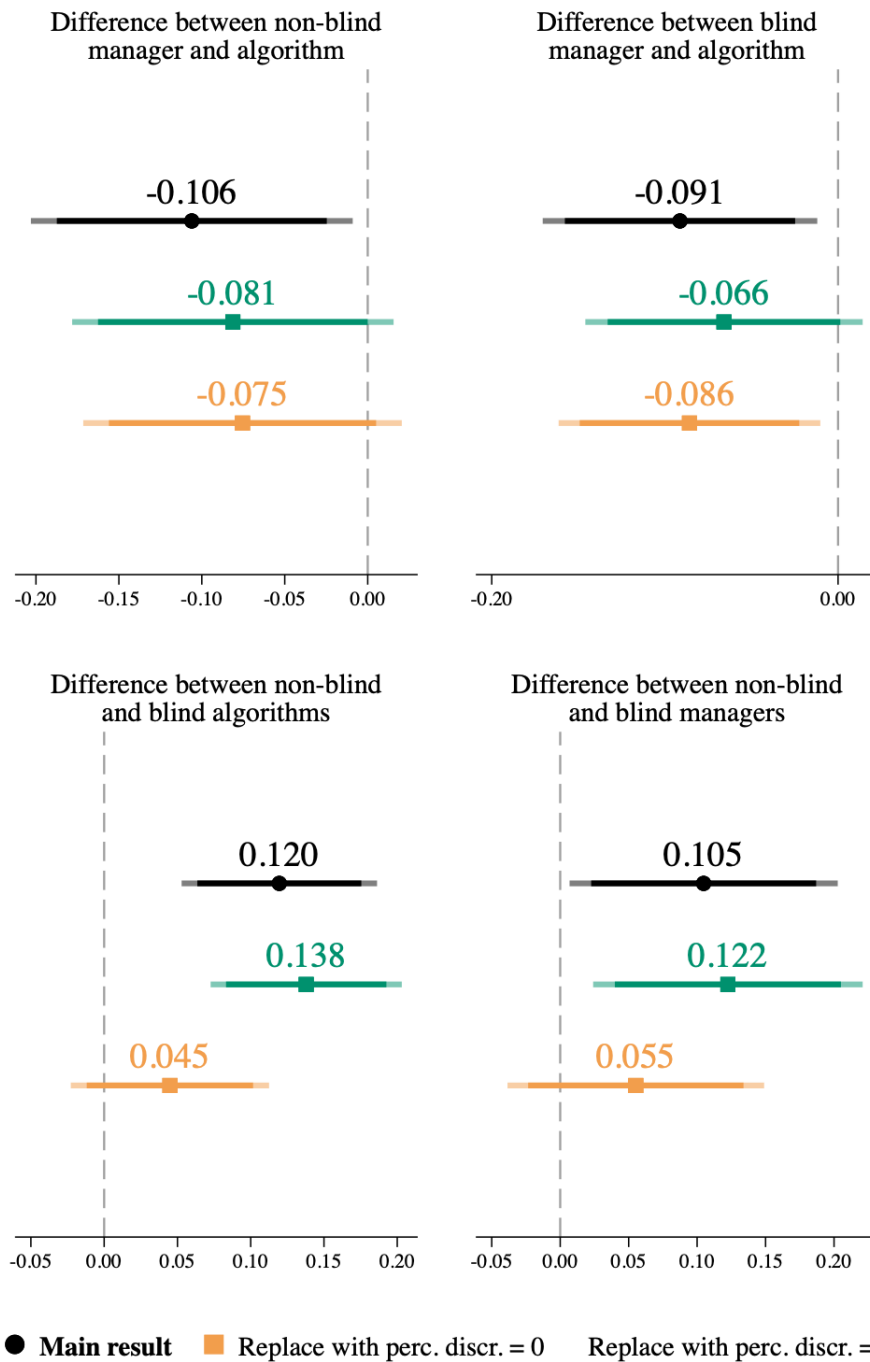
Note: This figure plots the treatment effects of learning the demographics of previously-promoted workers in the promotion experiment among workers evaluated by the demographic-blind algorithm with adjustments to account for possible experimenter demand effects, analogous to Appendix Figure A.51 for the secondary outcomes in Figure 1.7.

Figure A.53: Accounting for possible demand effects among workers who saw three versus two white men previously promoted



Note: This figure plots the effects of seeing that three white men were previously promoted from Figure 1.5, with adjustments to account for possible experimenter demand effects. These bounds replace the outcome variable (perceived discrimination) with a zero or one if workers thought that the study was about discrimination when they were asked after the proofreading task, willingness to pay elicitation, and at the end of the study (plotted first, second, and third, respectively). This accounts for workers who knew the study topic potentially changing their answer to the question about perceived discrimination based on what they thought the researcher “wanted” to hear, differentially by whether they saw that two or three white men were previously promoted. The specifications are otherwise identical to Figure 1.5. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Figure A.54: Accounting for possible demand effects in the hiring experiment



Note: This figure plots the differences in perceived discrimination between treatment arms in Figure 1.4, with adjustments to account for possible experimenter demand effects. These bounds replace the outcome variable (perceived discrimination) with a zero or one if workers thought that the study was about discrimination when they were asked just after the elicitation of the outcome variable and the reservation wage elicitation. The differences between treatment arms estimated using these alternate outcome variables are plotted below the main estimate. This accounts for workers who knew the study topic potentially changing their answer to the question about perceived discrimination based on what they thought the researcher “wanted” to hear, differentially by treatment group. The specifications are otherwise identical to Figure 1.4. 90 and 95 percent confidence intervals are calculated with standard errors robust to heteroskedasticity.

Table A.29: Attrition

	Starts experiment (baseline sample)	Finishes experiment (among starters)	Failed survey attention check
	(1)	(2)	(3)
<i>Panel A: Manager arms of the promotion experiment</i>			
Non-blind manager	-0.011 (0.018)	-0.002 (0.012)	-0.003 (0.024)
N	1735	1450	1389
Control mean	0.837	0.953	0.239
<i>Panel B: Algorithm arms of the promotion experiment</i>			
See avatars	-0.027 (0.028)	0.033* (0.018)	-0.044 (0.040)
N	859	727	691
Control mean	0.860	0.932	0.275
<i>Panel C: Non-blind manager arm of the promotion experiment</i>			
See three white men	-0.036 (0.026)	0.012 (0.014)	0.015 (0.037)
N	865	722	695
Mean, see two white men	0.851	0.954	0.223
<i>Panel D: Algorithm+avatars arm of the promotion experiment</i>			
See three white men	-0.040 (0.044)	-0.020 (0.024)	0.156** (0.062)
N	396	329	320
Mean, see two white men	0.844	0.977	0.213

Note: This table tests for differential attrition in the promotion experiment. In column 1, the sample is all workers who took the baseline survey and would not have been promoted under any procedure and the outcome is an indicator for starting the experimental survey. In column 2, the sample is all workers who start the experimental survey and would not have been promoted under any procedure, and the outcome is an indicator for finishing the experiment. In the third column, the sample is the main analysis sample of workers who finished the experimental survey and would not have been promoted under any procedure, and the outcome is an indicator for failing the attention check question near the end of the survey. Controls are the same as in all main analysis. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.30: Knowledge of study topic

	After proofreading	After WTP elicitations	End of survey
	(1)	(2)	(3)
<i>Panel A: Manager arms of the promotion experiment</i>			
Non-blind manager	0.030*** (0.011)	0.085*** (0.022)	0.101*** (0.029)
N	697	692	1389
Control mean	0.011	0.038	0.399
<i>Panel B: Algorithm arms of the promotion experiment</i>			
See avatars	0.039** (0.019)	0.042** (0.016)	0.053 (0.042)
N	353	338	691
Control mean	0.000	0.000	0.353
<i>Panel C: Non-blind manager arm of the promotion experiment</i>			
See three white men	0.065*** (0.023)	0.053 (0.047)	0.086* (0.046)
N	344	351	695
Mean, see two white men	0.005	0.132	0.449
<i>Panel D: Algorithm+avatars arm of the promotion experiment</i>			
See three white men	0.022 (0.051)	0.090** (0.045)	-0.009 (0.067)
N	161	159	320
Mean, see two white men	0.018	0.038	0.389

Note: This table tests for differential knowledge of the study topic in the promotion experiment. In column 1, the outcome is an indicator for thinking that the study is about discrimination after the proofreading job (before the reservation wage elicitation and other survey questions) in the random half of the sample that is asked at that time; in column 2, it is a similar indicator but measured after the reservation wage elicitation, in the other half of the sample; in column 3, it is an indicator for thinking the study is about discrimination at the end of the study (asked of the whole sample). Controls are the same as in all main analysis. Standard errors, in parentheses, are robust to heteroskedasticity. Significance at the 0.1, 0.05, and 0.01 levels indicated by *, **, and ***, respectively.

Table A.31: Comparing workers with sensical and non-sensical MPL elicitations

	Dropped from any MPL elicitation	Not dropped	<i>p</i> -value (1)=(2)
	(1)	(2)	(3)
Panel A: Means in the blind manager arm			
<i>Indicators of worse attention:</i>			
Non-standard open response	0.082	0.033	0.006
Failed att check 1 (screener)	0.087	0.062	0.244
Failed att check 2 (screener)	0.120	0.058	0.005
Failed att check (experiment)	0.361	0.184	0.000
<i>Education:</i>			
Less than high school	0.019	0.008	0.217
High school graduate	0.120	0.143	0.432
Some college but no degree	0.279	0.293	0.699
2 year college degree	0.106	0.099	0.792
4 year college degree	0.399	0.349	0.212
Professional or Masters degree	0.072	0.099	0.257
Doctorate	0.005	0.008	0.623
<i>Proofreading performance, effort:</i>			
Correct highlights, required paragraphs	12.471	15.806	0.000
Incorrect highlights, required paragraphs	4.346	3.017	0.000
Time on required paragraphs	4.872	4.992	0.130
N (blind manager arm)	208	484	
Panel B: Treatment effects (SEs)			
Number of paragraphs	0.234 (0.469)	-0.796 (0.281)	0.050
Proofread more than required	0.041 (0.034)	-0.037 (0.019)	0.032
Finished all paragraphs	0.011 (0.047)	-0.074 (0.028)	0.104
Time on required paragraphs	0.081 (0.102)	-0.068 (0.062)	0.191
Bonus in required paragraphs	0.038 (0.040)	-0.009 (0.024)	0.306
Beliefs about promotion, same mgr	-4.284 (1.980)	-1.917 (1.341)	0.301
Beliefs about promotion, cutoff	-4.701 (2.016)	0.304 (1.422)	0.034
Pr(same manager) - Pr(cutoff)	0.417 (1.251)	-2.221 (0.830)	0.066
Shared zero with manager	0.159 (0.051)	0.003 (0.033)	0.008
Amount shared with manager	-0.770 (0.436)	-0.033 (0.263)	0.134
N (both manager arms)	402	985	

Note: This table compares those who are dropped from the sample when looking at outcomes elicited using multiple price lists and those who are not, in the manager sample of the promotion experiment. Workers are dropped if they display multiple switching points or switch in the wrong direction (e.g. display downward-sloping labor supply) for any of the three reservation wage elicitation or the elicitation of willingness to pay to be able to choose one's own manager. 80 percent of those who are dropped only display non-sensical responses for their willingness to pay to choose their own managers (the elicitation of which switched "direction" from the three previous reservation wage elicitation). Panel A compares the means of measures of attention and skill/effort in the blind manager arm across the two groups. Panel B shows the estimated treatment effect of being in the non-blind manager arm in each sample, estimated in one regression. The specification is the same as Table 1.2. In panel B, standard errors are robust to heteroskedasticity (in parentheses).

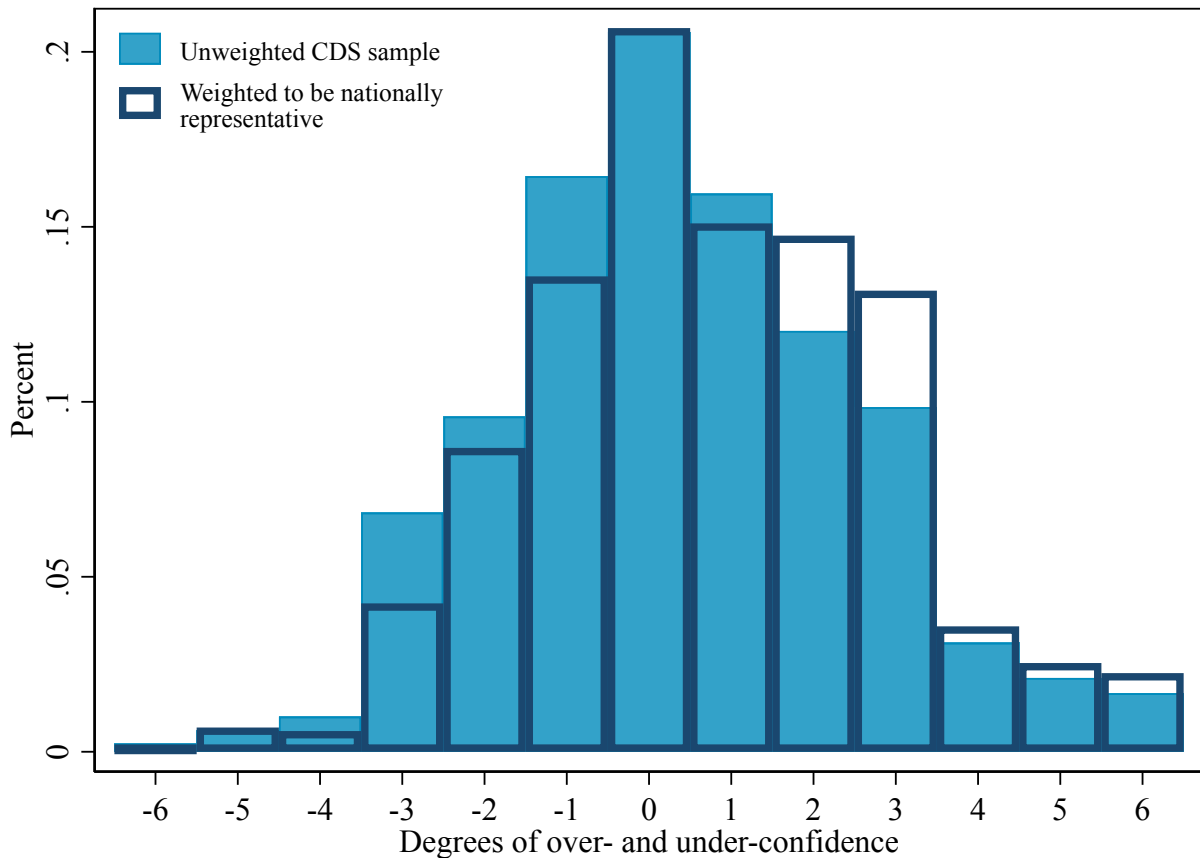
Appendix B

Appendix to Childhood Confidence, Schooling, and the Labor Market

Appendix B.1 contains supplementary tables and figures. Appendix B.2 describes the variables that make up each index used as an outcome or control variable in our main analysis and details index construction. Appendix B.3 describes how our measures of childhood over- and under-confidence in math correlate with a range of children's attitudes towards math and school. Appendix B.4 compares our results for over- versus under-confidence, and Appendix B.5 provides more detail on our measures of childhood personality, teacher and parent beliefs and investment, and elementary/middle school quality. Finally, Appendix B.6 outlines the alternate definitions of biased beliefs in math that we use in our robustness checks.

B.1 Supplementary Figures and Tables

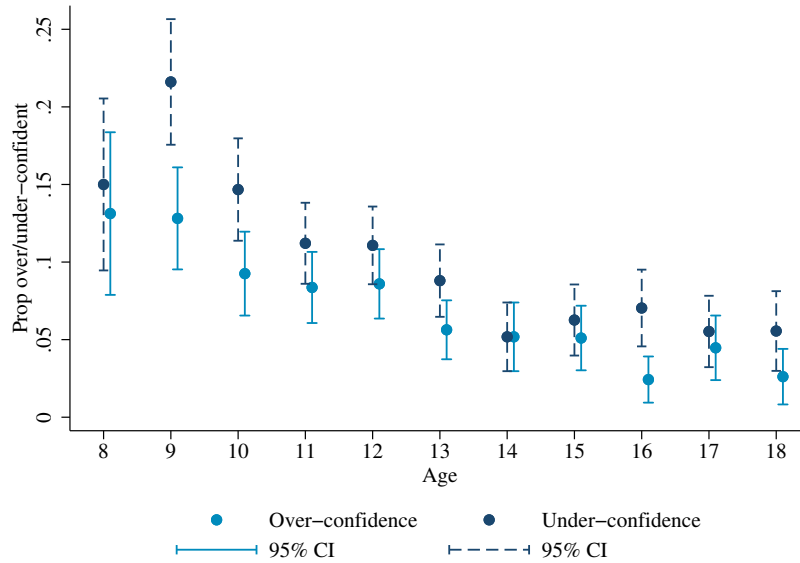
Figure B.1: Distribution of the degrees of over- and under-confidence measure



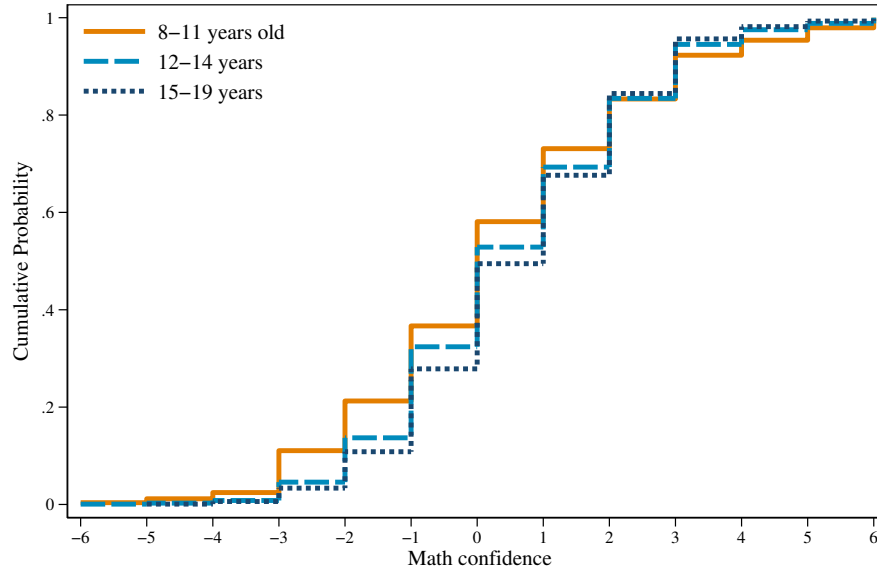
Note: This figure plots the distributions of our more continuous measure of confidence for our sample and when we use weights that make the sample nationally representative. The measure takes on value of integers from -6 (under-confident) to 6 (over-confident) and is calculated as the difference between children's self-assessed ability from 1-7 and the bin in which they should have placed themselves if they knew their score and the (uniform) national distribution of test scores. Weights are calculated using iterative proportional fitting (raking) on the original weights provided by the CDS so that our sample matches population shares in quintiles of income, in race categories, and in deciles of nationally-normed WJ-R math percentile scores.

Figure B.2: Patterns in over- and under-confidence by age

Panel A: Proportion over- and under-confident (binary measure)

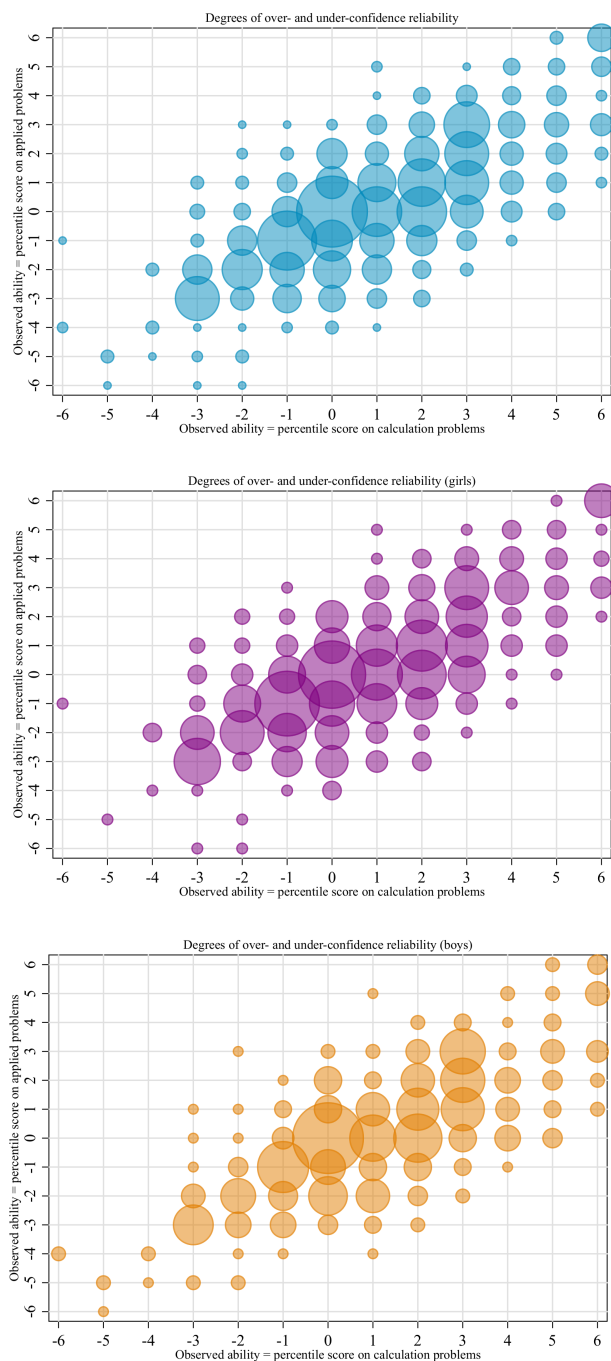


Panel B: CDF of degrees of over- and under-confidence



Note: Panel A plots the proportion of respondents that are over- and under-confident by age. Panel B plots the cumulative density function for the degrees of confidence measure, which takes on values from -6 to 6, separately for children in three age groups. We estimate these proportions using all observations of self-reported ability and test scores in math for our 2985 respondents, including two reports for the 60% of our sample with confidence measures in two CDS waves. In Panel A, we identify over- and under-confidence in math using gaps between children’s self-reported math ability and their performance on the WJ-R math test administered in the CDS. In particular, we classify a respondent as under-confident if she scored above the 75th percentile on the WJ-R math assessments and ranked herself at 1-4 on the 7-point scale of math ability, or if she scored above the 50th percentile and ranked herself at 1-3. Similarly, we identify any respondent as over-confident in math if she scored below the 25th percentile and rated herself at 6 or 7 on the response scale, or if she scored below the 50th percentile and rated herself at 7. In Panel B, we measure biased beliefs as the difference between children’s self-assessed ability (between 1 and 7) and the bin of the ability distribution in which they should have placed themselves if they had full information about the national distribution of scores and their place in it.

Figure B.3: Differences in over- or under-confidence classification using 2 subtests of the WJ-R



Note: This figure plots the joint distribution of children's degree of over- and under-confidence, which takes on values of integers from -6 (under-confident) to 6 (over-confident), when we use two different measures of demonstrated ability: percentile scores on the applied reasoning section of the WJ-R test (our main measure) and percentile scores on the calculation section of the WJ-R test (only administered in 1997).

Figure B.5: Specification chart for adolescent math scores

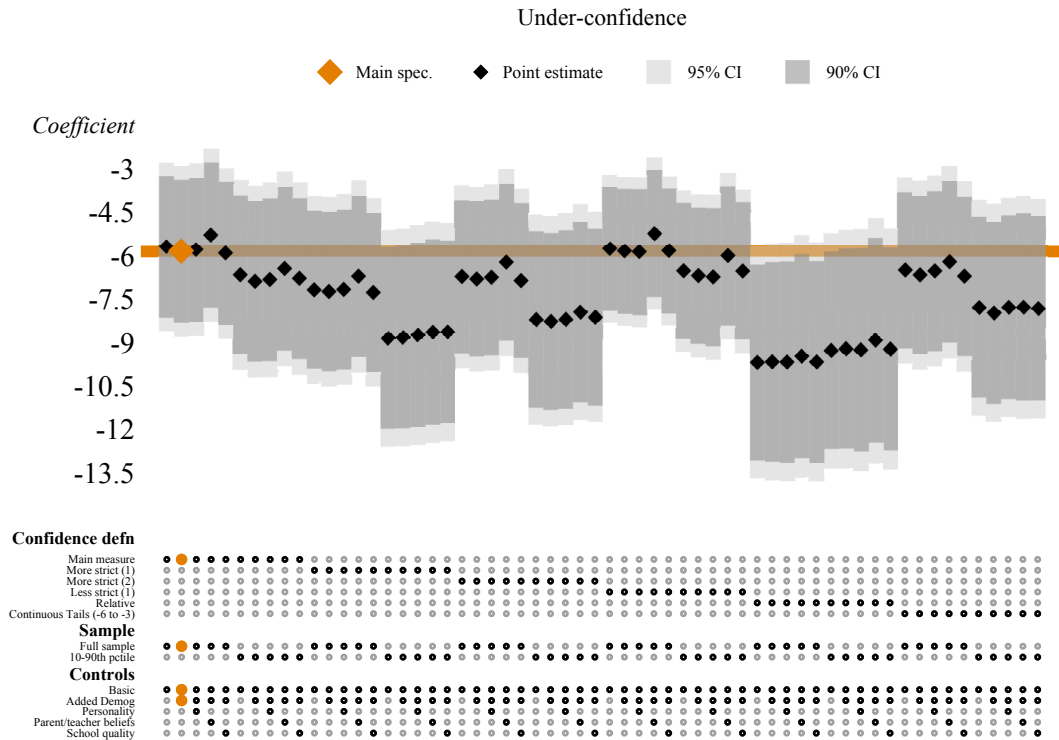
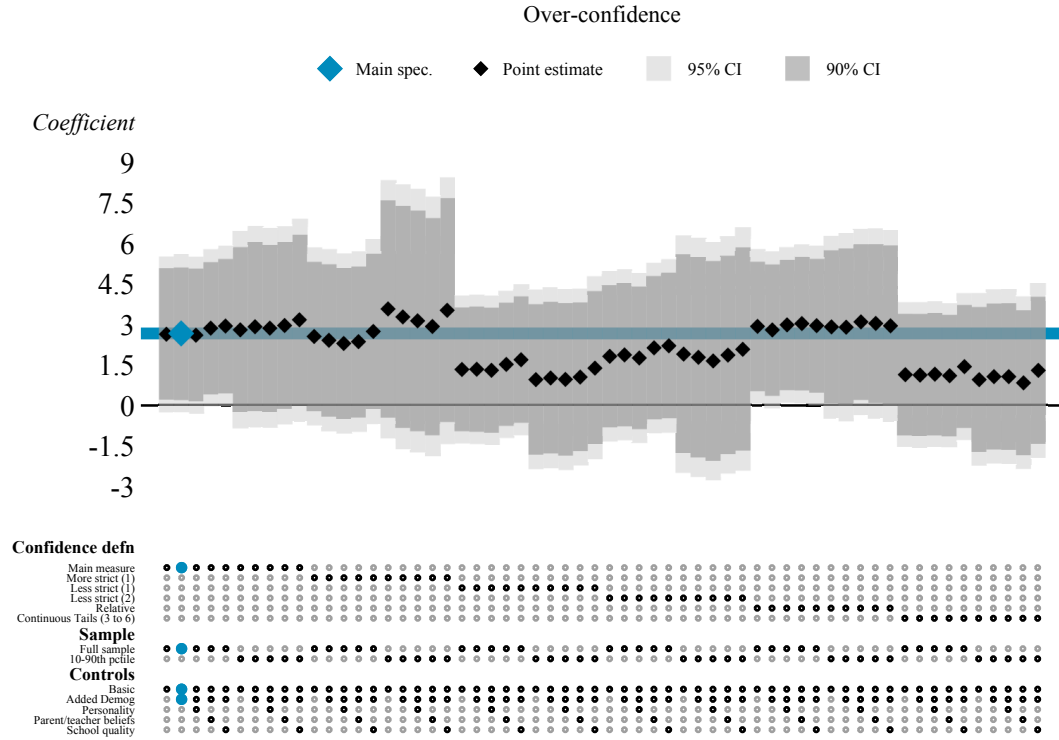


Figure B.6: Specification chart for adolescent reading scores

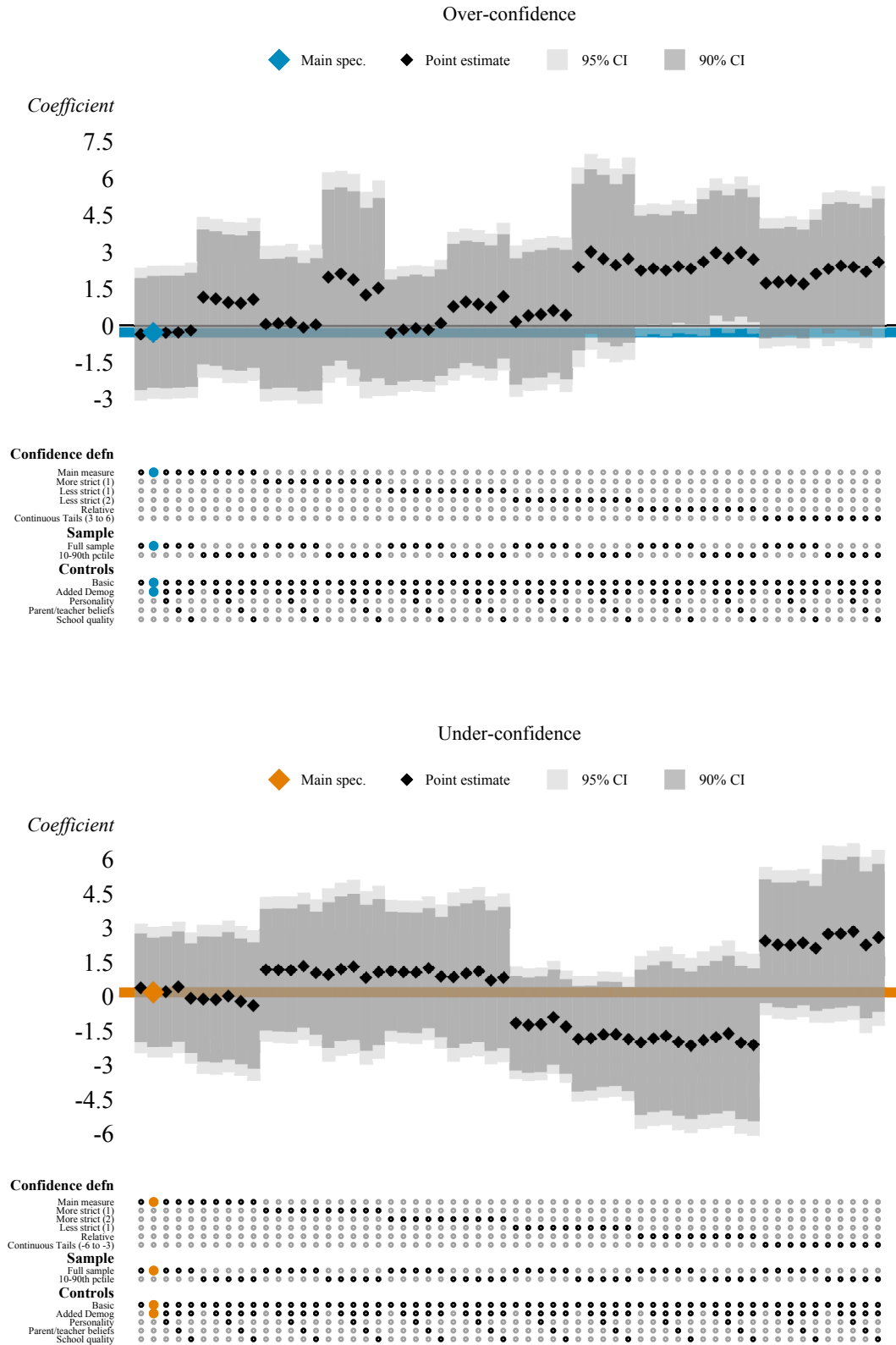
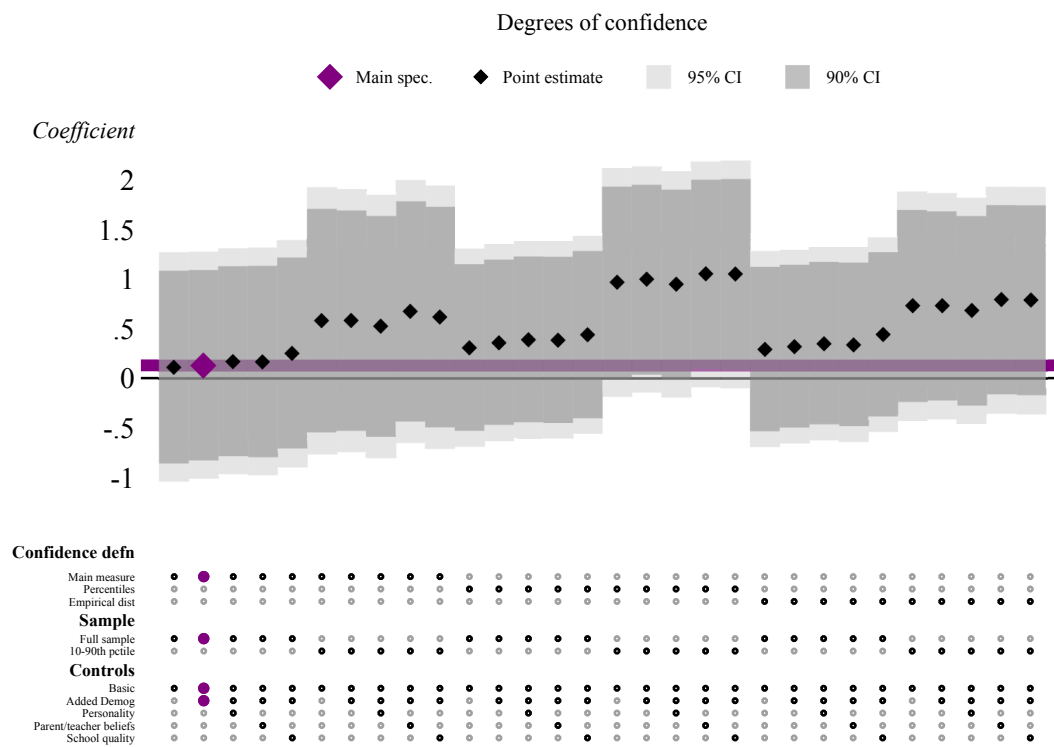


Figure B.6, continued



Note: This figure is analogous to Appendix Figure B.4, but the outcome is adolescent reading test scores.

Figure B.7: Specification chart for graduating from high school

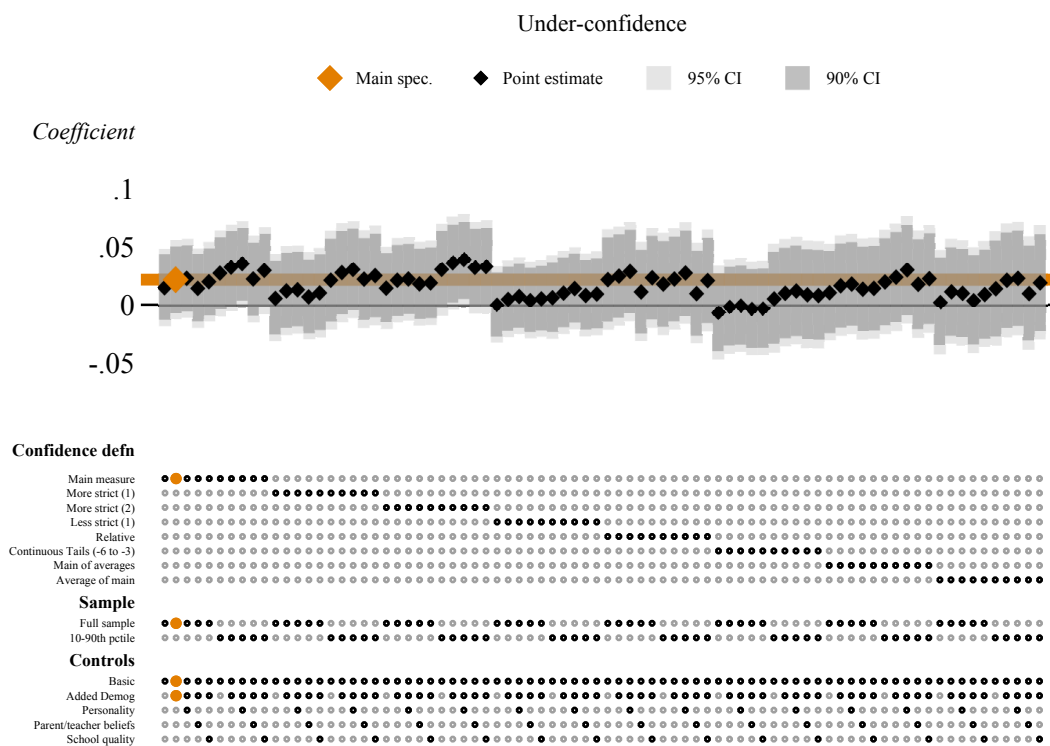
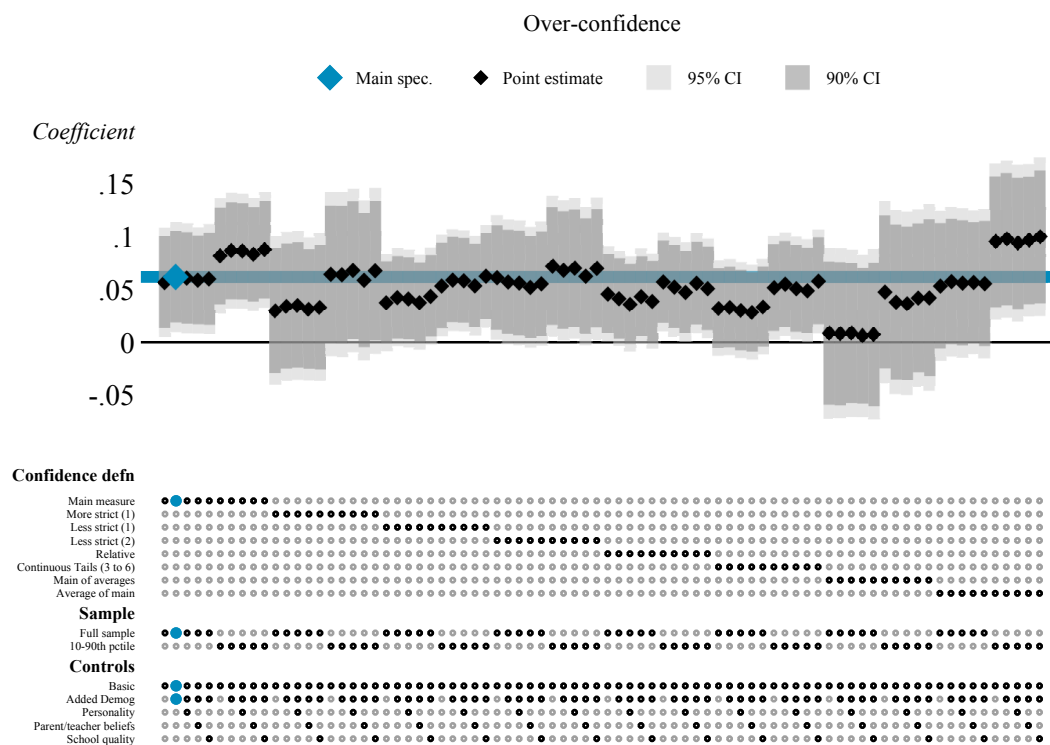
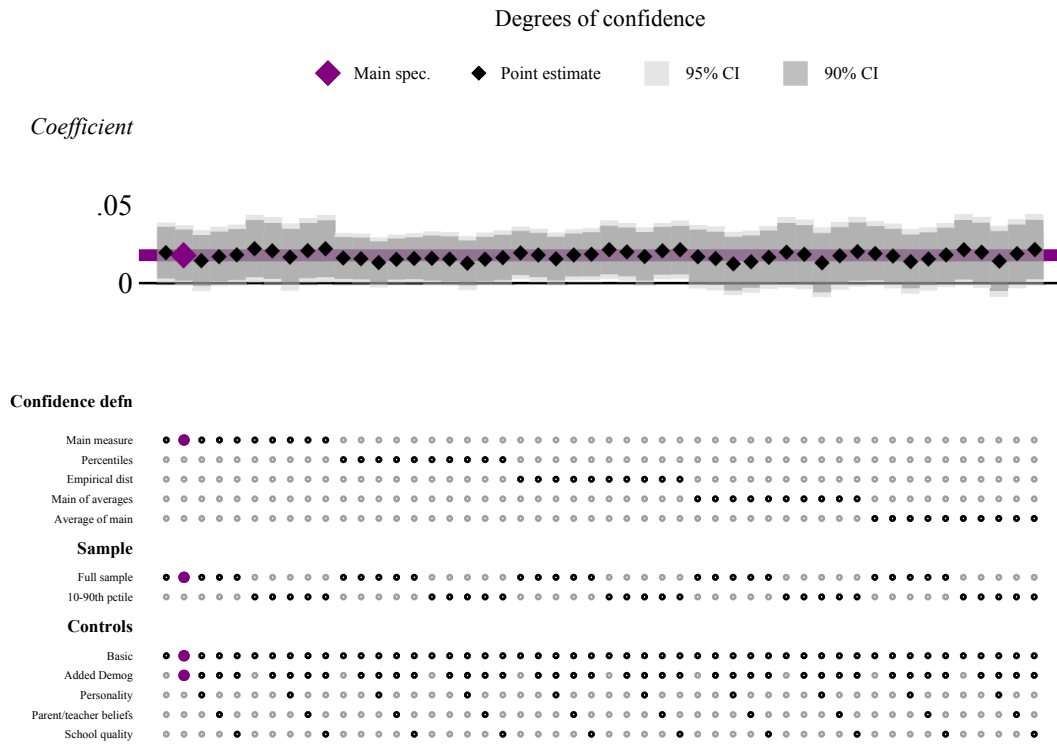


Figure B.7, continued



Note: This figure plots the coefficient of interest on either over- or under-confidence or our degrees of confidence measure for a large number of specification tests. The outcome is an indicator for graduating from high school. We test the relationship between childhood biased beliefs and high-school graduation when we (a) change our definitions of over- and under-confidence and the degrees of confidence measure, (b) drop the bottom and top ten percent of the ability distribution, since those children are the mechanically most likely to be over- or under-confident, respectively, and (c) iterate through each of five sets of control variables. Here, our alternate definitions of biased beliefs include measures of confidence that replicate the main measure but are based on information from multiple waves of the CDS. Appendix B.6 describes each alternate confidence definition in detail.

Figure B.8: Specification chart for graduating from college

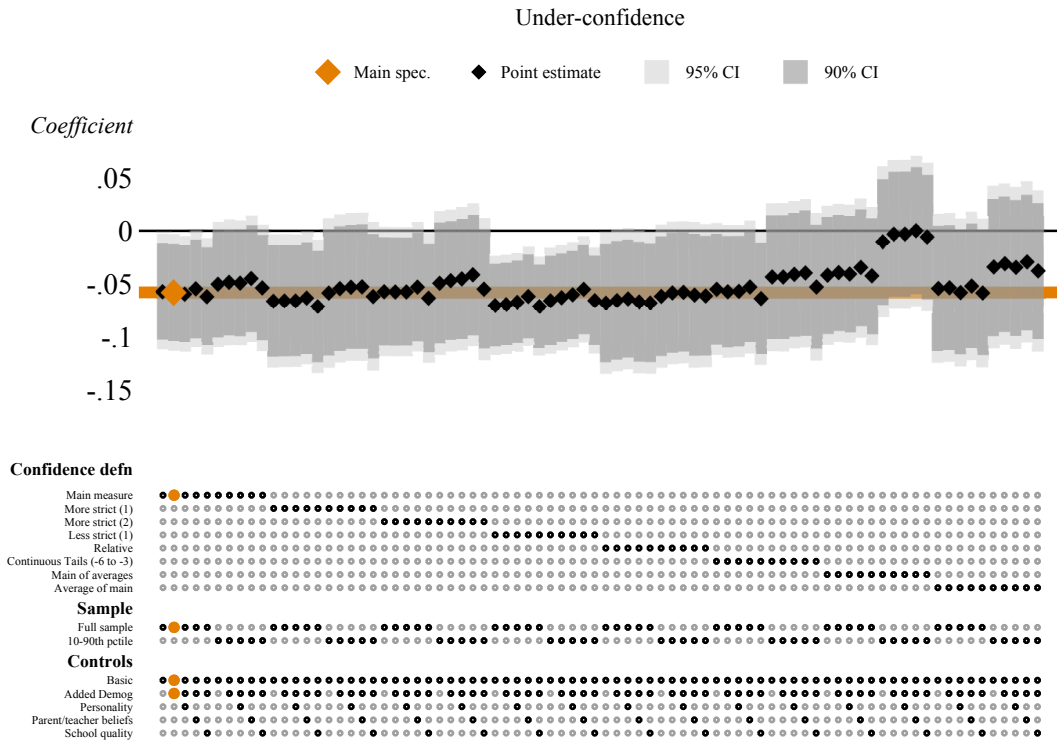
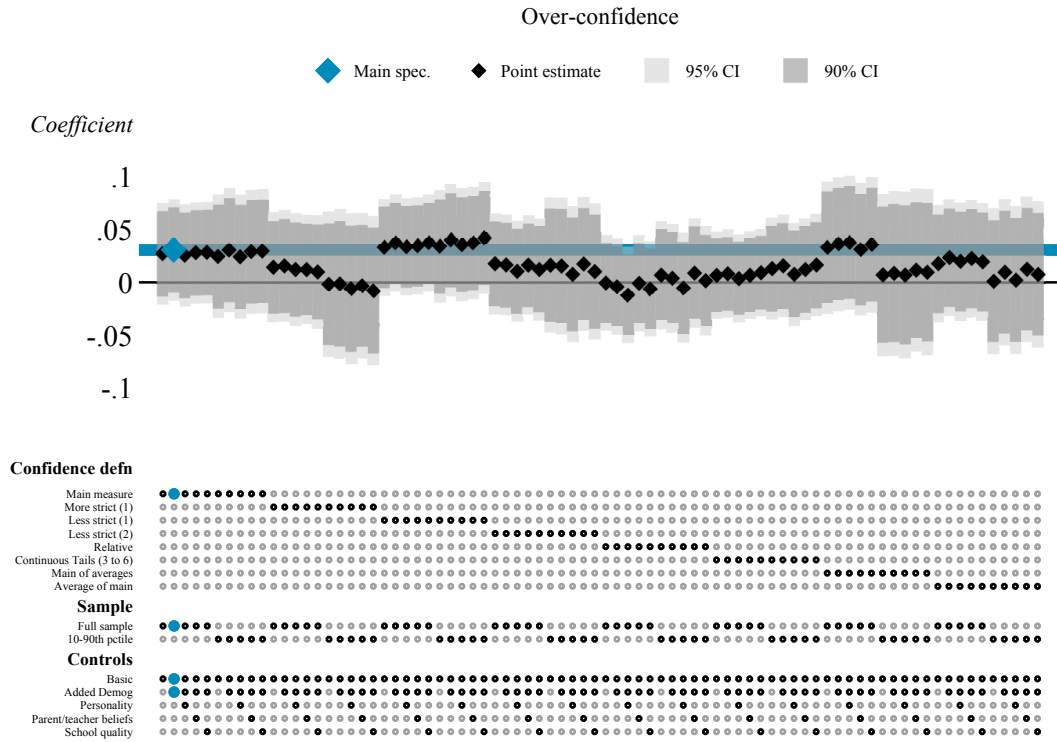


Figure B.9: Specification chart for college quality index

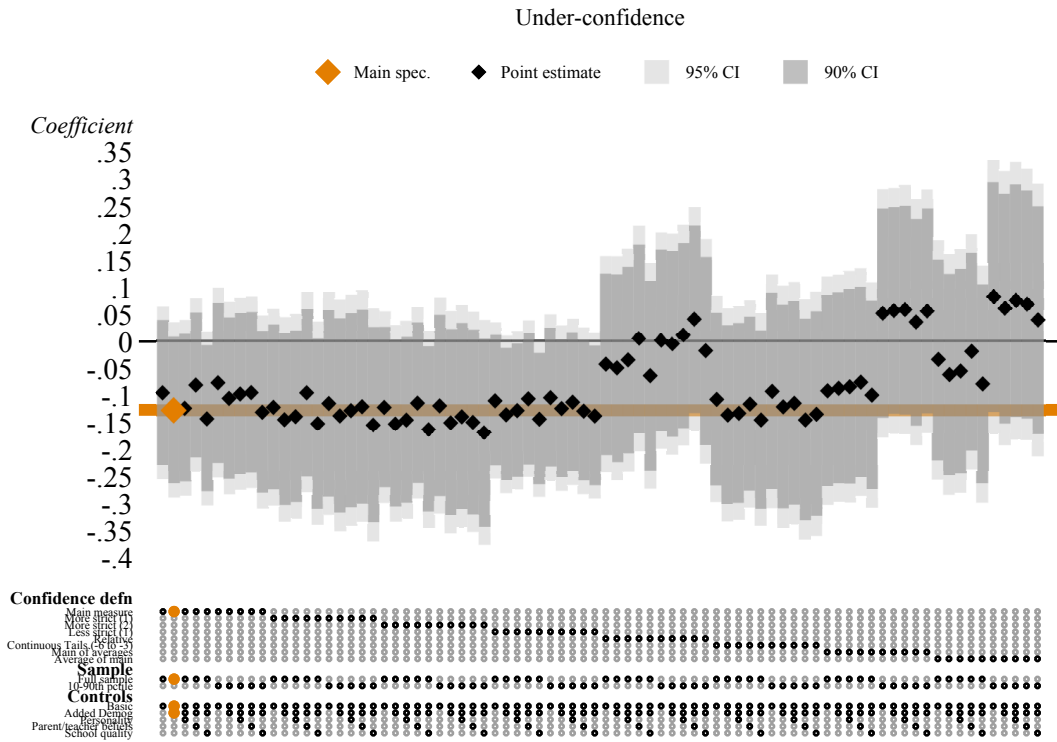
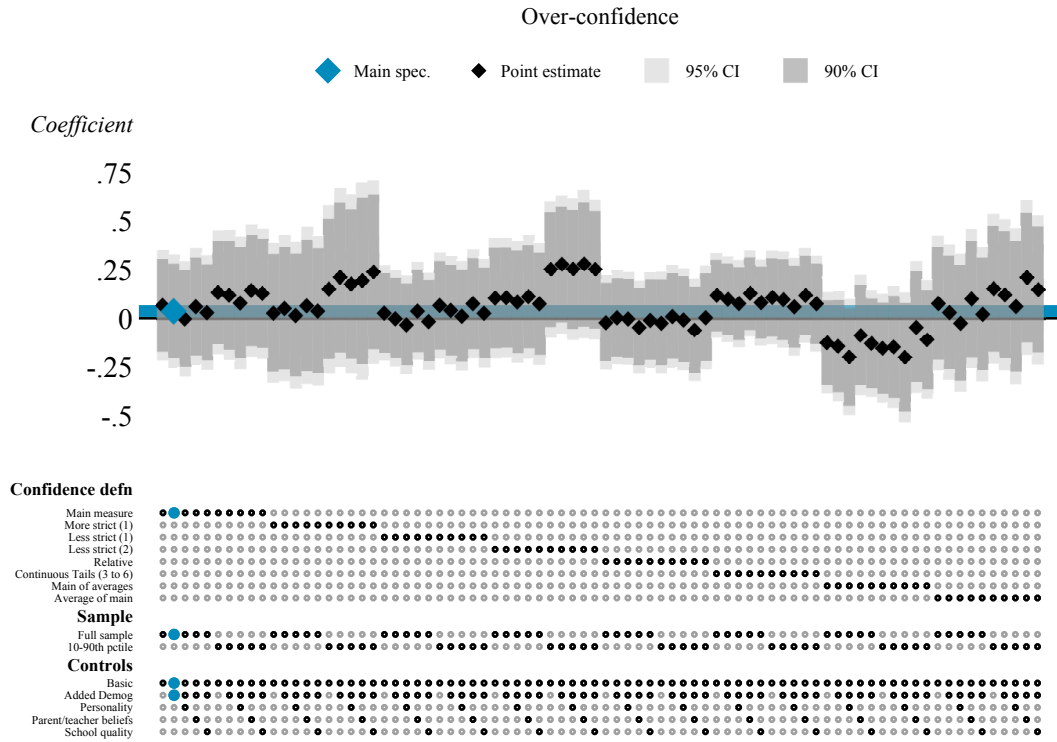
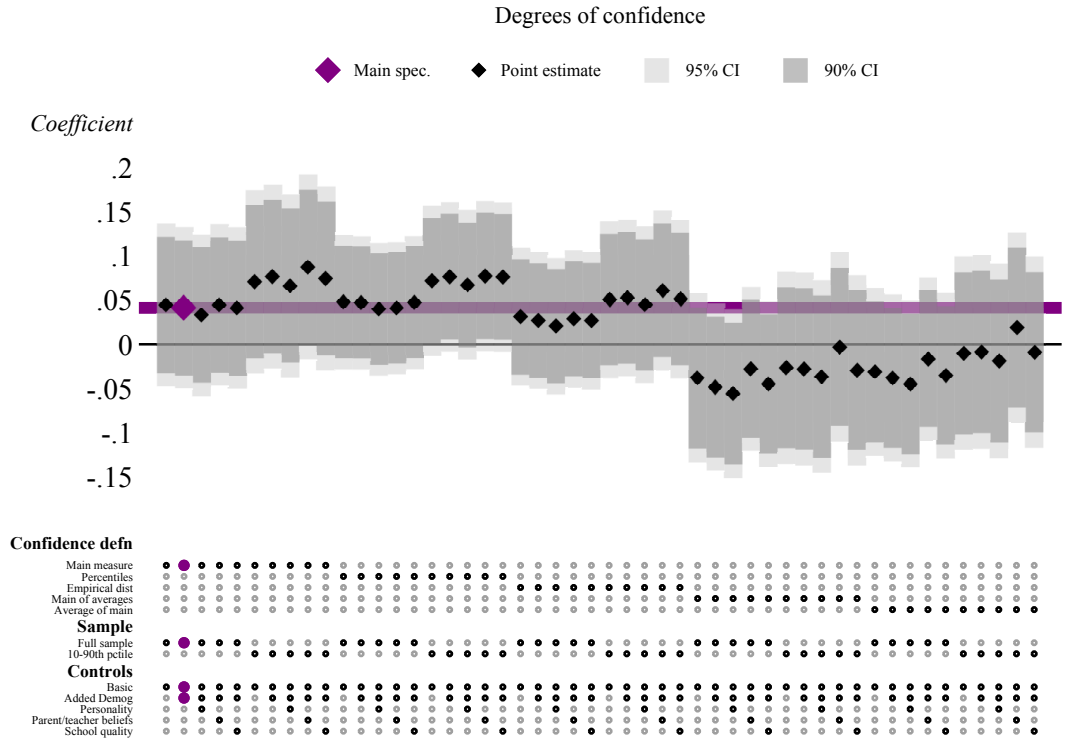


Figure B.9, continued



Note: This figure is analogous to Appendix Figure B.7, but the outcome is the index of college quality for a student's first college attended (conditional on going to college).

Figure B.10: Specification chart for college's 75th percentile math SAT score

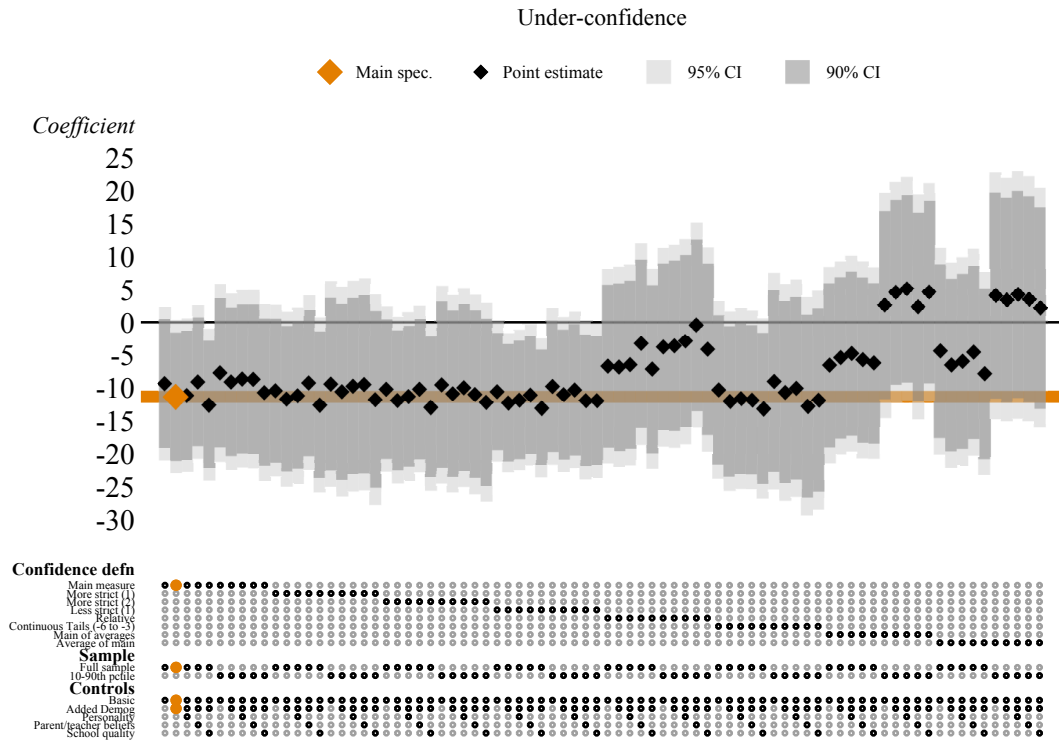
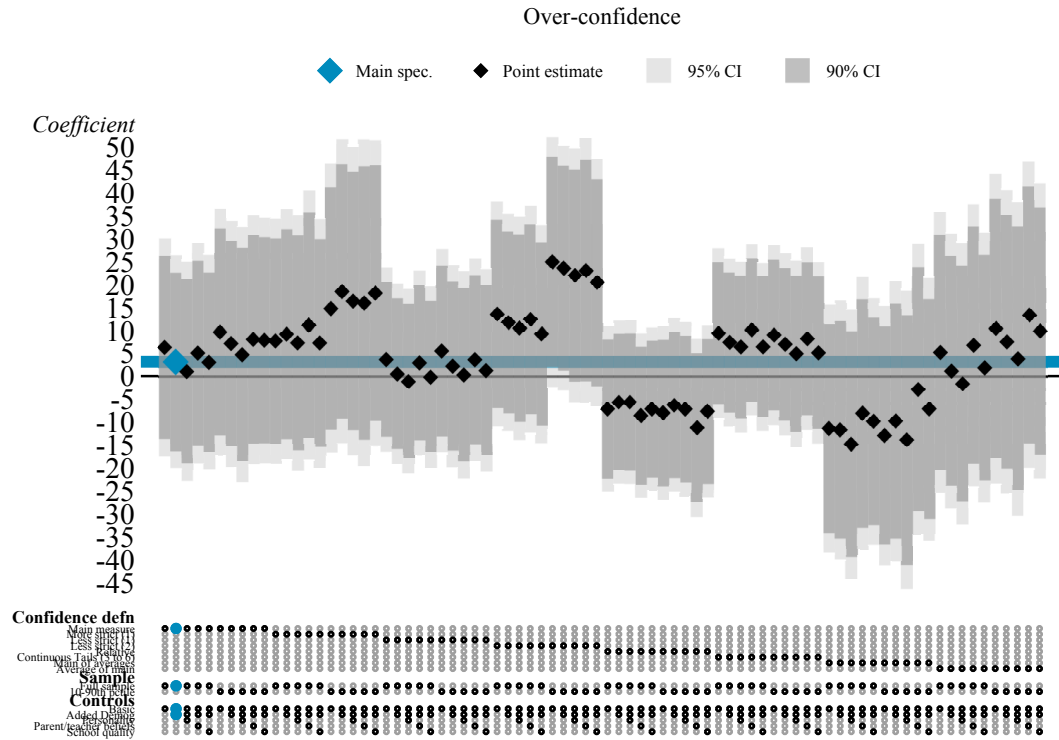
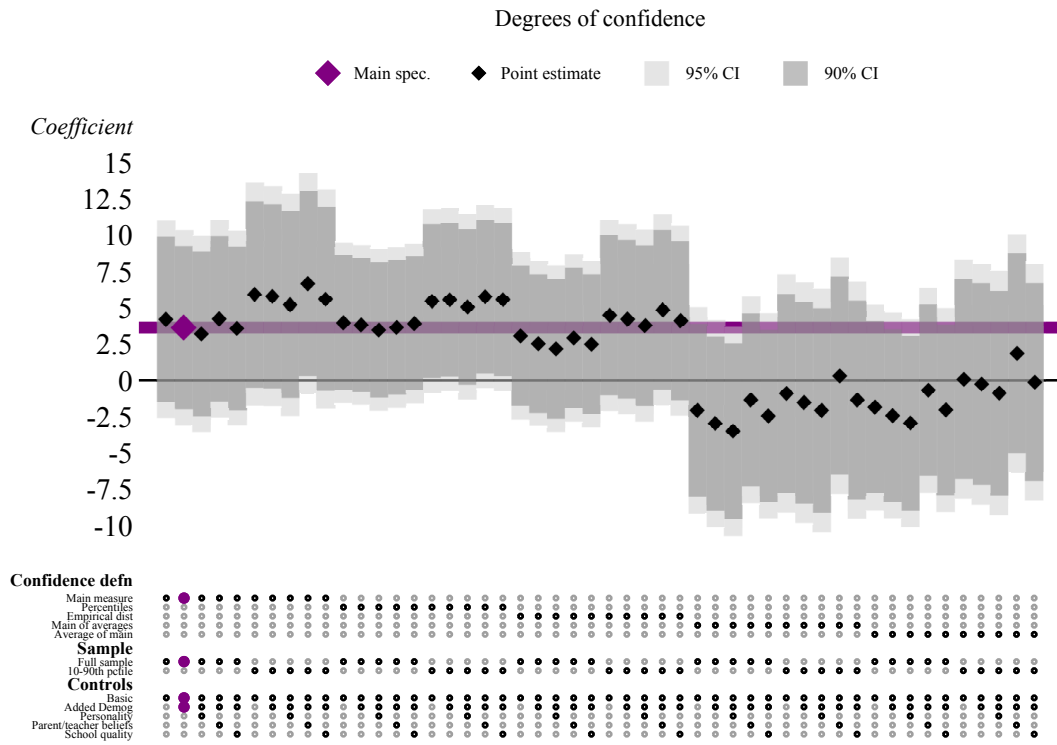


Figure B.10, continued



Note: This figure is analogous to Appendix Figure B.7, but the outcome is the 75th percentile math SAT score at a student's first college attended (conditional on going to college).

Figure B.11: Specification chart for majoring in STEM

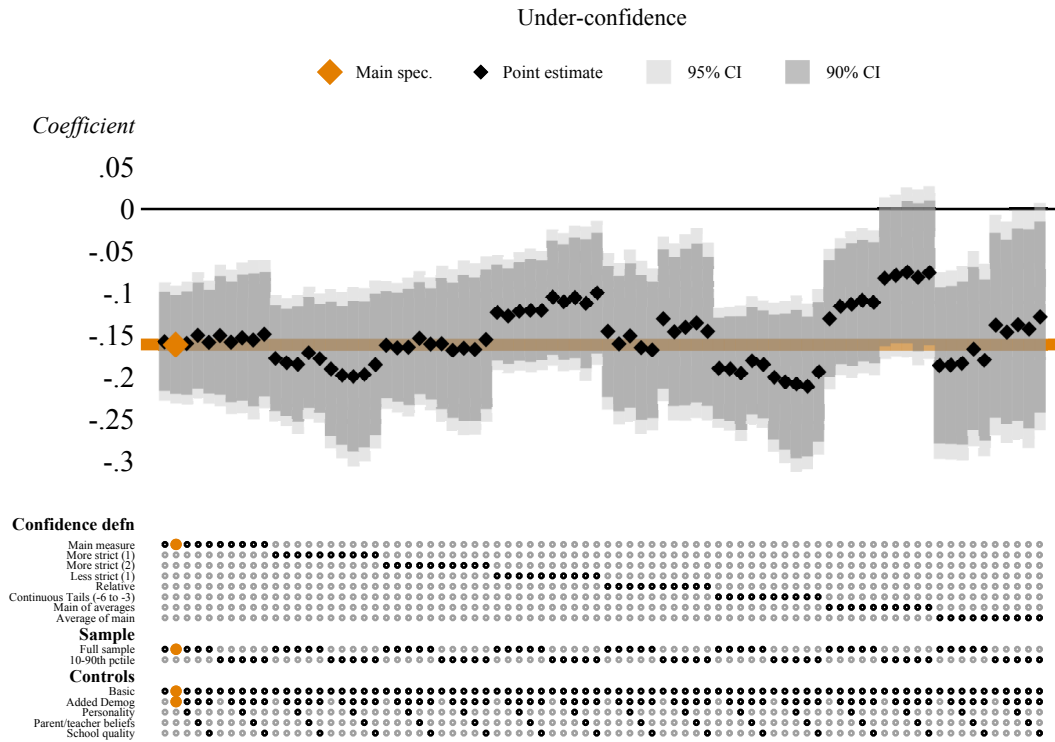
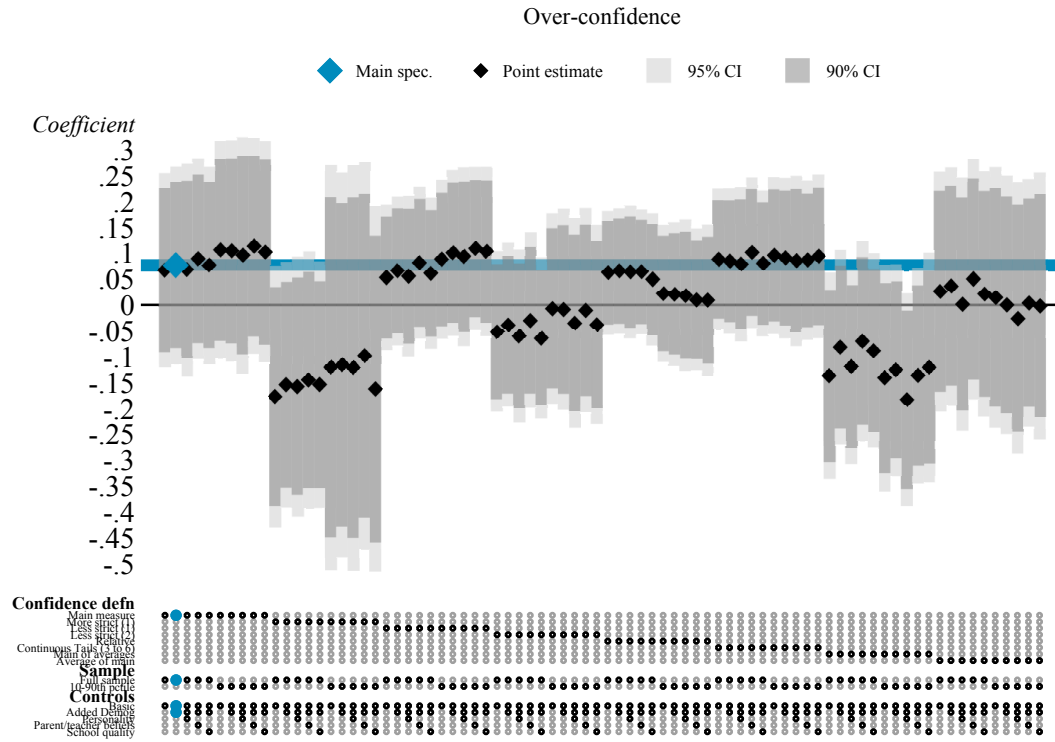


Figure B.12: Specification chart for having a graduate school degree

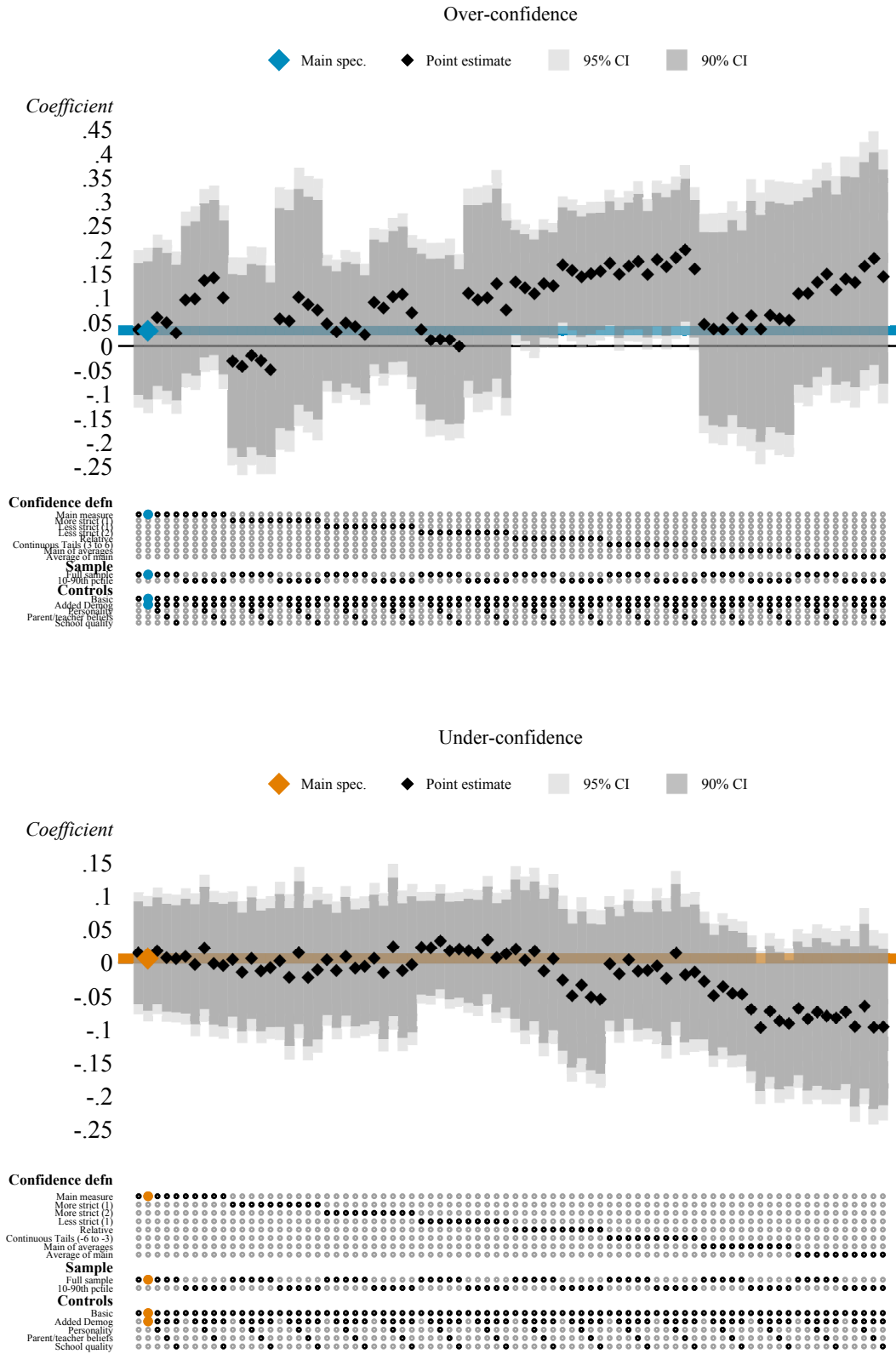


Figure B.13: Specification chart for working in STEM

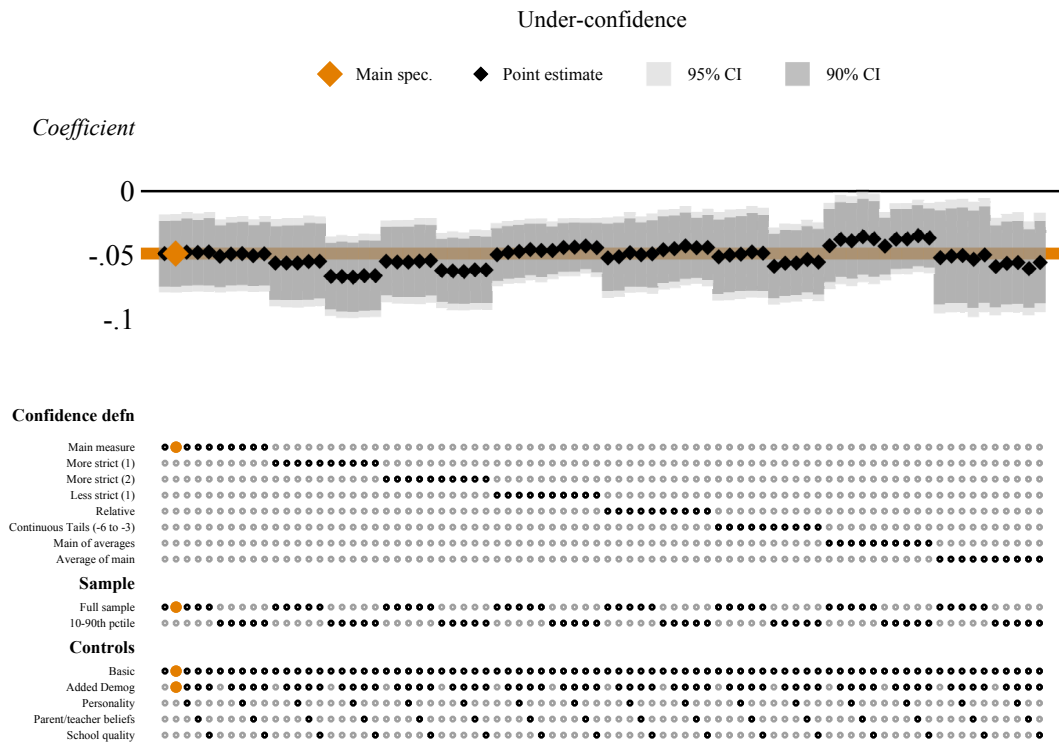
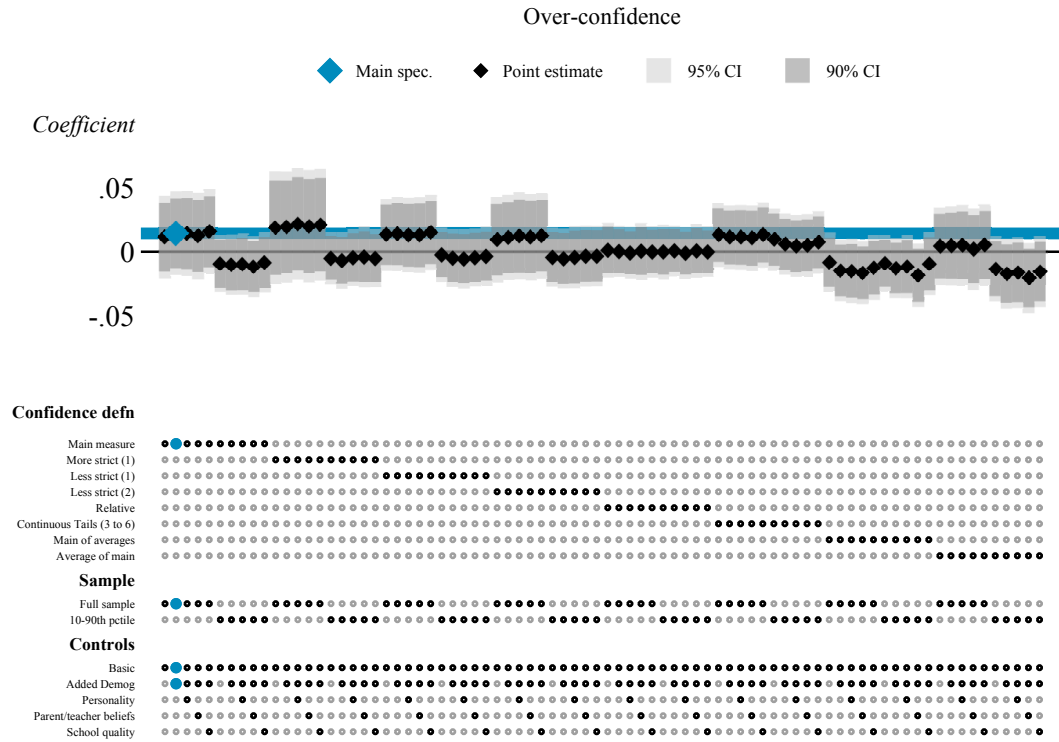
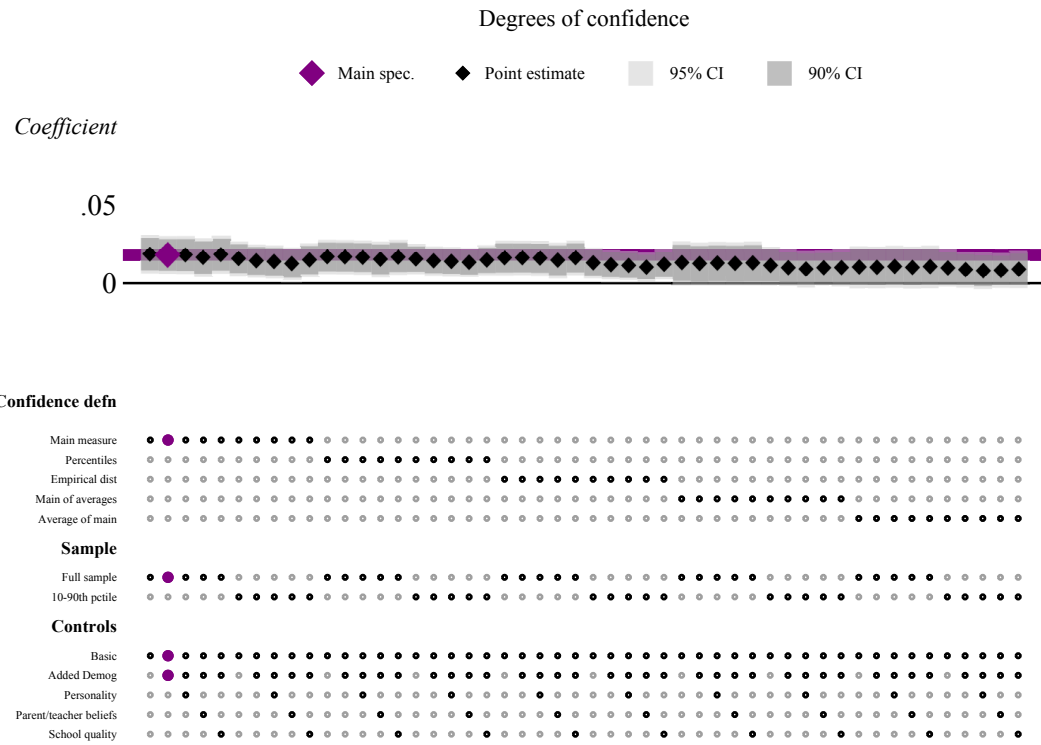


Figure B.13, continued



Note: This figure is analogous to Appendix Figure B.7, but the outcome is an indicator for working in STEM.

Figure B.14: Specification chart for working in non-STEM high-education occupation

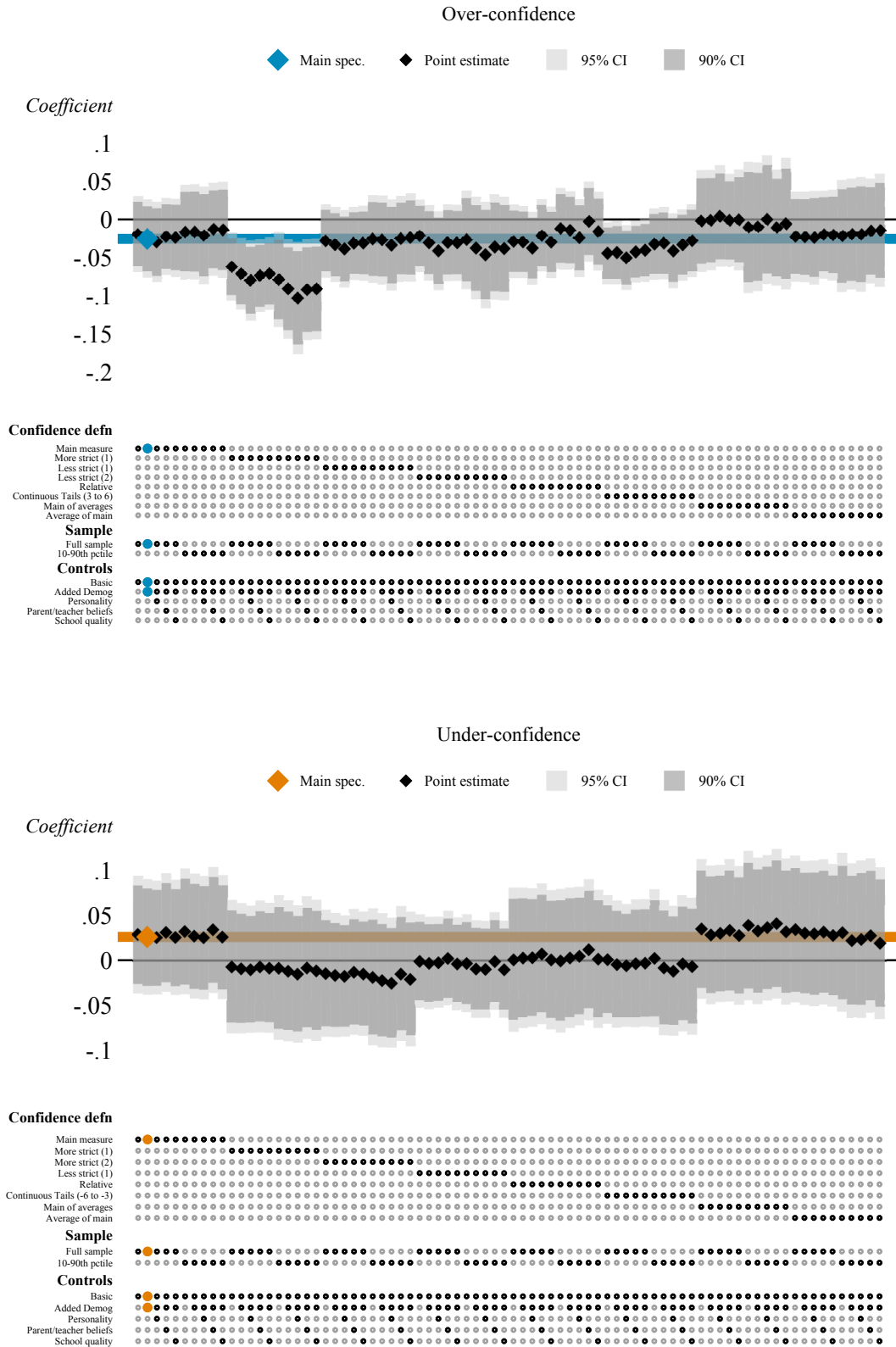


Figure B.15: Specification chart for $\ln(\text{earnings})$

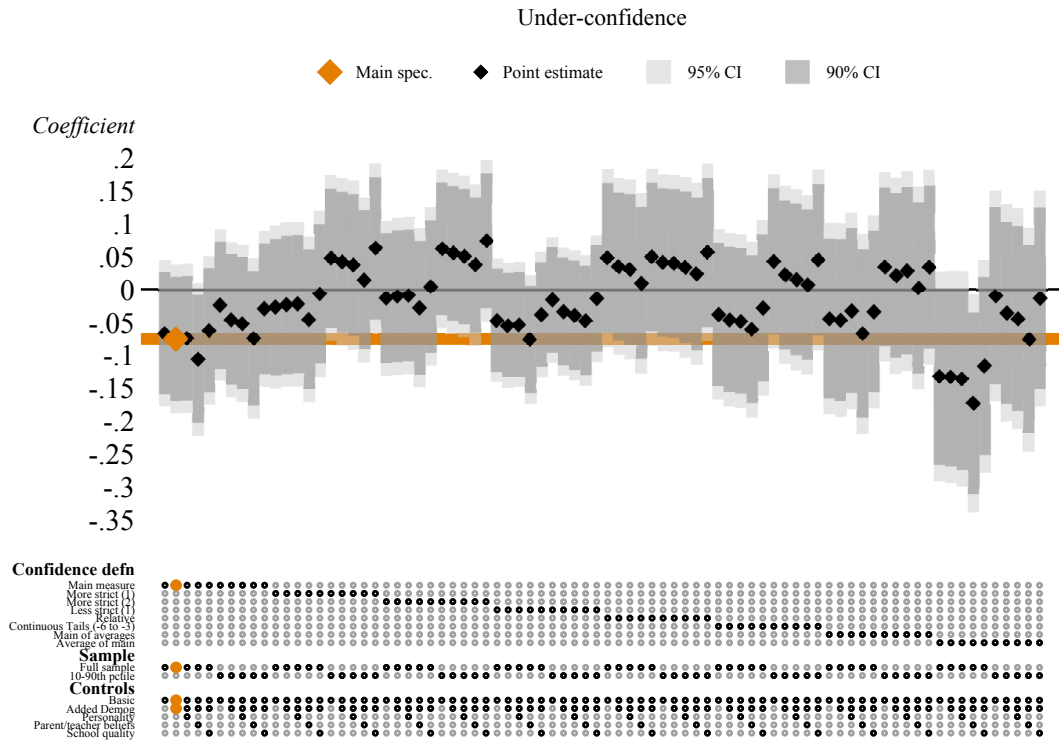
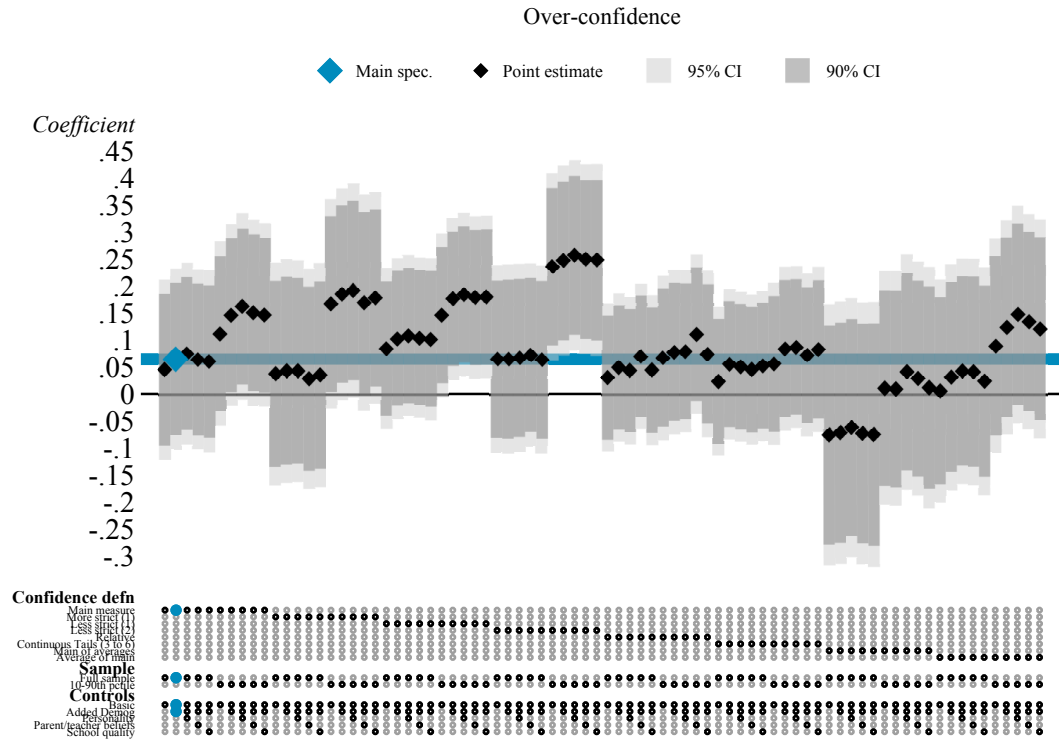
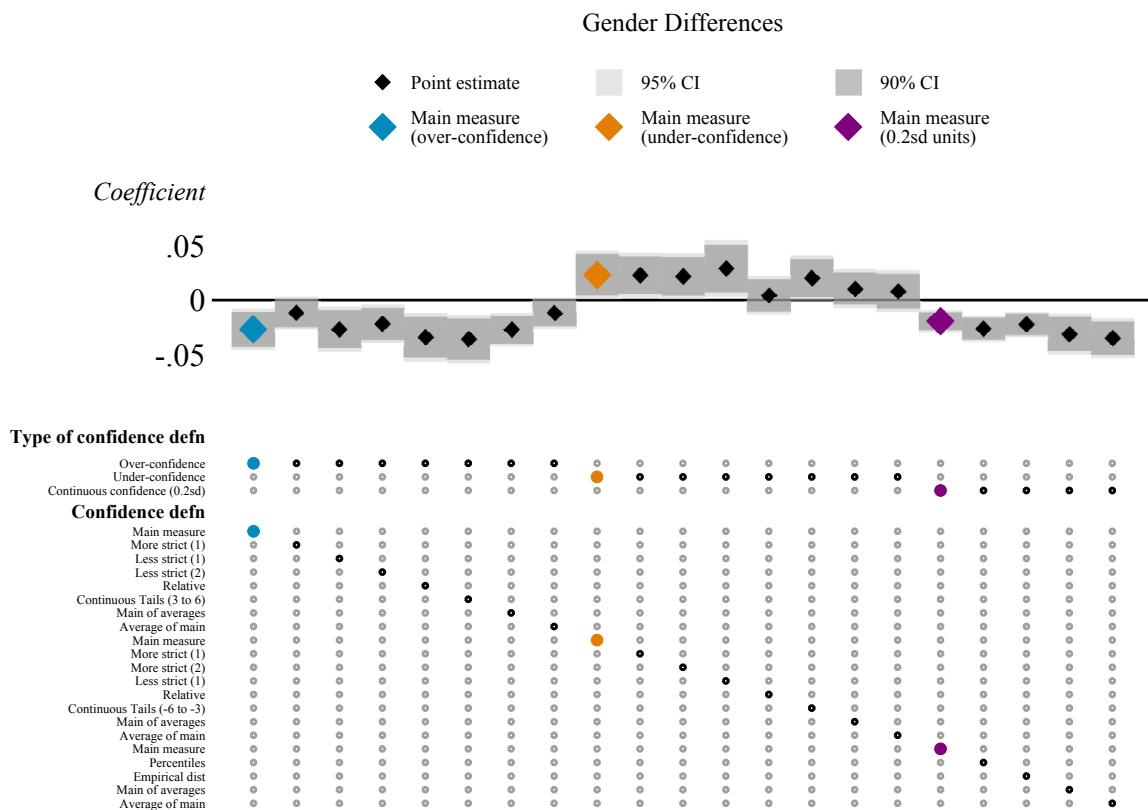
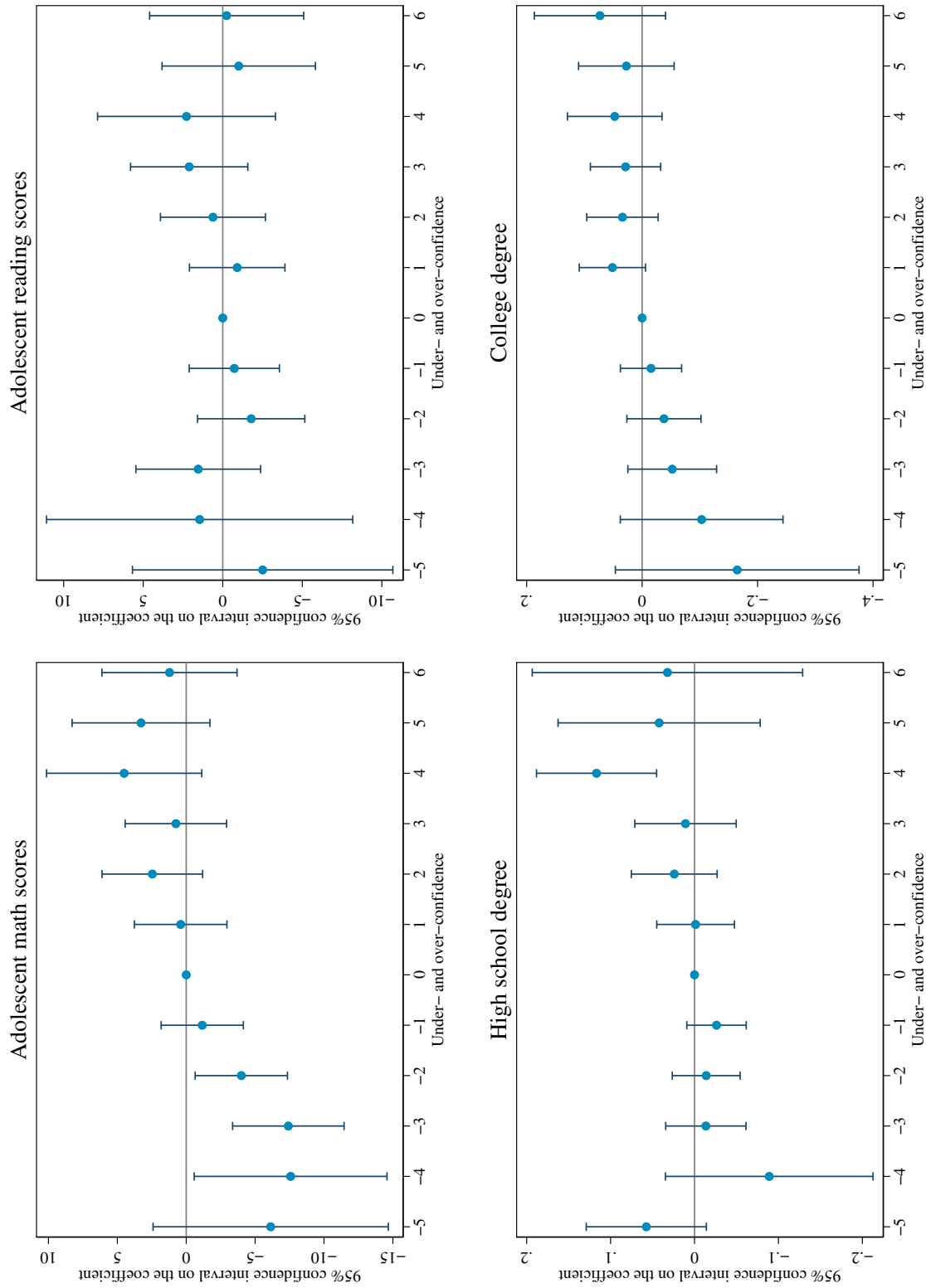


Figure B.17: Specification chart for gender differences



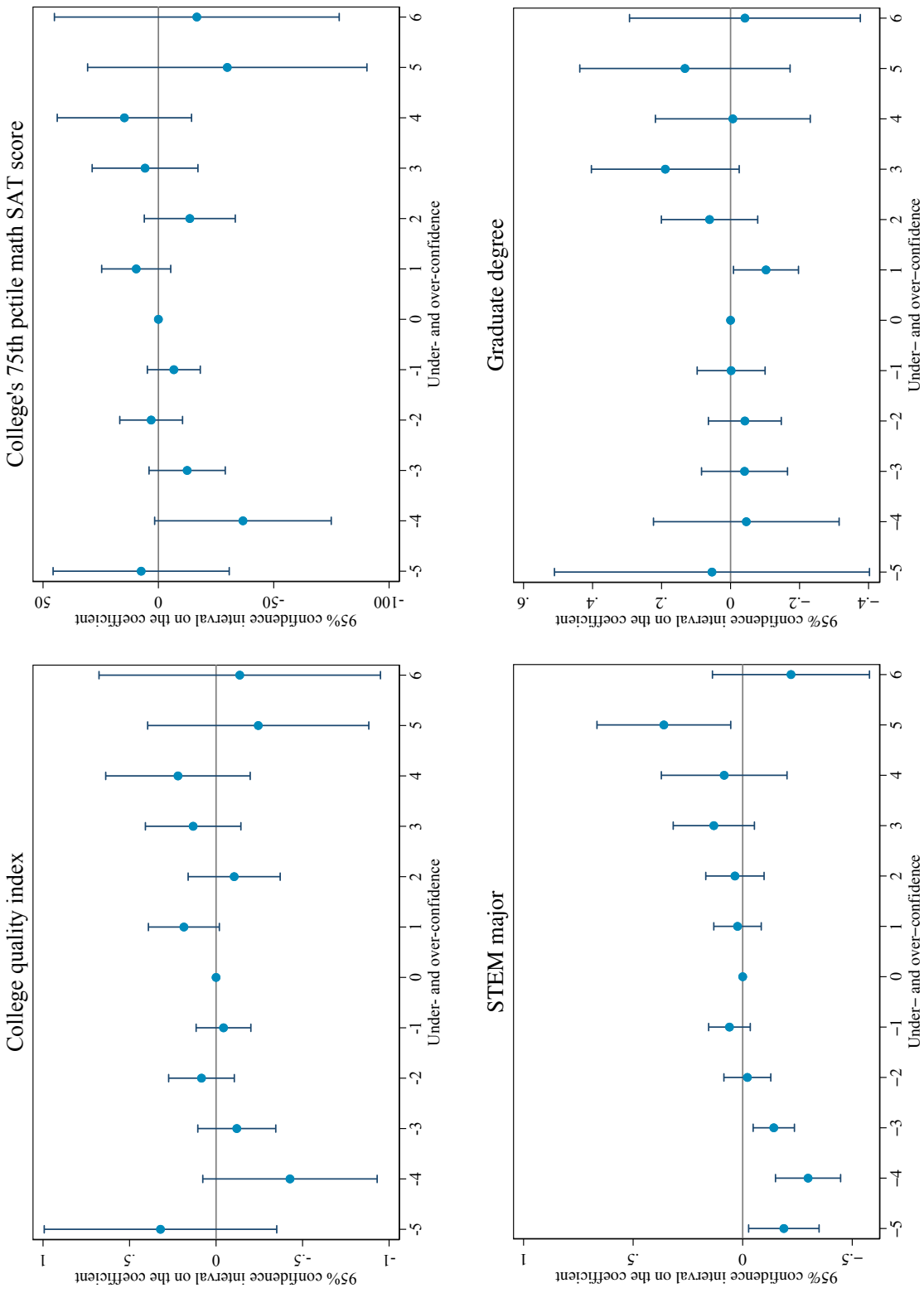
Note: This figure presents the gender gap in confidence for every measure of confidence we consider. For all measures, there is a robust gender gap: girls are less likely to be over-confident, more likely to be under-confident, and have lower degrees of confidence. Each point plots the coefficient on the female indicator when we replace the dependent variables in Table 2.2 with each of our alternate measures. Note that the more continuous measures of degrees of over- and under-confidence are all divided by 5 so that the resulting coefficient is on a similar scale as the coefficients when the outcome is an indicator for over- and under-confidence. This chart aims to communicate the stability of these coefficients, but one can obtain the gender gap in standard deviations by multiplying the coefficient for the more continuous measures by 5.

Figure B.18: Coefficients on each confidence level fixed effect (medium-term educational achievement and attainment)



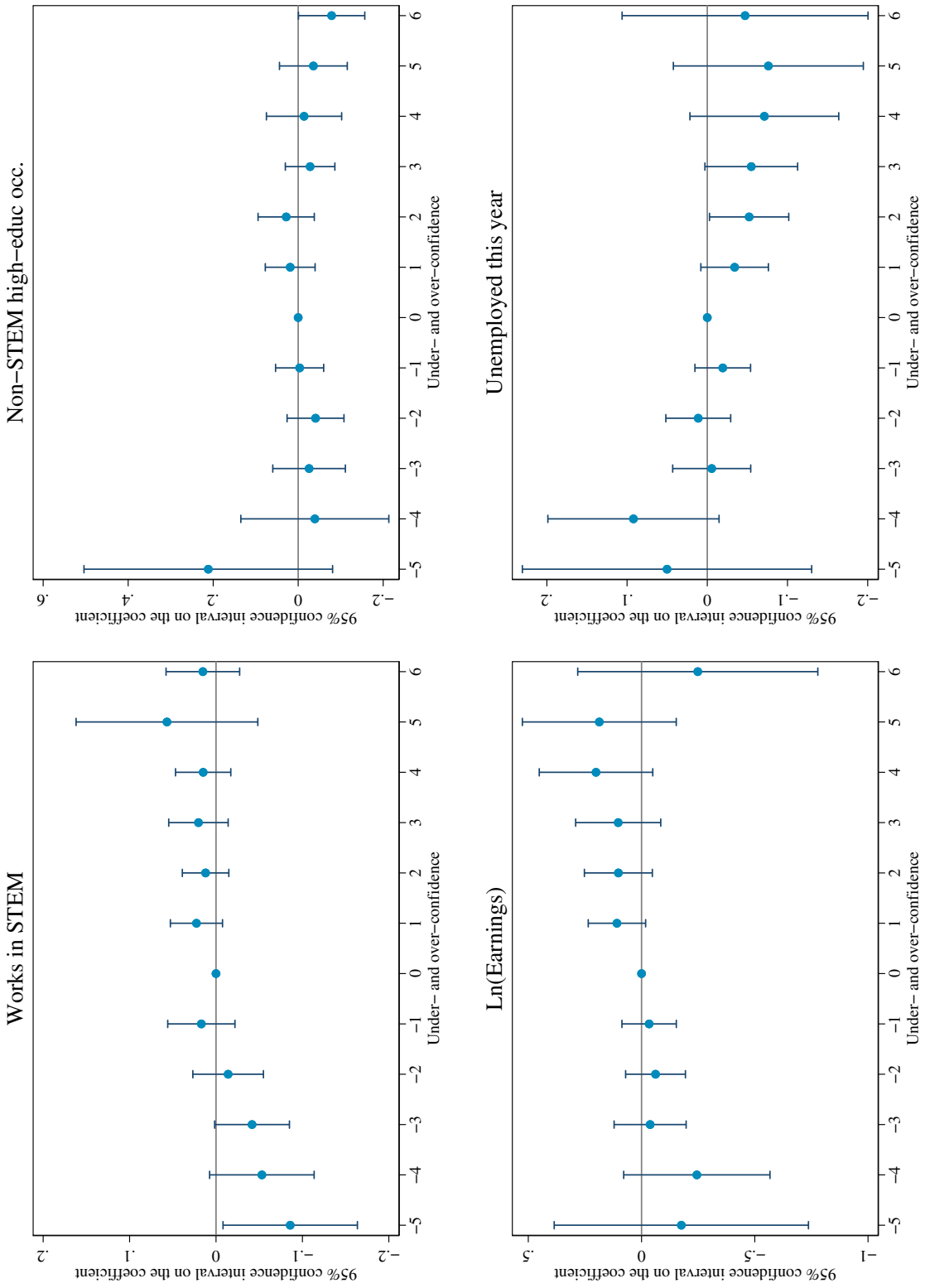
Notes: This figure plots results from a version of our main specification where we include fixed effects on the degrees of confidence measure, which takes on integer values from -6 to 6. It is measured as the difference between a child's self-assessed ability (from 1-7) and the bin from 1-7 in which they should have placed themselves if they knew the national score distribution and their place in it. The outcomes are the same as in Table 2.3, where the degrees of confidence measure enters linearly into the specification. These figures support that linearity assumption.

Figure B.19: Coefficients on each confidence level fixed effect (college quality, college major, and post-college schooling)



Notes: This figure parallels figure B.18, but for the outcomes presented in Table 2.4.

Figure B.20: Coefficients on each confidence level fixed effect (employment outcomes)



Notes: This figure parallels figure B.18, but for the outcomes presented in Table 2.5.

Table B.1: WJ-R Applied Problems section scores predict long-run outcomes

	Math Score (1)	Reading Score (2)	HS grad (3)	College grad (4)	STEM major (5)	STEM occup (6)	High-educ occup (7)	ln(Earnings) (8)	Unempl (9)
Math score percentile/10	4.953*** (0.231)	1.316*** (0.241)	0.008*** (0.003)	0.026*** (0.004)	0.023*** (0.008)	0.006*** (0.002)	0.008** (0.004)	0.058*** (0.010)	-0.011*** (0.003)
Reading score percentile/10	1.503*** (0.242)	6.371*** (0.242)	0.015*** (0.003)	0.021*** (0.004)	0.007 (0.008)	-0.001 (0.002)	0.014*** (0.004)	0.014 (0.010)	-0.002 (0.003)
N	1747	1745	2714	2725	736	4592	4592	4423	4975
Sample mean	50.808	48.231	0.876	0.297	0.189	0.046	0.163	10.185	0.167
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table regresses educational and employment outcomes on CDS math and reading scores. We drop our fixed effect controls for math and reading score deciles, replacing them with linear controls for math and reading score percentiles. These regressions include all controls included in Table 1 except for childhood over- and under-confidence in math. Basic controls also include year fixed effects when the outcome is observed in a panel. All controls that are indices are normalized relative to the weighted distribution. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.2: The persistence of reading over- and under-confidence

	(1)	(2)
Panel A: Reading over-confidence	0.194***	0.198***
	(0.037)	(0.037)
N	1732	
Sample mean	0.153	
Panel B: Reading under-confidence	0.098***	0.102***
	(0.034)	(0.034)
N	1732	
Sample mean	0.061	
Panel C: Reading confidence (SD units)	0.165***	0.166***
	(0.024)	(0.024)
N	1732	
Sample mean	-0.002	
Basic controls:	✓	✓
Added background controls:		✓

Notes: This table regresses adolescent confidence outcomes on childhood reading confidence with various controls. All controls are the same as described in Table 1. Standard errors are clustered by family, and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.3: Childhood reading confidence and medium-term educational achievement and attainment

Dependent variable:	Adolescent math scores		Adolescent reading scores		High school degree		College degree	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	-3.630*** (1.265)	-3.745*** (1.291)	1.668 (1.320)	1.564 (1.314)	-0.033 (0.025)	-0.032 (0.025)	0.008 (0.020)	0.008 (0.020)
Under-confidence	2.075 (1.835)	2.002 (1.830)	-5.720*** (1.707)	-5.811*** (1.749)	-0.004 (0.024)	-0.008 (0.023)	-0.040 (0.040)	-0.043 (0.040)
N	1734	1734	1732	1732	2698	2698	2709	2709
OC = -1*UC? p-value:	0.487	0.436	0.057	0.050	0.288	0.244	0.467	0.423
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	-1.773*** (0.609)	-1.750*** (0.623)	1.989*** (0.565)	2.008*** (0.577)	-0.014 (0.010)	-0.013 (0.010)	-0.006 (0.010)	-0.004 (0.010)
N	1734	1734	1732	1732	2698	2698	2709	2709
Sample mean of dep. var.	50.949	50.949	48.421	48.421	0.875	0.875	0.297	0.297
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses educational achievement and attainment outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.4: Childhood reading confidence and college quality, college major choice, and post-college schooling

Dependent variable:	College quality index		College's 75th pctile math SAT score		STEM Major		Graduate degree	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	-0.011 (0.112)	-0.005 (0.110)	-6.274 (9.595)	-5.052 (9.452)	0.057 (0.070)	0.065 (0.070)	0.009 (0.076)	0.016 (0.074)
Under-confidence	-0.221* (0.117)	-0.193* (0.114)	-13.598* (8.098)	-10.917 (7.932)	0.103* (0.062)	0.079 (0.064)	0.060 (0.062)	0.054 (0.062)
N	1103	1103	1112	1112	732	732	804	804
OC = -1*UC? <i>p-value:</i>	0.154	0.212	0.117	0.198	0.081	0.126	0.467	0.461
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	0.103** (0.042)	0.090** (0.042)	7.223** (3.259)	6.193* (3.238)	-0.045* (0.025)	-0.041 (0.027)	-0.010 (0.026)	-0.012 (0.026)
N	1103	1103	1112	1112	732	732	804	804
Sample mean of dep. var.	0.051		594.052		0.189		0.199	
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses college outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.5: Childhood reading confidence and employment outcomes

Dependent variable:	Works in STEM		Non-STEM high-educ occ.		Ln(Earnings)		Unemployed this year	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	0.003 (0.012)	0.005 (0.012)	0.012 (0.022)	0.005 (0.022)	-0.048 (0.071)	-0.063 (0.071)	-0.007 (0.022)	-0.003 (0.022)
Under-confidence	0.056* (0.033)	0.055* (0.033)	-0.006 (0.045)	-0.012 (0.045)	0.073 (0.071)	0.054 (0.070)	-0.037 (0.023)	-0.034 (0.022)
N	4564	4564	4564	4564	4395	4395	4943	4943
OC = -1*UC? p-value:	0.092	0.079	0.897	0.878	0.801	0.925	0.164	0.235
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	-0.005 (0.007)	-0.005 (0.007)	0.003 (0.012)	0.001 (0.012)	-0.054* (0.029)	-0.051* (0.029)	0.006 (0.009)	0.006 (0.009)
N	4564	4564	4564	4564	4395	4395	4943	4943
Sample mean of dep. var.	0.045	0.045	0.163	0.163	10.185	10.185	0.168	0.168
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses employment outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Basic controls also include year fixed effects when the outcome is observed in a panel. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.6: Summary statistics

	Mean	SD	Med.	Share Mi.
<i>Panel A: Child Demographics</i>				
Child is female	0.497	0.500	0	0.000
Child is white	0.458	0.498	0	0.000
Child is black	0.417	0.493	0	0.000
Child is hispanic	0.075	0.264	0	0.000
Child's birth year	1990.020	3.748	1990	0.000
<i>Panel B: Parent and Family Demographics</i>				
Father at least graduated high school	0.835	0.371	1	0.370
Father at least has bachelors	0.257	0.437	0	0.369
Mother at least graduated high school	0.817	0.387	1	0.110
Mother at least has bachelors	0.134	0.341	0	0.113
Mother works in STEM	0.018	0.133	0	0.157
Father works in STEM	0.072	0.259	0	0.321
Mother works in other high-educ field	0.169	0.375	0	0.157
Father works in other high-educ field	0.109	0.311	0	0.321
Total family taxable income (thous 2016 USD)	69.777	80.333	52.03	0.000
HH receives govt transfers	0.478	0.500	0	0.000
# Siblings in the HH	1.415	1.081	1	0.000
<i>Panel C: Parenting Practices and Beliefs</i>				
Father figure in HH	0.727	0.445	1	0.008
Two adults in HH	0.645	0.479	1	0.000
Parent says key thing for success is:				
to obey	0.278	0.448	0	0.023
to think for one's self	0.711	0.453	1	0.023
to work hard	0.284	0.451	0	0.023
to help others in need	0.174	0.380	0	0.023
At least once/week, parent:				
reads with child	0.386	0.487	0	0.004
does art with child	0.072	0.258	0	0.004
plays sports with child	0.157	0.364	0	0.005
does homework with child	0.634	0.482	1	0.005
plays board games with child	0.145	0.352	0	0.004
shows phys. affection to child	0.912	0.283	1	0.642
says <i>I love you</i> to child	0.894	0.308	1	0.005
Parent's traditional gender norms (index)	0.019	0.561	0	0.147
Parent's poor mental health (index)	-0.001	0.674	-0	0.116
Parent's self esteem (index)	0.011	0.992	0	0.146
Parent's self efficacy (index)	0.012	0.994	-0	0.144
Aggravation in parenting (index)	0.006	1.008	-0	0.143
Parent expectations for educ. attainment:				
Graduate degree	0.125	0.331	0	0.010
Bachelors' degree	0.493	0.500	0	0.010
High school degree	0.376	0.484	0	0.010
High school dropout	0.006	0.078	0	0.010

Table B.6: Summary statistics (continued)

	Mean	SD	Med.	Share Mi.
<i>Panel D: Other Child Characteristics</i>				
Child ever in gifted prog	0.243	0.429	0	0.031
Child ever in special ed prog	0.127	0.333	0	0.032
Child has repeated grade	0.122	0.327	0	0.021
Child qualifies for FRP lunch	0.598	0.490	1	0.256
Parent's rating of child health	0.014	1.000	1	0.005
Big 5 personality scores (indices)				
Conscientiousness	-0.003	0.684	0	0.007
Extroversion	0.005	0.618	-0	0.005
Neuroticism	-0.006	0.622	-0	0.009
Agreeableness	-0.003	0.644	0	0.008
Openness to experiences	-0.001	0.500	0	0.010
<i>Panel E: Teacher Beliefs</i>				
Perceptions of competence (stdized):				
Academic competence	0.003	0.992	-0	0.806
Social competence	-0.011	1.002	-0	0.805
Physical competence	0.003	1.006	-0	0.818
Teacher expectations for educ. attainment:				
Graduate degree	0.155	0.362	0	0.660
Bachelors' degree	0.358	0.480	0	0.660
High school degree	0.425	0.495	0	0.660
High school dropout	0.062	0.241	0	0.660
<i>Panel F: School Quality</i>				
Percent FRPL	0.000	0.000	0	0.000
Student-teacher ratio	0.000	0.000	0	0.000
Average math and reading achievement	0.000	0.000	0	0.000
Difference btwn math and reading achievement	0.000	0.000	0	0.000
Cohort slope of average achievement	0.000	0.000	0	0.000
<i>Panel G: Child Ability Measures</i>				
Math score percentile	58.477	29.214	60	0.000
Reading score percentile	55.386	28.907	54	0.004
Digit span score	14.246	3.718	14	0.041

Notes: All variables marked as indices are standardized to mean 0 and a standard deviation of 1 by year. All variables are taken from the first year in which we observe the child's over-confidence in reading or math. Except for the indicator that the child lives in a two-adult household, all variables in Panel C are reported by the child's primary caregiver. We identify two-parent households by whether a family has both a *head* and a *wife* in the main PSID.

Table B.7: The persistence of math over- and under-confidence (weighted)

	(1)	(2)
Panel A: Math over-confidence	0.179***	0.177***
	(0.042)	(0.041)
N	1747	
Sample mean	0.036	
Panel B: Math under-confidence	0.046	0.041
	(0.033)	(0.033)
N	1747	
Sample mean	0.065	
Panel C: Math confidence (SD units)	0.260***	0.256***
	(0.039)	(0.040)
N	1747	
Sample mean	-0.008	
Basic controls:	✓	✓
Added background controls:		✓

Notes: This table regresses adolescent confidence outcomes on childhood math confidence with various controls. All controls are the same as described in Table 1. Observations are weighted so that the analysis sample matches the racial makeup of the US population in the 1990 census and so that the distribution of math percentile scores is uniform by decile, and the distribution of income is uniform by quartile. All controls that are indices are normalized relative to the weighted distribution. Standard errors are clustered by family, and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.8: Childhood math confidence and medium-term educational achievement and attainment (weighted)

Dependent variable:	Adolescent math scores		Adolescent reading scores		High school degree		College degree	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	5.103** (2.349)	5.039** (2.318)	1.553 (2.082)	1.872 (2.028)	0.104*** (0.039)	0.110*** (0.039)	0.053 (0.035)	0.067* (0.035)
Under-confidence	-4.565** (1.995)	-4.108** (2.008)	1.046 (1.891)	0.901 (1.922)	0.042* (0.023)	0.042* (0.023)	-0.011 (0.043)	-0.007 (0.041)
N	1747	1747	1745	1745	2714	2714	2725	2725
OC = -1*UC? p-value:	0.861	0.759	0.355	0.302	0.002	0.001	0.446	0.264
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	2.871*** (0.826)	2.684*** (0.841)	-0.204 (0.832)	-0.020 (0.834)	0.027 (0.017)	0.026 (0.017)	0.017 (0.015)	0.016 (0.015)
N	1747	1747	1745	1745	2714	2714	2725	2725
Sample mean of dep. var.	46.836	46.836	46.068		0.868		0.270	
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses educational achievement and attainment outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Observations are weighted so that our sample matches population shares in quintiles of income, in race categories, and in deciles of nationally-normed WJ-R math percentile scores. All controls that are indices are normalized relative to the weighted distribution. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.9: Childhood math confidence and college quality, college major choice, and post-college schooling (weighted)

Dependent variable:	College quality index		College's 75th pctile math SAT score		STEM major		Graduate degree	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	0.035 (0.165)	0.083 (0.167)	11.390 (11.115)	10.273 (11.569)	0.167 (0.122)	0.211* (0.128)	-0.030 (0.060)	-0.029 (0.070)
Under-confidence	-0.118 (0.106)	-0.132 (0.106)	-12.123* (7.178)	-12.976* (6.990)	-0.121** (0.054)	-0.113** (0.050)	-0.046 (0.059)	-0.071 (0.061)
N	1107	1107	1117	1117	736	736	810	810
OC = -1*UC? p-value:	0.689	0.813	0.957	0.845	0.732	0.472	0.359	0.281
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	0.145** (0.066)	0.137** (0.064)	11.005*** (3.997)	10.256*** (3.889)	0.086** (0.035)	0.074** (0.035)	0.027 (0.032)	0.029 (0.030)
N	1107	1107	1117	1117	736	736	810	810
Sample mean of dep. var.	0.091		599.785		0.186		0.180	
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses college outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the early years observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Observations are weighted so that our sample matches population shares in quintiles of income, in race categories, and in deciles of nationally-normed WJ-R math percentile scores. All controls that are indices are normalized relative to the weighted distribution. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.10: Childhood math confidence and employment outcomes (weighted)

Dependent variable:	Works in STEM		Non-STEM high-educ occ.		Ln(Earnings)		Unemployed this year	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	0.067 (0.054)	0.064 (0.046)	-0.020 (0.035)	-0.020 (0.034)	-0.015 (0.135)	-0.006 (0.136)	-0.012 (0.036)	-0.009 (0.036)
Under-confidence	-0.063** (0.025)	-0.060*** (0.023)	0.018 (0.040)	0.021 (0.038)	-0.013 (0.074)	-0.018 (0.072)	-0.019 (0.020)	-0.024 (0.021)
N	4592	4592	4592	4592	4423	4423	4975	4975
OC = -1*UC? p-value:	0.949	0.949	0.965	0.990	0.859	0.873	0.449	0.436
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	0.042*** (0.016)	0.040*** (0.014)	-0.008 (0.016)	-0.006 (0.015)	-0.009 (0.047)	-0.001 (0.046)	-0.014 (0.013)	-0.013 (0.012)
N	4592	4592	4592	4592	4423	4423	4975	4975
Sample mean of dep. var.	0.049		0.151		10.180		0.136	
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses employment outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the early years observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Basic controls also include year fixed effects when the outcome is observed in a panel. Observations are weighted so that our sample matches population shares in quintiles of income, in race categories, and in deciles of nationally-normed WJ-R math percentile scores. All controls that are indices are normalized relative to the weighted distribution. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.11: Demographic predictors of over- and under-confidence (decile coefficients)

	Over-confidence	Under-confidence	Confidence (sd)
<i>Math score deciles</i>			
Decile 1	0.147*** (0.04)	-0.257*** (0.03)	1.918*** (0.09)
Decile 2	0.212*** (0.03)	-0.257*** (0.03)	1.767*** (0.07)
Decile 3	0.204*** (0.03)	-0.246*** (0.03)	1.564*** (0.07)
Decile 4	0.140*** (0.02)	-0.250*** (0.02)	1.231*** (0.06)
Decile 5	0.189*** (0.02)	-0.250*** (0.02)	1.075*** (0.05)
Decile 6	-0.001 (0.01)	-0.163*** (0.03)	0.945*** (0.06)
Decile 7	-0.003 (0.01)	-0.161*** (0.03)	0.675*** (0.06)
Decile 8	-0.004 (0.01)	0.028 (0.03)	0.250*** (0.05)
Decile 9	0.004 (0.01)	0.049 (0.03)	0.064 (0.05)
Decile 10	0.000 (.)	0.000 (.)	0.000 (.)
<i>Reading score deciles</i>			
Decile 1	0.160** (0.07)	-0.026 (0.08)	0.281 (0.21)
Decile 2	0.074 (0.07)	-0.021 (0.08)	0.114 (0.21)
Decile 3	0.001 (0.07)	-0.030 (0.08)	-0.006 (0.20)
Decile 4	0.021 (0.07)	-0.034 (0.08)	0.043 (0.20)
Decile 5	0.001 (0.07)	-0.001 (0.08)	-0.098 (0.20)
Decile 6	-0.012 (0.07)	0.006 (0.08)	-0.157 (0.20)
Decile 7	0.002 (0.07)	0.010 (0.08)	-0.197 (0.20)
Decile 8	-0.000 (0.07)	0.041 (0.08)	-0.240 (0.20)
Decile 9	-0.018 (0.07)	0.051 (0.08)	-0.277 (0.20)
Decile 10	-0.019 (0.07)	0.093 (0.08)	-0.301 (0.20)
Mean of dependent variable	0.085	0.121	0.000
N	2985	2985	2985
R-squared	0.21	0.21	0.57

Notes: This table shows the coefficients on math and reading test score decile fixed effects that are not included in Table 2.2 due to space constraints.

Table B.12: Benchmarking the relationships between confidence and long-run outcomes and test scores

	Math Score (1)	Reading Score (2)	HS grad (3)	College grad (4)	College quality ind. (5)	College SAT 75p (6)	STEM major (7)	Grad degree (8)	STEM occup (9)	High-educ occup (10)	ln(Earnings) (11)	Unempl (12)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>												
Math over-confidence	3.618** (1.464)	0.544 (1.319)	0.054** (0.026)	0.038 (0.024)	0.070 (0.135)	5.079 (11.253)	0.117 (0.095)	-0.007 (0.083)	0.023 (0.017)	-0.030 (0.027)	0.033 (0.083)	-0.029 (0.029)
Math under-confidence	-6.482*** (1.475)	0.797 (1.407)	0.021 (0.017)	-0.056** (0.027)	-0.136* (0.080)	-11.503** (5.706)	-0.149*** (0.035)	0.005 (0.046)	-0.049*** (0.015)	0.023 (0.032)	-0.089 (0.057)	0.007 (0.016)
Math pctile/10	5.291*** (0.235)	1.289*** (0.250)	0.008*** (0.003)	0.029*** (0.004)	0.056*** (0.014)	4.068*** (1.107)	0.030*** (0.009)	0.015* (0.008)	0.008*** (0.003)	0.005 (0.004)	0.060*** (0.011)	-0.012*** (0.003)
Reading pctile/10	1.597*** (0.240)	6.387*** (0.241)	0.015*** (0.003)	0.022*** (0.004)	0.048*** (0.016)	3.850*** (1.082)	0.007 (0.008)	0.013 (0.008)	0.000 (0.002)	0.013*** (0.004)	0.014 (0.010)	-0.002 (0.003)
N	1747	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>												
Math confidence	2.963*** (0.563)	0.223 (0.560)	0.017* (0.010)	0.037*** (0.011)	0.060 (0.044)	4.795 (3.237)	0.076*** (0.022)	0.015 (0.025)	0.018*** (0.006)	0.004 (0.011)	0.058** (0.028)	-0.021** (0.009)
Math pctile/10	5.583*** (0.260)	1.357*** (0.274)	0.011*** (0.004)	0.034*** (0.005)	0.062*** (0.016)	4.563*** (1.260)	0.038*** (0.009)	0.019** (0.009)	0.009*** (0.003)	0.008 (0.005)	0.068*** (0.012)	-0.016*** (0.004)
Reading pctile/10	1.655*** (0.240)	6.399*** (0.242)	0.016*** (0.003)	0.023*** (0.004)	0.050*** (0.016)	3.962*** (1.089)	0.009 (0.008)	0.013 (0.008)	0.000 (0.002)	0.013*** (0.004)	0.015 (0.010)	-0.002 (0.003)
N	1747	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975

Notes: This table presents the same regressions as the even-numbered columns of Tables 2.3, 2.4, and 2.5, but replaces the math and reading test score decile fixed effects with linear terms for math and reading percentile scores divided by 10 (so that the coefficients can be interpreted in terms of increasing test scores by one decile). We use this table to benchmark the relationships between math confidence versus math test scores and long-term outcomes. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.13: Robustness to potential confounders

	Math Score (1)	Reading Score (2)	HS grad (3)	College grad (4)	College quality ind. (5)	College math SAT 75p (6)	STEM major (7)	Grad degree (8)	STEM occup (9)	High-educ occup (10)	ln(Earnings) (11)	Unempl (12)
Section 1: Controlling for childhood Big 5 personality traits												
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>												
Over-confidence	2.600*	-0.286	0.061**	0.025	-0.001	1.075	0.068	0.060	0.014	-0.030	0.075	-0.033
	(1.483)	(1.393)	(0.026)	(0.025)	(0.150)	(12.165)	(0.104)	(0.087)	(0.017)	(0.027)	(0.086)	(0.030)
Under-confidence	-5.808***	0.187	0.023	-0.060**	-0.123	-11.110*	-0.160***	0.017	-0.048***	0.025	-0.074	0.007
	(1.513)	(1.464)	(0.017)	(0.028)	(0.083)	(5.961)	(0.037)	(0.048)	(0.016)	(0.032)	(0.058)	(0.017)
N	1747	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>												
Confidence	2.824***	0.168	0.014	0.029***	0.033	3.194	0.077***	0.014	0.018***	-0.002	0.054*	-0.022**
	(0.569)	(0.579)	(0.010)	(0.011)	(0.046)	(3.442)	(0.024)	(0.025)	(0.006)	(0.012)	(0.029)	(0.010)
N	1745	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975
Section 2: Controlling for parent and teacher expectations and investments												
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>												
Over-confidence	2.861*	-0.274	0.059**	0.028	0.061	5.025	0.089	0.050	0.012	-0.023	0.064	-0.037
	(1.487)	(1.385)	(0.025)	(0.024)	(0.153)	(12.230)	(0.099)	(0.087)	(0.017)	(0.026)	(0.085)	(0.030)
Under-confidence	-5.295***	0.398	0.015	-0.055**	-0.081	-9.065	-0.150***	0.007	-0.048***	0.031	-0.106*	0.011
	(1.520)	(1.460)	(0.017)	(0.028)	(0.081)	(5.946)	(0.038)	(0.049)	(0.015)	(0.033)	(0.059)	(0.018)
N	1747	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>												
Confidence	2.761***	0.165	0.017*	0.031***	0.044	4.228	0.078***	0.026	0.016***	0.001	0.065**	-0.025**
	(0.582)	(0.583)	(0.010)	(0.011)	(0.046)	(3.461)	(0.024)	(0.026)	(0.006)	(0.012)	(0.029)	(0.010)
N	1745	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975
Section 3: Controlling for elementary school quality												
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>												
Over-confidence	2.938*	-0.195	0.060**	0.028	0.030	3.039	0.076	0.028	0.016	-0.024	0.060	-0.033
	(1.511)	(1.408)	(0.026)	(0.025)	(0.150)	(11.908)	(0.097)	(0.086)	(0.017)	(0.026)	(0.086)	(0.030)
Under-confidence	-5.931***	-0.097	0.020	-0.062**	-0.143*	-12.553**	-0.159***	0.006	-0.047***	0.026	-0.062	0.005
	(1.499)	(1.460)	(0.018)	(0.028)	(0.082)	(5.892)	(0.037)	(0.049)	(0.016)	(0.033)	(0.058)	(0.017)
N	1747	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>												
Confidence	2.944***	0.251	0.018*	0.034***	0.041	3.577	0.073***	0.018	0.018***	0.001	0.054*	-0.022**
	(0.574)	(0.582)	(0.010)	(0.011)	(0.046)	(3.426)	(0.023)	(0.025)	(0.006)	(0.012)	(0.029)	(0.010)
N	1745	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975

Notes: This table presents the robustness of our main results to adding controls for potential confounders. In the first section, we add controls for measurements of children's big 5 personality traits taken at the same time as the confidence measurements. In the second section, we add controls for parent and teacher expectations and investment: how teachers rate the child's social, physical, and academic competency; whether parents report reading, playing sports, doing homework, playing games, expressing physical affection, and saying *I love you* more than once per week; and separate indicators for whether parents and teachers think the child will get a high school or bachelors degree. Finally, the third section adds controls for elementary school quality at the time of confidence measurement: the student-teacher ratio, the percent of students qualifying for free or reduced-price lunch, and three measures of school achievement from 2009-2018: the average math and reading score, the difference between math and reading scores, and the cohort slope on the average math and reading score. Each of these sets of controls is individually added to our main specification in the even-numbered columns of Tables 2.3, 2.4, and 2.5. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.14: Correlations between childhood confidence and personality measures

	Math over-confidence		Math under-confidence		Math confidence (sd)	
	(1)	(2)	(3)	(4)	(5)	(6)
Conscientiousness	-0.023** (0.011)	-0.024** (0.011)	0.025** (0.012)	0.028** (0.012)	-0.226*** (0.038)	-0.232*** (0.038)
Extroversion	0.000 (0.010)	-0.002 (0.009)	-0.001 (0.010)	0.000 (0.010)	0.027 (0.032)	0.022 (0.032)
Neuroticism	0.014 (0.012)	0.013 (0.012)	0.005 (0.013)	0.004 (0.013)	0.046 (0.042)	0.046 (0.042)
Agreeableness	-0.006 (0.015)	-0.009 (0.015)	-0.011 (0.014)	-0.007 (0.014)	0.075 (0.048)	0.068 (0.049)
Openness	-0.017 (0.013)	-0.018 (0.013)	0.010 (0.011)	0.011 (0.011)	-0.160*** (0.041)	-0.163*** (0.041)
General Confidence	—	0.020** (0.008)	—	-0.033*** (0.009)	—	0.070** (0.028)
R-squared	0.009	0.023	0.005	0.013	0.026	0.029
N	2985	2985	2985	2985	2985	2985

Note: This table shows the relationship between math confidence and measures of childhood personality and general confidence. In columns 1 and 2 the outcome is our main binary measure of over- or under-confidence, respectively. In column 3, the outcome is our measure of the degrees of confidence that takes on values of -6 to 6, *standardized to have mean zero and standard deviation one in our sample*. All independent variables are recoded to zero if missing and we include a missing indicator (coefficient not shown). Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.15: Parent and teacher predictors of math over- and under-confidence

	Math over-confidence			Math under-confidence			Confidence (sd)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<i>Teacher Perceptions</i>									
Academic competence	0.015 (0.01)	0.017 (0.01)	0.016 (0.01)	-0.019 (0.02)	-0.019 (0.02)	-0.020 (0.02)	-0.006 (0.04)	-0.002 (0.04)	-0.008 (0.04)
Social competence	-0.019 (0.01)	-0.016 (0.01)	-0.015 (0.01)	0.017 (0.02)	0.016 (0.02)	0.015 (0.02)	-0.029 (0.04)	-0.024 (0.04)	-0.020 (0.04)
Physical competence	-0.015 (0.01)	-0.019 (0.01)	-0.018 (0.01)	-0.002 (0.02)	-0.000 (0.02)	-0.001 (0.02)	-0.026 (0.04)	-0.038 (0.04)	-0.037 (0.04)
<i>Teacher Expectations</i>									
Expects grad degree	0.002 (0.02)	0.003 (0.02)	0.002 (0.02)	-0.073* (0.04)	-0.073* (0.04)	-0.052 (0.04)	0.153** (0.07)	0.157** (0.07)	0.134* (0.07)
Expects bachelors' degree	0.010 (0.02)	0.009 (0.02)	0.012 (0.02)	0.041 (0.03)	0.040 (0.03)	0.048* (0.03)	0.047 (0.06)	0.049 (0.06)	0.051 (0.06)
<i>Parent Investment</i>									
Reads to child	0.008 (0.01)	0.008 (0.01)	0.006 (0.01)	0.024* (0.01)	0.024* (0.01)	0.025* (0.01)	-0.016 (0.03)	-0.014 (0.03)	-0.016 (0.03)
Art with child	-0.035* (0.02)	-0.028 (0.02)	-0.031 (0.02)	-0.015 (0.02)	-0.020 (0.02)	-0.019 (0.02)	-0.060 (0.05)	-0.043 (0.05)	-0.048 (0.05)
Sports with child	0.006 (0.01)	0.004 (0.01)	0.006 (0.01)	-0.030* (0.02)	-0.026 (0.02)	-0.027* (0.02)	0.077** (0.04)	0.068* (0.04)	0.076** (0.04)
Homework with child	0.007 (0.01)	0.005 (0.01)	0.006 (0.01)	0.025* (0.01)	0.026* (0.01)	0.026* (0.01)	-0.022 (0.03)	-0.028 (0.03)	-0.029 (0.03)
Games with child	0.005 (0.02)	0.004 (0.02)	0.003 (0.02)	-0.030* (0.02)	-0.032** (0.02)	-0.033** (0.02)	0.070* (0.04)	0.071* (0.04)	0.068* (0.04)
Physical affection to child	0.028 (0.03)	0.029 (0.03)	0.027 (0.03)	-0.039 (0.04)	-0.039 (0.04)	-0.037 (0.03)	0.074 (0.09)	0.083 (0.09)	0.078 (0.09)
Says I love you to child	-0.016 (0.02)	-0.011 (0.02)	-0.012 (0.02)	0.010 (0.02)	0.009 (0.02)	0.012 (0.02)	-0.008 (0.04)	0.010 (0.05)	0.009 (0.05)
<i>Parent Expectations</i>									
Expects grad degree	-0.016 (0.02)	-0.015 (0.02)	-0.015 (0.02)	-0.043** (0.02)	-0.044** (0.02)	-0.020 (0.02)	0.036 (0.04)	0.036 (0.04)	0.019 (0.05)
Expects bachelors' degree	-0.003 (0.01)	-0.001 (0.01)	0.000 (0.01)	0.003 (0.01)	0.003 (0.01)	0.014 (0.01)	0.024 (0.03)	0.032 (0.03)	0.029 (0.03)
Digit span score	0.000 (0.00)	0.000 (0.00)	-0.000 (0.00)	-0.004* (0.00)	-0.003* (0.00)	-0.003* (0.00)	0.009** (0.00)	0.008** (0.00)	0.009** (0.00)
General confidence	0.037*** (0.01)	0.037*** (0.01)	0.038*** (0.01)	-0.054*** (0.01)	-0.056*** (0.01)	-0.054*** (0.01)	0.208*** (0.02)	0.212*** (0.02)	0.210*** (0.02)
Added demographic controls		✓	✓		✓	✓		✓	✓
Added all other Table 2 controls			✓			✓			✓
N	2985	2985	2985	2985	2985	2985	2985	2985	2985
R-squared	0.20	0.21	0.22	0.20	0.21	0.23	0.57	0.57	0.58

Notes: All variables are taken from the first year in which we observe the child's confidence in math. Teacher and parent expectations are indicators for each adult's expected educational attainment for each child, and the omitted category is expecting a child to obtain a high school degree or less. Parent investment controls are indicators for doing each activity more than once per week. Teacher perceptions of competence in each domain are standardized to have mean zero and standard deviation one based on teacher reports of whether a child is extremely competent to not at all competent on a four-point scale. Additional controls in all columns include math and reading test score decile fixed effects, birth year, birth quarter, state, and age at which confidence was measured fixed effects. All controls are recoded to be zero if missing and the regressions include missing indicators for each variable (not shown). Standard errors are clustered by family.

Table B.16: Robustness to definitions of confidence

	Math Score (1)	Reading Score (2)	HS grad (3)	College grad (4)	College quality (5)	College math SAT 75p (6)	STEM major (7)	Grad degree (8)	STEM occup (9)	High-educ occup (10)	ln(Earn) (11)	Unempl (12)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>												
Main measure:												
Over-confidence	2.666*	-0.286	0.062**	0.031	0.037	3.133	0.076	0.032	0.014	-0.025	0.064	-0.035
	(1.496)	(1.385)	(0.026)	(0.024)	(0.148)	(11.829)	(0.097)	(0.087)	(0.017)	(0.026)	(0.085)	(0.030)
Under-confidence	-5.860***	0.162	0.022	-0.058**	-0.127	-11.312*	-0.162***	0.006	-0.049***	0.026	-0.075	0.005
	(1.497)	(1.452)	(0.017)	(0.028)	(0.082)	(5.925)	(0.036)	(0.048)	(0.016)	(0.033)	(0.057)	(0.017)
More strict (1):												
Over-confidence	2.412	0.091	0.034	0.016	0.052	9.116	-0.154	-0.042	0.019	-0.071***	0.042	-0.020
	(1.717)	(1.609)	(0.036)	(0.027)	(0.190)	(13.872)	(0.121)	(0.115)	(0.022)	(0.024)	(0.105)	(0.038)
Under-confidence	-7.246***	1.150	0.012	-0.066**	-0.144	-11.660*	-0.183***	-0.014	-0.056***	-0.009	-0.023	0.004
	(1.668)	(1.629)	(0.020)	(0.032)	(0.096)	(6.749)	(0.039)	(0.054)	(0.018)	(0.037)	(0.064)	(0.020)
More strict (2):												
Over-confidence	2.631*	-0.290	0.062**	0.030	0.038	3.173	0.075	0.032	0.014	-0.025	0.063	-0.035
	(1.495)	(1.384)	(0.026)	(0.024)	(0.148)	(11.827)	(0.097)	(0.087)	(0.017)	(0.026)	(0.085)	(0.030)
Under-confidence	-6.820***	1.046	0.021	-0.058*	-0.152	-11.870*	-0.165***	-0.012	-0.055***	-0.016	-0.010	0.001
	(1.622)	(1.595)	(0.019)	(0.031)	(0.093)	(6.464)	(0.039)	(0.052)	(0.017)	(0.035)	(0.061)	(0.019)
Less strict (1):												
Over-confidence	1.344	-0.161	0.043*	0.037	-0.000	0.537	0.066	0.030	0.014	-0.033	0.102	-0.028
	(1.406)	(1.300)	(0.024)	(0.023)	(0.126)	(9.982)	(0.073)	(0.076)	(0.015)	(0.025)	(0.076)	(0.026)
Under-confidence	-5.859***	-1.266	0.005	-0.069***	-0.135*	-12.247**	-0.128***	0.022	-0.048***	-0.003	-0.055	0.004
	(1.311)	(1.255)	(0.017)	(0.024)	(0.074)	(5.400)	(0.035)	(0.043)	(0.012)	(0.026)	(0.049)	(0.016)
Less strict (2):												
Over-confidence	1.873	0.411	0.057*	0.016	0.106	11.714	-0.038	0.012	0.011	-0.031	0.064	-0.015
	(1.607)	(1.570)	(0.030)	(0.024)	(0.156)	(12.051)	(0.081)	(0.102)	(0.017)	(0.025)	(0.089)	(0.031)
Under-confidence	-5.895***	-1.265	0.004	-0.071***	-0.136*	-12.345**	-0.128***	0.021	-0.048***	-0.002	-0.057	0.005
	(1.309)	(1.253)	(0.017)	(0.024)	(0.074)	(5.408)	(0.035)	(0.043)	(0.012)	(0.026)	(0.048)	(0.015)
Relative:												
Over-confidence	2.795*	2.308*	0.041*	-0.004	0.004	-5.672	0.066	0.121*	0.000	-0.030	0.049	-0.023
	(1.480)	(1.361)	(0.023)	(0.023)	(0.122)	(9.029)	(0.062)	(0.067)	(0.010)	(0.024)	(0.070)	(0.022)
Under-confidence	-9.674***	-1.848	0.025	-0.066*	-0.049	-6.820	-0.160***	0.003	-0.051***	0.003	0.035	-0.017
	(2.081)	(1.956)	(0.022)	(0.034)	(0.105)	(7.755)	(0.049)	(0.064)	(0.016)	(0.040)	(0.069)	(0.022)
Continuous tails:												
Over-confidence	1.127	1.751	0.033	0.008	0.098	7.333	0.084	0.149*	0.011	-0.043**	0.055	-0.031
	(1.372)	(1.339)	(0.023)	(0.021)	(0.112)	(9.301)	(0.071)	(0.081)	(0.013)	(0.021)	(0.071)	(0.024)
Under-confidence	-6.680***	2.244	-0.001	-0.057*	-0.136	-12.011*	-0.191***	-0.017	-0.050***	-0.004	-0.046	0.017
	(1.664)	(1.655)	(0.021)	(0.032)	(0.100)	(6.948)	(0.038)	(0.052)	(0.018)	(0.036)	(0.067)	(0.021)
Main of averages:												
Over-confidence	-	-	0.008	0.036	-0.143	-11.712	-0.082	0.036	-0.015	-0.001	-0.071	0.002
			(0.041)	(0.032)	(0.144)	(14.373)	(0.095)	(0.122)	(0.012)	(0.033)	(0.124)	(0.043)
Under-confidence	-	-	0.017	-0.040	-0.087	-5.400	-0.116***	-0.049	-0.038**	0.028	-0.047	-0.035*
			(0.022)	(0.033)	(0.107)	(7.458)	(0.042)	(0.055)	(0.018)	(0.039)	(0.075)	(0.021)
Average of main:												
Over-confidence	-	-	0.057	0.023	0.031	1.166	0.036	0.110	0.005	-0.023	0.031	-0.016
			(0.036)	(0.028)	(0.177)	(15.192)	(0.117)	(0.120)	(0.016)	(0.030)	(0.107)	(0.038)
Under-confidence	-	-	0.012	-0.054	-0.062	-6.511	-0.186***	-0.084	-0.051**	0.030	-0.133	-0.019
			(0.023)	(0.036)	(0.113)	(7.963)	(0.057)	(0.065)	(0.023)	(0.041)	(0.082)	(0.022)

Table B.16: Robustness to definitions of confidence (continued)

	Math Score (1)	Reading Score (2)	HS grad (3)	College grad (4)	College quality (5)	College math SAT 75p (6)	STEM major (7)	Grad degree (8)	STEM occup (9)	High-educ occup (10)	ln(Earn) (11)	Unempl (12)
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>												
Main measure	2.827*** (0.569)	0.128 (0.580)	0.018* (0.010)	0.033*** (0.011)	0.041 (0.046)	3.631 (3.417)	0.078*** (0.023)	0.022 (0.025)	0.018*** (0.006)	0.001 (0.012)	0.059** (0.029)	-0.023** (0.009)
Percentiles	2.631*** (0.528)	0.356 (0.505)	0.016** (0.008)	0.027*** (0.010)	0.047 (0.038)	3.814 (2.803)	0.062*** (0.019)	0.006 (0.020)	0.017*** (0.006)	-0.005 (0.011)	0.045* (0.025)	-0.012 (0.008)
Empirical dist	2.525*** (0.487)	0.318 (0.497)	0.018** (0.008)	0.027*** (0.009)	0.027 (0.039)	2.517 (2.895)	0.065*** (0.020)	0.018 (0.022)	0.016*** (0.006)	0.006 (0.010)	0.050** (0.025)	-0.019** (0.008)
Main of averages	–	–	0.016 (0.010)	0.019* (0.011)	-0.048 (0.048)	-2.972 (3.633)	0.042* (0.025)	0.047* (0.026)	0.012* (0.007)	-0.007 (0.012)	0.001 (0.031)	-0.009 (0.010)
Average of main	–	–	0.017* (0.010)	0.018 (0.011)	-0.038 (0.048)	-2.434 (3.609)	0.047* (0.025)	0.040 (0.026)	0.010 (0.007)	-0.008 (0.012)	0.012 (0.031)	-0.009 (0.010)

Notes: This table presents the robustness of our main results to changing our definitions of our main measures of confidence. All regressions estimate our main specification in the even-numbered columns of Tables 2.3, 2.4, and 2.5, replacing the main measure of over-confidence and under-confidence or the main measure of degrees of confidence with alternate definitions. Sample sizes for each regression are the same as in each main table. In Panel A, we iterate over our definitions of binary over- and under-confidence variables. Each pair of over- and under-confidence measures are estimated in the same regression. The definitions labeled 'more strict' or 'less strict' change the self-assessment and percentile score cutoffs in our main measure. The relative measure identifies children who score in the top or bottom 25 percent of test scores within each self-assessment bucket as under- or over-confident, respectively. The 'continuous tails' measure identifies over-confident children as those whose more continuous measure of confidence is between 3 and 6, and under-confident children as those whose continuous measure of confidence is between -6 and -3. Finally, the last two measures combine data over the two waves of the CDS where we observe confidence measurements, when available. The first averages test scores and self-reports over the two waves and then applies our main cutoffs, and the second averages the main measure over the two waves. In Panel B, we iterate over our definitions of the more continuous measure of biased beliefs. The one labeled 'percentiles' differences the percentile of children's self-assessment and their percentile score, and the one labeled 'empirical dist' assumes that children knew the empirical distribution of self-reports and should have correspondingly reported their self-assessments (instead of assuming a uniform distribution). Again, the last two measures combine data over the two waves where possible: the first averages test scores and self-reports over the two waves and then applies the transformation to the same scale, and the second averages the main measure over the two waves. Further iteration is presented in specification charts for each outcome, found in Appendix Figures B.5-B.16. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.17: Sample means by whether missing confidence variables in the CDS

	Sample (1)	Non-Sample (2)	<i>p</i> -value (1)=(2)
<i>Panel A: Demographic Characteristics</i>			
Child is female	0.497 (2985)	0.458 (578)	0.089
Child is white	0.458 (2985)	0.481 (578)	0.303
Child is black	0.417 (2985)	0.370 (578)	0.036
Child is hispanic	0.075 (2985)	0.076 (578)	0.928
Child's birth order	1.625 (2645)	1.521 (457)	0.067
Child's birth year	1990.020 (2985)	1991.478 (573)	0.000
<i>Panel B: Parent and Family Characteristics</i>			
Father at least graduated high school	0.835 (1882)	0.811 (365)	0.256
Father at least has bachelors	0.257 (1883)	0.260 (365)	0.880
Mother at least graduated high school	0.817 (2657)	0.735 (499)	0.000
Mother at least has bachelors	0.134 (2648)	0.123 (496)	0.490
Total taxable family income (thous 2016 USD)	69.777 (2985)	61.977 (578)	0.032
HH lives in public housing	0.061 (2985)	0.076 (577)	0.178
HH receives food stamps	0.198 (2985)	0.213 (577)	0.404
Two adults in HH	0.645 (2985)	0.666 (578)	0.329
<i>Panel C: Other Child Characteristics</i>			
Child ever in gifted prog	0.243 (2893)	0.077 (568)	0.000
Child ever in special ed prog	0.127 (2888)	0.067 (568)	0.000
Child has repeated grade	0.122 (2921)	0.033 (568)	0.000
Child qualifies for FRP lunch	0.598 (2220)	0.510 (288)	0.005
Parent's rating of child health	0.014 (2969)	0.025 (572)	0.802
# Siblings in the HH	1.415 (2985)	1.178 (578)	0.000
<i>Big 5 personality scores (indices)</i>			
Conscientiousness	-0.003 (2964)	0.014 (153)	0.754
Extroversion	0.005 (2970)	0.060 (154)	0.282
Neuroticism	-0.006 (2957)	0.040 (152)	0.374
Agreeableness	-0.003 (2962)	0.025 (153)	0.605
Openness to experiences	-0.001 (2955)	-0.069 (151)	0.102
<i>Panel D: Child Ability Measures</i>			
Math score percentile	58.477 (2985)	49.256 (156)	0.000
Reading score percentile	55.386 (2973)	52.587 (63)	0.449
Digit span score	14.246 (2863)	7.403 (149)	0.000

Notes: This table regresses an indicator for whether a child is in our final sample on child characteristics. The sample is all 3563 children in the CDS survey. 578 children are dropped from our analysis sample. These are children for whom we never observe *both* a self-assessed and observed ability measure. Of those, 99 percent are missing a self-assessed measure and 73 percent are missing a math test score.

Table B.18: Correlations between math confidence and other attitudes

	Math			General Conf
	Over-Conf	Under-Conf	Confidence (sd)	
Panel A: Other Math Attitudes				
Math skill relative to peers	0.275	-0.307	0.534	0.242
Expected performance in math this year	0.189	-0.219	0.403	0.209
How good at learning new thing in math	0.152	-0.190	0.316	0.242
How easy is math for you	0.078	-0.095	0.166	0.044
How useful is what you learn in math	0.055	-0.056	0.130	0.207
Being good in math is important	0.063	-0.057	0.146	0.199
Working on math is interesting	0.157	-0.120	0.291	0.167
How much do you like math	0.240	-0.189	0.404	0.142
Panel B: Social and School Performance				
Do you feel like part of your school	-0.000	-0.036	0.054	0.213
Do you feel close to people at your school	0.001	-0.051	0.049	0.247

Note: This table shows the partial correlations between over- and under-confidence in math and general confidence and children's other attitudes towards math and social experiences at school after partialling out the relationship with math test score deciles.

Table B.19: Childhood math confidence and average employment outcomes from age 28-33

Dependent variable:	Works in STEM (non-health)		Non-STEM high-educ occ.		Ln(Earnings)		Unemployed this year	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>								
Over-confidence	0.011 (0.017)	0.015 (0.017)	-0.034 (0.030)	-0.041 (0.032)	0.122 (0.109)	0.132 (0.109)	-0.031 (0.035)	-0.035 (0.035)
Under-confidence	-0.039** (0.019)	-0.040** (0.019)	0.022 (0.039)	0.020 (0.039)	-0.027 (0.073)	-0.030 (0.075)	0.025 (0.022)	0.021 (0.022)
N	1301	1301	1301	1301	1269	1269	1364	1364
OC = -1*UC? p-value:	0.262	0.318	0.803	0.687	0.463	0.436	0.881	0.735
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>								
Confidence	0.017** (0.007)	0.017** (0.007)	0.003 (0.013)	0.004 (0.014)	0.084** (0.038)	0.089** (0.039)	-0.024* (0.013)	-0.024* (0.013)
N	1301	1301	1301	1301	1269	1269	1364	1364
Sample mean of dep. var.	0.043		0.167		10.227		0.141	
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓		✓		✓		✓

Notes: This table regresses employment outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.20: Comparing predictiveness of biased beliefs and Big-Five traits

	Math Score (1)	Reading Score (2)	HS grad (3)	College grad (4)	College quality ind. (5)	College SAT 75p (6)	STEM major (7)	Grad degree (8)	STEM occup (9)	High-educ occup (10)	ln(Earnings) (11)	Unempl (12)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>												
Math over-confidence	2.485* (1.484)	-0.440 (1.387)	0.061** (0.026)	0.023 (0.025)	-0.004 (0.147)	0.517 (12.009)	0.053 (0.102)	0.062 (0.087)	0.013 (0.017)	-0.033 (0.027)	0.072 (0.086)	-0.033 (0.030)
Math under-confidence	-5.633*** (1.495)	0.473 (1.465)	0.023 (0.017)	-0.055** (0.028)	-0.107 (0.081)	-10.609* (5.941)	-0.155*** (0.038)	0.021 (0.048)	-0.049** (0.016)	0.025 (0.032)	-0.078 (0.057)	0.008 (0.017)
Conscientiousness	-0.944 (0.913)	-0.026 (0.924)	0.021 (0.014)	0.051*** (0.015)	0.130** (0.064)	6.120 (4.777)	0.009 (0.039)	0.037 (0.037)	0.002 (0.010)	0.038** (0.016)	0.030 (0.045)	-0.004 (0.014)
Extroversion	-0.109 (0.712)	0.020 (0.695)	0.001 (0.011)	0.007 (0.012)	-0.069 (0.053)	-3.263 (4.101)	-0.057** (0.029)	0.011 (0.029)	-0.016** (0.007)	0.012 (0.013)	0.029 (0.033)	-0.019* (0.010)
Agreeableness	-0.611 (1.086)	-0.342 (1.084)	0.039** (0.018)	0.020 (0.017)	-0.052 (0.072)	-4.402 (5.848)	-0.041 (0.051)	0.039 (0.045)	-0.006 (0.011)	0.002 (0.019)	0.063 (0.048)	-0.002 (0.015)
Openness	-0.789 (0.922)	0.706 (0.950)	-0.036** (0.014)	-0.036** (0.015)	-0.035 (0.063)	-0.270 (4.818)	0.021 (0.038)	-0.078** (0.034)	-0.019** (0.009)	-0.027* (0.016)	-0.023 (0.042)	0.036*** (0.013)
Neuroticism	-0.640 (0.933)	0.549 (0.964)	0.015 (0.015)	0.051*** (0.017)	0.019 (0.066)	-1.238 (4.904)	-0.069** (0.032)	0.016 (0.034)	-0.005 (0.008)	0.009 (0.019)	-0.084* (0.045)	0.006 (0.014)
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>												
Math confidence	2.775*** (0.568)	0.080 (0.581)	0.014 (0.010)	0.027** (0.011)	0.030 (0.046)	3.262 (3.442)	0.076*** (0.025)	0.014 (0.025)	0.019*** (0.006)	-0.004 (0.012)	0.054* (0.028)	-0.023** (0.009)
Conscientiousness	-1.007 (0.914)	-0.042 (0.923)	0.021 (0.014)	0.051*** (0.016)	0.128** (0.064)	5.933 (4.749)	0.004 (0.040)	0.036 (0.038)	0.003 (0.010)	0.037** (0.016)	0.030 (0.045)	-0.003 (0.014)
Extroversion	-0.191 (0.713)	0.017 (0.697)	0.000 (0.011)	0.007 (0.012)	-0.070 (0.053)	-3.356 (4.100)	-0.059** (0.030)	0.010 (0.029)	-0.016** (0.007)	0.013 (0.013)	0.028 (0.032)	-0.018* (0.010)
Agreeableness	-0.684 (1.086)	-0.315 (1.082)	0.039** (0.018)	0.020 (0.017)	-0.053 (0.072)	-4.516 (5.796)	-0.042 (0.051)	0.039 (0.045)	-0.007 (0.012)	0.003 (0.019)	0.061 (0.048)	-0.002 (0.015)
Openness	-0.718 (0.925)	0.715 (0.950)	-0.036** (0.014)	-0.034** (0.015)	-0.035 (0.062)	-0.246 (4.826)	0.020 (0.038)	-0.076** (0.033)	-0.018** (0.009)	-0.028* (0.016)	-0.023 (0.042)	0.035*** (0.013)
Neuroticism	-0.700 (0.935)	0.552 (0.964)	0.015 (0.015)	0.051*** (0.017)	0.019 (0.066)	-1.215 (4.924)	-0.069** (0.032)	0.017 (0.034)	-0.005 (0.008)	0.009 (0.019)	-0.085* (0.045)	0.006 (0.014)

Notes: This table presents the estimates from Panel A of Table B.13 but includes the coefficients on the childhood personality measures. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.21: Math confidence and young adult social outcomes

Dependent variable:	In a romantic relationship		Mental health		Social anxiety		Drinks alcohol often		Dangerous behavior	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
<i>Panel A: Independent variables are binary measures of over- and under-confidence</i>										
Over-confidence	-0.060** (0.026)	-0.063** (0.026)	-0.038 (0.048)	-0.023 (0.048)	-0.040 (0.055)	-0.043 (0.055)	0.020 (0.014)	0.019 (0.014)	0.048 (0.043)	0.051 (0.044)
Under-confidence	-0.004 (0.023)	-0.001 (0.023)	-0.034 (0.034)	-0.036 (0.034)	-0.001 (0.044)	-0.000 (0.044)	0.013 (0.015)	0.015 (0.015)	-0.022 (0.029)	-0.022 (0.029)
N	10389	10389	10360	10360	10374	10374	10360	10360	10277	10277
OC = -1*UC? p-value:	0.064	0.066	0.218	0.314	0.553	0.535	0.113	0.102	0.621	0.584
<i>Panel B: Independent variable is degrees of over- and under-confidence in standard deviation units</i>										
Confidence	-0.005 (0.011)	-0.005 (0.010)	0.008 (0.017)	0.010 (0.017)	-0.016 (0.021)	-0.015 (0.021)	-0.001 (0.006)	-0.002 (0.006)	0.018 (0.015)	0.018 (0.015)
N	10389	10389	10360	10360	10374	10374	10360	10360	10277	10277
Sample mean of dep. var.	0.551	0.551	0.001	0.000	0.000	0.000	0.105	0.105	-0.002	-0.002
Basic controls:	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Added background controls:		✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table regresses social (placebo) outcomes on childhood biased beliefs with various controls. Biased beliefs are measured in the earliest observed wave in the CDS with non-missing test scores and self-assessed ability. In Panel A, the outcome is regressed on an indicator for over-confidence, an indicator for under-confidence and our basic set of controls (in odd-numbered columns) and our extended set of controls (in even-numbered columns). The p-value listed tests whether the coefficient on the over-confidence indicator is equal to -1 times the coefficient on the under-confidence indicator. In Panel B, the outcome is regressed on our more continuous measure of biased beliefs which has been standardized to have mean zero and standard deviation one in our sample and the same sets of controls. All controls are the same as described in Table 2.1, minus the controls for adolescent test score deciles. Basic controls also include year fixed effects when the outcome is observed in a panel. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table B.22: Heterogeneity by over- and under-confidence using the degrees of confidence measure

	Math Score (1)	Reading Score (2)	HS grad (3)	College grad (4)	College quality (5)	College math SAT 75p (6)	STEM major (7)	Grad degree (8)	STEM occup (9)	High-educ occup (10)	ln(Earnings) (11)	Unempl (12)
Math confidence (sd)	3.100*** (1.042)	0.905 (1.018)	0.022 (0.016)	0.047*** (0.019)	-0.002 (0.065)	-0.714 (4.979)	0.015 (0.039)	0.009 (0.037)	0.012 (0.010)	0.034* (0.020)	0.100*** (0.042)	-0.029** (0.014)
Confidence (sd)*Over	-1.662 (1.875)	-3.046 (1.889)	0.011 (0.046)	-0.023 (0.035)	-0.091 (0.234)	-6.059 (18.494)	-0.193 (0.142)	-0.232* (0.132)	-0.001 (0.024)	-0.055* (0.030)	-0.198 (0.146)	0.025 (0.044)
Confidence (sd)*Under	-0.254 (3.621)	-1.621 (4.425)	-0.067 (0.041)	0.008 (0.084)	-0.279 (0.277)	-12.113 (17.160)	0.108 (0.089)	-0.097 (0.159)	0.010 (0.030)	-0.133 (0.097)	-0.033 (0.176)	-0.022 (0.062)
Math over-confidence	-0.389 (2.999)	5.185* (3.074)	-0.010 (0.069)	-0.016 (0.052)	0.225 (0.334)	16.374 (27.290)	0.346 (0.218)	0.472** (0.227)	-0.002 (0.034)	-0.003 (0.042)	0.226 (0.220)	-0.033 (0.066)
Math under-confidence	-3.173 (6.610)	0.584 (7.752)	-0.091 (0.072)	0.015 (0.147)	-0.620 (0.471)	-33.718 (29.471)	0.013 (0.144)	-0.172 (0.270)	-0.018 (0.058)	-0.193 (0.171)	0.014 (0.302)	-0.055 (0.106)
N	1747	1745	2714	2725	1107	1117	736	810	4592	4592	4423	4975
$Slope_{Over} = Slope_{Under}$	0.712	0.760	0.181	0.727	0.589	0.802	0.065	0.500	0.762	0.428	0.454	0.522
$Slope_{Over} = Slope_{Under} = Slope_{Neither}$	0.674	0.266	0.233	0.801	0.574	0.752	0.171	0.192	0.943	0.101	0.396	0.786

Notes: This table estimates our main specification for our more continuous degrees of confidence measure in standard deviation units (found in the even-numbered columns of Panel B in Tables 2.3, 2.4, and 2.5), but adds indicators for being over- and under-confident according to this measure and the interactions between the degrees of confidence measure and the indicators. Any student whose difference between their self-assessed bin from 1-7 and the bin they should have reported given their test scores is greater than two is considered over-confident and any student whose difference is less than negative two is considered under-confident. Standard errors are clustered at the family level and included in parentheses below each estimate. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

B.2 Constructing Indices

College quality measures:

Using restricted data from the TAS, we identify the first college that each child in our sample attended, if they attended college. We then match those schools to college quality data from the first year that they attended the college. Following [Cohodes and Goodman \(2014\)](#), we construct an index of college quality by taking the first component from a principal component analysis of colleges' 75th-percentile math SAT scores among incoming freshmen, graduation rates, and per-pupil instructional expenditures, separately by year from 2005-2019.

We impute SAT scores where possible to increase our sample size; some schools report 75th-percentile math ACT scores but not SAT scores. For those schools, we impute 75th-percentile math SAT scores as predicted values from a regression of 75th-percentile SAT scores on 75th-percentile ACT scores among schools with both measures. We also use the 6-year graduation rate rather than the 4-year graduation rate because the 6-year rate is available for more schools and the two measures are highly correlated.

Depending on the year, the first principal component captures 70-80 percent of the variation between these three variables and assigns nearly equal weights to all three variables in all years. We standardize the first principal component to have mean 0 and standard deviation 1 in the full sample of four-year colleges in the US by year, and call this our college quality index.

Secondary outcome variables:

To minimize the number of outcomes and controls include in our analysis, we create many indices of similar variables. Here we list each index with its underlying variables. All underlying variables are scales from 1-3, 1-4, 1-5, or 1-7. When the variable applies to parents, we standardize by year before taking the average. When the variable applies to children, we standardize by year and age group (8-11, 12-14, and 15-19).

Adult math confidence:

- How good would you be in a career that required you to use math?
- How good would you be in a career that required you to use physical science or technology?

Adult reading confidence:

- How good would you be in a career that required you to read and write a lot?

Adult general academic confidence:

- How good would you be in a career that required you to be creative?
- How good are you at solving problems you encounter?
- How good are you at logical, analytic thinking?
- How intelligent are you, compared to others?
- How good are you at listening to and understanding others?
- How good are you at teaching and explaining to others?

Adult career confidence:

- How successful do you think you could be in the type of job you most want?
- How likely do you think you are to end up in the job you most want at age 30?

Adult general confidence: (variables marked with a * are flipped so that a higher score means more confident)

- How confident are you, compared with others?

- How decisive are you, compared with others?
- How independent are you, compared with others?
- How good are you at being a leader?
- How good are you at supervising others?
- How often do you feel discouraged about the future*
- How often, in the last month, did you feel that you had something important to contribute to society?
- How often, in the last month, did you feel good at managing responsibilities of daily life?
- How often, in the last month, did you feel confident to think or express your own ideas and opinions?
- How often, in the last month, did you feel that your life had a direction or purpose?

Adult Big-Five personality measures: (variables marked with a * are flipped so that a higher score means more conscientious, agreeable, etc.)

- Conscientiousness:
 - You are someone who does a thorough job. Does this describe you not at all, a little, some, or a lot?
 - You are someone who tends to be lazy. Does this...?*
 - You are someone who does things efficiently. Does this...?
- Agreeableness:
 - You are someone who is sometimes rude to others. Does this...?*
 - You are someone who has a forgiving nature. Does this...?
 - You are someone who is considerate and kind to almost everyone. Does this...?
- Extroversion:
 - You are someone who is talkative. Does this...?
 - You are someone who is outgoing, sociable. Does this...?
 - You are someone who is reserved. Does this...?*
- Neuroticism:

- You are someone who worries a lot. Does this...?
 - You are someone who gets nervous easily. Does this...?
 - You are someone who is relaxed, handles stress well. Does this...?*
- Openness to experience:
 - You are someone who is original, comes up with new ideas. Does this...?
 - You are someone who values artistic experiences. Does this...?
 - You are someone who has an active imagination. Does this...?

Dangerous behavior index:

- How often in the last 6 months did you do something you knew was dangerous just for the thrill of it?
- How often in the last 6 months did you damage public or private property?
- How often in the last 6 months did you get into a physical fight?
- How often in the last 6 months did you drive when you were drunk or high on drugs?
- How often in the last 6 months did you ride with a driver who had too much to drink?

Control variables:

Child general confidence:

- Does the statement never, sometimes, always apply to you ... I do things as well as most people
- ... When I do something, I do it well
- ... I'm as good as most other people
- ... A lot of things about me are good
- ... I have a lot to be proud of

Child Big-Five personality measures, reported by primary caregiver: (variables marked with a * are flipped so that a higher score means more conscientious, agreeable, etc.)

- Conscientiousness:

- According to [child’s] behavior, [he/she] cheats or tells lies*
 - ... [he/she] has difficulty concentrating, cannot pay attention for long*
 - Thinking about [child], tell me if [child] waits [his/her] turn in games and other activities
 - ... tell me if [child] does neat, careful work
 - ... tell me if [child] usually does what you tell [him/her] to do
- Agreeableness:
 - ... [he/she] argues too much*
 - ... [he/she] bullies or is cruel or mean to others*
 - ... [he/she] is disobedient*
 - ... [he/she] has trouble getting along with other children*
 - ... [he/she] is stubborn, sullen, or irritable*
 - ... [he/she] breaks things on purpose or deliberately destroys [his/her] own or another’s things*
 - ... tell me if [child] is cheerful, happy
 - ... tell me if [child] gets along well with other children
 - ... tell me if [child] is admired and well-liked by other children
- Extroversion:
 - ... [he/she] is withdrawn, does not get involved with others*
 - ... [he/she] demands a lot of attention
- Neuroticism:
 - ... [he/she] has sudden changes in mood or feeling
 - ... [he/she] is rather high strung, tense and nervous
 - ... [he/she] is too fearful or anxious
 - ... [he/she] has a lot of difficulty getting [his/her] mind off certain thoughts
 - ... [he/she] feels others are out to get [him/her]
 - ... [he/she] worries too much
 - ... tell me if [child] can get over being upset quickly*
- Openness to experience:

- ... [he/she] is impulsive, or acts without thinking
- ... [he/she] clings to adults*
- ... [he/she] hangs around with kids who get into trouble
- ... tell me if [child] is curious and exploring, likes new experiences

Parent adherence to traditional gender norms: (variables marked with a * are flipped so that a higher score means more traditional gender norms)

- Most of the important decisions in the life of the family should be made by the man of the house
- Women are much happier if they stay at home and take care of their children
- There is some work that is men's and some that is women's and they should not be doing each other's
- It is much better for everyone if the man earns the living and the woman takes care of the home and family
- It is more important for a wife to help her husband's career than to have one herself
- Preschool children are likely to suffer if their mother is employed
- An employed mother can establish as warm and secure a relationship with her children as a mother who is not employed*
- Parents should encourage just as much independence in their daughters as their sons*
- A father should be as heavily involved in the care of his child as the mother*
- If a husband and wife both work full-time, they should share household tasks equally*

Parent aggravation in parenting:

- Thinking about [child], there are some things that [he/she] does that really bother me a lot
- ... I find myself giving up more of my life to meet [child's] needs than I ever expected
- ... I often feel angry with [child]
- Thinking about my child[ren], being a parent is harder than I thought it would be
- ... I feel trapped by my responsibilities as a parent
- ...I find that taking care of my child[ren] is much more work than pleasure

- ...I often feel tired, worn out, or exhausted from raising a family

Parent self-esteem: (variables marked with a * are flipped so that a higher score means higher self-esteem)

- I feel that I'm a person of worth, at least on an equal basis with others
- I feel that I have a number of good qualities
- All in all, I am inclined to feel that I am a failure*
- I am able to do things as well as most other people
- I feel I do not have much to be proud of*
- I take a positive attitude toward myself
- On the whole, I am satisfied with myself
- I wish I could have more respect for myself*
- I certainly feel useless at times*
- At times I think I am no good at all*

Parent self-efficacy: (variables marked with a * are flipped so that a higher score means higher self-efficacy)

- There is really no way I can solve some of the problems I have*
- Sometimes I feel that I'm being pushed around in life*
- I have little control over the things that happen to me*
- I can do just about anything I really set my mind to
- I often feel helpless in dealing with the problems of life*
- What happens to me in the future most depends on me
- There is little I can do to change many of the important things in my life*

B.3 Biased Beliefs and Other Attitudes Towards School

Our confidence measures consistently correlate with children's other attitudes towards math in ways we would expect. Appendix Table B.18 shows the pairwise correlations between children's attitudes towards math and school, our measures of over- and under-confidence in math, and the general confidence index. Over-confidence in math is positively correlated with children's self-assessed ability relative to their peers, their expected performance in math that year, how good they think they are at learning a new skill in math, how interesting they think math is, and how much they like math ($\rho \in [0.15, 0.28]$). All of the same correlations are negative and of similar magnitude for children who are under-confident in math. The correlations between over- and under-confidence and how easy, useful, or important math is are much smaller in magnitude but have the expected signs. There are very similar patterns using the more continuous measure of math confidence. General confidence is also positively correlated with these attitudes ($\rho \in [0.14, 0.24]$), except for how easy a child thinks math is.

On the other hand, whether children report feeling like part of their school community or close to their peers are both uncorrelated with math over- or under-confidence ($\rho < |0.05|$), but are positively correlated with our index of general confidence ($\rho \approx 0.23$). Together, these patterns suggest that our measures isolate over- and under-confidence in the particular domain of math, but our regressions also control for general confidence and other measures of child ability to further isolate the relationship between children's biased beliefs about their math ability and their medium- and long-run outcomes.

B.4 Over- versus under-confidence

One ex-ante strength of our binary measures of biased beliefs is that they offer a clear way to test whether over- and under-confidence correlate with later-life outcomes with symmetric magnitudes; we display p-values for all of these comparisons at the bottom of Panel A in Tables 2.3, 2.4, and 2.5. In practice, we find that the coefficient magnitudes for over- and under-confidence are only significantly different for two of our twelve outcomes: high-school graduation and working in STEM. Over-confidence predicts high-school graduation significantly more strongly than does under-confidence, while only under-confidence predicts working in STEM.

We also test for heterogeneity in the direction of biased beliefs using our more continuous measure of degrees of confidence. In Appendix Table B.22, we allow the coefficient on this measure to differ by whether a child is over-confident (assessing one's ability at least 3 bins, or 42 percentiles, too high), under-confident (assessing one's ability at least 3 bins too low), or neither. We cannot reject that the slope of the outcome with respect to the degrees of confidence variable is equal across these groups for any outcome, though we are likely under-powered to do so. This result supports the functional-form assumptions we make in Panel B of each of our main tables, where degrees of confidence enter linearly for all outcomes. More broadly, these results and those using our binary measures of over- and under-confidence suggest that over- and under-confidence largely predict similarly-sized, oppositely-signed gaps in long-term educational and employment outcomes.

B.5 Measuring key confounders

Big-Five Personality Traits

In Section 2.7.1, we show that our results are robust to controlling for children’s Big-Five personality traits. The CDS did not measure these traits using standard psychometric scales, so we approximate them using parents’ reports of child behavior. See Appendix B.2 for the variables that make up the index for each trait.

While our proxies for these traits may be noisy, they do correlate with other variables in expected ways. First, the TAS did collect standard psychometric scales to measure Big-5 traits among young adults, and our childhood measures correlate with these adult measures at levels similar to other estimates of the longitudinal persistence of the Big-Five traits (Hampson and Goldberg, 2006; Edmonds et al., 2013). The intercorrelations of our childhood Big-Five personality measures are also broadly similar to those found in studies that use more standard scales to measure these traits (van den Akker et al., 2014; Soto, 2016). Finally, if we regress contemporaneous math and reading cognitive test scores on our childhood Big-Five measures while controlling for IQ, race, and gender, the coefficients on the Big-Five characteristics follow similar patterns as those reported in Almlund et al. (2011) (results available upon request).

We also consider the extent to which the Big-Five traits predict long-term outcomes in our data. We present the coefficients on each personality trait in the specifications above in Appendix Table B.20. Some correlations are consistent with prior estimates of the contemporaneous links between personality and economic outcomes (Almlund et al., 2011; Heckman et al., 2019), but we find fewer significant relationships than expected. These null results may reflect noise in our constructed measures of personality, or they could reflect that childhood personality traits only moderately persist into adulthood (Hampson and Goldberg, 2006; Edmonds et al., 2013).

Parent and Teacher Beliefs and Investment

In Section 2.7.1, we also test that our main results are robust to controlling for measures of parents' and teachers' investments and beliefs. We construct these controls using data from the CDS. We measure caregiver investment from self-reports of how often they do certain activities with their child (e.g. do homework, play games), and we observe both caregiver and teacher reports of the level of educational attainment they expect the child to achieve. Our data also include teacher ratings of the student's academic, social, and physical competence on a scale from 1 (extremely competent) to 4 (not at all competent); we standardize these ratings by year and age group as a measure of teacher perceptions, which likely relate to teacher investment. See Appendix Table B.6 for summary statistics on these variables.

We have relatively low data coverage for teacher reports because the CDS only interviewed elementary school teachers, while the CDS sample includes older children, and because questionnaires were mailed to teachers and had relatively low response rates. In total, 54 percent of our final sample had a teacher respond in any wave of the CDS. We observe teacher predictions of educational attainment in the same year in which we observe biased beliefs for 34 percent of our sample, and we observe teacher reports of student competence for 20 percent of our sample (this variable was only recorded in the 1997 CDS). In contrast, we observe caregiver reports of investment and predicted educational attainment for more than 99 percent of our sample.

These measures of teacher and parent beliefs and investments correlate with children's beliefs in math in our sample, making them potential confounders of the main associations we estimate. Appendix Table B.15 regresses childhood over- and under-confidence in math on our variables for teacher perceptions and expectations, parent investment and expectations, and child test scores. First, teacher expectations of educational attainment predict children's biased beliefs: children that teachers think are going to get a graduate

degree are more confident, and in particular are less likely to be under-confident. Next, parent investment predicts children's under-confidence but not over-confidence: children whose parents read or do homework with them more than once per week are (marginally significantly) more likely to be under-confident in math, whereas we find suggestive evidence that children with parents who play sports or games with them are less likely to be under-confident. Similar to the results for teacher expectations of educational attainment, children whose parents think they are likely to get a graduate degree are less likely to be under-confident.

Overall, these results show that our measures of children's over- and under-confidence in math are correlated with parent and teacher beliefs and investment in largely expected ways, even when we control for children's ability and general confidence. This suggests that one mechanism through which childhood over- and under-confidence could relate to long-term outcomes could be through parent and teacher behavior. However, adding controls for these adult beliefs and behaviors does not change the relationship between children's over- and under-confidence and long-run outcomes – if anything, children's biased beliefs become *more* predictive of long-run outcomes when we condition on these variables.

School quality when confidence is measured

Finally, Section 2.7.1 tests that our results are robust to controlling for the quality of the school that children were attending when we observe their first measures of over- and under-confidence in math. We match respondents with school IDs using restricted data from the CDS.

Then, we collect data on free or reduced-price lunch and student-teacher ratios from the NCES, while we collect data on testing achievement from the Stanford Education Data Archive (SEDA; Fahle et al., 2021). The measures are scaled relative to national grade- and subject-specific test score distributions. SEDA's data for school test scores pools data

from 2009-2018 and is unavailable in earlier years. The students in our sample attended these schools in 1997, 2003, or 2007; we are forced to assume that relative school quality was similar in the decade before we observe testing data. 60 percent of our sample attends a school where we observe test scores in 2009-2019; 80 (50) percent of students attend a school where we observe the student-teacher ratio (percent FRPL) in the year in which we observe confidence. We also include an indicator for missing an NCES School ID in the CDS data.

B.6 Alternate definitions of childhood biased beliefs

This section describes the alternate definitions of over-confidence, under-confidence, and more continuous degrees of confidence to which we test robustness in Section 2.7 above. Throughout the following definitions, p refers to children's score percentiles in math and r refers to children's self-reported math ability from 1 to 7. The names referring to each definition match those used in the specification charts given in Appendix Figures B.4-B.16.

Section A. Over-confidence:

1. Main measure:

- Over-confident if
$$\begin{cases} p < 25 & r \in \{6, 7\} \\ p < 50 & r = 7 \end{cases}$$

2. Main - more strict (1):

- Over-confident if
$$\begin{cases} p < 15 & r \in \{6, 7\} \\ p < 40 & r = 7 \end{cases}$$

3. Main - less strict (1):

- Over-confident if
$$\begin{cases} p < 35 & r \in \{6, 7\} \\ p < 60 & r = 7 \end{cases}$$

4. Main - less strict (2):

- Over-confident if
$$\begin{cases} p < 15 & r \in \{5, 6, 7\} \\ p < 40 & r \in \{6, 7\} \end{cases}$$

- Estimated with under-confidence measure *Original - less strict*

5. Relative:

- Over-confident if $p <$ the 25th percentile of people who report the same self-reported ability (r) in the same age bucket and if $r < 5$

6. Continuous tails (3 to 6):

- Over-confident if *Main* degrees of confidence measure (Section C #1) ≥ 3

7. Main of averages:

- Take average of first and second self-reported ability (r) and first and second percentile scores (p)
- Apply cutoffs of *Main* measure (Section A #1) to these averages

8. Average of main:

- Take the average of the first- and second-observed *Main* (Section A #1) over-confidence measures

Section B. Under-confidence

1. Main measure::

- Under-confident if $\begin{cases} p > 75 & r \in \{1, 2, 3, 4\} \\ p > 50 & r \in \{1, 2, 3\} \end{cases}$

2. Main - more strict (1):

- Under-confident if $\begin{cases} p > 85 & r \in \{1, 2, 3, 4\} \\ p > 60 & r \in \{1, 2, 3\} \end{cases}$

3. Main - more strict (2):

- Under-confident if $\begin{cases} p > 85 & r \in \{1, 2, 3, 4\} \\ p > 60 & r \in \{1, 2, 3\} \end{cases}$
- Estimated with over-confidence measure *Original - more strict*

4. Main - less strict (1):

- Under-confident if
$$\begin{cases} p > 65 & r \in \{1, 2, 3, 4\} \\ p > 40 & r \in \{1, 2, 3\} \end{cases}$$

5. Relative:

- Under-confident if $p >$ the 75th percentile of people who report the same self-reported ability (r) in the same age bucket and if $r > 3$

6. Continuous tails (-6 to -3):

- Over-confident if *Main* degrees of confidence measure (Section C #1) ≤ -3

7. Main of averages:

- Take average of first and second self-reported ability (r) and first and second percentile scores (p)
- Apply cutoffs of *Main* measure (Section B #1) to these averages

8. Average of main:

- Take the average of the first- and second-observed *Main* (Section B #1) under-confidence measures

Section C. Degrees of confidence

- Main measure:

- Assume that kids with accurate beliefs would have reported $r^* = 1$ if $p \in \{1, 14\}$, $r^* = 2$ if $p \in \{15, 28\}$, ... $r^* = 7$ if $p \in \{86, 100\}$.
- Confidence measure is self-reported ability ($r \in \{1, \dots, 7\}$) minus what they would have reported if they had accurate beliefs ($r^* \in \{1, \dots, 7\}$). This variable has range -6 to 6.

- Percentiles:

- Convert empirical distribution of self-reports (r) into percentiles from 0 to 100 (p_r)
- Degree of confidence = $p_r - p$, or percentile of self-reported ability minus actual score percentile in our sample
- Empirical distribution:
 - Assume that the empirical distribution of self-reported ability is correct, but kids may be wrong about their place in it. In other words, if the full sample had accurate beliefs, the bottom 4% of scorers in our sample would report $r^* = 1$, the next 2% would report $r^* = 2$, and the top 22% of scorers would report $r^* = 7$. These values come from the empirical distribution of r , graphed in Figure 2.1.
 - Degree of confidence = $r - r^*$, or self-reported ability minus what children would have reported if they had accurate beliefs by this measure. This variable has range -6 to 6.
- Main of averages:
 - Take average of first and second self-reported ability (r) and first and second percentile scores (p)
 - Apply the same rule as the *Main* measure (Section C #1) to these averages
- Averages of main:
 - Take the average of the first- and second-observed *Main* measures (Section C #1) of degrees of confidence.
- **To make the specification charts, we standardize all of these measures of degrees of confidence to have mean 0 and standard deviation 1 in our analysis sample so that they can be compared on the same scale.**

Appendix C

Appendix to *The Narrative of Policy*

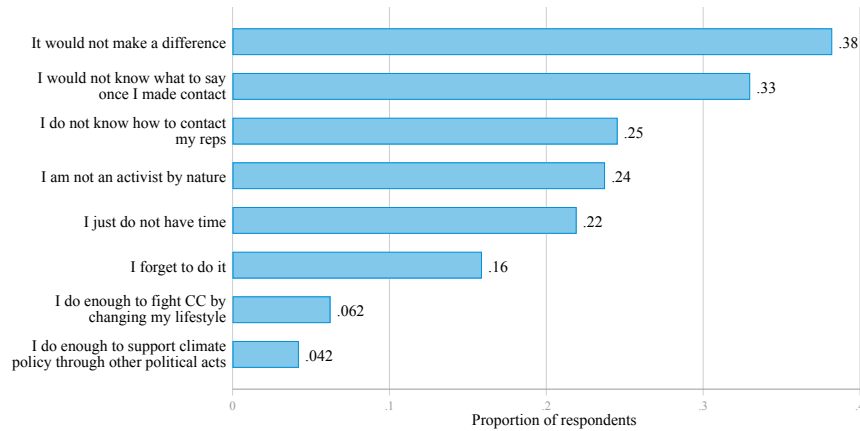
Change

Appendix C.1 contains supplementary tables and figures. Appendix C.2 discusses additional details of study recruitment. Appendix C.3 describes the obfuscation process for the follow-up survey. Appendix C.4 describes the production of the story. Appendix C.5 describes the 5 minutes of filler questions that half of all participants who did not watch the climate-advocacy story were randomly assigned to complete. Appendix C.6 describes the comprehension questions used to assess attention to all videos. Finally, Appendix C.7 describes additional details on how the variables used in our main analysis were measured and defined.

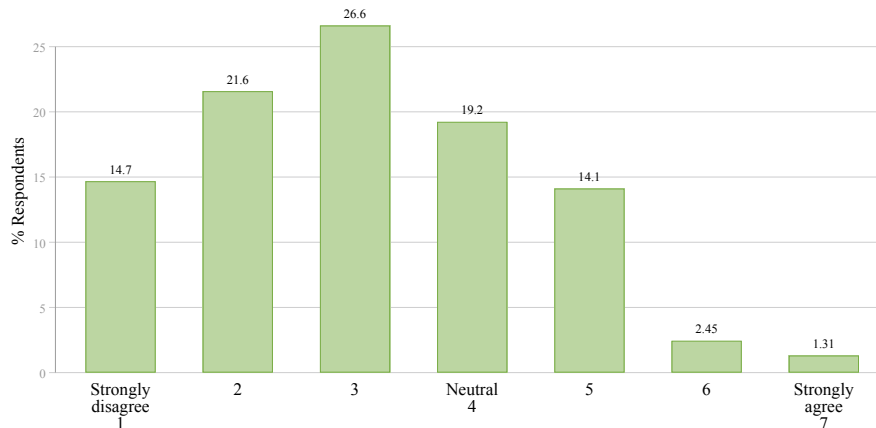
C.1 Supplementary tables and figures

Figure C.1: Political efficacy among those who want more climate policy

Panel A. June 2022 Prolific survey: Share citing each option as top-2 reason for not previously contacting Congress about climate change



Panel B. Screening sample for main study: Agreement that “When groups of citizens push for policy on issues like climate change, the US government responds to their demands.”



Note: Panel A plots responses from a sample of 445 Prolific participants recruited in June 2022. These participants were split evenly by gender, live in the 48 contiguous United States, and were between the ages of 18 and 25. All of these participants reported that they had not phoned, emailed, or called Congress about climate change in the previous 12 months. We asked participants to rank each of the 8 reasons in Panel A from most important (1) to least important (8) in preventing them from contacting Congress; participants could leave any reason that was not at all relevant out of their ranking. Panel A plots the share of participants who ranked each reason among their top-2 most important reasons for not contacting Congress. Panel B plots the distribution of responses to one of the qualitative political-efficacy questions elicited in the screening survey for this experiment, among our experimental sample (N = 5,879). In particular, it plots participants’ agreement from 1 (Strongly disagree) to 7 (Strongly agree) with the following statement: “When groups of citizens push for policy on issues like climate change, the US government responds to their demands.”

Figure C.2: Research Design

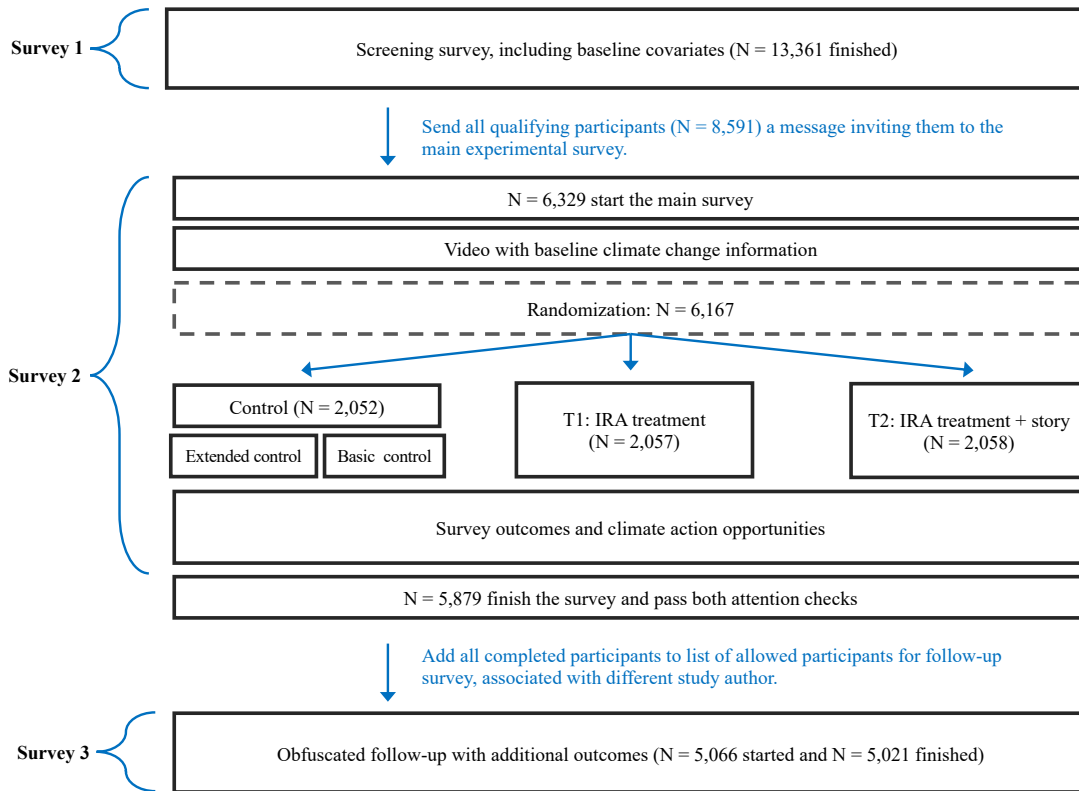


Figure C.3: Baseline desire for government climate action

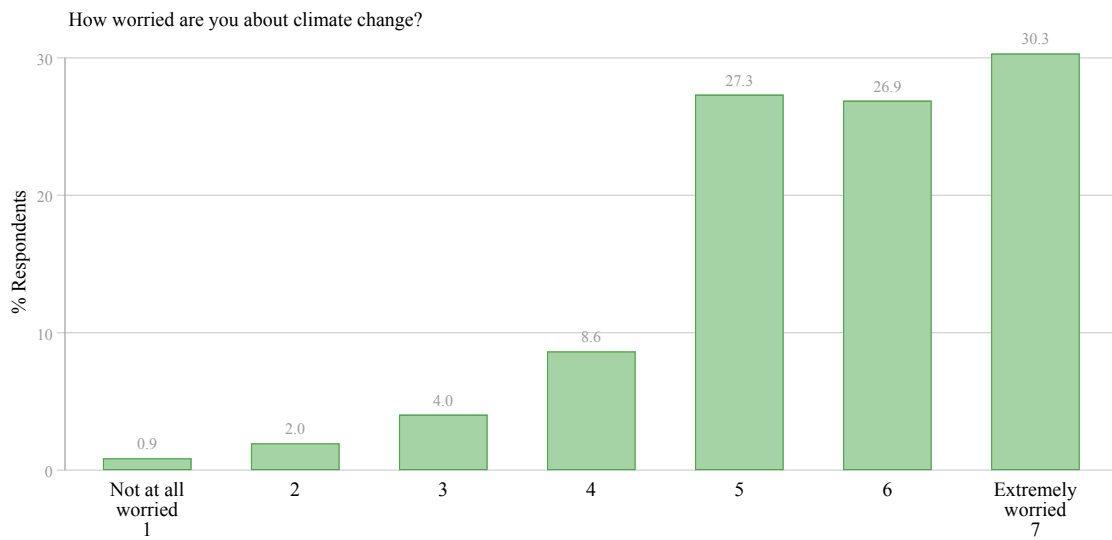
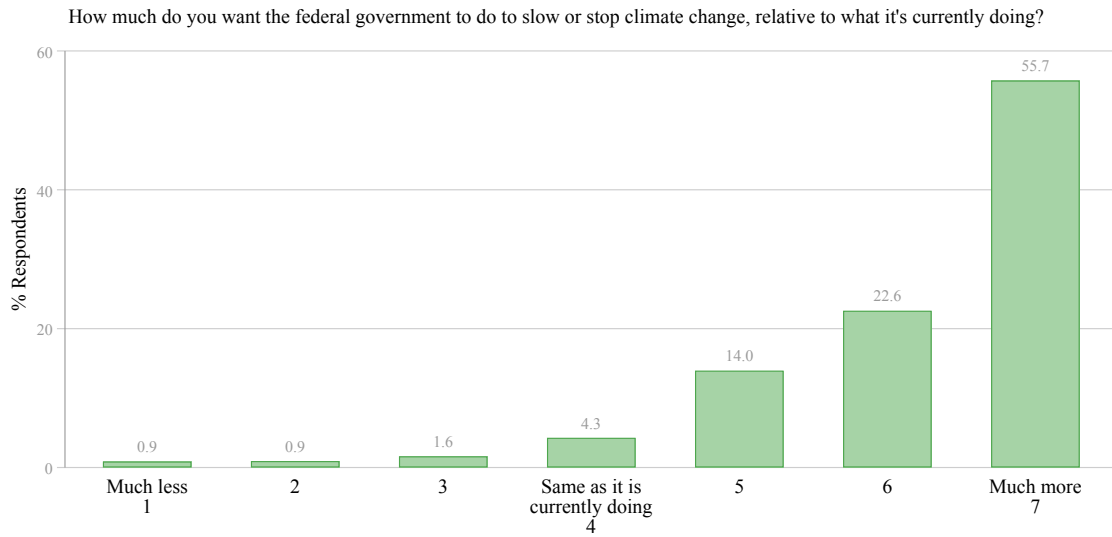
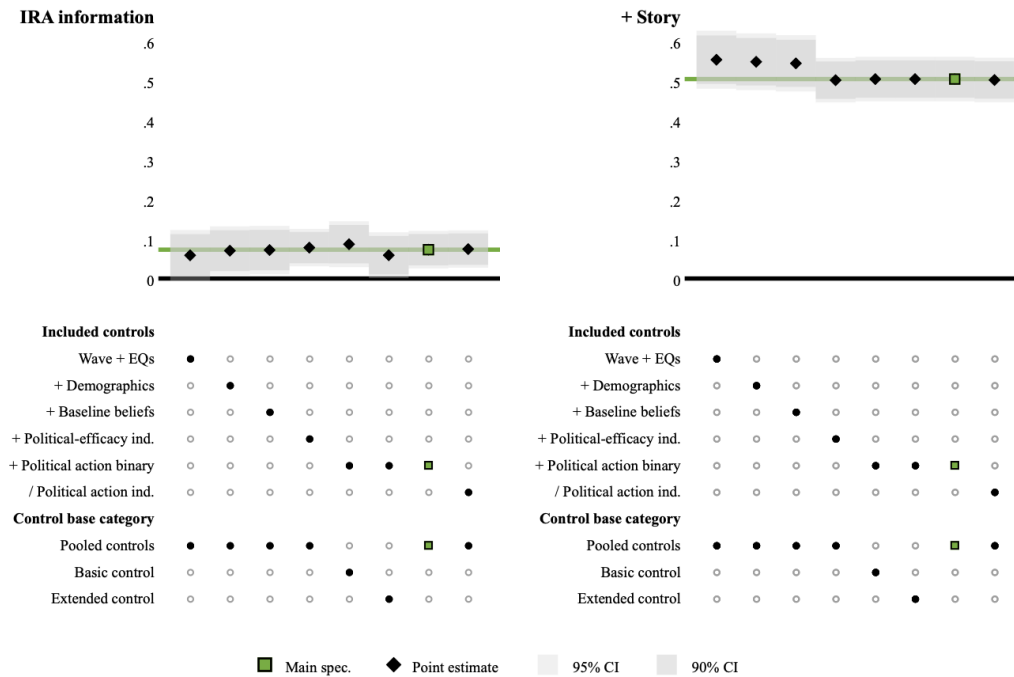
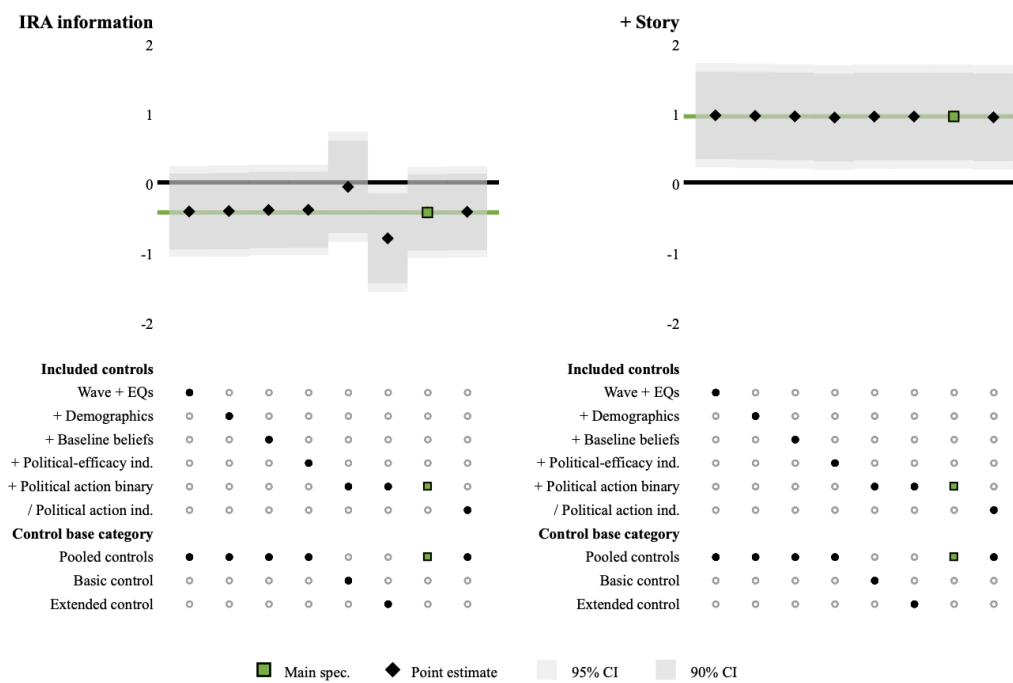


Figure C.4: Specification chart: Standardized political-efficacy index, main survey



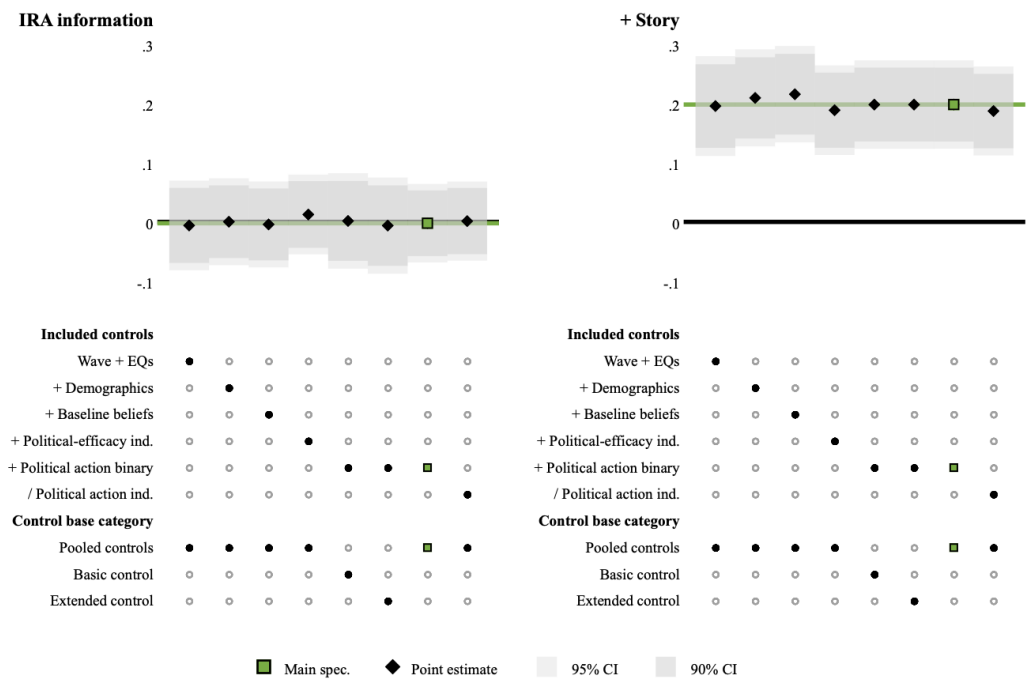
Note: This figure plots the impacts of the IRA information and story treatment on the main-survey political-efficacy index under a range of regression specifications. Appendix Section C.7 describes this outcome variable in detail. Each regression follows the same basic structure as that presented in Table 3.1, column 4, and the colored line and colored squares reproduce the estimates from that main specification. The other specifications presented in this chart test the robustness of these estimates to (a) restricting the sample to those who pass both attention-check questions, (b) iteratively add control variables, and (c) define the omitted category for the IRA information regression coefficient to be the basic control arm, the extended control arm, or a pooled control arm.

Figure C.5: Specification chart: Gradient of bill passage with respect to citizen calls



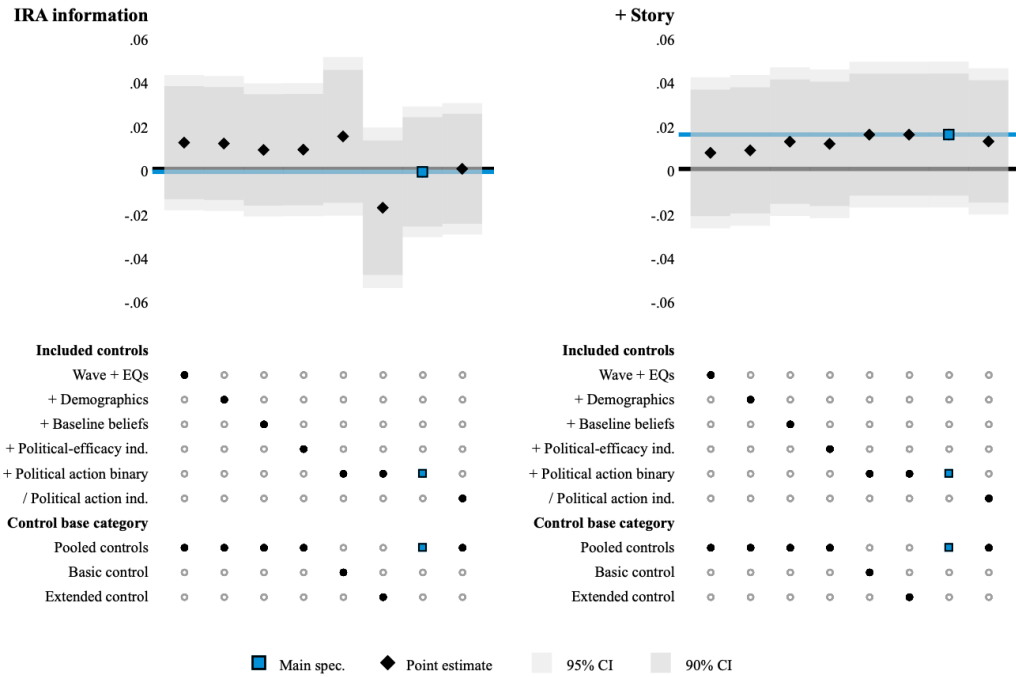
Note: This figure is analogous to C.4, but the outcome here is the gradient in the likelihood that a climate bill would be passed if 10% of Americans called to support it rather than 2%. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 5 of Table 3.1. Appendix Section C.7 describes this outcome variable in detail.

Figure C.6: Specification chart: Standardized political-efficacy index, follow-up survey



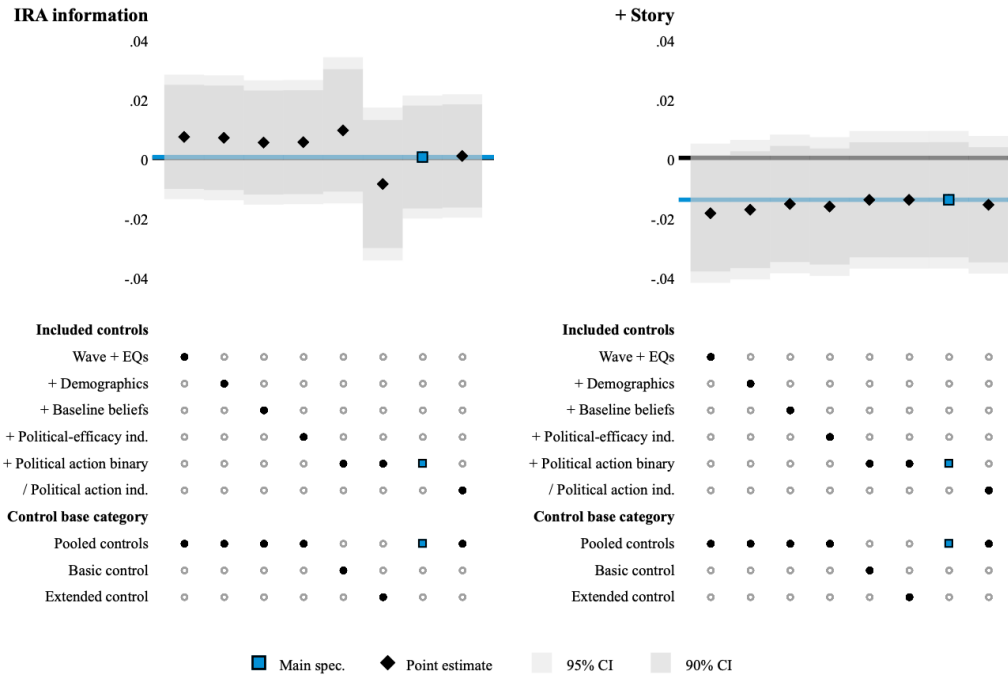
Note: This figure is analogous to C.4, but the outcome here is the standardized index of political efficacy measured in the obfuscated follow-up survey. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 9 of Table 3.1. Appendix Section C.7 describes this outcome variable in detail.

Figure C.7: Specification chart: Started process of writing to Congress



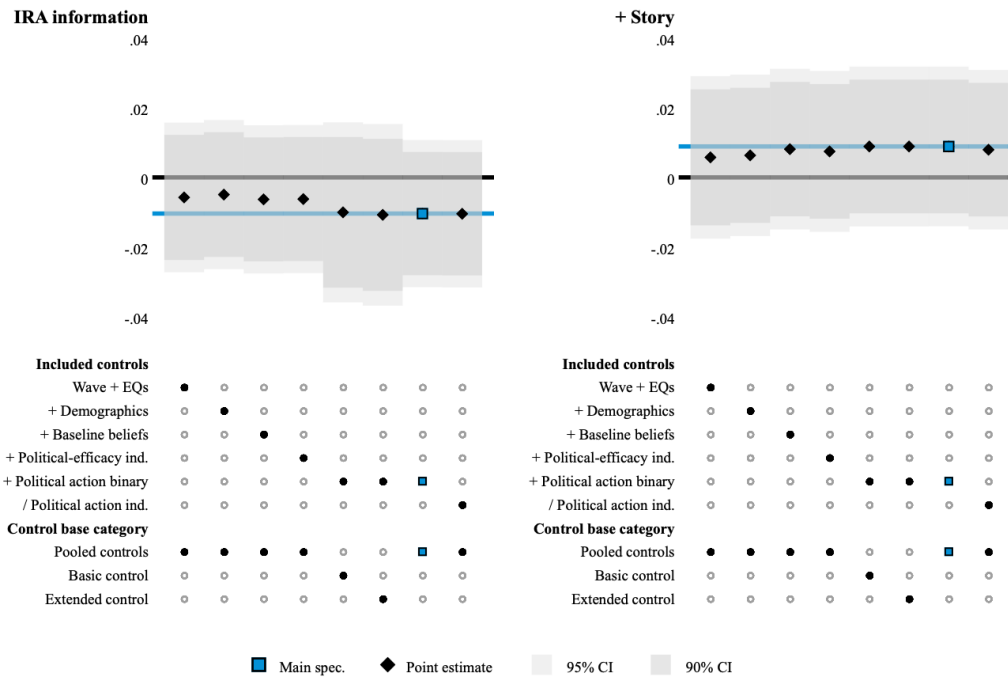
Note: This figure is analogous to C.4, but the outcome here is whether participants opted into the process of emailing Congress. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 1 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.8: Specification chart: Wrote custom text for letter to Congress



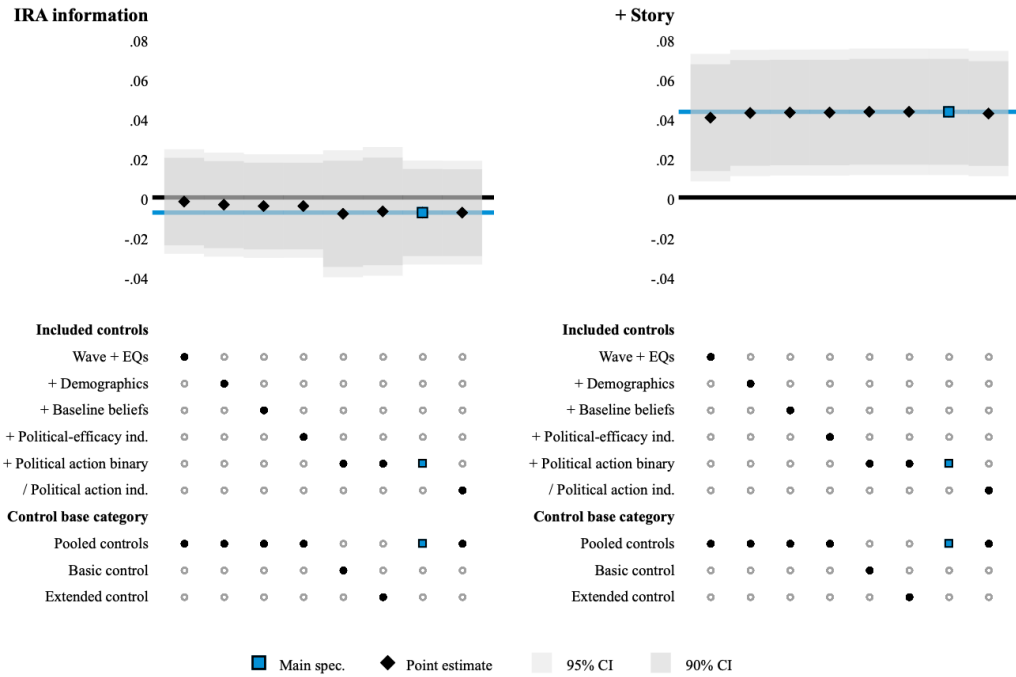
Note: This figure is analogous to C.4, but the outcome here is whether participants wrote out custom text to send to Congress. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 2 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.9: Specification chart: Clicked to send letter to Congress



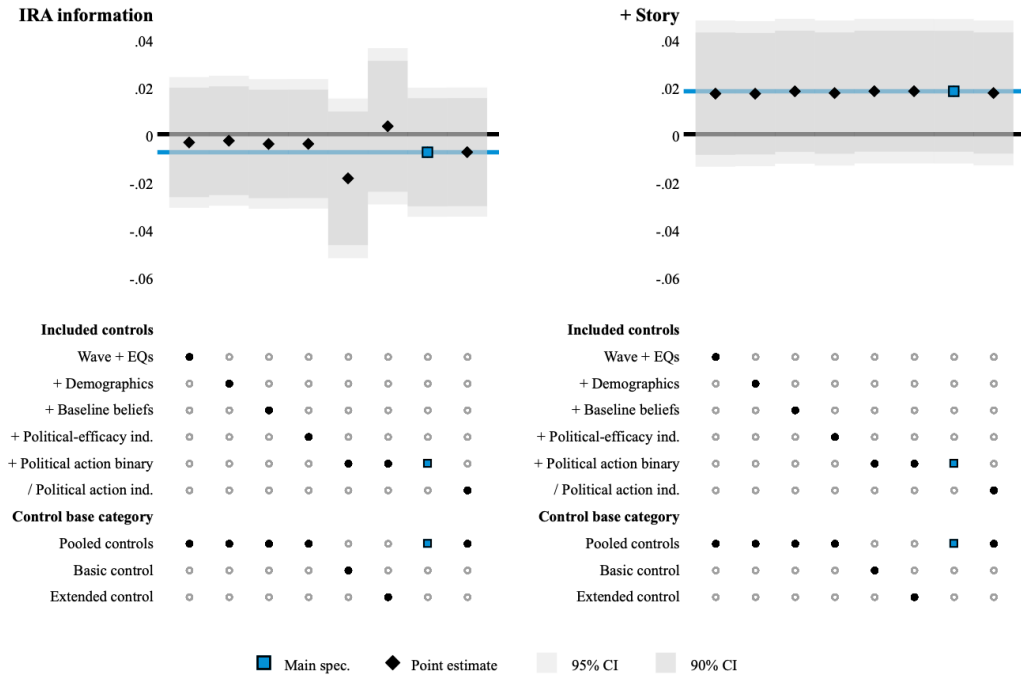
Note: This figure is analogous to C.4, but the outcome here is whether participants clicked a link to a portal from which to email Congress. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 3 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.10: Specification chart: Clicked link for climate marches



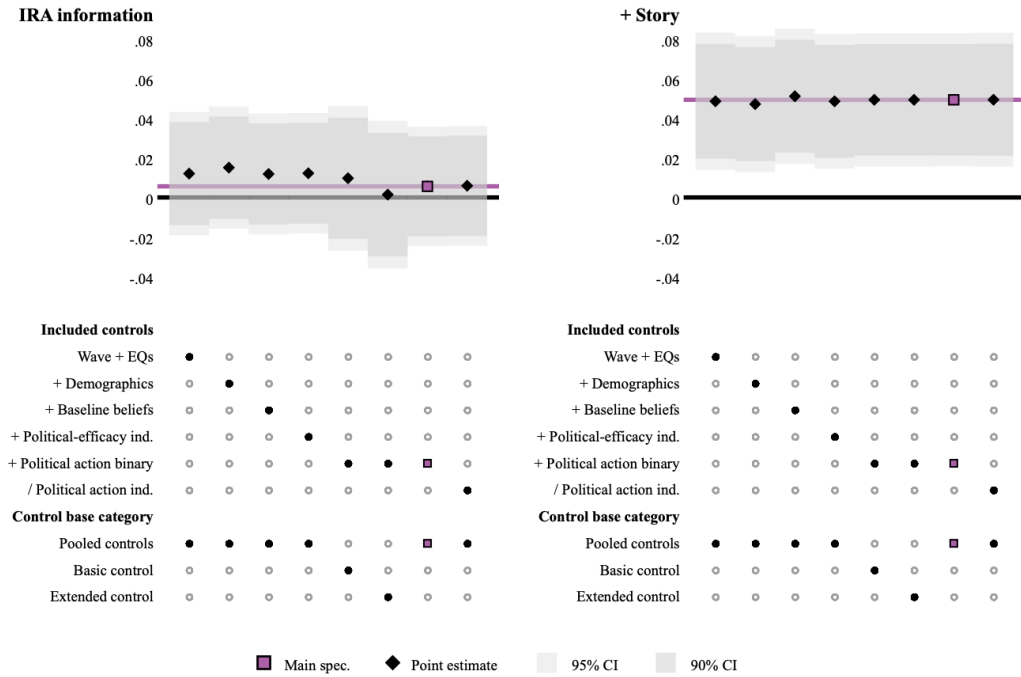
Note: This figure is analogous to C.4, but the outcome here is whether participants clicked a link for information about nearby climate marches. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 4 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.11: Specification chart: Downloaded guide for contacting Congress



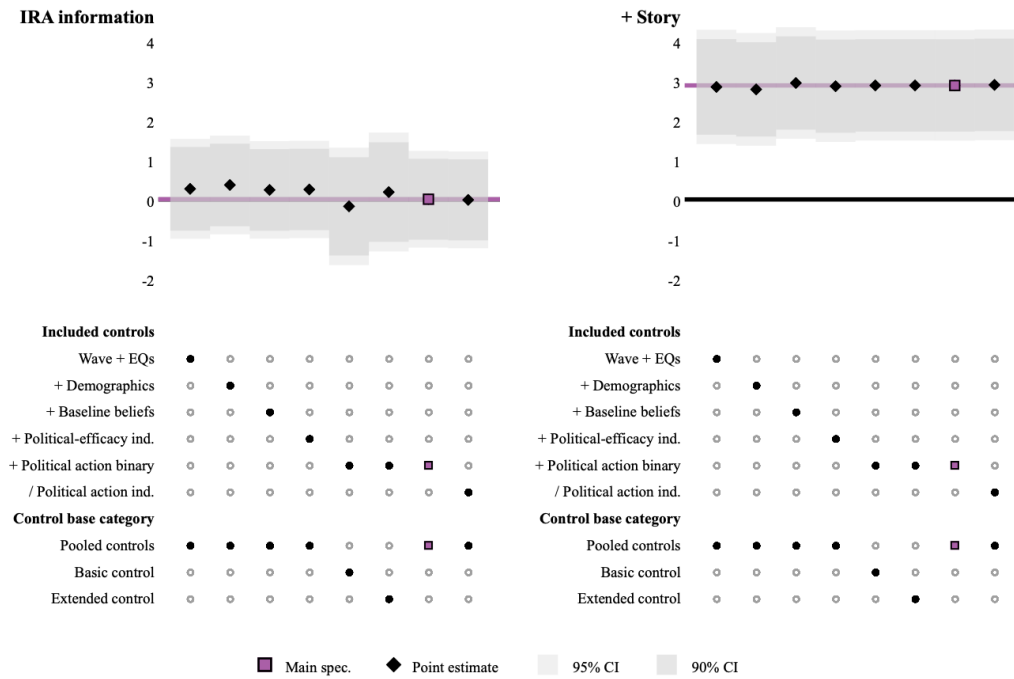
Note: This figure is analogous to C.4, but the outcome here is whether participants downloaded the guide for contacting Congress in the follow-up survey. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 5 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.12: Specification chart: Whether donated to climate organization in main survey



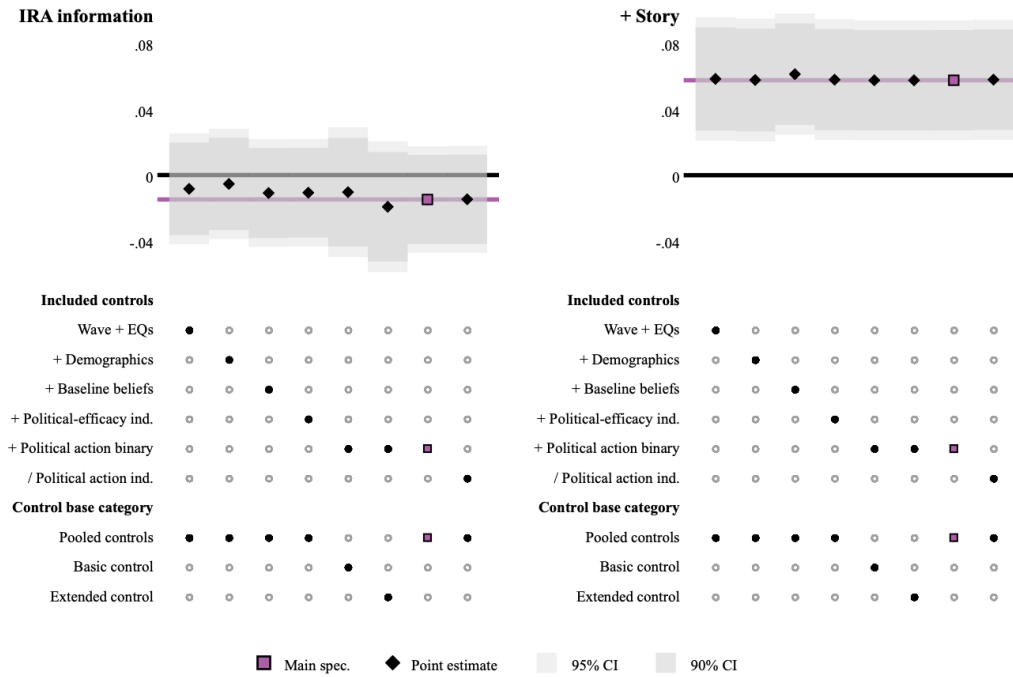
Note: This figure is analogous to C.4, but the outcome here is whether participants donated to a climate organization during the main experimental survey. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 6 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.13: Specification chart: Amount donated to climate organization in main survey



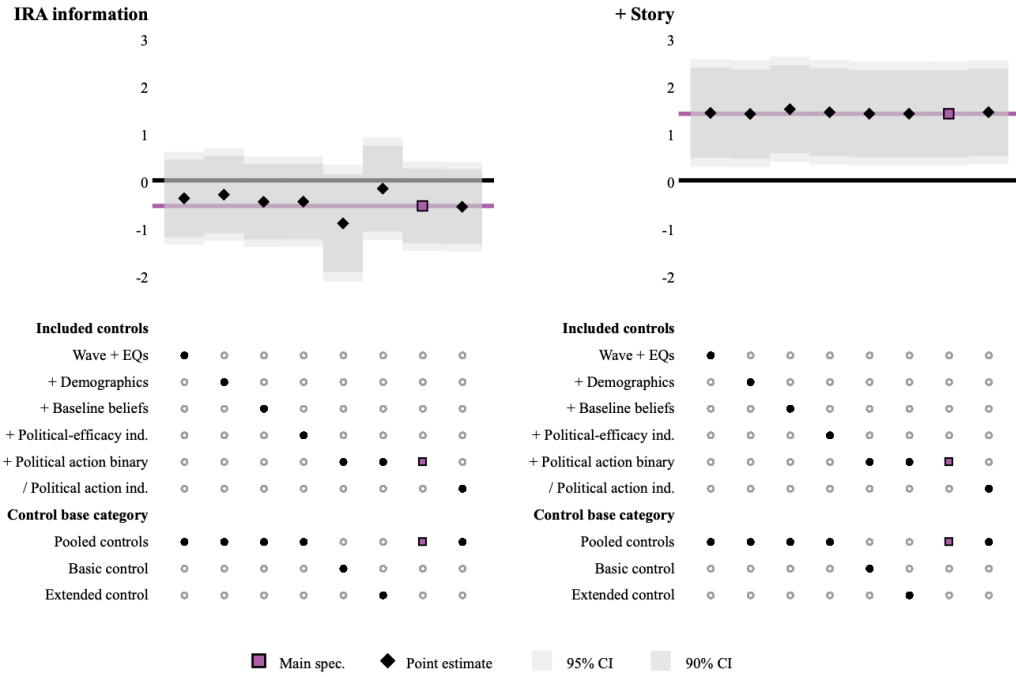
Note: This figure is analogous to C.4, but the outcome here is the amount that participants donated to a climate organization during the main experimental survey. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 7 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.14: Specification chart: Whether donated to climate organization in follow-up survey



Note: This figure is analogous to C.4, but the outcome here is whether participants donated to the climate organization during the obfuscated follow-up survey. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 8 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Figure C.15: Specification chart: Amount donated to climate organization in follow-up survey



Note: This figure is analogous to C.4, but the outcome here is the amount that participants donated to a climate organization during the obfuscated follow-up survey. Our main specification for this outcome (highlighted here in the colored markers) is also presented in column 9 of Table 3.2. Appendix Section C.7 describes this outcome variable in detail.

Table C.1: Descriptive statistics and sample balance

	Mean: Full sample (1)	Δ Extended control (2)	Δ IRA (3)	Δ IRA + story (4)	Δ Extra questions (5)
Surveyed Wave 2	0.442	0.004 (0.022)	0.002 (0.019)	-0.001 (0.021)	0.007 (0.016)
Female	0.526	-0.009 (0.023)	0.002 (0.019)	0.001 (0.021)	-0.007 (0.016)
Age	37.184	-0.106 (0.606)	-0.456 (0.517)	-0.385 (0.555)	-0.382 (0.424)
Ethnic groups:					
Asian	0.076	0.006 (0.012)	-0.009 (0.010)	0.001 (0.011)	0.009 (0.008)
Black	0.066	0.018 (0.011)	0.015* (0.009)	0.002 (0.010)	0.003 (0.008)
White	0.738	-0.037* (0.020)	-0.016 (0.017)	-0.017 (0.018)	-0.016 (0.014)
Other	0.009	0.007 (0.005)	-0.002 (0.003)	0.002 (0.004)	0.003 (0.003)
Missing	0.112	0.006 (0.014)	0.013 (0.012)	0.011 (0.013)	0.001 (0.010)
Whether has 4 year college degree	0.555	-0.027 (0.022)	-0.009 (0.019)	0.001 (0.021)	0.003 (0.016)
Political affiliation:					
Democrat	0.587	0.013 (0.022)	-0.013 (0.019)	-0.008 (0.021)	-0.024 (0.016)
Republican	0.088	0.010 (0.013)	-0.003 (0.011)	0.008 (0.012)	-0.008 (0.009)
Independent	0.277	-0.035* (0.020)	0.007 (0.018)	-0.012 (0.019)	0.018 (0.014)
Other	0.047	0.012 (0.009)	0.009 (0.008)	0.012 (0.009)	0.014** (0.007)
Political engagement index (std)	0.040	0.035 (0.044)	0.136*** (0.039)	0.112** (0.044)	0.004 (0.033)
Prev. contacted elected reps	0.246	-0.011 (0.019)	0.024 (0.017)	0.002 (0.018)	0.005 (0.014)
Prev. donated	0.390	0.028 (0.022)	0.038** (0.019)	0.050** (0.020)	-0.002 (0.015)
Prev. canvassed	0.016	0.004 (0.005)	0.006 (0.004)	0.012** (0.005)	0.002 (0.004)
Prev. signed petition	0.591	0.012 (0.022)	0.057*** (0.019)	0.024 (0.021)	0.001 (0.016)
Prev. phonebanked	0.028	0.000 (0.007)	0.012** (0.006)	0.007 (0.007)	-0.003 (0.005)
Climate worry (std)	0.005	-0.016 (0.045)	0.032 (0.038)	-0.017 (0.041)	-0.036 (0.031)
Desire for climate action (std)	-0.017	-0.024 (0.046)	0.004 (0.040)	-0.037 (0.042)	-0.039 (0.033)
External efficacy index (std)	0.026	-0.051 (0.044)	-0.048 (0.039)	0.031 (0.042)	0.047 (0.032)

Column 1 of this table presents summary statistics of baseline characteristics for the full experiment sample, with N = 6,001. Age data are missing for 29 participants. Columns 2 through 5 then present the results of regressions testing each characteristic for balance across the randomized treatment arms. In particular, we regress each characteristic on indicators for participants' assignment to the Extended Control group, the IRA Information group, or the IRA Information + Story group, as well as an indicator for being assigned to answer the extra filler questions. (Recall that these extra questions are cross-randomized within the control groups and IRA Information group.) Robust standard errors are given below in parentheses each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively. Appendix section C.7.2 defines these baseline traits.

Table C.2: Impacts of treatments on policy knowledge

	(1)	(2)	(3)	(4)
	Have heard of the IRA	Did govt make substantial progress on climate change in 2022?		
		Yes	Don't know	No
IRA info	0.259*** (0.013)	0.245*** (0.014)	-0.037*** (0.013)	-0.208*** (0.015)
+ Story	0.008 (0.012)	0.008 (0.017)	0.002 (0.015)	-0.010 (0.017)
N	5879	5879	5879	5879
Control mean	0.634	0.216	0.248	0.536

Note: This table estimates the impact of IRA information and the fictional story on participants' climate-policy knowledge. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. Column 1 estimates impacts on whether participants check off that they've heard of the IRA on a list of four recent bills, elicited at the end of the experimental survey. Columns 2 through 4 present impacts on whether participants answer "Yes," "I don't know," or "No," respectively, when asked at the end of the experimental survey whether the US government made substantial progress on climate change during 2022. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.3: Attrition by treatment groups

	(1)	(2)	(3)	(4)	(5)	(6)
	Finished main experimental survey		Finished follow-up survey:			
			<i>Unconditional</i>		<i>If finished main</i>	
IRA info	-0.004 (0.006)	-0.004 (0.006)	0.009 (0.012)	0.011 (0.011)	0.006 (0.011)	0.008 (0.011)
+ Story	-0.019*** (0.007)	-0.019*** (0.007)	-0.023* (0.013)	-0.024* (0.013)	-0.006 (0.012)	-0.006 (0.012)
<i>Control variables:</i>						
Wave and EQ	✓	✓	✓	✓	✓	✓
Full controls		✓		✓		✓
Control mean	0.967	0.967	0.835	0.835	0.848	0.848
N	6167	6167	6167	6167	6001	6001

Note: This table documents differential attrition across treatment arms. In each column, we regress the attrition outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. In columns 1 and 2, we test whether participants differentially finished the main experimental survey and passed at least one attention check—thus qualifying for our main sample—by treatment arm. In columns 3 through 6, we test whether participants differentially showed up to and completed the obfuscated follow-up survey by treatment arm. Columns 3 and 4 test for differential attrition through the obfuscated follow-up survey without conditioning on completing the main experimental survey, while columns 5 and 6 test whether participants who finished the main experimental survey differentially completed the follow-up survey. Columns 1, 3, and 5 only control for wave number and whether participants were assigned to complete the extra questions, while columns 2, 4, and 6 also control for demographics, climate attitudes, political efficacy, and political engagement. The only difference between this set of “full” controls and those listed in the note for Table 3.1 is that here we exclude controls for college attainment and political affiliation. We elicited these variables at the end of the main experimental survey, so they are missing for those who did not complete that survey. We detail all control variables in Appendix Section C.7.2. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.4: Impacts of treatments on political efficacy: Lower Lee (2009) bounds

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
	Main survey:				Follow-up survey:					
	Agreement that:					Agreement:		How effective on		
	People like me have no say	Lobbyists have more power	Gov't responds citizens	Index (all +)	Δ P(Pass bill) if 2% -10% call	Citizen movements make change	How effective on govt policy?	Marches	Contacts	Index
IRA info	-0.065** (0.027)	-0.070*** (0.026)	0.143*** (0.028)	0.109*** (0.024)	0.381 (0.305)	0.107*** (0.034)	0.096*** (0.034)	0.087** (0.035)	0.094*** (0.033)	
+ Story	-0.327*** (0.032)	-0.343*** (0.029)	0.377*** (0.031)	0.465*** (0.028)	0.162 (0.359)	0.136*** (0.038)	0.050 (0.038)	0.006 (0.040)	0.102*** (0.037)	
N	5840	5840	5840	5840	5840	3829	3829	3829	3829	
Control mean	0.000	0.000	0.000	-0.000	9.029	0.000	-0.000	-0.000	-0.000	

Note: This table presents lower bounds for the regression coefficients presented in Table 3.1, accounting for differential attrition between those who were or were not assigned to watch the fictional climate-action story. As described in Lee (2009), we estimate lower-bound treatment effects of the climate story for each outcome by selectively dropping “control” participants—who received IRA information but did not watch the climate story—to equalize attrition across the IRA-Info and IRA-Info-plus-Story groups. For each outcome, we drop control participants with the lowest or highest residualized outcomes when our main story-treatment estimate for that outcome is positive or negative, respectively. We estimate differential attrition for each outcome, residuals from our main treatment regressions, and these attrition-adjusted regressions including the full set of controls included in columns 2, 4, and 6 of Table C.3. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.5: Impacts of treatments on climate donations and citizen advocacy: Lower Lee (2009) bounds

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Climate donation outcomes:				Direct-action outcomes:				
	<i>Main survey:</i>		<i>Follow-up:</i>		<i>Main survey:</i>		<i>Follow-up:</i>		
	Y/N	Amount	Y/N	Amount	Said interested	Wrote letter	Clicked to send	Clicked for march	Downloaded guide
IRA info	0.022 (0.015)	0.586 (0.624)	0.004 (0.016)	-0.088 (0.483)	0.011 (0.015)	-0.019* (0.010)	-0.004 (0.011)	0.001 (0.014)	0.004 (0.014)
+ Story	0.035** (0.017)	2.348*** (0.712)	0.040** (0.019)	0.991* (0.563)	0.002 (0.017)	0.004 (0.011)	0.002 (0.012)	0.035** (0.016)	0.006 (0.016)
N	5840	5840	4976	4976	5840	5840	5828	2573	4976
Control mean	0.511	14.944	0.438	8.552	0.426	0.126	0.145	0.079	0.210

Note: This table presents lower bounds for the regression coefficients presented in Table 3.2, accounting for differential attrition between those who were or were not assigned to watch the fictional climate-action story. Our approach in these regressions is analogous to that in Appendix Table C.4. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.6: Effects on donations to each cause in the follow-up survey

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	<i>Any cause</i>		<i>Reproductive</i>		<i>Gun control</i>		<i>Free-market</i>	
	Y/N	Amount	Y/N	Amount	Y/N	Amount	Y/N	Amount
IRA info	0.011 (0.016)	0.920 (0.995)	0.031* (0.016)	1.394*** (0.476)	-0.002 (0.016)	0.292 (0.335)	-0.008 (0.014)	-0.226 (0.251)
+ Story	0.047*** (0.018)	3.015*** (1.120)	0.024 (0.018)	0.904 (0.573)	0.028 (0.018)	0.091 (0.364)	0.033** (0.016)	0.615** (0.272)
N	5021	5021	5021	5021	5021	5021	5021	5021
Control mean	0.558	23.341	0.402	7.239	0.325	4.647	0.226	2.903

Note: This table estimates the impact of IRA information and the fictional story on participants' total donations and donations to non-climate causes in the obfuscated follow-up survey. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. Columns 1 and 2 present whether participants donated to any cause and how much they donated in total (including to the climate organization). Columns 3 through 8 then estimate impacts on whether and how much participants donated to advocacy groups focusing on reproductive rights, gun control, and free market policy. Appendix Section C.7 defines all of these outcome variables in detail. We stratify these regressions by whether participants took the follow-up survey 0-1 days (Panel A), 2-4 days (Panel B), or 5+ days (Panel C) after the main experimental survey. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively. The last two rows of the table present p-values testing whether we can reject that the treatment effects of the IRA information and fictional story are equal across panels.

Table C.7: Correlations between action index and emotions in the control group

	(1)	(2)	(3)
	Motivation-related	Other positive	Other negative
Hope / strength	0.023 (0.021)	Happiness -0.018 (0.011)	Sadness 0.025 (0.024)
Motivation	0.067*** (0.021)	Peacefulness -0.030** (0.014)	Anger 0.073*** (0.024)
Pessimism	0.072*** (0.027)	Connectedness 0.034* (0.018)	Anxiety 0.093*** (0.023)
Apathy / fatigue	-0.086*** (0.018)	Yearning -0.003 (0.023)	Surprise / doubt 0.039 (0.024)
			Guilt -0.000 (0.023)
Sample size:	1968		

Note: This table presents bivariate correlations between an index of climate action and each emotion outcome, estimated in the pooled Basic and Extended Control groups. We construct an index of climate action as the standardized sum of standardized variables for each climate-action outcome included in Table 3.2. We then separately regress this index on standardized measures of how strongly participants reported each of the emotion categories described in Appendix Section C.7. This table presents the estimated coefficients for motivation-related outcomes in column 1, for other positive emotions in column 2, and for other negative emotions in column 3. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.8: Impacts of treatments on climate worry and desire for action

	(1)	(2)	(3)	(4)	(5)
	Main survey:				Follow-up:
	Worry about climate	Desire for govt action	Priority on climate in Congress	Summary index	Hope that Congress focuses on climate
IRA info	-0.009 (0.020)	-0.105*** (0.024)	0.001 (0.028)	-0.046** (0.020)	-0.041 (0.030)
+ Story	0.086*** (0.021)	0.156*** (0.028)	0.073** (0.030)	0.129*** (0.022)	0.074** (0.032)
N	6001	6001	6001	6001	5125
Control mean	-0.000	0.000	-0.000	-0.000	-0.000

Note: This table estimates the impact of IRA information and the fictional story on participants' climate worry and desire for government action. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. Columns 1 through 4 present impacts on outcomes collected during the main experimental survey: worry about climate change, desire for additional government climate action, desire for Congress to prioritize climate change relative to other issues, and an index combining these measures. Column 5 presents impacts on how much participants in the obfuscated follow-up survey state that they want the current Congress to focus on climate change. All of these outcomes are standardized, and Appendix Section C.7 defines them in detail. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.9: Effects on beliefs about support and advocacy for climate policy

	(1)	(2)	(3)
	# of Americans that would say climate change is a prob for govt	Of those, # that would call to support bill	Share concerned that would call
IRA info	0.608 (0.542)	0.046 (0.459)	-0.002 (0.009)
+ Story	-0.041 (0.600)	1.206** (0.528)	0.025** (0.010)
N	5879	5879	5879
Control mean	55.923	16.497	0.302

Note: This table estimates the impact of IRA information and the fictional story on participants' beliefs about other Americans' support and action on climate policy. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. Column 1 presents impacts on participants' beliefs about the number of Americans that would say climate change is a problem that the US government should take action to solve. Column 2 presents impacts on participants' beliefs about the number of Americans who would call or email their national representatives to support a climate bill if it were proposed in the next few months. Column 3 then combines Columns 1 and 2 by presenting impacts on participants' implied beliefs for the share of Americans who would contact Congress among those who support government action on climate change. Appendix Section C.7 defines all of these outcome variables in detail. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.10: Effects of on political efficacy by wave

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Main survey:			Follow-up survey:					
	Agreement that:			Agreement:			How effective on		
	People like me have no say	Lobbyists have more power	Gov't responds citizens	Index (all +)	$\Delta P(\text{Pass}$ bill) if 2% to 10% call	Citizen movements make change	Marches	Contacts	Index
<i>Panel A: Wave 1</i>									
IRA info	-0.001 (0.036)	-0.040 (0.037)	0.104*** (0.038)	0.063* (0.033)	-0.534 (0.444)	-0.016 (0.055)	-0.066 (0.054)	-0.071 (0.055)	-0.060 (0.053)
+ Story	-0.410*** (0.044)	-0.391*** (0.039)	0.423*** (0.042)	0.532*** (0.038)	1.133** (0.519)	0.252*** (0.060)	0.229*** (0.060)	0.095 (0.062)	0.227*** (0.058)
Extra Qs	-0.066* (0.037)	-0.132*** (0.037)	0.015 (0.038)	0.092*** (0.033)	0.418 (0.444)	0.002 (0.056)	-0.006 (0.054)	-0.133** (0.056)	-0.054 (0.053)
N	3280	3280	3280	3280	3280	1731	1731	1731	1731
Control mean	0.035	0.064	-0.038	-0.059	8.165	-0.005	-0.001	0.046	0.016
<i>Panel B: Wave 2</i>									
IRA info	-0.058 (0.042)	-0.017 (0.039)	0.097** (0.042)	0.075** (0.036)	-0.287 (0.510)	0.021 (0.045)	0.050 (0.047)	0.030 (0.048)	0.040 (0.044)
+ Story	-0.302*** (0.049)	-0.376*** (0.045)	0.406*** (0.048)	0.471*** (0.044)	0.688 (0.581)	0.213*** (0.052)	0.107** (0.054)	0.116** (0.055)	0.172*** (0.050)
Extra Qs	-0.047 (0.042)	0.039 (0.039)	0.044 (0.042)	0.023 (0.036)	-0.467 (0.508)	0.082* (0.045)	0.032 (0.047)	0.035 (0.047)	0.059 (0.044)
N	2599	2599	2599	2599	2599	2168	2168	2168	2168
Control mean	-0.045	-0.081	0.048	0.076	10.128	0.003	0.001	-0.036	-0.012
<i>p-val: IRA info</i>	0.294	0.665	0.903	0.802	0.713	0.605	0.104	0.162	0.143
<i>p-val: + Story</i>	0.102	0.796	0.792	0.293	0.565	0.620	0.126	0.799	0.471

Note: This table estimates impacts of the IRA information and fictional story on the political-efficacy outcomes, stratified by wave of participant recruitment. Wave-1 participants were recruited to the main experimental survey from November 2 through November 9, 2022, and Wave-2 participants were recruited from January 13 through February 8, 2023. Appendix Section C.2 details these recruitment waves. All outcomes match those reported in Table 3.1. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. In addition to reporting the coefficients on the IRA-information and climate-story treatment indicators, we also report coefficients on an indicator that participants were assigned to the extra filler questions in each wave. Appendix Section C.5 describes these filler questions, which differed across recruitment waves 1 and 2. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively. The last two rows of the table present p-values testing whether we can reject that the treatment effects of the IRA information and fictional story are equal across recruitment waves.

Table C.11: Effects of on climate action by wave

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Climate donation outcomes:					Direct-action outcomes:				
	<i>Main survey:</i>		<i>Follow-up:</i>		<i>Main survey:</i>				
	Y/N	Amount	Y/N	Amount	Sending letter to interested	Wrote letter	Clicked to send	Clicked for march	<i>Follow-up:</i> Downloaded guide
<i>Panel A: Wave 1</i>									
IRA info	0.012 (0.020)	0.438 (0.825)	-0.025 (0.022)	0.176 (0.643)	0.005 (0.020)	0.014 (0.014)	-0.026* (0.014)		-0.021 (0.019)
+ Story	0.013 (0.023)	2.002** (0.934)	0.051** (0.024)	0.910 (0.739)	0.001 (0.023)	-0.027* (0.016)	0.014 (0.016)		-0.001 (0.021)
Extra Qs	0.032 (0.020)	2.318*** (0.830)	0.032 (0.022)	1.245* (0.648)	-0.027 (0.020)	-0.012 (0.014)	0.009 (0.014)		-0.007 (0.019)
N	3280	3280	2853	2853	3280	3280	3273		2853
Control mean	0.512	14.186	0.422	7.345	0.419	0.115	0.143		0.231
<i>Panel B: Wave 2</i>									
IRA info	-0.009 (0.023)	-0.653 (0.957)	-0.005 (0.025)	-1.429** (0.726)	-0.008 (0.023)	-0.018 (0.016)	0.007 (0.016)	-0.008 (0.013)	0.010 (0.021)
+ Story	0.095*** (0.026)	3.934*** (1.094)	0.068** (0.029)	1.982** (0.859)	0.035 (0.026)	0.003 (0.017)	0.004 (0.018)	0.043*** (0.016)	0.046* (0.024)
Extra Qs	0.024 (0.023)	0.061 (0.948)	0.034 (0.025)	-0.464 (0.728)	0.003 (0.023)	-0.007 (0.016)	0.010 (0.016)	0.001 (0.013)	0.025 (0.021)
N	2599	2599	2168	2168	2599	2599	2596	2595	2168
Control mean	0.509	15.907	0.461	10.152	0.435	0.140	0.149	0.079	0.183
<i>p-val: IRA info</i>	0.493	0.385	0.564	0.096	0.669	0.141	0.128		0.258
<i>p-val: + Story</i>	0.018	0.177	0.643	0.341	0.326	0.205	0.662		0.129

Note: This table estimates impacts of the IRA information and fictional story on the climate-action outcomes, stratified by wave of participant recruitment. This table is fully analogous to Appendix Table C.10 above. All outcomes match those reported in Table 3.2, though we only observe whether participants click for climate-march information in the second recruitment wave. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

Table C.12: Effects on probabilistic beliefs about passing climate policy

	(1) Prob that US meets 2030 goal	(2) Prob that limit warming to 1.5°	(3) Prob pass climate bill if: if 2% call	(4) Prob pass climate bill if: if 10% call	(5) Δ Prob pass climate bill
IRA info	5.735*** (0.561)	1.767*** (0.606)	3.008*** (0.642)	2.576*** (0.708)	-0.431 (0.333)
+ Story	0.441 (0.653)	1.717** (0.688)	1.640** (0.733)	2.588*** (0.805)	0.948** (0.383)
N	5879	5879	5879	5879	5879
Control mean	28.691	31.207	40.308	49.337	9.029

Note: This table estimates the impact of IRA information and the fictional story on participants’ beliefs about the probability that we hit global and national climate goals and that the US would pass hypothetical climate policy. In each column, we regress the outcome variable on an indicator for receiving IRA information and an indicator for additionally watching the fictional climate story. We include the same control variables listed in the note for Table 3.1 and detailed in Appendix Section C.7.2. Columns 1 and 2 present impacts on participants’ estimates for the probability that the US will meet its 2030 emissions commitment under the Paris Agreement and that globally we will limit warming to 1.5° C. Columns 3 and 4 present impacts on participants’ estimates for the probability that the US Congress would pass a climate bill if it were proposed in the next few months and if 2% and 10% of Americans contacted their representatives to support it, respectively. Column 5 presents impacts on participants’ estimates for how much more likely Congress would be to pass the climate bill if 10% rather than 2% of Americans contacted them in support. (We also present these estimates in Table 3.1.) Appendix Section C.7 defines all of these outcome variables in detail. Robust standard errors are given in parentheses below each coefficient. *, **, and *** indicate significance at the 0.1, 0.05, and 0.01 percent level, respectively.

C.2 Study recruitment

Participants were recruited to the study in two “waves,” first in November 2022 and again in January 2023. While we originally planned to recruit the full sample at once before the 2022 midterm elections, recruitment was slower than anticipated. Thus, we paused the study just after the 2022 midterm elections due to concerns that the change in political representation could affect our results. We updated our pre-analysis plan on January 11, 2023 to describe this change in plans ([link](#)). The following information describes the same information as in the updated PAP, with some additional context.

Specifically, we waited to resume data collection until the new Congress was sworn

in, which occurred several days before we posted the PAP update. The delay allowed us to ensure that uncertainty in Congressional leaders' status would not depress the rates at which participants contacted their legislators. In the interim, we presented the results with the first half of data collection at an internal MIT seminar, which led us to make two changes to the study.

First, recall that half of the sample in the IRA-only and control groups were randomly assigned to answer additional filler questions that were timed to take 5 minutes, the length of the fictional story video, in order to ensure that the additional length of the survey was not causing its own effects or differential attrition between groups. After halting the first recruitment wave, we observed that the open-ended filler questions designed to control for the duration of the climate action story were producing potentially large priming effects, so we decided to re-adjust this condition to control only for time effects. To do so, we changed the items from open-ended questions about themes related to the story to multiple choice questions about scientific topics. These questions do not refer to climate change or any adjacent topics (temperature, erosion, etc.). We describe the filler questions themselves in more detail in Appendix C.5.

Second, we saw that while the story was affecting participants' beliefs, there were no significant effects on whether participants contacted Congress or donated to climate organizations.¹ One possible explanation for this gap was that the action outcomes were not close enough to the behaviors represented in the story: namely, the story focuses on citizen marches, rather than contacting legislators. If the story inspires participation in immediately-related forms of pro-climate action but not others, our previously identified outcomes might miss these effects. Thus, we added an additional secondary outcome to see whether the story affects participants' interest in participating in climate marches or demonstrations.

Our main results control for the wave in which a participant completed the survey.

¹This latter result became significant when we had recruited the full pre-registered sample.

Appendix Tables C.10 and C.11 show our main results separately by wave and the coefficients on an indicator for being in the group randomized to receive the filler questions. (We include this indicator as a control in our main specifications, but we do not include its estimate coefficient in the main tables).

C.3 Obfuscating the follow-up survey

We design the obfuscated follow-up so that participants cannot connect it to the main experimental survey. Specifically, the first two surveys were posted on Prolific under Lucy Page's name, while the obfuscated follow-up survey was posted on Prolific under Hannah Ruebeck's name. The follow-up used a different survey font, header, consent-form layout, and color scheme than the earlier surveys and was advertised as being about general political activity, while the earlier surveys were listed as studying climate change. The follow-up survey was much shorter, and even questions that measured the same construct as in the main survey were formatted differently. All of the questions in the obfuscated follow-up referred to multiple other policy issues in addition to climate change.

When we re-contacted participants between the screening and main surveys, they were sent a direct Prolific message with a link to the main survey. Participants were never invited via a direct message to the follow-up survey; instead, they were simply added to a list of eligible Prolific accounts and saw the obfuscated follow-up as one of any number of available Prolific surveys. 85 percent of participants in the main sample completed the obfuscated follow-up survey; we attribute the high return rate to its very short duration (2 minutes). The only information that could link the follow-up survey with the earlier surveys is that all were fielded by researchers from MIT Economics. However, no participants indicated that they connected the obfuscated follow-up survey with the earlier surveys.

C.4 Story production

The 5-minute fictional story video was animated by an animation firm based in the UK and voiced by professional voice actors. Before getting the story animated, we asked small samples of Prolific users to read and react to several variations of its main text.

In a first survey with 31 respondents, we asked participants to read two different stories and compare them – one centered on Annie organizing a climate march on her own, while the second focused on Annie’s conversation with an older man who explained why he was organizing a climate march. We selected elements from each draft story to include in the final version, based on pilot participants’ written responses about why they liked each, which would be a better story when it was illustrated and animated, and what they thought could be improved to make each story more enjoyable and effective at motivating action. In a second survey with 45 respondents, we provided participants with the text of a story that was very close to the story used in the main experiment, but randomly varied the ending. One version ended with the Gilbert March shown in the final story, another ended with a senator who was influenced by the march and eventually confirmed that she would help to draft a climate bill in response, and a final version with lawmakers coming together to actually pass a bill. Again, the final video included a combination of these candidate endings, compiled based on participants’ emotional responses and open-ended reactions to the story – what they found boring, memorable, unrealistic, etc. We also asked if the story would change the way they felt about the likelihood that the US can address climate change.

Our analysis of both surveys was purely qualitative, and we used participants’ reactions to make sure the story was as natural, interesting, and moving as possible. While we originally developed the story before the passage of the IRA, aiming solely to build political efficacy, we adapted the very end of the story in August 2022. Our revised ending accounted for the passage of the IRA and positioned the Gilbert March as a quasi-backstory

to the bill's passage.

We provided a narrative script to the animation firm Cut The Mustard, which they adapted to be appropriate for a 5-minute video. We iterated with them on character sketches, storyboards, color schemes, and music before they produced the final product. The research team recruited two voice actors on Fiverr and provided their recordings to Cut The Mustard. We contracted with Cut The Mustard in late June 2022 and they provided the final product at the end of October 2022.

C.5 Filler questions

The fictional climate story has a duration of about 5 minutes. To ensure that any treatment effects of the fictional story do not derive just from a longer survey, we also cross-randomize half of all participants not assigned to watch the story to answer additional “filler questions” ensuring that their surveys also take five minutes longer. Initially, these were a series of open-ended questions with minimum-time timers that focused on events and themes similar to those referenced in the story, helping us to also rule out the possibility that the story acts simply as a prime. However, as discussed in Appendix C.2, we changed the filler questions before launching the second wave of study recruitment because the questions themselves seemed to have large priming effects; we designed the filler questions in the second round of data collection to control only for duration effects. In the second wave, we asked multiple choice questions about scientific topics (without any reference to climate change) with timers to ensure that participants spent exactly 5 minutes answering them.

C.5.1 Open-ended filler questions

We introduced the open-ended questions to participants as chance to hear their thoughts about climate change and politics; we described the time restrictions (1 minute per ques-

tion) as encouragement to think carefully about each question. The questions were as follows:

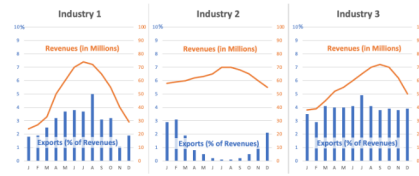
1. As a warm-up question, think about any childhood pets that your family had. Did you have pets? If so, what were they like?
2. Next, think about whenever in your life you first learned about climate change. Roughly how old were you when you learned about climate change, and in what context? For example, did you learn about climate change in school? How did you feel about climate change when you first learned about it?
3. Next, to what extent do you feel like you're personally seeing the impacts of climate change in the world, maybe through changes in weather or natural disasters?
4. Next, some people choose to personally advocate for climate policies by calling their Senators, showing up to marches, or writing opinion pieces in their local newspapers. We might call those people "climate activists." In your mind, what kind of people tend to be climate activists?
5. Finally, some people think that engaging with politics on issues like climate change (for example, by calling your Senators or going to climate marches) is useless. Do you agree with that? Why or why not?

Panel A in each of Appendix Tables C.10 and C.11 show the effects of being randomly assigned to answer these questions on our outcomes of interest. Answering these questions increased the main-survey political-efficacy index by 0.09sd while reducing participants' estimates of the effectiveness of emailing Congress in the follow-up survey by 0.13sd. The filler questions substantially increased climate donations in the main survey by \$2.32 and by \$1.25 in the follow-up survey.

C.5.2 Multiple choice filler questions

In the second wave, participants randomly assigned to answer filler questions took a “general science knowledge” quiz with 20 multiple-choice questions. Participants could only progress to the next page of the survey after 4 minutes and 50 seconds had elapsed, and the page automatically advanced after 5 minutes had elapsed. We asked participants not to look up the answers to any questions, which were as follows:

These graphs show the monthly revenues and exports as a percent of revenues for three industries over the course of one year. Based on the graphs, which industry has the greatest annual range of monthly revenues?



A scientist is conducting a study to determine how well a new medication treats ear infections. The scientist tells the participants to put 10 drops in their infected ear each day. After two weeks, all participants' ear infections had healed.

Which of the following changes to the design of this study would most improve the ability to test if the new medication effectively treats ear infections?

- Create a second group of participants with ear infections who use 15 drops a day
- Have participants put ear drops in both their infected ear and healthy ear
- Have participants use ear drops for only one week
- Create a second group of participants with ear infections who do not use any ear drops

Which of the following is an example of genetic engineering?

- Growing a whole plant from a single cell
- Finding the sequences of bases in plant DNA
- Attaching the root of one type of plant to the stem of another type of plant
- Inserting a gene into plants that makes the plant resistant to insects

Many diseases have an incubation period. Which of the following best describes what an incubation period is?

- The recovery period after being sick
- The period during which someone has an infection, but is not showing symptoms
- The period during which someone builds up immunity to a disease
- The effect of a disease on babies

Which of these is a major concern about the overuse of antibiotics?

- It can lead to antibiotic-resistant bacteria
- There will be an antibiotic shortage
- Antibiotics can cause secondary infections
- Antibiotics will get into the water system

What are frogs classified as?

- Amphibians
- Reptiles
- Gastropods
- Anihozoa

Which scientist developed the theory of evolution?

- Isaac Newton
- Sigmund Freud
- Charles Darwin
- Marie Curie

What are the building blocks that make up everything on Earth called?

- Nuclei
- Atoms
- Mitochondria
- Gravitrons

What is the main cause of seasons on the Earth?

- The tilt of the Earth's axis in relation to the sun
- The distance between the Earth and the sun
- Changes in the amount of energy coming from the sun
- The speed that the Earth rotates around the sun

What molecule absorbs sunlight during photosynthesis?

- Thylacoid
- Chlorophyll
- Sucrose
- Oxygen

A car travels at a constant speed of 40 miles per hour. How far does the car travel in 45 minutes?

- 25 miles
- 30 miles
- 35 miles
- 40 miles

Which parent's genes determine a baby's sex?

- Male parent
- Female parent
- Both male and female parent
- Neither

In geometry, how many sides are in a heptagon?

- 3
- 5
- 7
- 9

What feature of a sound wave determines how loud the sound is?

- Amplitude
- Frequency
- Length
- Range

The time a computer takes to start has increased dramatically. One possible explanation for this is that the computer is running out of memory. This explanation is a scientific...

- Experiment
- Hypothesis
- Observation
- Conclusion

An antacid relieves an overly acidic stomach because the main components of antacids are ...

- Acids
- Neutral
- Isotopes
- Bases

What is the atomic symbol for lead?

- Pb
- La
- Ld
- Rh

At what temperature does water freeze?

- 0 degrees C
- 0 degrees F
- 10 degrees C
- 40 degrees F

What does a light-year measure?

- Time
- Velocity
- Distance
- Mass

Which of these planets is earth's closest neighbor?

- Venus
- Neptune
- Mercury
- Jupiter

Panel B in each of Appendix Tables C.10 and C.11 show the effects of being randomly assigned to answer these questions on our outcomes of interest. Answering these questions has no effect on political efficacy in the main survey, though it increases agreement that citizen movements can make change (elicited in the follow-up survey) by 0.08sd ($p = 0.07$). The extra questions have no statistically-significant effect on any form of climate action. Thus, we conclude that the story’s additional duration cannot explain its impacts on political efficacy or climate action.

C.6 Comprehension questions

Before all informational videos and the story video, we ask participants to watch the videos carefully because we will ask several comprehension questions afterwards. We emphasize that we will randomly choose 10 participants and pay them \$5 for each comprehension question that they answer correctly. Immediately after all participants watch the baseline informational, they first answer the following comprehension questions:

- Comprehension question 1: Under the Paris Agreement, to what level does the international community hope to limit warming? [1 degree C; 1.5 degrees C; 2 degrees C; 2.5 degrees C]. 97% of the sample answered this correctly.
- Comprehension question 2: By how much have temperatures already risen, on average, from pre-industrial levels? [0.8 degrees C; 0.9 degrees C; 1 degree C; 1.2 degrees C]. 77% of the sample answered this correctly.

Before participants watch the next video (either the basic control, extended control, or IRA-treatment video), we reiterate that it will be followed by additional comprehension questions subject to the same incentives. Participants assigned to watch the **basic-control** video answer just one additional question:

- Comprehension question 3: The US commitment under the Paris Agreement is to

reduce emissions to what percent of 2005 emissions levels by 2030? [50%; 55%; 60%; 65%]. 97% of the sample answered this correctly.

Participants assigned to watch the **extended-control** video answer three additional questions:

- Comprehension question 3: What is the baseline year that the US emissions reductions commitments reference? In other words, we have committed to reducing emissions by a certain percentage below emissions levels in what year? [2005; 2006; 2009; 2010]. 94% of the sample answered this correctly.
- Comprehension question 4: The US commitment under the Paris Agreement is to reduce emissions to what percent of 2005 emissions levels by 2030? [45%; 50%; 55%; 60%; 65%]. 95% of the sample answered this correctly.
- Comprehension question 5: What are emissions commitments under the Paris Agreement called? [Nationally-determined contributions (NDCs); Country emissions standards (CES); Voluntary emissions levels (VELs)]. 94% of the sample answered this correctly.

Participants assigned to watch the **IRA-treatment** video answer three additional questions:

- Comprehension question 3: The US commitment under the Paris Agreement is to reduce emissions to what percent of 2005 emissions levels by 2030? [45%; 50%; 55%; 60%; 65%]. 86% of the sample answered this correctly.
- Comprehension question 4: Comprehension question 4: What is the name of the recent climate bill signed into law? [Inflation Reduction Act; Infrastructure Investment and Jobs Act; Emissions Reduction Act]. 88% of the sample answered this correctly.

- Comprehension question 5: According to projections, what share of the remaining emissions reductions cuts required to hit the United States' 2030 target will the Inflation Reduction Act achieve? [40%; 65%; 70%; 80%]. 71% of the sample answered this correctly.

Finally, we also asked several comprehension questions after the **climate-advocacy story**. Again, we told participants in advance that 10 participants would be randomly chosen to win \$5 for each question they answered correctly. These questions were as follows:

- What was the dog's name in the story? [Rufus; Milo; Gilbert; Charlie]. 97% of the sample answered this correctly.
- Which of the following social movements did the story not reference? [The disability-rights movement; The civil-rights movement; The movement for women's right to vote; The labor-rights movement]. 85% of the sample answered this correctly.

C.7 Variable definitions

C.7.1 Outcome variables

Political efficacy.

- **Main survey:** The main experimental survey captures both qualitative and quantitative measures of political climate efficacy.
 - Qualitative measures elicit participants' agreement with the following statements from 1 (Strongly disagree) to 7 (Strongly agree):
 1. People like me don't have any say about what the federal government does about issues like climate change;
 2. Fossil fuel companies and their lobbyists have more power than citizens in determining what the US government does about climate change;

3. When groups of citizens push for policy on issues like climate change, the US government government responds to their demands.

We standardize these variables to have mean zero and standard deviation one in the control group, and present results separately for agreement with each statement as well as for an index constructed from all three statements. We calculate this index by summing the standardized component variables, flipping the sign of agreement with the first and second statements, then standardizing this sum to have mean zero and standard deviation one in the control group

- The quantitative measure of political efficacy elicits participants' guess for the probability that another climate bill would pass if it were introduced to Congress in the next few months. Participants who completed the survey in October or November were asked to estimate the probability (on a slider from 0 to 100, with labels for "Definitely not" and "Definitely yes" at either end, with "Fairly low chance" and "Fairly high chance" centered at 35 and 65, respectively) that a hypothetical climate bill would pass if it were proposed in January, assuming that Democrats maintained control of both houses of Congress. Participants who completed the survey in January or February were asked to consider a hypothetical climate bill that would be proposed in April. We ask participants to separately guess the probability that such a bill would pass if 2% or 10% of Americans contacted their national representatives to support it. The difference between participants' guesses in each of these cases provides a numeric measure of external collective efficacy: the impact of additional citizen pressure on government action. Specifically, the two slider questions read as follows:

- * Imagine that a bill pushing for climate action were introduced to Congress in (January) April 2023. Now imagine that 2% of Americans contacted their national representatives to support the climate bill. That would be

about 15,000 people per district in the House of Representatives. What do you think is the probability that Congress would pass the bill?

* Now imagine that 10% of Americans contacted their national representatives to support the climate bill. That would be about 76,000 people per district in the House of Representatives. What do you think is the probability that Congress would pass the bill in that case? (Recall that you thought there would be a [Previous Answer]% chance if 2% of Americans contacted their representatives.)

• **Obfuscated follow-up:** We elicit three measures of political efficacy in the follow-up survey after participants have the chance to download the Call the Halls guide.

1. To what extent do you agree with the following statement? Statement: "Citizen movements on issues like gun control and climate can make real change." A slider from 0 (Disagree completely) to 7 (Agree extremely strongly).
2. How effective do you think marches / rallies are in affecting government policy? A Likert scale from 0 (Not effective at all) to 6 (Extremely effective)
3. How effective do you think contacting politicians (for example by phone or email) is in affecting government policy? A Likert scale from 0 (Not effective at all) to 6 (Extremely effective)

We standardize these variables to have mean zero and standard deviation one in the control group, and present results separately for agreement with each statement as well as for an index constructed as the sum of these standardized variables and then itself standardized to have mean zero and standard deviation one in the control group.

Donations to climate advocacy organizations.

- **Main survey:** Before participants have the chance to engage in action (donations, or personal advocacy, with the order randomized) we say: “The United States still has lots of work to do to meet its 2030 emissions reductions commitments under the Paris Climate Agreement. That means that it’s important that we continue to push for ambitious climate action at the federal, state, and local levels.” When participants get to the donation outcome (either immediately or after the action outcomes described below), we say (with additional spacing), “[One/Another] important way to push for climate policy is to support climate advocacy organizations like the [Natural Resource Defense Council](#), the [Sunrise Movement](#), and the [Citizens’ Climate Lobby](#). Remember that one respondent will be randomly chosen to win a bonus of \$80. You can choose now to give some amount that, if you win, we will subtract from your lottery reward and instead donate to the climate organization of your choice. You are entirely free to keep the \$80 prize for yourself; please don’t feel pressured to donate.” We ask them if they’d like to donate to any of the three organizations (Yes or No), and if they say yes, we ask them which organization they would like to donate to (they may only choose one), repeating the links to each group’s website. We then ask how much they’d like to donate, on a slider ranging from 0 to 80. We define outcomes as whether and how much participants donate to one of these organizations.
- **Obfuscated follow-up:** After eliciting hope that the new Congress will focus on various issues, and before participants are offered the guide to contacting politicians, we say (with additional spacing), “One important way to advocate for policy you support is by donating money to effective advocacy organizations. You might remember that one participant in this survey is going to be randomly chosen to win a Prolific bonus of \$100. On the next page, you can decide if you want to donate any of that money, if you win it, to any of the following top-rated advocacy organizations:

- [Violence Policy Center](#), which studies and advocates for solutions to gun violence in the US.
- [NARAL Pro-Choice America Foundation](#), which advocates to expand abortion access in the US.
- [Environmental Defense Action Fund](#), which advocates for ambitious climate policy in the US.
- [The Heritage Foundation](#), which advocates for free-market policies and individual liberty in the US.

You could split the bonus between multiple organizations, donate some to just one organization, or keep the full bonus. Anything you choose is fine! Below, please decide how much to keep yourself versus donating to each organization, if you win the \$100 bonus. (Your answers must sum to \$100.)”

We included the Heritage Foundation in order to reduce the survey’s partisan slant towards stereotypically liberal causes. The order of each choice in the following question where participants enter their donation amounts (with a fifth option labeled “Amount you take home”) is randomized. Our main outcome of interest is whether and how much they donate to the Environmental Defense Action Fund, though we also define secondary outcomes for total amount donated and whether/how much they donate to each non-climate cause.

Citizen advocacy.

- **Main survey:** We observe two measures of revealed interest and engagement in direct citizen advocacy:
 - *Contacting Congress about climate change:* Again, before participants have the change to engage in action (donations or personal advocacy, with the order

randomized) we say: “The United States still has lots of work to do to meet its 2030 emissions reductions commitments under the Paris Climate Agreement. That means that it’s important that we continue to push for ambitious climate action at the federal, state, and local levels.” When they get to the letter-writing outcome (either immediately or after the donation outcomes described above), we say (with additional spacing), “[One/Another] crucial way to help enact strong climate policy is to directly tell your representatives in Congress that you support climate action. If you want, we’ll link you in a few pages to a portal hosted by the Natural Resource Defense Council where you can email your legislators. Don’t worry, you don’t have to be an expert to contact Congress! Are you interested in being linked to contact your legislators?” Participants can then answer yes or no; this determines our outcome for whether participants opt into the process of emailing Congress. If they answer yes, they see the following: “Great! The portal we’ll link you to will include a form letter that you could send, but your email will be much more effective if you personalize it. On this page, you can write out a personalized message you’d like to send on the next page. (It will go to your Senators, House Representative, and President Biden.) If you don’t write out a letter, you can still move ahead and just send the form letter. Here are some tips. The best messages:

- * Give a specific reason for why climate change matters to you or has impacted you personally.
- * Are three or more sentences long.
- * State that whether those politicians act on climate change will affect whether you will vote for them in the future.”

We provide participants with an essay-style (multiple line) text box in which to draft a letter. Finally, on the next page, they see the following: “Here is the letter you wrote out on the last page, if you did so: [Previous Answer]

Click here for a link to the contact portal, hosted by the Natural Resource Defense Council. You'll have an option to click "Read more and personalize your letter." To make your letter as effective as possible, click that and then paste in the letter you wrote out here!" (Note: the letter-writing campaign that we were directing participants to has closed. Below are screenshots of the portal components.)

We define outcomes for whether participants initially said they were interested in emailing Congress, whether they wrote out a personalized email in our text box, and whether they clicked the link to the NRDC portal to send a letter.

YOUR INFORMATION

SALUTATION
- Select -

FIRST NAME
LAST NAME

E-MAIL ADDRESS

ADDRESS

ZIP CODE

CITY

STATE
- Select -

PHONE

GET TEXT MESSAGE ALERTS
By entering my mobile number and checking this box, I agree to receive urgent NRDC Alerts. Message frequency varies. Text HELP to 61636 for help; text STOP to 61636 to end. Msg& data rates may apply. I understand that I'm not required to opt-in as condition for taking action or donating. By leaving this box unchecked, I will not be opted-in to SMS messages at this time. Read our [Privacy Policy](#) and [Terms & Conditions](#).

YOUR MESSAGE

SUBJECT
Prioritize climate action in 2023

Dear President Biden, Senators, and Representative:

(Consider adding your own thoughts — personalized messages are especially effective)

MESSAGE BODY

Climate action can't wait — I'm counting on you to prioritize the following to secure our clean energy future:

* Sufficiently fund the implementation of the Inflation Reduction Act, the strongest climate action ever taken in U.S. history

Read more and personalize your letter >

SEND YOUR MESSAGE

When you take action, you'll become an NRDC member. We will keep you informed with the latest action alerts and campaign updates.

- *Seeking information about climate marches:* In our second wave of data collection (collected in January and February 2023), we added an additional outcome to the main experimental survey to capture participants' interest in specifically march-related climate action. We observe whether participants click a link to a map of upcoming climate marches published by Fridays for Future, a decentralized group begun by Greta Thunberg that organizes climate marches around the world. We

define an outcome as whether participants click on this link. We introduced this secondary outcome in an amendment to our pre-registration posted on January 11, 2023 before starting our second round of data collection. The survey presents this link after the two donation and letter-writing outcomes, so the addition of this outcome does not change the interpretation of the donation or letter-writing outcomes. Specifically, we provide the following (with additional spacing): “Another important way to push for policy change is through marches and other kinds of public demonstrations that make clear to governments and other people around us that we care about climate action. One of the main groups that organizes climate marches is called Fridays for Future. It’s a global movement with climate marches in more than 200 countries and across many US states. If you’d like to find an upcoming climate march near you, click [here](#) for a map showing all of Fridays for Future’s upcoming events.”

- **Obfuscated follow-up:** “Call the Halls” is a guide to contacting legislators written by Emily Ellsworth, a former Congressional staffer. We provide participants a link to download the file, and observe whether they do so as our outcome of interest. Specifically, the page read as follows (with additional spacing): “Donating money to organizations is great, but arguably an even more impactful way that you can support action on political and social issues that you care about is by directly demanding action from politicians at the local, state, and national levels. Politicians’ jobs are to represent citizen preferences, so one of the best ways to make change is to communicate what’s important to you. You don’t have to be an expert to do so! It can be intimidating to get started with contacting elected officials if you’ve never done so before. Below, we’re attaching “Call the Halls,” an excellent guide to contacting your legislators written by Emily Ellsworth, a former Congressional staffer. The guide is meant to be read and shared. It will explain what to say in a message to legislators, how to choose who to contact, and the most effective ways

to make contact.

--> Click [here](#) to download the guide! <--”

Emotions.

- **Main survey:** We test the impacts of the IRA information and climate-action story on participants’ emotional states. We elicit participants’ emotions immediately after the experimental treatments. We ask them to list (writing out whatever they want to) at least one (and up to three) emotions that they were currently feeling, with a note to list the first thing(s) that comes to mind. On the next page of the survey, we then ask participants to rate how strongly they’re feeling each of the emotions they listed on a scale from 1 (Very weakly) to 6 (Extremely strongly).

Two authors hand-coded emotions into categories from a treatment-blind list, generating the classification scheme below. First, one author cleaned the text responses, equating free-responses that were written differently but had the same meaning. This included summarizing a sentence as one emotion (e.g. “determined to make a difference” became “determined”), changing equivalent emotions to the same tense (e.g. sympathy and sympathetic, annoyed and annoyance, pride and proud), and fixing spelling mistakes. 48 responses (out of 16,180) were changed to missing because they did not reference an emotion (e.g. “gilbert” or “children”). This resulted in 607 unique words describing emotions.

A different author categorized those 607 words into the 13 categories presented in the paper (plus “other”, 2.6% of all emotions, and missing, 0.2% of all responses). The table below shows the component emotion words that are included in each category; below each emotion category is the percent of the 16,096 total responses (excluding missing) that fall in that category. In addition to defining dummy variables for whether each participant reporting feeling an emotion in a given category,

we also defined standardized variables for the strength with which they felt that emotion. If participants listed multiple emotions in one category, we use the strength of the emotion that they felt most strongly. We standardize their strongest emotion in each category to have mean zero and standard deviation one in the control group.

- We do not measure emotional responses to the topic of climate change in the obfuscated follow-up.

Emotion category	Emotion words
Hope/strength (7.6%)	ability, accomplished, achievable, ambition, brave, competence, confident, courageous, elevated, empowered, encouraged, expectant, faith, good, grit, hopeful, lucky, optimism, patriotic, positive, potential, powerful, progress, strength, strong, success, trust
Motivation (7.3%)	action, actionable, activated, active, adrenaline, alert, alerted, aroused, called, challenged, commitment, compelled, competitive, convicted, creative, dedicated, determined, driven, eager, emboldened, energetic, engaged, enlightened, enthusiastic, excited, fierce, focused, galvanized, hastened, helpful, hyped, influenced, initiative, inspired, intent, invested, invigorated, involved, justice, moral, motivated, moved, opportunity, passion, persistence, pro action, proactive, productive, protective, pumped, ready, resolve, responsible, revolutionary, righteous, rushed, solidarity, steadfast, stimulated, stirred, stubborn, urge, urgency, vibrant, vindication, willing, woke, zeal, zoned-in
Pessimism (3.9%)	afflicted, beaten, bleak, cringe, cynical, defeated, demoralized, difficulty, discouraged, disenfranchised, disheartened, disillusioned, dismay, division, done, doomed, doubtful, failure, fatalism, fruitless, futility, hopeless, impotence, inadequate, ineffectual, inevitability, insignificant, jaded, judgement, negative, nihilistic, pessimism, pointless, powerless, skeptical, small, stagnant, stoic, unamused, unconvicted, underwhelmed, unrealistic, unsurprised, useless, weak
Apathy/fatigue (8.5%)	aloof, ambivalence, apathy, blah, blank, blase, bored, complacency, demotivated, detached, disinterest, distanced, drained, drowsy, ennui, exhausted, flat, impassive, indifference, lackluster, lazy, lethargic, listless, meh, overworked, passive, resigned, sleepy, slow, sluggishness, spent, tired, uncaring, unfocused, unmotivated, unmoved
Happiness (5.8%)	admiration, amazed, amused, appreciation, awe, blessed, cheerful, content, delighted, elated, enjoyment, entertained, euphoric, exhilaration, ecstatic, fulfilled, glad, grateful, happy, impressed, joyful, laughter, nice, overjoyed, playful, pleasant, pleased, proud, refreshed, thankful, upbeat, uplifted
Peacefulness (5.2%)	acceptance, at ease, attuned, balanced, benign, calm, centered, comfortable, contemplative, docile, ease, easygoing, euthymic, grounded, harmony, lax, mellow, mindful, nonchalant, peaceful, placated, quiet, reflective, relaxed, relief, rested, safe, satisfied, serene, serenity, soothed, stable, still, tranquility, unbothered, well
Compassion/ connection (1.4%)	attentive, camaraderie, caring, collective, compassion, condolence, connected, emotional, empathy, generosity, gentle, gracious, heart, heartwarmed, humanity, impacted, kindness, love, loving, open, patience, poignancy, sensitive, sentimental, sympathy, tolerance, touched, understanding, united, warm
Yearning (0.4%)	desire, dissatisfied, impatience, longing, nostalgic, unfinished, wishful, yearning

Emotion category	Emotion words (continued)
Sadness (18.1%)	aching, alone, anguish, bad, bittersweet, blue, bothered, bummed, deflated, dejected, depressed, despair, despondent, devastated, disappointed, discontent, distant, distraught, distressed, down, drab, empty, forlorn, gloomy, grief, heartache, heartbroken, horrible, hurt, ill, isolated, lonely, loss, lost, malaise, melancholy, misunderstood, monotone, moody, morose, mournful, numb, pain, pitiful, pity, reticent, sad, shitty, solemn, somber, sorrow, strained, subdued, tearful, ugh, unhappy, unloved, unpleasant, upset, weary, weltshmerz, wistful, withdrawn, woeful
Anger (13.0%)	aggravated, angry, annoyed, appalled, betrayed, bitter, condemnation, consternation, contempt, critical, deceived, defensive, derision, devious, disdain, disgruntled, disgust, dislike, displeased, disrespected, enraged, exasperated, frustrated, furious, fury, grumbly, grumpy, hatred, hostility, incensed, indignation, infuriated, injusticed, irritated, jealousy, manipulated, murderous, offended, off-put, outraged, peeved, pissed, rage, resentment, revenge, ridicule, unsatisfied
Anxiety (19.3%)	afraid, agitated, alarmed, angsty, anticipation, antsy, anxious, apprehension, awake, cautious, concern, crazy, dangerous, desperate, discomfort, disturbed, dread, eerie, existential, fearful, fret, fright, guarded, helpless, hesitant, horror, insecure, jittery, meek, nauseous, nervous, on edge, panicked, paranoid, perturbed, pressed, restless, scared, stressed, tense, tension, terrified, trapped, troubled, turmoil, uncomfortable, unease, unnerved, unprepared, unrest, unsettled, vulnerable, wary, watchful, worry
Shock/ questioning (6.0%)	aghast, astounded, baffled, befuddlement, bemused, bewildered, blindsided, captivated, conflicted, confoundedness, confusion, curious, dazed, disbelief, distracted, fascination, flustered, imaginative, incredulous, indecision, inquisitive, interest, intrigued, introspective, investigative, overstimulated, overwhelmed, pensive, perplexed, ponderous, preoccupied, puzzled, questionable, questioning, quizzical, realization, reminiscent, retrospective, shock, startled, stunned, surprise, suspicion, thoughtful, uncertainty, unclear, unknowledgeable, unsure, winded, wondering
Guilt (0.8%)	ashamed, avoidant, behind, careless, dumb, embarrassed, guilty, humbled, naive, pathetic, regret, remorse, shame, sheepish, stupid, wasteful
Other (2.6%)	absurdist, agreeable, artistic, aware, blarged, broke, bullshititude, busy, change, cheesy, cold, collected, concentrating, confession, confirmed, congested, conscious, cool, decent, decisive, dejavu, dissolution, dreamish, dutiful, eco-communist, educated, environconscientious, fair, favored, food, forgetful, full, future awareness, gassy, green, headache, heat, horny, hot, humor, hungry, hurried, impoverished, in tune, in-between, informed, innocent, insightful, intelligent, intense, intentional, knowledgeable, logical, memory loss, move on, movement, need, needy, neutral, nosey, observing, old, pandered, pragmatic, present, progressive, rational, realism, recession, reluctant, reserved, sane, sated, serious, sick, sleeplessness, smart, smirk, smug, snuggly, sore, stretched, studious, stuffy, sweetness, thirsty, treading water, unique, witty

Desire for climate policy.

- **Main survey:** We measure three variables capturing desire for policy change after we elicit emotions (i.e. after the treatment and before any action is taken).
 1. How worried are you about climate change? A Likert scale from 1 (Not at all worried) to 7 (Extremely worried)
 2. How much do you want the federal government to do to slow or stop climate change, relative to what it's currently doing? A Likert scale centered at 4 (The same as it's currently doing) and extending to 1 (Much less) and to 7 (Much more)
 3. Please rank these issues (click and drag to re-order) based on how much you would like Congress to prioritize them in legislation moving forward. The issue ranked at (1) should be the issue you think Congress should prioritize most.
Options: Climate change, reproductive rights, reducing inflation, combatting terrorism, and racial justice

We standardize each response to have a mean zero and standard deviation one in the control group and create an index by first summing the three standardized variables and then standardizing this sum to have mean zero and standard deviation one in the control group

- **Obfuscated follow-up:** The first question in the obfuscated follow-up asks participants about their political priorities, framed in the context of the soon-to-be- or newly-elected Congress (in wave 1 and 2, respectively), with the context that all of the seats in the House of Representatives and 35 of the 100 seats in the Senate were up for reelection. Specifically, “The new Congress could focus on a range of policy issues, including the economy, climate change, abortion rights, or gun policy. To what extent do you hope that the newly-elected Congress will focus on the following

issues?” *Options:* Gun control, climate change, reducing inflation, and reproductive rights/abortion access. Each had a Likert scale from 1 (Not at all) to 6 (Very much so). We standardize the Likert response for climate change to have mean zero and standard deviation one in the control group.

Second-order beliefs.

- **Main survey:** We ask two questions to measure participants beliefs about other Americans’ support for climate policy and willingness to contact their political representatives:

1. Out of 100 Americans, how many do you think would say that they think climate change is a problem the US government should take action to solve? A slider labeled “# of people” from 0 to 100.
2. In the last question, you guessed that [Previous Answer] Americans out of 100 would say that climate change is a problem the US government should take action to solve. How many of those [Previous Answer] Americans do you think would actually call or email their national representatives to support a climate bill if it were proposed in January 2023? A slider labeled “# of people” from 0 to 100.

The answer to each question is a secondary outcome of interest, along with the share of those they think are worried who they think will call.

- We do not measure beliefs about support for climate policy in the obfuscated follow-up

Knowledge of the IRA.

- **Main survey:** The last questions in the experimental survey measure our outcome for our “first stage” (for the effects of the IRA information treatment) and are used to test whether the story affects recollection of the IRA. We ask two questions:

1. To your knowledge, did the US government make substantial progress on climate change during 2022? This could include things you've learned about in this survey. (Please don't look anything up. We're interested in your honest best guess, and it's totally fine if you don't know.) *Options:* Yes, No, and I don't know.
2. Have you heard of any of the following recent bills, including during this survey? Please select any that you've heard of. *Options:* Inflation Reduction Act, Honoring our PACT Act, Affordable Insulin Now Act, and Infrastructure Investment and Jobs Act.

The outcomes of interest are whether they have heard of the IRA and whether they answer Yes to the question about making substantial progress on climate change.

- We do not measure knowledge of the IRA in the obfuscated follow-up

C.7.2 Control variables

Unless otherwise indicated below, all control variables were elicited during the 1-minute screening survey (which participants took at least 1 day before they took the experimental survey). After the two screening questions, participants answered questions which provided the following:

Demographic controls.

- Sex, Age, and Ethnicity come from merging our data with Prolific's provided demographic data using participants' Prolific IDs. We control in our main regressions for whether participants identify as male or female, age bins {18-20, 21-25, 26-30, 31-35, 36-40, ..., 71-75, over 75, missing}, and ethnicity categories {Asian, Black, White, Other, Missing}.

- Education: Do you have a 4-year college degree? (If you are currently in college, please answer “No”) *Note: this question was asked at the very **end** of the experimental survey.* We combined this variable with age to create dummy variables for the interactions of being over age 25 (or missing age) and having a 4-year college degree.
- Political affiliation: In politics today, do you consider yourself a Republican, Democrat, or Independent? (They were also offered options of other, with a fill-in-the-blank, or prefer not to answer). *Note: this question was asked at the very **end** of the experimental survey.* We control in our main regressions for separate indicators that participants identified as a Democrat, Republican, Independent, or Other.

Baseline climate worry.

- How worried are you about climate change? A Likert scale from 1 (Not at all worried) to 7 (Extremely worried). In analysis, we standardize this variable to have mean zero and standard deviation 1 in the control group.

Baseline desire for climate action.

- How much do you want the federal government to do to slow or stop climate change, relative to what it’s currently doing? A Likert scale from 1 (Much less) to 7 (Much more), centered at 4 (The same as it’s currently doing). In analysis, we standardize this variable to have mean zero and standard deviation 1 in the control group.

Baseline external political efficacy.

- Participants’ agreement with the following statements from 1 (Strongly disagree) to 7 (Strongly agree):
 1. People like me don’t have any say about what the federal government does about issues like climate change;

2. Fossil fuel companies and their lobbyists have more power than citizens in determining what the US government does about climate change;
3. When groups of citizens push for policy on issues like climate change, the US government government responds to their demands.

We standardize these variables to have mean zero and standard deviation one in the control group and construct an index as the sum of these standardized variables, flipping the sign of agreement with the first and second statements as those indicate negative efficacy. We then standardize this sum to have mean zero and standard deviation one in the control group.

Baseline political engagement.

- We elicit participants' baseline political engagement with the following framing:
“Some people get directly involved in social and political issues, while others don't have the time or interest. In the last two years, have you engaged in any of the following forms of civic engagement? (In other words, since October 2020). Please select all that apply:
 - Contacted an elected representative about a social or political issue
 - Donated money to an organization working on a social or political issue
 - Canvassed door-to-door about a political or social issue
 - Signed a petition about a political or social issue
 - Phone-banked for a political or social issue”

We create an index for political engagement by standardizing indicators for each of the above to have mean zero and standard deviation one in the main sample, adding these together, and then standardizing the sum to have mean zero and standard de-

viation one in the control group. *Note: this question was asked as the first question in the experimental survey, not in the initial screening survey.*

Appendix Figures C.4 through C.15 show the robustness of our main results to our choice of control variables.