Essays on Incentives and Human Capital Formation

Citation

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029622

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Essays on Incentives and Human Capital Formation

A dissertation presented
by
Ashley Cooper Craig
to
The Department of Economics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Economics
Harvard University
Cambridge, Massachusetts
March 2019
Abstract

I study policies designed to affect the production of human capital, either to improve efficiency or reduce inequality. Each chapter provides insights into how human capital accumulation is affected by the structure of the labor market, information, and policy interventions themselves.

In Chapter 1, I study optimal income taxation when workers’ human capital investment is imperfectly observable by employers. Bayesian employer inference about worker productivity drives a wedge between the private and social returns to human capital investment. The resulting positive externality from worker investment implies lower optimal marginal tax rates. I calibrate my model to match empirical moments from the United States. Taking into account the spillover from human capital investment introduced by employer inference reduces optimal marginal tax rates by 13 percentage points at around 100,000 dollars of income, with little change in the tails of the income distribution.

In Chapter 2 (joint with Roland Fryer), I develop a model of two-sided statistical discrimination, in which firms try to infer whether workers have made investments required for them to be productive, and simultaneously, workers try to deduce whether firms have made investments necessary for them to thrive. Complementarity between worker and firm beliefs complicates both empirical analysis designed to detect discrimination and policy meant to alleviate it. We propose “two-sided investment insurance” as an alternative tool that is weakly more effective than any alternative. The chapter concludes by proposing a way to identify statistical discrimination when beliefs are complementary.
In Chapter 3 (joint with David Martin), I study the causal effect of student suspension on test scores. Our identification comes from quasi-experimental variation in the impact of a policy change in 2012 in New York City, which eliminated suspensions for minor offenses such as smoking or using obscene language. For the majority of schools, in which suspensions for minor infractions were used extremely rarely, the new suspension policy necessarily had no impact. However, it led to a sharp reduction in the total suspension rate in schools that had previously used them. Our estimates indicate that more relaxed student discipline can be beneficial for students.
Contents

Abstract ................................................................. iii
Acknowledgments ......................................................... xv

Introduction ................................................................. 1

1 Optimal Income Taxation with Spillovers from Employer Learning ........................................... 5
  1.1 Introduction ............................................................. 5
    1.1.1 Connections in the Income Taxation Literature ...................... 10
  1.2 A Model of Optimal Taxation with Employer Learning ......................... 11
    1.2.1 Building Blocks ...................................................... 11
    1.2.2 Worker and Firm Payoffs ............................................ 12
    1.2.3 Worker and Firm Strategies ......................................... 12
    1.2.4 Equilibrium ........................................................... 13
    1.2.5 Optimal Firm and Worker Behavior .................................. 13
    1.2.6 Characterizing and Selecting Equilibria ............................... 14
    1.2.7 The Social Planner .................................................... 15
  1.3 A Simple Example with Linear Taxation ........................................ 16
    1.3.1 Additional Assumptions .............................................. 17
    1.3.2 Equilibrium ........................................................... 17
    1.3.3 Optimal Taxation ..................................................... 18
    1.3.4 Graphical Demonstration ............................................. 20
    1.3.5 Special Cases ........................................................ 22
    1.3.6 Tagging and Statistical Discrimination ................................. 22
  1.4 Non-linear Taxation .................................................. 24
    1.4.1 Perturbation of the Tax Schedule .................................... 24
    1.4.2 Regularity Assumptions .............................................. 25
    1.4.3 Mechanical Effect .................................................... 27
    1.4.4 Fiscal Externality ..................................................... 28
    1.4.5 Belief Externality ..................................................... 28
    1.4.6 “U” Shaped Tax Schedules and The Importance of Incidence .......... 30
    1.4.7 A Necessary Condition for Optimality ................................. 32
1.5 Quantitative Analysis ........................................... 34
  1.5.1 Evidence on Employer Learning ......................... 34
  1.5.2 New Evidence on Heterogeneity in Learning .......... 36
  1.5.3 Evidence on the Response of Productivity to Returns . 38
  1.5.4 Calibration to the United States Economy .......... 39
  1.5.5 Solving for Optimal Taxes ............................. 42
  1.5.6 A Naïve Benchmark for Comparison .................. 43
  1.5.7 Simulation Results ................................... 43
  1.5.8 Approximately Optimal Taxation ...................... 46
1.6 Unproductive Signaling ...................................... 47
  1.6.1 Unproductive Signaling: Example with Linear Taxation . 47
  1.6.2 Non-Linear Taxation: Signaling with Observable Investment . 49
  1.6.3 Non-Linear Taxation: Imperfectly Observable Investment .......... 51
1.7 Conclusion .................................................... 52

2 Complementary Bias: A Model of Two-Sided Statistical Discrimination 54
  2.1 Introduction ................................................ 54
  2.2 A Brief Review of the Literature .......................... 59
    2.2.1 Models of Discrimination ................................ 59
    2.2.2 Strategic Complementarities .......................... 61
  2.3 The Basic Model ........................................... 62
    2.3.1 Building Blocks ...................................... 62
    2.3.2 Payoffs .............................................. 63
    2.3.3 Strategies .......................................... 64
    2.3.4 Expected Payoffs ..................................... 64
    2.3.5 Bayesian Nash Equilibrium ............................. 66
    2.3.6 Dynamics ............................................. 70
    2.3.7 An Example with Uniform Cost and Signal Distributions .. 71
  2.4 Extending the Basic Model: One-Sided Policies .......... 73
    2.4.1 Equality in Offers ..................................... 74
    2.4.2 Equality in Employment .............................. 79
    2.4.3 Wage and Employment Subsidies ....................... 81
    2.4.4 An “Impossibility” Result ............................ 82
  2.5 Extending the Basic Model: Two-Sided Policies .......... 84
    2.5.1 Two-Sided Investment Insurance ..................... 84
    2.5.2 Affirmative Action .................................... 86
  2.6 Interpreting Group Differences in the Presence of Two-Sided Statistical Discrimination .................. 87
2.6.1 Detecting Employer Discrimination .................................................. 89
2.7 Conclusion ............................................................................................. 93

3 The Impact of Student Suspension on Human Capital Accumulation ................. 95
3.1 Introduction ........................................................................................... 95
3.2 Suspensions in the Literature and in Practice .............................................. 97
  3.2.1 Effects on Individual Outcomes .......................................................... 98
  3.2.2 Spillovers to Other Students .............................................................. 99
  3.2.3 Racial Discipline Gap ....................................................................... 99
  3.2.4 Zero Tolerance Policies .................................................................. 100
3.3 Institutional Context and Data .................................................................. 101
  3.3.1 Discipline in NYC Schools ................................................................. 101
  3.3.2 Evolution of Disciplinary Policy .......................................................... 102
  3.3.3 Data .................................................................................................. 105
3.4 Empirical Strategy ................................................................................... 108
  3.4.1 Fixed Effects Strategy: Estimating the Direct Effects of Suspensions .... 110
  3.4.2 Quasi-experiment: Harnessing a Change in Suspension Policy .......... 111
  3.4.3 Predicting Suspension Rates .............................................................. 113
3.5 Results ..................................................................................................... 113
  3.5.1 Fixed Effects Strategy ....................................................................... 113
  3.5.2 Quasi-experiment ............................................................................ 115
  3.5.3 Discussion ......................................................................................... 120
3.6 Conclusion ............................................................................................... 122

References .................................................................................................. 123

Appendix A Appendix to Chapter 1 ................................................................ 132
A.1 Generalized Contracts ........................................................................... 132
A.2 Continuity and Stability ......................................................................... 133
  A.2.1 Continuity of Investment Responses ............................................... 133
  A.2.2 Invertibility of $J_{f,x}$ .................................................................. 134
  A.2.3 Stability of Equilibria .................................................................... 135
A.3 Beyond the First Order Approach ............................................................ 136
  A.3.1 Special Case of the Model with Binary Investment ......................... 137
  A.3.2 Optimal Taxation with Binary Investment ....................................... 139
  A.3.3 Limitations of the First Order Approach ......................................... 141
A.4 Multiple Groups and Self-fulfilling Disparities .......................................... 141
A.5 Approximately Optimal Taxation ............................................................ 144
A.6 Unproductive Signaling with Observable Investment .................................. 145

vii
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.7 Proofs and Derivations</td>
<td>146</td>
</tr>
<tr>
<td>A.8 Simulation of the Model</td>
<td>162</td>
</tr>
<tr>
<td>A.8.1 Evaluation of a Single Perturbation</td>
<td>162</td>
</tr>
<tr>
<td>A.8.2 Decomposition of a Perturbation</td>
<td>163</td>
</tr>
<tr>
<td>A.8.3 Solving for the Optimal Tax Schedule</td>
<td>163</td>
</tr>
<tr>
<td>A.8.4 Recovery of Fundamentals</td>
<td>164</td>
</tr>
<tr>
<td>A.8.5 Additional Figures</td>
<td>164</td>
</tr>
<tr>
<td>A.8.6 Alternative Parameter Values</td>
<td>165</td>
</tr>
<tr>
<td>Appendix B Appendix to Chapter 2</td>
<td>170</td>
</tr>
<tr>
<td>B.1 Proofs</td>
<td>170</td>
</tr>
<tr>
<td>B.2 Stability</td>
<td>180</td>
</tr>
<tr>
<td>B.3 Derivations</td>
<td>182</td>
</tr>
<tr>
<td>B.3.1 Normalization of Worker Payoffs</td>
<td>182</td>
</tr>
<tr>
<td>B.3.2 Returns in Our Example</td>
<td>182</td>
</tr>
<tr>
<td>B.4 Further Extensions of the Model</td>
<td>183</td>
</tr>
<tr>
<td>B.4.1 Endogenous Wages</td>
<td>183</td>
</tr>
<tr>
<td>B.4.2 Marginal Firm Investment Costs</td>
<td>189</td>
</tr>
<tr>
<td>Appendix C Appendix to Chapter 3</td>
<td>191</td>
</tr>
<tr>
<td>C.1 Adjustment for 2013 Testing Waiver</td>
<td>191</td>
</tr>
<tr>
<td>C.2 Alternative Prediction Period</td>
<td>192</td>
</tr>
<tr>
<td>C.3 Supplementary Figures</td>
<td>192</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Heterogeneity in Employer Learning ............................................. 37
1.2 Calibrated and Implied Objects ...................................................... 40

3.1 Timeline of School Discipline Reform in New York City .................. 104
3.2 Most Common Infractions Leading to Suspension ............................. 107
3.3 Summary Statistics ...................................................................... 108
3.4 Most Common Level 2 Infractions Leading to Suspension ................. 112
3.5 Regressions of Standardized Math Scores on Suspensions ............... 114
3.6 Regressions of Standardized ELA Scores on Suspensions ................. 115
3.7 Change in Relationship Between Predicted Suspensions and Test Score Gains
   – Sub-samples ............................................................................... 121
3.8 Change in Relationship Between Predicted Suspensions and Math Score Gains
   – Sub-samples ............................................................................... 121

A.1 Summary Statistics for High and Low AFQT Workers ...................... 165
List of Figures

1.1 Timeline of the game ................................................................. 12
1.2 Effects of a reduction in the tax rate $\tau$ on wages and utility at each productivity level in the linear taxation example, calibrated to achieve $s = 0.75$ and match the United States wage distribution (see Appendix A.8 for details). The utility impacts in panel (b) are scaled by the productivity density so that the area under each curve is proportional to the average impact. ........................................... 21
1.3 The effect of a stylized perturbation to $T$: the hypothetical marginal tax change applies in the shaded region, lowering the slope of the relationship between after-tax and before-tax income. ................................................................. 25
1.4 Areas in which the productivity effect and redistributive effects are largest: the productivity effect is proportional to the height of $f(\hat{\theta}|q)$, while the redistributive effect is proportional to the change in $f(\hat{\theta}|q)$. ................................................................. 30
1.5 The effects on wages of a reduction in the marginal tax rate on income between $60,000$ and $61,000$, with a small belief externality (top panel) or a large externality (bottom panel). The baseline economy is the linear taxation example with $s = 0.75$, calibrated to match the United States wage distribution. The effect at each productivity level is scaled by the productivity density so that the area under each curve is proportional to the average wage change accounted for by that component. The gray-shaded bar shows the wage range that is directly affected by this perturbation. The productivity and income distributions are the same in both panels. ................................................................. 31
1.6 Effects on utility of a rise in the marginal tax rate on income between $60,000$ and $61,000$. The baseline economy is the linear taxation example, calibrated to achieve $s = 0.75$ in equilibrium and match the United States wage distribution. The effect at each productivity level is scaled by the productivity density so that the area under each curve is proportional to the aggregate impact. The gray-shaded bar shows the wage range that is directly affected by this perturbation. ................................................................. 33
1.7 Implications of the calibration procedure for the simulation described in Section 1.5. Panel (a) compares the empirical (target) and approximate (simulated) wage distributions. Panel (b) shows the relationship between expected wages and productivity in the baseline economy.

1.8 The optimal non-linear tax schedule in the simulation described in Section 1.5. The solid red line in the top panel shows the optimal tax schedule, while the dashed blue line shows a tax schedule that would be accepted by a naïve social planner who sets the sum of the mechanical effect and fiscal externality equal to zero. The lower panel shows a decomposition of small marginal tax cuts in each tax bracket. The tax function in this simulation is discretized into $20,000 brackets. Details of the procedure are available in Appendix A.

1.9 A comparison of the utility levels of agents at each productivity level under naïve and optimal taxation in the simulation described in Section 1.5. Further details are included in Appendix A.

2.1 Timing of actions in the model

2.2 This figure shows the nature of equilibria in the two-sided model. The left panel shows solutions to equation 2.5 when $\delta = \hat{\delta}$. Varying $\delta = \hat{\delta}$ traces out $\pi^*(\delta)$ in the right panel. A similar exercise can be conducted for equation 2.5 to yield $\delta^*(\pi)$. Finally, the three points of intersection between $\pi^*(\delta)$ and $\delta^*(\pi)$ – Z, U and S – are equilibria in the two-sided model.

2.3 Equilibria in the clear / unclear example

3.1 This figure provides a simplified outline of the set of decision points that lead to a suspension.

3.2 This figure shows the long-term trend in the total number of in-school and out-of-school suspensions used per year in the New York City public school system, including all suspensions in all grades. Data for this figure come from reports by the New York Civil Liberties Union (NYCLU).

3.3 This figure shows the proportion of students who are black, among students who are suspended for an infraction at each level, and across all students.

3.4 This figure shows the long-term trend in the number of suspensions for Level 2 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system. Similar figures for other suspension levels are available in Appendix C.
3.5 This figure shows the relationship between suspension rates and test scores in our sample of middle school students (grades 6-8). Students are grouped into bins based on percentiles of their math and ELA exam scores. The figure shows suspension rates in each bin plotted against the average in each bin. Graphs of suspension rates versus standardized scores are available in Appendix C.3.

3.6 This figure plots the coefficients, $\rho_s$, from regression 3.7. The results for Math are shown in the top panel, and ELA below. Each point measures the change in the relationship between test score growth and suspension rates across school-grades, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. Analogous graphs using test score percentiles are available in Appendix C.3.

3.7 This figure plots the coefficients, $\rho_{sk}$, from regression 3.8. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. Analogous graphs using test score percentiles are available in Appendix C.3.

A.1 This figure shows an example economy with binary investment. In the left panel, the aggregate rate of investment implied by worker and firm optimization, $G(\beta(\pi))$, is plotted against the employer prior, $\pi$. Any intersection between this line and the 45 degree line is an equilibrium. The arrows show the direction in which each equilibrium moves as $\tau$ rises. The right panel shows the set of equilibria over a range of values of $\tau$. Pareto dominant equilibria are shown by the black line segments.

A.2 This figure shows the results of the simulation described in Appendix A.8. The solid red line shows the optimal tax schedule, the dashed blue line shows the naive schedule, and the dotted black line shows a schedule what would be accepted by a planner who implemented equation A.17.

A.3 This figure compares the fiscal externality in each tax bracket under naive and optimal taxation, in the simulation described in Section 1.5.

A.4 This figure compares the belief externality in each tax bracket under naive and optimal taxation, in the simulation described in Section 1.5.
A.5 This figure compares the mechanical effect in each tax bracket under naïve and optimal taxation, in the simulation described in Section 1.5.

A.6 This figure plots the change in marginal social welfare weights starting from naïve taxation and transitioning to optimal taxation, for the simulation described in Section 1.5.

A.7 This figure shows the results of exercises similar to that described in Section 1.5 but with alternative parameter values. The top panel shows optimal and naïve tax schedules when $\varepsilon_l = 0.2$ and $\beta = 0.25$, which still yields $\varepsilon_w^{LR}/\varepsilon_z^{LR} = 0.6$ but with $\varepsilon_z^{LR} = 0.8$. The lower panel shows the adjustment to marginal tax rates in: (i) the baseline exercise; (ii) the case with $\varepsilon_z^{LR} = 0.8$; (iii) a third case with $\varepsilon_w^{LR}/\varepsilon_z^{LR} = 0.5$ but $\varepsilon_z^{LR}$ unchanged; and (iv) a fourth with $dE(w)/dq = 0.9$, which implies a smaller belief externality.

C.1 This figure shows the share of grade 8 students who take the Regents exam in mathematics, with and without taking the regular grade 8 exam.

C.2 This figure plots the coefficients, $\rho_{sk}$, from regression 3.8. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. In this version of the graph, we exclude all students who go on to sit the Regents exam for Math in grade 8.

C.3 This figure plots the coefficients, $\rho_{sk}$, from regression 3.8. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. In this version of the graph, we define pre-period suspension rates based on 2006-2009 data.

C.4 This figure shows the relationship between suspension rates and test scores in our sample of middle school students (grades 6-8). Students are grouped into bins based on their standardized math and ELA exam scores. The figure shows suspension rates in each bin plotted against the average in each bin.
C.5 This figure shows the accuracy of in-sample predictions of school-grade suspension rates, depending on which covariates are used. Available covariates include the race and gender composition of each school-grade, average age, the share of students who are English Language Learners (ELL), the share of students qualifying for free or reduced-price lunch, and the share of students who have previously repeated a grade.

C.6 This figure shows the long-term trend in the number of suspensions for Level 3 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system.

C.7 This figure shows the long-term trend in the number of suspensions for Level 4 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system.

C.8 This figure shows the long-term trend in the number of suspensions for Level 5 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system.

C.9 This figure plots the coefficients, $r_s$, from regression 3.7 using test score percentiles as an outcome rather than $z$-scores. The results for Math are shown in the top panel, and ELA below. Each point measures the change in the relationship between test score growth and suspension rates across school-grades, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period.

C.10 This figure plots the coefficients, $r_{sk}$, from regression 3.8 using test score percentiles as an outcome rather than $z$-scores. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period.
Acknowledgments

I am deeply grateful to my advisors, Roland Fryer, Lawrence Katz, David Laibson and Stefanie Stantcheva. All four provided invaluable advice, support and encouragement throughout graduate school. David started as my first year advisor, and I benefited from his words of wisdom from the beginning. Roland has been extraordinarily helpful and generous with his time; I cannot count the insightful conversations we have had over the years. Larry’s advice and encyclopedic knowledge have been instrumental in thinking through nearly every research idea I have ever had. Stefanie’s comments have never failed to be insightful, and her support genuine, especially as I moved toward the job market.

In addition to my committee, I am grateful to several other mentors, including Edward Glaeser, Benjamin Golub, Louis Kaplow, Casey Rothschild and Matthew Weinzierl. I also thank Roshan Verma for his help with background research for my third chapter.

Most importantly of all, I would like to thank my amazing wife Sophia for her support, patience and encouragement. I could not have done this with without her. Thank you also to Sebastian for being adorable, and for distracting me at all the right moments.

Finally, financial support from the Harvard Kennedy School Women and Public Policy Program, and a James M. and Cathleen D. Stone Ph.D Scholarship from the Multidisciplinary Program in Inequality & Social Policy at Harvard University is gratefully acknowledged.
To Sebastian and Sophia
Introduction

How should governments intervene in the production of human capital? Analyzing and evaluating policies that affect human capital investment is crucial, given its significant role in determining labor market outcomes and material well-being. In this dissertation, I provide new insights into how human capital accumulation is affected by the structure of the labor market, information, and changes in policy. In the process, I characterize optimal policy, which trades off the dual goals of correcting for inefficiencies and reducing inequality.

In Chapter 1, I study optimal income taxation when human capital investment is imperfectly observable by employers. In my model, Bayesian employer inference about worker productivity drives a wedge between the private and social returns to human capital investment by compressing the wage distribution. The resulting positive externality from worker investment implies lower optimal marginal tax rates, all else being equal.

To quantify the significance of this externality for optimal taxation, I calibrate my model to match empirical moments from the United States, including new evidence on how the speed of employer learning about new labor market entrants varies over the worker productivity distribution. Taking into account the spillover from human capital investment introduced by employer inference reduces optimal marginal tax rates by 13 percentage points at around 100,000 dollars of income, with little change in the tails of the income distribution. The welfare gain from making this adjustment is equivalent to raising every worker’s consumption by one percent.

If employers can categorize workers based on exogenous characteristics such as race or gender, an additional implication of my model is that they will statistically discriminate
in any situation in which the equilibrium productivity distribution varies by group. Discrimination may in turn motivate the planner to set group-specific marginal tax rates if differences in the size of the belief externality cause the private return to attaining a higher level of productivity to differ across groups.

Finally, I extend the model to allow inherently more able workers to have systematically higher or lower investment costs. Investment then serves two roles: increasing human capital, and revealing innate ability. Residual uncertainty about a worker’s productivity continues to produce wage compression due to employer belief formation, but there is also a signaling component of the private return that may be either positive or negative. In the extreme case in which productivity is entirely innate, the net externality from investment is negative. More generally, however, less accurate employer information reinforces the positive externality that arises from wage compression, but dampens the signaling externality.

In Chapter 2 (co-authored with Roland Fryer), I develop a model of two-sided statistical discrimination, in which firms try to infer whether workers have made investments required for them to be productive, and simultaneously, workers try to deduce whether firms have made investments necessary for them to thrive. Specifically, workers make costly and imperfectly observable investments in human capital, while firms make investments to create a work environment conducive to workers (e.g. flexible work hours for women, or affinity groups for minority workers), which are imperfectly observable to workers.

Complementarity between worker and firm beliefs in our model complicates policy meant to alleviate discrimination. We begin our analysis of policy by considering affirmative action in the sense of a requirement that firms make job offers to members of both groups with equal probability. In classic one-sided models, such a requirement leads to homogeneous employer beliefs when lower hiring standards do not undermine worker investment, but negative stereotypes about minorities may persist if low standards are too demotivating (Coate and Loury, 1993). Affirmative action can have the same issues in our model. But worse, affirmative action can undermine firm investment incentives and trigger zero investment by minority workers. Such severe inequality can be sustained indefinitely
in our model despite affirmative action. Moreover, it may be impossible for affirmative action to eliminate discrimination because firms have less incentive to invest in a numerical minority if investment costs are fixed.

Based on our model, we propose a new policy – which we label “investment insurance” – as a simple solution to statistical discrimination. We imagine that the government can observe a noisy version of the signals employers and workers receive, and offers them contracts. If the government believes an individual invested, it will subsidize them. The same assurance is provided to employers regarding their investment. This provides assurance to both workers and firms that their investments will pay off. Investment insurance can never be harmful to minority workers. Quite to the contrary, there is always a policy of this type that leads immediately to full equality.

The paper concludes by deriving a model-based empirical test for statistical discrimination by employers. The test is designed to be robust to the confounds of worker belief formation and complementarity between firm and worker investment. Our analysis focuses directly on the mechanism through which rational stereotyping affects incentives: pessimistic employers shrink their estimates of worker productivity toward the group mean, causing a flattening of the relationship between productivity and wages. Based on this insight, we propose examining workers who switch firms. Under the assumption that firms gain some private information about a worker’s ability with tenure, we demonstrate that wage profiles should flatten more for minority than majority workers when they move.

In Chapter 3 (joint with David Martin), I study the causal effect of student suspension on human capital accumulation – a controversial topic for education policy-makers who seek to balance deterrence and the need for a peaceful classroom against keeping at-risk students in class where they can learn. Using administrative data from public middle school students in New York City, we first document the negative unconditional relationship between suspensions and student test scores. Second, we present a fixed effects strategy, which shows that much of this negative association can be attributed to variation across schools and fixed differences between students. Nonetheless, there remains a negative
conditional relationship between suspensions and test scores.

Our final step is to harness quasi-experimental variation from the elimination in 2012 of suspensions for minor offenses such as smoking or using obscene language. The key to our identification is that such suspensions were used extremely rarely by the majority of schools; for these schools, the 2012 policy change had no impact. However, it led to a sharp reduction in the total suspension rate in schools that had used them extensively. Our results suggest that the abolition of suspensions was beneficial, on average, for students in the school-grades that were affected by them. This suggests that the recent focus by policymakers on alternative approaches to school discipline may indeed be justified.

In addition to providing input into to an active policy debate surrounding school discipline, our results contribute toward a better understanding of the human capital production function more generally. While it is well-established that test scores are causally improved by some schools (Angrist et al., 2013) and by better teachers (Chetty et al., 2011), it is much less clear what specific policies and methods are important in driving these gains. Our results suggest that the implementation of harsher discipline alone does not help students in traditional public schools, at least in New York City.
Chapter 1

Optimal Income Taxation with Spillovers from Employer Learning

1.1 Introduction

Employers base hiring and remuneration decisions on imperfect information. When evaluating workers, they rely on noisy correlates of productivity such as references, academic transcripts, and job market papers. Although employers’ beliefs about a given worker become more accurate over time, there can be a substantial delay before the worker’s wage reflects her marginal product (Farber and Gibbons, 1996; Altonji and Pierret, 2001; Lange, 2007; Kahn and Lange, 2014). Until then, employer inference based on imperfect information compresses the wage distribution, which drives a wedge between the present discounted private and social returns to raising one’s productivity.

In this way, rational inference by employers introduces a positive externality from human capital investment. Intuitively, a student who studies harder obtains higher future wages by improving her test scores, recommendations, and other indications of her ability. But with imperfect employer information, she also benefits from the hard work of other similar students: if her peers were to invest more, employers would tend to look more favorably on
her as well.\textsuperscript{1} Her peers do not internalize this spillover when choosing how hard to work, and invest less than is socially optimal. This principle applies to learning by any worker while at high school or college, and to investments later in life.

I study the role of income taxation to correct this type of externality. First, I develop a model of optimal taxation with imperfectly observable human capital investment. Next, I show with a simple example how Bayesian inference by employers compresses the wage distribution. This drives a wedge between the private and social returns to investment and lowers the optimal tax rate. Third, I generalize to non-linear taxation, and show that the downward adjustment to marginal tax rates is concentrated at intermediate levels of income. Finally, I calibrate the model to match empirical moments from the United States, introducing new evidence on how employer learning varies over the productivity distribution. Taking into account the spillover introduced by employer inference reduces optimal marginal tax rates by 13 percentage points at around 100,000 dollars of income, with little change in the tails of the income distribution. The welfare gain from this adjustment is equivalent to raising every worker’s consumption by one percent.

After observing her investment cost, each worker in my model makes an imperfectly observable investment in human capital, which determines her productivity.\textsuperscript{2} Employers cannot directly observe the worker’s true productivity level. Instead, they infer it based on a noisy but informative signal, combined with a prior belief. As a direct consequence of Bayesian inference by employers, every worker’s equilibrium wage is a weighted average of her own productivity and the productivity of other similar workers. An increase in investment by one group of workers therefore has the side effect of altering employers’ perceptions of other workers who send similar signals.

Taxation in this model has an effect on welfare that is not present in classic models of

\textsuperscript{1}This suggests that an encounter with one worker will affect assessments of other observably similar workers. Sarsons (2018) shows this occurs, although her results are hard to reconcile with full rationality.

\textsuperscript{2}Viewed through the lens of the model, obtaining a formal qualification is an imperfect signal of having raised one’s productivity. This is a useful approximation of the world to the extent that marginal increases in human capital accumulation require costly effort rather than simply arising from attendance at school.
income taxation (e.g., Mirrlees, 1971). When investment in human capital is depressed by higher taxes and productivity falls, employers become less optimistic, and pay workers a lower wage in equilibrium given the same information about their productivity.\(^3\) Individual workers do not take this into account. This is in addition to the usual fiscal externality, which arises because workers ignore the effects of their decisions on government revenue. Since the externality introduced by imperfect employer inference adds to the cost of taxation, taking it into account pushes toward lower marginal tax rates.

The core insights of my model apply more generally. For example, asymmetric employer learning leads to monopsony power for firms, which gives them an incentive to invest in their workers (Acemoglu and Pischke, 1998); but imperfect employer information still leads to underinvestment in skills.\(^4\) Similarly, introducing a motive for employers to screen their workers using contracts specifying both labor supply and a wage (Stantcheva, 2014) causes utility compression rather than wage compression, but nonetheless undermines the incentive for workers to invest in human capital.

If employers can categorize workers based on exogenous characteristics such as race or gender, my model implies that they will statistically discriminate in any situation in which the equilibrium productivity distribution varies by group.\(^5\) Discrimination may in turn motivate the planner to set group-specific marginal tax rates if differences in the size of the belief externality cause the return to increasing one’s productivity to differ across groups: for example, there is some evidence to suggest a lower return to skill for black workers than white workers (Bertrand and Mullainathan, 2004; Pinkston, 2006).

Using a simple example with linear taxation, I demonstrate how rational employer inference based on imperfect signals causes compression of workers’ wages toward the

---

\(^3\)In Section 1.4, I show how investment can also hurt others in some cases, although not in simple examples.

\(^4\)Asymmetric firm information may also indirectly undermine worker investment by affecting firms’ incentives to promote workers (Milgrom and Oster, 1987).

\(^5\)My model is a generalization of classic models of statistical discrimination. Pioneered by Phelps (1972) and Arrow (1973), such models rely on imperfect observability of productivity to explain employers’ use of a worker’s group identity. Contributions include Aigner and Cain (1977), Coate and Loury (1993), Moro and Norman (2004), Lang and Manove (2006), and Fryer (2007). See Fang and Moro (2011) for a review.
average level of productivity. This flattens the relationship between productivity and remuneration, introducing a wedge between the private and social returns to investment. Relative to a model with perfect employer information, the optimal tax is therefore lower; this correction is larger if employers have less precise information about their workers’ productivity, or if productivity is more responsive to taxation. In the special case in which all agents receive equal social welfare weight, the optimal tax is always negative, reflecting only the efficiency motive for intervention.

When I generalize to non-linear taxation, imperfect employer information introduces a novel effect of a change to the tax schedule, which I refer to as the belief externality: every worker who changes her investment decision also shifts employers’ beliefs, which in turn affects the wages and welfare of others. Less accurate employer information makes this externality larger, and pushes toward lower taxes. This is in addition to the two classic effects of income taxation: the mechanical effect from the transfer of consumption from high income workers to low income workers; and the fiscal externality, which arises because individuals ignore the impact on government revenue of re-optimization of their human capital investment and labor supply decisions.

The welfare impact of the belief externality is greatest at intermediate incomes, which contributes to a “U” shape of the optimal marginal tax schedule. There are two steps to understand this result. First, a given spillover in wages has a larger effect on consumption for higher-income workers, because they supply more labor. Second, as incomes rise even further, social welfare weights decline toward zero. In turn, this means that a given change in consumption has little effect on social welfare at the highest incomes.

My results also highlight how the belief externality can be decomposed into two components of opposite sign and different incidence. When a worker invests more, her higher productivity raises the wages of workers who send signals most similar to her own. Ho-

---

6The standard trade-off between equality and efficiency already produces a “U” shape, given the shape of the income distribution typically estimated (Diamond, 1998). Raising the marginal tax rate at a given income transfers resources from those above to those below that level, but distorts decisions locally. The “U” shape arises because the efficiency cost increases at low incomes as the density of income rises; it then decreases at high incomes, as the density falls. This shape is amplified by the forces in my model.
However, she hurts workers whose signal distributions are concentrated in regions where her own distribution changes the most. The reason for this negative effect is that she becomes more likely to send high signals where her productivity lowers the average, and less likely to send low signals where she had previously raised the average.

To quantify the importance of the belief externality, I calibrate my model to match the United States wage and income distributions, evidence on the gap between the private and social returns to productivity, and estimates of the elasticities of wages and labor supply (e.g., Blomquist and Selin, 2010). I calibrate the belief externality in two steps. First, I infer its overall size from existing estimates of the speed of employer learning (Lange, 2007; Kahn and Lange, 2014). Second, I use data from the National Longitudinal Survey of Youth (NLSY79) to show that there is stronger evidence of learning among low-productivity than high-productivity workers. As a proxy for worker productivity in my empirical work, I use scores on the Armed Forces Qualification Test (AFQT) from before each worker entered the labor market.

Taking into account the belief externality significantly reduces optimal marginal tax rates for most workers. Moreover, the welfare gain from adopting the optimal tax schedule is notable – equivalent to increasing every worker’s consumption by around one percent. As predicted by my theoretical results, the downward adjustment to taxes is concentrated at moderate-to-high levels of income, with little change to the marginal tax rates faced by workers with the lowest and highest incomes.

Finally, I extend the model to allow inherently more able workers to have systematically higher or lower investment costs à la Spence (1973). Investment then serves two roles: increasing human capital, and revealing innate ability. Residual uncertainty about a worker’s productivity continues to produce wage compression due to employer belief formation, but there is also a signaling component of the private return that may be either positive or negative. In the extreme case in which productivity is entirely innate, the net externality from investment is negative. More generally, however, less accurate employer information reinforces the positive component of the externality that arises from wage compression, but
dampens the signaling externality.

### 1.1.1 Connections in the Income Taxation Literature

This paper builds on a rich literature studying optimal income taxation, the modern analysis of which began with Mirrlees (1971). In these models, a social planner seeks to redistribute resources from high skill to low skill workers. A trade-off between equity and efficiency arises because workers’ skill levels are not directly observable by the planner. Redistribution must therefore occur via a tax on earnings, which distorts labor supply choices. Subsequent work (Diamond, 1998; Saez, 2001) has enriched our understanding of Mirrlees’ original results, and has extended them to incorporate extensive margin labor supply responses (Saez, 2002), lifecycle concerns Albanesi and Sleet (2006); Farhi and Werning (2013); Golosov et al. (2016), rent-seeking effects (Piketty et al., 2014), occupational choice (Gomes et al., 2018), and migration (Simula and Trannoy, 2010; Lehmann et al., 2014).7

My model most closely relates to the strand of this literature in which workers’ skills are attained through investment in human capital (Bovenberg and Jacobs, 2005; Jacobs, 2005, 2007; Boháček and Kapička, 2008; Maldonado, 2008; Kapička, 2015), which includes models with risky human capital (da Costa and Maestri, 2007; Stantcheva, 2017; Findeisen and Sachs, 2016), overlapping generations (Krueger and Ludwig, 2016) and ongoing learning (Best and Kleven, 2013; Makris and Pavan, 2017). In fact, when employer information becomes arbitrarily accurate in my model, it becomes isomorphic to the version of Bovenberg and Jacobs (2005) in which the social planner cannot observe human capital. Away from this limit, the externality introduced in my model by imperfect employer information suggests that marginal taxes should be lower, all else being equal.

There has been less attention paid to taxation with human capital investment that is imperfectly observable by employers. Where this has occurred, it has been limited to the case of purely unproductive signaling. For example, Andersson (1996) analyzes taxation in a two-type pure signaling model. Similarly, Spence (1974) discusses the case of pure

---

7See Piketty and Saez (2013) for a review of the literature on optimal labor income taxation.
signaling with perfectly inelastic labor supply. In related work, Stantcheva (2014) analyzes
the two-level screening problem that arises when labor disutility is directly related to a
worker’s productivity, so that willingness to work long hours signals high ability. However,
productivity is immutable in her model. Most similar in spirit, Hedlund (2018) analyzes
bequest taxation in a model that features a similar belief externality to mine, but in which
investment is binary and there is no redistributive motive for taxation.

The paper also connects to the literature on optimal income taxation with general
equilibrium externalities (e.g., Stiglitz, 1982; Rothschild and Scheuer, 2013, 2016; Sachs et al.,
2016; Lockwood et al., 2017), and to the broader literature on human capital externalities
in production (Moretti, 2004; Kline and Moretti, 2014). However, the aggregate production
function remains linear in my model. More importantly, the belief externality that is the focus
of this paper tends to have local incidence: rather than complementarities in production
between dissimilar types, the spillovers here arise because each worker’s investment decision
changes perceptions by employers about others who are observably similar. This distinction
is important in determining the shape of the optimal tax schedule.

1.2 A Model of Optimal Taxation with Employer Learning

1.2.1 Building Blocks

Let there be a fixed tax schedule $T$. This induces a game between a single worker and
several identical firms, indexed by $j \in J$ with $|J| \geq 2$. The timeline is shown in Figure 1.1.
Nature first distributes a cost of investment $k \in K \subseteq \mathbb{R}_{++}$ to the worker, with cumulative
distribution $G(k)$. After observing $k$, the worker invests $x \in \mathbb{R}_+$ at utility cost $kx$, yielding
productivity $q = Q(x)$ where $Q'(x) > 0$, $Q''(x) < 0$, $Q(0) = 0$ and $\lim_{x \to 0} Q'(x) = \infty$.\(^8\)

Nature then distributes a signal of productivity to the worker and all firms. Specifically,
let $\theta \in \Theta \subseteq \mathbb{R}_+$ be a non-contractible signal with conditional density $f(\theta|q)$, which is twice

\(^8\)The existence of a utility cost of investment is supported by the findings of Heckman et al. (2006a) and
Heckman et al. (2006b), and is appropriate to model unobservable investment.
Nature distributes investment cost
Firms see signal
Worker invests
Firms offer wages
Worker accepts highest wage
Worker supplies labor
Payoffs realized

Figure 1.1: Timeline of the game

continuously differentiable in $q$, and has full support for every $q$. Define $\overline{\theta} = \sup(\Theta)$ and let $f (\theta)$ be the marginal distribution of $\theta$. Assume that $f (\theta|q)$ is differentiable with respect to $\theta$, and that it satisfies the monotone likelihood ratio property: i.e., \[ \frac{d}{d\theta} \left( \frac{f(\theta|q_H)}{f(\theta|q_L)} \right) > 0 \] for all $q_H > q_L$. Based on $\theta$ and a prior $\pi (q)$, each firm forms a posterior belief about the worker’s productivity.

Next, each of $|J| \geq 2$ firms simultaneously offers a wage $w_j \in \mathbb{R}_+$ to the worker.\(^9\) The worker accepts her preferred offer, choosing firm $j \in J$, and supplies labor $l \in \mathbb{R}_+$. Of her pre-tax income $z$, the worker consumes $c = z - T (z)$ where the function $T \in \mathcal{T} \subseteq \mathcal{C} (\mathbb{R}_+, \mathbb{R})$ is the exogenous tax system set by the social planner.\(^10\)

1.2.2 Worker and Firm Payoffs

The worker receives utility $u (z - T (z), l) - kx$, where: $u_c > 0$, $u_l < 0$, $u_{cc} \leq 0$ and $u_{ll} < 0$. I further assume that $u_c$ is finite for all $c > 0$ and that $\lim_{l \to \infty} u_l = -\infty$ and $\lim_{l \to 0} u_l = 0$. Firms are risk neutral and obtain benefit $q$ per unit of supplied labor.

1.2.3 Worker and Firm Strategies

I focus on pure strategy equilibria. The worker’s strategy is a set of three functions – an investment decision, an acceptance rule and a labor supply decision. These can be written as: $x : K \times \mathcal{T} \to \mathbb{R}_+$; $A : K \times \mathcal{T} \times \Theta \times \mathbb{R}_+^{||l||} \to J$; and $L : K \times \mathcal{T} \times \Theta \times \mathbb{R}_+^{||l||} \to \mathbb{R}_+$. Each

\(^9\)Appendix A.1 shows that this is isomorphic to a model with contracts specifying labor supply and income, because the marginal rate of substitution between labor supply and consumption is independent of labor quality. If this is relaxed, employers may be able to use contracts to screen workers (Stantcheva, 2014).

\(^10\)I use $C (A, B)$ to denote the space of continuous functions mapping from $A$ to $B$. 

12
employer’s strategy maps signals and tax systems to wage offers $O_j : \Theta \times \mathcal{T} \rightarrow \mathbb{R}_+$. 

1.2.4 Equilibrium

An equilibrium of the game induced by a given tax schedule is a Perfect Bayesian Equilibrium (PBE). This requires that firms’ beliefs are rationally formed using Bayes rule whenever it applies, and that all strategies satisfy sequential rationality.

1.2.5 Optimal Firm and Worker Behavior

Each firm chooses the wage $w_j$ to maximize its expected profit $\bar{P}_j$. Letting $\Pr (A_j = 1|w_j)$ be the probability that the worker accepts firm $j$’s offer, its expected profit is:

$$\bar{P}_{j, \theta} = E[P_j|\theta, \pi, w_j] = \Pr (A_j = 1|w_j) \times (E[q|\theta, \pi, A_j = 1] - w_j) \times l$$

where $l$ is the quantity of labor supplied by the worker.

Given the assumptions above, every firm earns zero expected profit, and each worker receives a wage $w(\theta|\pi)$ equal to her expected marginal product $E[q|\theta, \pi]$ given the signal $\theta$ and the equilibrium distribution of productivity.

Lemma 1. Fix a value of $\theta$ and assume $E[q|\theta, \pi]$ is strictly positive and finite given beliefs $\pi(q)$. In any pure-strategy equilibrium, all firms $j \in J$ earn zero expected profit, and the wage offered to each worker by each firm is her expected marginal product $E[q|\theta, \pi]$.

All technical proofs are presented in Appendix A.7.

After accepting a wage offer, the worker supplies labor $l(\theta|\pi, T)$ as follows.\(^{11}\)

$$l(\theta|\pi, T) \in L^* = \arg\max_{ l \in \mathbb{R}_+ } u\left( w(\theta|\pi) \bar{I} - T \left( w(\theta|\pi) \bar{I} \right) , \bar{I} \right)$$

In turn, this implies that her income is $z(\theta|\pi, T) = w(\theta|\pi) l(\theta|\pi, T)$. Knowing this, the

\(^{11}\)Throughout the paper, optimal choices of labor supply and investment will be unique.
worker can calculate her expected utility $v(\theta|\pi, T)$ for any signal realization.

$$v(\theta|\pi, T) = u\left(z(\theta|\pi, T) - T(z(\theta|\pi, T))\right)$$

(1.2)

Evaluating the expectation of $v(\theta|\pi, T)$ by integrating over $\theta$, investing $x$ leads to expected utility $V(Q(x)|\pi, T) = E_{\theta}[v(\theta|\pi, T)|Q(x)] - kx$. At the investment stage, a worker with cost $k$ takes the function $V(q|\pi, T)$ as given, and optimally invests $x(k|\pi, T)$, which solves problem 1.3. This yields productivity $q(k|\pi, T)$.

$$x(k|\pi, T) \in X^* = \arg\max_{x \in R^+} \int_{\Theta} v(\theta|\pi, T) f(\theta|Q(x)) d\theta - k\bar{x}$$

(1.3)

In turn, these investment decisions collectively suffice to characterize the expected marginal product, and thus the wage, of an individual with signal realization $\theta$.

$$w(\theta|\pi) = \frac{\int_K q(k|\pi, T) f(\theta|q(k|\pi, T)) dG(k)}{\int_K f(\theta|q(k|\pi, T)) dG(k)}$$

(1.4)

The monotone likelihood ratio property ensures that the equilibrium wage is strictly increasing in $q$, and that $V(q|\pi, T)$ increases with $q$.

### 1.2.6 Characterizing and Selecting Equilibria

Equations 1.1, 1.3 and 1.4 comprise a fixed point at which worker investment decisions and employer beliefs are consistent. Each employer has a correct prior belief $\pi(q)$, and rationally updates it upon observing a signal. Perfect competition ensures that every firm offers the worker a wage equal to her expected marginal product. Combined with the signal distribution, this wage schedule then pins down the worker’s expected utility at each productivity level. Finally, the worker’s choices of productivity levels induce a productivity distribution that must coincide with every employer’s prior belief in equilibrium.

For any tax schedule $T$, there is a set of equilibria $E(T)$. I consider a selection of these equilibria, defined by choosing one equilibrium $E^+(T) \in E(T)$ for each $T$.$^{12}$ The expected

---

$^{12}$An alternative is simply to assume that the initial equilibrium is stable. This ensures that the economy does not switch equilibria in response to a small change in $T$. I define stability in Appendix A.2.
utility of a worker with investment cost \( k \) is then defined as her expected utility given the tax schedule and this selection: \( \bar{V} (k, T) = V (k, E^\dagger (T), T) \). For example, one possibility is to assume that agents always coordinate on one of the social planner’s preferred equilibria. I assume that this is the case when presenting the results for non-linear taxation in Section 1.4. However, my approach is equally valid for other selections.

Note 1. The game here is described as one between a single worker and a set of firms, with the worker’s type \( k \) drawn from \( G (k) \). An alternative interpretation is that there is a continuum of workers whose investment costs have distribution \( G (k) \) in the population. I adopt this terminology throughout much of the paper.

1.2.7 The Social Planner

Having established the nature of an equilibrium, I now introduce the social planner. The role of this planner is to choose a tax schedule \( T \) to maximize social welfare \( W (T) \), which is defined as the average across types of the worker’s expected utility levels after they have been transformed by a social welfare function \( W \).\(^{13}\)

\[
\max_{T \in T} W (T) = \int_k W (\bar{V} (k, T)) \, dG (k)
\]

The social welfare function \( W \) is assumed to be increasing, concave and differentiable.

The choice of \( T \) must satisfy two constraints. First, it can be a direct function only of realized income \( z \). Second, the planner must raise enough tax revenue to cover an exogenously fixed revenue requirement, \( R \). In some examples, I further restrict \( T (z) \) to be linear in \( z \) (Section 1.3, Appendix A.3 and Appendix A.4).

The planner’s problem can be written as a choice of a tax system to maximize welfare, subject to the resource constraint, individual optimization and rational belief formation.

\[
\max_{T \in T} W (T) = W (\bar{V} (k, T)) \, dG (k)
\]

\(^{13}\)I omit profits from social welfare because they are zero in expectation.
where:
\[
V(k, T) = \int_{\Theta} \left( v(\theta | \pi, T) - kx(k, \pi, T) \right) f(\theta, q(k | \pi, T)) \, d\theta
\]

subject to:
\[
x(k | \pi, T) \in \arg\max_{\tilde{x} \in \mathbb{R}_+} \int_{\Theta} v(\theta | \pi, T) f(\theta | Q(\tilde{x})) \, d\theta - k\tilde{x} \tag{1.6}
\]
\[
l(\theta | \pi, T) \in \arg\max_{\tilde{l} \in \mathbb{R}_+} \left( w(\theta | \pi) \tilde{l} - T \left( w(\theta | \pi) \tilde{l} \right) , \tilde{l} \right) \tag{1.7}
\]
\[
w(\theta | \pi) = \frac{\int_{k} q(k | \pi, T) f(\theta | q(k | \pi, T)) \, dG(k)}{\int_{k} f(\theta | q(k | \pi, T)) \, dG(k)} \tag{1.8}
\]
\[
R = \int_{\Theta} T(z(\theta | \pi, T)) f(\theta) \, d\theta \tag{1.9}
\]

In summary, the planner’s choice of a tax system \( T \) alters the set of equilibria in the economy. Given a selection from this equilibrium correspondence – for example, the planner’s preferred equilibrium for each tax schedule – the planner maximizes welfare. Changes in the tax schedule shift the worker’s incentives to invest and her willingness to supply labor. Due to imperfect employer information, the worker’s investment decisions also affect equilibrium wages – an effect she ignores when she invests.\(^{14}\)

### 1.3 A Simple Example with Linear Taxation

I begin with an example in which the planner is restricted to choosing a linear tax, \( \tau \).\(^{15}\) Each worker’s consumption is then an average of her own income, \( z \), and the mean income, \( \bar{z} \):
\[
c = (1 - \tau) z + \tau \bar{z}.
\]
This example highlights an effect that is not present in classic models of income taxation: when productivity falls in response to higher taxes, employers become less optimistic, and pay workers lower wages given the same information about their productivity. Taking this effect into account leads to a lower optimal tax.

---

\(^{14}\)This can be thought of as a problem with inner and outer components à la Rothschild and Scheuer (2013), with rational belief formation serving as the consistency constraint. A difference is that Rothschild and Scheuer (2013) re-write the social planner’s problem as a direct choice over allocations.

\(^{15}\)For simplicity, I also assume that the government’s revenue requirement, \( R \), is zero.
1.3.1 Additional Assumptions

For convenience, I assume that workers have quasilinear isoelastic utility and that the production function for investment is also isoelastic.\textsuperscript{16}

\[ u = c - l^{1 + \frac{1}{\eta}} \left( 1 + \frac{1}{\varepsilon l} \right) \]

I also make assumptions about the cost and signal distributions, which jointly yield a tractable signal extraction problem for employers. First, the relationship between the signal \( \theta \) and productivity \( q \) is: \( \ln \theta = \ln q + \ln \xi \), where \( \ln \xi \sim N(0, \sigma^2_\xi) \). Secondly, investment costs \( k \) are distributed log-normally: \( k \sim LN(\ln \mu_k - \sigma^2_k/2, \sigma^2_k) \).

1.3.2 Equilibrium

Given any linear tax rate \( \tau \), there is an equilibrium in which productivity and income are both log-normally distributed; and in this equilibrium, a worker’s wage is a weighted geometric average of her own productivity \( q \), average productivity \( \mu_q \), and idiosyncratic noise. The weight on a worker’s own productivity is the fraction of the variance of the signal that arises due to variation in productivity rather than noise, \( s = \sigma^2_q / (\sigma^2_q + \sigma^2_\xi) \). Intuitively, the signal is only useful to employers to the extent that variation in it reflects differences in productivity rather than noise.

**Proposition 1.** For any fixed tax rate \( \tau \), there exists an equilibrium in which productivity and income are both log-normally distributed.

\[ \ln q \sim N \left( \ln \mu_q - \frac{\sigma^2_q}{2}, \sigma^2_q \right) \]

A worker’s wage is \( w = q^s \mu^{1-s}_q \xi^s \) where \( s = \frac{\sigma^2_q}{\sigma^2_q + \sigma^2_\xi} \in (0, 1) \).

The weight on a worker’s own productivity is a measure of the wedge between the private and social returns to investment. If a worker of a given cost type were to unilaterally

\textsuperscript{16}I assume \( \beta (1 + \varepsilon l) < 1 \) so that investment returns are concave and choices finite (see Appendix A.7).
increase her productivity by one percent, her expected wage would increase by $s < 1$ percent. If most of the variance in the signal $\theta$ comes from noise ($\sigma^2_\theta$ large), $s$ is close to zero, and employers place little weight on the signal when setting a worker’s wage. There is then little private return to investment. Alternatively, if $\sigma^2_\theta$ is small, then $s$ is close to one, and the private return to investment is close to the social return.

The simplicity of this example stems from the fact that the elasticities of investment and income with respect the retention rate, $1 - \tau$, are constant and independent of $s$. This may seem surprising, since more noise (lower $s$) flattens the relationship between a worker’s log productivity and her log wage. However, there is a second effect of lower $s$: employers place more weight on average productivity, which strengthens a social multiplier in the model. In response to a fall in $\tau$, workers invest more, and $\mu_q$ rises; this further increases investment returns, and amplifies the response of productivity to taxation. These two effects of noise cancel out in this example, leaving the elasticities unaffected.

**Lemma 2.** Assume that the log-normal equilibrium from Proposition 1 is played. The elasticities of productivity ($\epsilon_q$) and income ($\epsilon_z$) with respect to the retention rate $1 - \tau$ are:

$$
\epsilon_q = \frac{\beta (1 + \epsilon_l)}{1 - \beta (1 + \epsilon_l)}
$$

$$
\epsilon_z = \frac{\epsilon_l + \beta (1 + \epsilon_l)}{1 - \beta (1 + \epsilon_l)}
$$

### 1.3.3 Optimal Taxation

Building on Lemma 2, Proposition 2 provides a formula for the optimal linear tax. First, let $\psi_k = W' (\nabla (k, \tau))$ be the marginal social welfare weight placed on an individual with cost $k$, and let $\bar{\psi}$ be the average welfare weight. Similarly, define $\bar{z}_k$ as the average income for individuals with cost $k$, and let $\bar{z}$ be the population average income. The first-order condition for the optimal tax is given by equation 1.10.

**Proposition 2.** Assume that the log-normal equilibrium described in Proposition 1 is played. The
first-order condition for the optimal linear tax $\tau^*$ is:

$$\frac{\tau^*}{1 - \tau^*} = \frac{1 - \gamma}{\varepsilon_z^{\text{Standard}}} - \gamma (1 - s) \varepsilon_q^{\text{New}}$$ \hspace{1cm} (1.10)

where: $\gamma = E_k \left\{ \frac{\psi_k z_k}{\psi z_k} \right\} \geq 0$.

Equation 1.10 is similar to the optimal tax formula in the standard case with perfect employer information. Indeed, the first term captures the usual trade-off between redistribution and distortion. The second term is new, and captures the intuition that workers who become more productive impose a positive externality on others by making employers more optimistic, and raising the wage paid for any given signal realization.

The formula in Proposition 2 can be derived by combining the three effects of slightly raising the tax rate. First, there is a mechanical effect ($\text{ME}$), which is the welfare gain from taking money from high-income individuals and redistributing it.

$$\text{ME} = \overline{\psi z} - E_k \left( \overline{\psi z_k} \right)$$

This transfer raises social welfare to the extent that workers with high income have lower-than-average welfare weight: $E_k \left( \overline{\psi z_k} \right) < \overline{\psi z}$. If welfare weights decline rapidly with income, $\gamma$ is close to zero and $\tau^*$ is high. Conversely, a social planner with only a weak preference for redistribution has $\gamma$ close to one, which implies a low value of $\tau^*$.

The second traditional effect of taxation is the fiscal externality ($\text{FE}$), which captures the impact of changes in labor supply and investment decisions on the government budget.

$$\text{FE} = -\tau \overline{\psi \varepsilon z} \frac{\bar{z}}{1 - \tau}$$

When workers re-optimize in response to a change in $\tau$, the effect of this on their own welfare is second-order (by the envelope theorem). However, there is a first-order effect on government revenue, which is returned to workers. In classic income taxation models, the fiscal externality is a sufficient statistic for the cost of taxation Feldstein (1999).

With imperfect employer information, there is a new effect which I refer to as the belief
externality (BE). When workers increase their investment, they do not take into account their effects on employer beliefs, which translate into changes in the equilibrium wage paid for each signal realization. This constitutes a second externality.

\[
BE = -E_k (\psi_k z_k) (1 - s) \epsilon_q
\]

The new effect pushes toward a lower tax rate. Its magnitude depends on three factors. First, it rises with the size of the wedge between private and social returns, \(1 - s\). Second, it scales with \(\epsilon_q\), because the externality arises from workers becoming more productive. Third, the welfare impact scales with \(E_k (\psi_k z_k)\), because higher income individuals – who supply more labor – are more affected by a given change in their wage.

1.3.4 Graphical Demonstration

The effects of a small reduction in the linear tax rate are shown in Figure 1.2. Panel (a) shows the change in wages at each level of productivity, and decomposes it into the direct effect from a worker’s own re-optimization and the indirect effect via employer beliefs. I assume that \(s = 0.75\), which is a value that aligns with the evidence on employer learning (see Section 1.5). This implies that 25 percent of the change in the average wage is not internalized by the workers who respond to the tax change.

Panel (b) shows the utility impacts of the mechanical effect, fiscal externality and belief externality. The effects are weighted by the density of productivity so that the area between each curve and the zero line is the average utility impact. Since the tax rate has been reduced, there is a mechanical transfer of utility from low productivity to high productivity workers. Second, there is a positive fiscal externality, as incomes rise and the government collects more revenue. Finally, there is a positive belief externality, as employers become more optimistic and pay workers a higher wage given any signal realization.
Figure 1.2: Effects of a reduction in the tax rate $\tau$ on wages and utility at each productivity level in the linear taxation example, calibrated to achieve $s = 0.75$ and match the United States wage distribution (see Appendix A.8 for details). The utility impacts in panel (b) are scaled by the productivity density so that the area under each curve is proportional to the average impact.
1.3.5 Special Cases

It is instructive to consider three special cases of the optimal tax formula. First, if employers perfectly observe productivity \((s = 1)\), equation 1.10 collapses to the standard case.

\[
\frac{\tau^*}{1 - \tau^*} \bigg|_{s=1} = \frac{1 - \gamma}{\varepsilon_z} \tag{1.11}
\]

While it is critical that \(\varepsilon_z\) incorporates the long-run response of human capital in any calibration, this equation is otherwise the same as that which arises in a model with fixed productivity types and perfect employer information.\(^{17}\)

In general, however, there is an efficiency motive to intervene, and this is reflected by the formula that arises when the planner has no redistributive motive (i.e., \(\psi_k = 1\forall k\)).

\[
\frac{\tau^*}{1 - \tau^*} \bigg|_{\psi_k = 1\forall k} = -\frac{(1 - s) \varepsilon_q}{\varepsilon_z}
\]

In this case, the planner simply aims to align private and social returns. This can be contrasted with a Rawlsian social planner, who cares only about the highest-cost worker. The Rawlsian tax rate, \(\frac{\tau}{1 - \tau} \bigg|_{\tau} = \frac{1}{\varepsilon_z}\), maximizes government revenue, and is unchanged from the standard case because the highest-cost worker is unaffected by the belief externality.

1.3.6 Tagging and Statistical Discrimination

If employers can observe exogenous characteristics about a worker such as race, gender or disability status, this model implies that they will statistically discriminate in any situation in which groups of workers differ in their equilibrium productivity distributions. The logic is simple: if the productivity distributions differ, then employers rationally make different assessments of a worker’s productivity given the same signal.

For example, suppose there is an advantaged group \(A\) and a disadvantaged group \(D\), which are identical except that group \(D\)’s costs are proportionally higher than group \(A\)’s \((\mu^D_k > \mu^A_k)\). This implies that the equilibrium wage and income distributions of \(D\) workers

\(^{17}\)In this case, both the condition for optimal taxes and the elasticity \(\varepsilon_z\) are identical to the results of Bovenberg and Jacobs (2005) under the assumption that the planner cannot directly subsidize investment.
are shifted down relative to those of A workers. An audit study in this economy would reveal a positive wage gap between A and D workers with identical signals.

\[
\ln \left( \frac{w(\theta | \pi_A)}{w(\theta | \pi_D)} \right) = (1 - s) \ln \left( \frac{\mu_k^A}{\mu_k^D} \right) \quad (1.12)
\]

Specifically, with \( s = 0.75 \), discrimination would appear to "account for" around one quarter of the overall wage gap between the two groups.

This raises the question of whether discrimination motivates the planner to set different tax rates for each group. There are two traditional reasons for doing so. First, the elasticity of taxable income, \( \varepsilon_z \), may differ between groups; and second, the covariance between incomes and welfare weights, \( \gamma \), may differ.\(^{18}\) In this model, there is an additional tagging motive: the size of the belief externality may differ across groups.

A key result is that the size of the externality depends on the dispersion but not the level of investment. As a result, a cost disadvantage as in the example above does not affect the externality, and does not provide a motive to differentiate between groups. In this sense, statistical discrimination does not necessarily imply that a group-specific subsidy is optimal. This contrasts with subsidy programs suggested based on classic models of purely "self-fulfilling" statistical discrimination (Coate and Loury, 1993).\(^{19}\)

**Corollary 1.** The standard deviation of log wages (\( \sigma_q^2 \)) and signal-to-noise ratio \( s = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_x^2} \) are pinned down by the following condition.

\[
\sigma_q^2 = \left( \frac{\beta}{\beta s (1 + \varepsilon_l)} \right)^2 \sigma_k^2
\]

The optimal tax rate is thus independent of the level of costs \( \mu_k \), and does not directly depend on the level of log productivity \( \mu_q \).

---

\(^{18}\)For discussions of "tagging", see Akerlof (1978), Kaplow (2007), and Mankiw and Weinzierl (2010). See also Fryer and Loury (2013), who analyze policies designed to improve the outcomes a disadvantaged group in a job assignment model with endogenous investment in skills.

\(^{19}\)In simple models of self-fulfilling statistical discrimination, investment is binary and the social benefit is constant. Any equilibrium with less investment must then feature a larger externality. The question these papers ask is also different: they focus on equilibrium selection, which I discuss in Appendix A.4.
The underlying reason for statistical discrimination is critical, however. For example, suppose that the cost distributions of groups \( A \) and \( D \) are identical but that the signal is noisier for group \( D \); this has been posited as a reason for statistical discrimination (see Phelps, 1972; Aigner and Cain, 1977; Lundberg and Startz, 1983). In this case, the belief externality is larger for the disadvantaged group, and \( s \) is lower. The social planner therefore has a motive to set \( t_D < t_A \). Alternatively, statistical discrimination may be self-fulfilling in the sense of Arrow (1973)\(^{20}\), which is an issue that I take up in Appendix A.4.

If the reason for a disparity is unknown, the right question to ask from the point of view of the social planner is whether the return to increasing one’s productivity differs across groups – i.e., whether the belief externality is larger. For example, there is some evidence that the return to skill is either lower for black workers than white workers (Bertrand and Mullainathan, 2004; Pinkston, 2006), or roughly equal (Neal and Johnson, 1996). Combined with assumptions about group differences in \( z \), \( q \) and \( g \), such evidence provides guidance as to whether group-specific tax rates are theoretically warranted.

### 1.4 Non-linear Taxation

#### 1.4.1 Perturbation of the Tax Schedule

I now relax the restrictive assumptions of Section 1.3, and derive a necessary condition for optimal non-linear taxation by studying a small perturbation to the tax schedule. Specifically, I consider raising the marginal tax by \( d\tau \) over a small range of incomes between \( z \) and \( z + dz \), where \( d\tau \) is second-order compared to \( dz \).\(^{21}\) This is accompanied by a change in the intercept of the tax schedule – a uniform increase in the consumption of all workers – to ensure that the resource constraint still holds with equality.

An example of such an experiment is shown in Figure 1.3. Studying the effects of this perturbation leads to a tax formula that bears a conceptually close relationship to the

---

\(^{20}\)See also Coate and Loury (1993), Moro and Norman (2004), and Craig and Fryer (2017).

\(^{21}\)Requiring \( d\tau \) to be small abstracts from bunching and gaps from introducing a kink in the tax schedule.
Figure 1.3: The effect of a stylized perturbation to $T$: the hypothetical marginal tax change applies in the shaded region, lowering the slope of the relationship between after-tax and before-tax income.

standard one that arises when workers simply receive their marginal product (Mirrlees, 1971; Diamond, 1998; Saez, 2001). As in the example above, there are three effects: a mechanical effect (ME), a fiscal externality (FE) and – new to this model – a belief externality (BE).

1.4.2 Regularity Assumptions

In deriving a condition that characterizes the optimal tax, I take a continuously differentiable tax schedule $T$ and the social planner’s preferred equilibrium given that tax schedule, $E^*(T)$.\textsuperscript{22} I adopt regularity assumptions, which jointly ensure that a worker’s income responds smoothly to small changes in her wage or the tax schedule around this initial point, and that there is – generically, for an arbitrarily chosen tax schedule – a locally unique

\textsuperscript{22}Although I assume for concreteness that the planner can implement her preferred equilibrium, my approach is equally valid for any other locally continuous selection of equilibria.
Fréchet differentiable function mapping tax schedules to investments.\textsuperscript{23}

First, I make the standard single crossing assumption, which is that the marginal rate of substitution between income and consumption is decreasing in the wage (Assumption 1). Second, I assume that individuals’ second-order conditions for labor supply hold strictly (Assumption 2). As discussed by Saez (2001), this requires that $1 - T'(z) + \varepsilon'_zzT''(z) > 0$, where $\varepsilon'_z$ is the compensated elasticity of taxable income with respect to her wage. Assumption 2 can be viewed as a restriction on the curvature of the tax schedule. It always holds in my simulations, and must hold if $T''(z) \geq 0$.\textsuperscript{24}

**Assumption 1 (Single Crossing).** The marginal rate of substitution between income and consumption, $-\frac{u_l(c, z)}{w u(c, z)}$, is decreasing in $w$.

**Assumption 2 (Labor Supply SOC).** The second derivative of the tax schedule $T''(z)$ is bounded strictly below by $-\frac{1}{\varepsilon'_z}[1 - T'(z)]$.

Third, I assume that investment returns are strictly concave, which implies that workers’ second-order conditions for investment hold strictly (Assumption 3). This is a joint restriction on the tax schedule, cost distribution $G(k)$ and investment technology $Q(x)$. For any income, wage and productivity distributions, and any tax schedule, there exist underlying cost distributions and investment technologies such that Condition 1.13 holds. It can also be relaxed, with the key requirement being that workers are not indifferent between two investment levels. With finitely many cost types, this is satisfied generically for equilibrium investment choices; and with a continuum of cost types, the analysis is unchanged if it is violated for countably many types.

\textsuperscript{23}I discuss the existence of such a unique selection of equilibria in Appendix A.2. Appendix A.3 then discusses why the planner may in some cases choose to locate at a singularity where these conditions break down.

\textsuperscript{24}Failure of Assumption 2 implies bunching of workers with different wages at the same level of income. Accounting for bunching is conceptually straightforward, but unnecessarily complicates the exposition.
Assumption 3 (Investment SOC). Investment returns are strictly concave for all \( x \).

\[
- \frac{Q''(x)}{Q'(x)^2} > \frac{\int_\Theta v(\theta|\pi, T) \left. \frac{\partial^2 f(\theta|q)}{\partial q^2} \right|_{q=Q(x)} d\theta}{\int_\Theta v(\theta|\pi, T) \left. \frac{\partial f(\theta|q)}{\partial q} \right|_{q=Q(x)} d\theta}
\]

(1.13)

1.4.3 Mechanical Effect

Subject to these regularity assumptions, there are three effects of this perturbation. I begin with the mechanical effect (ME), which is isomorphic to the standard model in which workers are paid their marginal product. Raising the marginal tax at income \( z \) collects revenue from workers with income greater than \( z \) and redistributes it equally to all workers by raising the intercept of the tax schedule.

Assumptions 1 and 2 jointly ensure that income is strictly increasing in \( q \). As a result, \( z(\theta|\pi, T) \) can be inverted to obtain \( \theta(z|\pi, T) \). Defining \( G(k|\theta) \) as the distribution of \( k \) conditional on \( \theta \), and letting \( \psi(k) = W'(V(k, T)) \), the mechanical gain in welfare is:

\[
\int_{\Theta} u_c(\theta) \int_{K} \psi(k) dG(k|\theta) d\theta \times \int_{\Theta(z|\pi, T)} f(\theta) d\theta - \int_{\Theta(z|\pi, T)} u_c(\theta) \int_{K} \psi(k) dG(k|\theta) f(\theta) d\theta
\]

Value of transfer to average worker

\[
\int_{\Theta(z|\pi, T)} u_c(\theta) \int_{K} \psi(k) dG(k|\theta) f(\theta) d\theta
\]

Loss due to transfer from high income workers

To simplify this expression, let \( H(z) = \int_0^z h(v) dv \) be the CDF of income. Secondly, let \( \psi_z(z) \) be the normalized marginal social welfare weight of a worker with income \( z \).

\[
\psi_z(z) = \frac{u_c(\theta(z|\pi, T)) \int_{K} \psi(k) dG(k|\theta(z|\pi, T))}{\int_{\Theta} u_c(\theta) \int_{K} \psi(k) dG(k|\theta) f(\theta) d\theta}
\]

Finally, define \( \Psi(z) = \int_0^z \psi_z(v) h(v) dv \) as the cumulative welfare weight of workers with income less than \( z \). Using these definitions, the mechanical gain can be written as:

\[
\text{ME}(z) = d\tau dz \times \{ \Psi(z) - H(z) \}.
\]

Since \( \psi_z(z) \) is decreasing in \( z \), the welfare weight below any finite level of income is higher than the population weight.\(^{25}\) This in turn implies that \( \text{ME}(z) > 0 \). Intuitively,

\(^{25}\)Since \( \psi(k) \) is increasing in \( k \), the assumptions on \( f(\theta|q) \) guarantee that \( \int_{K} \psi(k) dG(k|\theta) \) is decreasing in \( \theta \).
transferring income from relatively rich individuals to the broader population of workers mechanically raises social welfare for a planner with a taste for redistribution.

1.4.4 Fiscal Externality

The second effect of the perturbation is a fiscal externality, which arises because workers ignore the effects of their decisions on government revenue. Although this effect appears in every income taxation model, it is more complicated with employer belief formation. Not only do all agents respond directly, but each response in turn changes the investment incentives of other workers. The fiscal externality is thus governed by the evolution of a fixed point at which workers’ investment decisions are optimal given employers’ beliefs and employers’ beliefs are rational given workers’ investment decisions.

The total fiscal externality is given by equation 1.15, and is comprised of two effects: changes in the level of income corresponding to each signal realization, and changes in the marginal density of the signal.\textsuperscript{26} In turn, the income response for a given signal realization captures two types of reaction: (i) direct responses of labor supply to the policy change; and (ii) changes in both wages and labor supply due to shifts in employer beliefs. Changes in the marginal density of the signal, \( f(q) \), capture workers’ investment responses.

\[
\text{FE}(z) = -d \tau dz \int_{\Theta} \left\{ T'(z(\theta|\pi, T)) \left( \frac{dz(\theta|\pi, T)}{d \left[ 1 - T'(z) \right]} \right) f(\theta) + T(z(\theta|\pi, T)) \frac{df(\theta)}{d \left[ 1 - T'(z) \right]} \right\} d\theta \\
\text{Response for given } \theta \text{ realization} \quad \text{Investment response}
\]

1.4.5 Belief Externality

The final effect of the perturbation is new to this model. When individuals re-optimize their investment decisions, they disregard their effects on the equilibrium wage paid for a given signal realization, \( w(\theta|\pi) \). Taking any signal realization \( \theta \), this wage externality is

Finally, \( u_c(\theta) \) is weakly decreasing and \( z(\theta|\pi, T) \) strictly increasing, so \( \psi_z(z) \) is strictly decreasing.

\textsuperscript{26}The derivatives in equation 1.15 are causal responses to this perturbation to the tax schedule. They cannot therefore be directly related to properties of the utility function.
comprised of two components, corresponding to the two effects of investment: an increase in productivity, and a shift in the signal distribution.

\[
\frac{dw(\theta | \pi)}{d \left[1 - T'(z)\right]} f(\theta) = \int_{k} \left( \frac{dq(k | \pi, T)}{d \left[1 - T'(z)\right]} \right) f(\theta | q(k | \pi, T))
\]

Productivity effect

\[
+ \left[ q(k | \pi, T) - E(q | \theta, \pi) \right] \left( \frac{\partial f(\theta | q)}{\partial q} \right)_{q=q(k | \pi, T)} dG(k)
\]

Redistributive effect

The first component of equation 1.16 is the productivity effect. A worker who invests more shifts employers beliefs upward, and causes the wage paid to individuals with signal $\tilde{q}$ to rise despite their qualification decisions being unchanged.

The second component is the redistributive effect, and is generally negative. It arises because a worker who raises her investment has higher productivity than the group she leaves, but lower productivity than the group she joins. This manifests in the effect of investment on the distribution of signals observed by employers. If $f(\tilde{\theta} | q)$ increases, $w(\tilde{\theta} | \pi)$ rises if $q > E(q | \tilde{\theta}, \pi)$ and falls if $q < E(q | \tilde{\theta}, \pi)$. The opposite occurs if $f(\tilde{\theta} | q)$ decreases. The reason the effect tends to be negative is that an increase in $q$ increases $f(\tilde{\theta} | q)$ at high values of $\tilde{\theta}$ where $q < E(q | \tilde{\theta}, \pi)$ and decreases $f(\tilde{\theta} | q)$ at low values of $\tilde{\theta}$ where $q > E(q | \tilde{\theta}, \pi)$.

The productivity and redistributive effects differ in both sign and incidence, as shown in Figure 1.5. When workers re-optimize, the productivity effect raises the equilibrium wages of workers whose signal distributions overlap most with those who increased their productivity. In contrast, the redistributive effect reduces the wages of workers who send signals in regions where the signal distribution changes most.

The importance of these differences in incidence are apparent in Figure 1.6, which starts from the linear tax example in Section 1.3 and shows the simulated effects on wages of a reduction in the marginal tax rate on income between $60,000 and $61,000. When the marginal tax rate falls and productivity rises, panel (a) shows that there is a large positive externality on workers around the epicenter of the productivity response, but also a negative effect on workers who are further way. If the overall externality is larger as in panel (b), the effects are dispersed more widely, and the positive productivity effect outweighs the
Figure 1.4: Areas in which the productivity effect and redistributive effects are largest: the productivity effect is proportional to the height of $f(\tilde{\theta}|q)$, while the redistributive effect is proportional to the change in $f(\tilde{\theta}|q)$.

The aggregate belief externality is calculated as follows. First, the effect on consumption is obtained by scaling the wage effect by labor supply and the retention rate, $1 - T'(z)$. Second, the effect on social welfare is obtained by multiplying by the welfare weight, $\psi_z$. Third, the total impact can be calculated by integrating over the signal distribution.

$$\text{BE}(z) = -d\tau dz \left\{ \int_\Theta \psi_z(z(\tilde{\theta}|\pi, T)) \left[ 1 - T'(z(\tilde{\theta}|\pi, T)) \right] l(\tilde{\theta}|\pi, T) \left( \frac{dw(\tilde{\theta}|\pi)}{d \left[ 1 - T'(z) \right]} \right) f(\tilde{\theta}) d\theta \right\}$$

1.4.6 “U” Shaped Tax Schedules and The Importance of Incidence

An intuitive special case arises when the tax rate is initially flat, there is no redistributive motive for taxation and labor supply is perfectly inelastic. In this case, incidence is irrelevant and the belief externality is proportional to the difference between the average productivity...
Figure 1.5: The effects on wages of a reduction in the marginal tax rate on income between $60,000 and $61,000, with a small belief externality (top panel) or a large externality (bottom panel). The baseline economy is the linear taxation example with $s = 0.75$, calibrated to match the United States wage distribution. The effect at each productivity level is scaled by the productivity density so that the area under each curve is proportional to the average wage change accounted for by that component. The gray-shaded bar shows the wage range that is directly affected by this perturbation. The productivity and income distributions are the same in both panels.
increase and the average private gain from investment.

\[-\text{BE}(z) \propto \frac{d\bar{q}}{d(1 - T'(z))} - \int K \left[ \int \Phi \left( \frac{df(\theta|q)}{d(1 - T'(z))} \bigg|_{q=q(k|\pi,T)} \right) w(\theta|\pi) d\theta \right] dG(k)\]

Social benefit

Average private gain of responders

In the general case, wage changes due to the belief externality are re-weighted in a way that is important in driving the shape of the optimal tax schedule. The weights are given by: \( \Omega(\theta) = \psi_z(z(\theta|\pi, T)) [1 - T'(z(\theta|\pi, T))] I(\theta|\pi, T) \). For two reasons, this pushes toward a “U” shape of the tax schedule. First, a given wage change is more important if it affects workers who supply a large amount of labor, and who receive significant social welfare weight. This implies larger weights at intermediate incomes. Second, the weights are proportional to the retention rate, which amplifies other forces in the model; in particular, this compounds the milder “U” shape that already arises from the usual trade-off between the mechanical effect of taxation and the fiscal externality Diamond (1998).

### 1.4.7 A Necessary Condition for Optimality

Bringing everything together, the perturbation leads to three effects: \( \text{ME}(z), \text{FE}(z) \) and \( \text{BE}(z) \). Figure 1.6 shows these effects graphically. Just as in the example in Section 1.3, the mechanical effect from a reduction in the marginal tax rate transfers utility from workers with low productivity to those with high productivity, and the fiscal externality raises the utility of all workers. For most workers, the belief externality is also positive.

If \( T \) is optimal, the sum of the three effects must be equal to zero for all \( z \). Otherwise, there exists a change to the tax schedule that raises welfare. This is summarized in Proposition 3, which also rewrites the belief externality in terms of the income distribution.

**Proposition 3.** Consider an arbitrarily small perturbation that raises the marginal tax rate by \( d\tau \) between income \( z \) and \( z + dz \), with \( d\tau \) second order compared to \( dz \). The effect on social welfare is:

\[
\text{ME}(z) + \text{FE}(z) - d\tau dz \int \bar{z} \psi_z(\bar{z}) \left( \frac{1 - T'(\bar{z})}{1 - T'(z)} \right) \tilde{e}^\tau_{\nu(\bar{z})1-T'(z)} dH(\bar{z}) \tag{1.17}
\]

BE(\( z \))
Figure 1.6: Effects on utility of a rise in the marginal tax rate on income between $60,000 and $61,000. The baseline economy is the linear taxation example, calibrated to achieve $s = 0.75$ in equilibrium and match the United States wage distribution. The effect at each productivity level is scaled by the productivity density so that the area under each curve is proportional to the aggregate impact. The gray-shaded bar shows the wage range that is directly affected by this perturbation.

where:

$$
π(z) = \frac{dw(θ(z) | π)}{d[1 − T(z)]} \times \frac{1 − T'(z)}{w(θ(z) | π)}.
$$

Except at a discontinuity, $\text{ME}(z) + \text{FE}(z) + \text{BE}(z) = 0$ for all $z$ if $T$ is optimal.

If condition 1.17 is zero, there is no first-order gain from perturbing the tax schedule and moving to an equilibrium near the status quo. An alternative way of writing this, following Hendren (2016), is as a requirement that the cost and benefit of a policy change be equated.

$$
\frac{1 − Ψ(z)}{1 − H(z)} = 1 + \frac{\text{BE}(z) + \text{FE}(z)}{1 − H(z)}
$$

(1.18)

Here, a one dollar reduction in the tax rate at income $z$ mechanically provides one dollar of consumption to workers with income greater than $z$. The direct benefit of this change is
$1 + \Psi (z) > 1 - H (z)$ per dollar of mechanical expenditure, while the cost in addition to the mechanical expenditure is the sum of the two externalities.

A caveat to Proposition 3 is that equation 1.17 is necessary for optimality at an equilibrium around which there exists a continuous selection from the equilibrium correspondence, but not at points of discontinuity. Although Assumptions 1 to 3 ensure continuity for a generic tax schedule, one can construct examples in which the planner chooses to locate at such a discontinuous point if one exists. I take up this issue in Appendix A.3.

1.5 Quantitative Analysis

1.5.1 Evidence on Employer Learning

My next step is to quantitatively assess the importance of the belief externality using new and existing empirical evidence. First, I look to the literature on employer learning for evidence on the accuracy of employer beliefs. The dominant approach – pioneered by Farber and Gibbons (1996) and Altonji and Pierret (2001) – posits that the econometrician has access to a correlate of productivity from before workers entered the labor market, which employers cannot directly observe. In most cases, this is a pre-market score on the Armed Forces Qualification Test (AFQT).

These studies involve estimation of a version of the following regression using data from the National Longitudinal Survey of Youth (NLSY).

$$\ln w = a_0 + \rho_0 \text{AFQT} + \rho_1 \text{AFQT} \times \text{Experience}$$
$$+ \gamma_0 \text{Education} + \gamma_1 \text{Education} \times \text{Experience}$$
$$+ \lambda_0 \text{Experience} + \lambda_1 \text{Experience}^2 + \lambda_2 \text{Experience}^3 + X'\beta + \epsilon \quad (1.19)$$

The typical finding is that $\rho_1$ is strictly positive. This is interpreted as evidence that employers do not initially reward workers fully for their productivity, but that the reward increases over time. A simultaneous finding that $\gamma_1 < 0$ further supports this conclusion, with the argument being that employers initially use education to gauge unobservable
productivity; but that over time, they obtain more direct information about productivity and reduce their reliance on pre-existing correlates such as education.

Building on this approach, Lange (2007) estimates the speed of employer learning. He finds that employers’ expectation errors take three years to decline to half their original values and five years to reach 36 percent. It then takes 26 years to reduce the remaining errors to less than 10 percent of their initial values. There is thus a long delay before a worker is fully rewarded for her productivity, as reflected by her AFQT score. In turn, this implies a substantial wedge between the private and social returns to improving it.27

A limitation of Lange’s (2007) approach is that the evidence is confounded if productivity evolves heterogeneously over the lifecycle, since this would itself explain why the weight on AFQT increases with experience. Recognizing this, Kahn and Lange (2014) measure employer learning using a different method. Their key insight is that employer learning predicts that innovations in pay correlate more with past than future innovations in performance, because firms rely on past information to set pay.

Using a structural model and a panel dataset with information about both wages and performance reviews, Kahn and Lange (2014) find that workers capture between 60 and 90 percent of the social return to an innovation in their productivity during the first 15 years of their careers – although they capture a much smaller fraction in the later years.28 This implies that 10 to 40 percent of the social return accrues to others, which is exactly the type of statistic required to calibrate my model.

Many other studies also suggest that employers imperfectly observe worker productivity. For example, MacLeod et al. (2017) study the introduction of college exit exams in Colombia. They show that when more information about productivity becomes available, wages begin to more strongly reflect individual ability rather than college reputation. In a similar vein, there is evidence from online marketplaces that information is imperfect (Stanton and

27The evidence also suggests that education has a causal impact on AFQT scores (Neal and Johnson, 1996; Hansen et al., 2004), implying that they do not simply measure innate ability.

28These data come from a firm in the United States, first analyzed by Baker et al. (1994).
Thomas, 2016), and that additional information can improve outcomes (Pallais, 2014; Pallais and Sands, 2016).

Furthermore, numerous studies uncover evidence of statistical discrimination, which implies imperfect information. For example, Blair and Chung (2018) find that occupational licensing reduces reliance on race and gender; and drug testing is shown by Wozniak (2015) to positively impact black employment. Conversely, Agan and Starr (2018) and Doleac and Hansen (2016) show that racial discrimination increases when employers are banned from asking about criminal histories; and Shoag and Clifford (2016) find that banning the use of credit checks leads to relative increases in employment in low credit score census tracts, and more demand for other information about productivity.

1.5.2 New Evidence on Heterogeneity in Learning

The results above help calibrate the relationship in my model between productivity $q$ and expected wages $E(w)$, one minus the slope of which is the external wage effect of investment. However, there is only limited evidence on how employer learning varies with productivity. For example, Arcidiacono et al. (2010) find faster learning for college graduates than other workers, which suggests a larger externality at the low end; and Lindqvist and Westman (2011) show that non-cognitive skills – likely the hardest for employers to learn – are most important at low levels of income.

Here, I provide more direct evidence on how learning varies over the productivity distribution. Taking AFQT as a proxy for productivity, I adapt equation 1.19 by interacting the variables of interest with indicators $I_A = \mathbb{1}(\text{AFQT} > m)$ and $I_B = \mathbb{1}(\text{AFQT} \leq m)$ for whether a worker’s AFQT score is above or below the median, $m$.

$$\ln w = \sum_{j \in \{A,B\}} \left\{ \rho_{0,j} \text{AFQT} + \rho_{1,j} \text{AFQT} \times \text{Experience} + \gamma_{0,j} \text{Education} + \gamma_{1,j} \text{Education} \times \text{Experience} + \lambda_{0,j} + \lambda_{1,j} \text{Exper.} + \lambda_{2} \text{Exper.}^2 + \lambda_{3} \text{Exper.}^3 \right\} \times I_j + X'\beta + \varepsilon \quad (1.20)$$

I then estimate equations 1.19 and 1.20 using data from the NLSY79 survey.
Table 1.1: Heterogeneity in Employer Learning

<table>
<thead>
<tr>
<th></th>
<th>12 or 16 Years Education</th>
<th>Full Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFQT</td>
<td>2.63*</td>
<td>2.90**</td>
</tr>
<tr>
<td></td>
<td>(1.49)</td>
<td>(1.32)</td>
</tr>
<tr>
<td>AFQT × Experience</td>
<td>0.94***</td>
<td>0.87***</td>
</tr>
<tr>
<td></td>
<td>(0.19)</td>
<td>(0.16)</td>
</tr>
<tr>
<td>Education</td>
<td>11.09***</td>
<td>8.02***</td>
</tr>
<tr>
<td></td>
<td>(0.87)</td>
<td>(0.64)</td>
</tr>
<tr>
<td>Education × Experience</td>
<td>−0.30**</td>
<td>−0.21**</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Below median AFQT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFQT</td>
<td>5.14**</td>
<td>2.51</td>
</tr>
<tr>
<td></td>
<td>(2.56)</td>
<td>(2.12)</td>
</tr>
<tr>
<td>AFQT × Experience</td>
<td>1.11***</td>
<td>1.21***</td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>Education</td>
<td>10.10***</td>
<td>7.21***</td>
</tr>
<tr>
<td></td>
<td>(1.42)</td>
<td>(0.87)</td>
</tr>
<tr>
<td>Education × Experience</td>
<td>−0.35*</td>
<td>−0.25**</td>
</tr>
<tr>
<td></td>
<td>(0.20)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>Above median AFQT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFQT</td>
<td>6.55</td>
<td>5.89*</td>
</tr>
<tr>
<td></td>
<td>(4.02)</td>
<td>(3.51)</td>
</tr>
<tr>
<td>AFQT × Experience</td>
<td>−0.05</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>(0.54)</td>
<td>(0.45)</td>
</tr>
<tr>
<td>Education</td>
<td>11.33***</td>
<td>8.18***</td>
</tr>
<tr>
<td></td>
<td>(0.95)</td>
<td>(0.75)</td>
</tr>
<tr>
<td>Education × Experience</td>
<td>−0.27**</td>
<td>−0.17*</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>Observations</td>
<td>15884</td>
<td>15884</td>
</tr>
<tr>
<td>Clusters</td>
<td>2553</td>
<td>25659</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.30</td>
<td>0.30</td>
</tr>
</tbody>
</table>

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes. Dependent variable is the worker’s log hourly wage multiplied by 100. AFQT is a worker’s score on the armed forces qualification test, standardized by age to have zero mean and unit standard deviation. Education and experience are measured in years. Standard errors, shown in parentheses, are clustered at the worker level. All regressions include an indicator for urban vs. rural, race, race × experience, and region and year fixed effects. Data are from the National Longitudinal Survey of Youth (NLSY79). The sample is restricted to working black and white men who have wages between one and one hundred dollars, at least eight years of schooling and fewer than 13 years of experience. NLSY sample weights are used.
The sample follows Arcidiacono et al. (2010). It restricts to black and white men who are employed, have wages between one and one hundred dollars, and at least eight years of education. Following Altonji and Pierret (2001), I also limit the analysis to workers with fewer than 13 years of experience – measured as the number of years a worker has spent outside of school. Employment in the military, at home or without pay is excluded.

Table 1 shows the results. The dependent variable is the log of each worker’s real hourly wage, multiplied by 100; and AFQT scores are standardized to have mean zero and unit standard deviation for each age at which the test was taken. The coefficient on AFQT is therefore approximately the percentage wage gain associated with a one standard deviation higher AFQT score. The coefficient on the interaction of AFQT with experience is the number of percentage points that this increases by with each year of experience.

Below the median, there is strong evidence of learning: the weight on AFQT rises steeply with experience, and the weight on education falls comparatively quickly. There is less evidence of learning above the median, where the coefficient on the interaction between AFQT and experience is very close to – and statistically indistinguishable from – zero. The large direct effect of AFQT in the upper half of the distribution suggests that these results are not simply driven by AFQT being unimportant at the high end; and the less negative interaction between education and experience supports the conclusion that learning is driving the results. All of these conclusions are robust to restricting the sample to workers who have exactly twelve or sixteen years of education.

1.5.3 Evidence on the Response of Productivity to Returns

The second piece of evidence I require is an estimate of the relative responsiveness of productivity, compared to taxable income. Although there is less evidence available on the productivity response than there is for labor supply, a precise short-run estimate is provided by Blomquist and Selin (2010) using a difference-in-difference approach applied to a tax...
reform in Sweden. Blomquist and Selin’s (2010) results suggest that around three quarters of the response of taxable income comes through wages.\textsuperscript{31} This is consistent with a calibration by Trostel (1993) which suggests that 60 to 80 percent of the long run response of income to taxation comes from changes in labor productivity.

There is also qualitative evidence that longer-run human capital investments respond. First, Abramitzky and Lavy (2010) study the reduction in effective marginal tax rates that occurred when Israeli kibbutzim shifted from equal-sharing to productivity-based wages.\textsuperscript{32} They find that the reform led to sharply higher graduation rates and test scores. Second, Kuka et al. (2018) study the introduction of the Deferred Action for Childhood Arrivals (DACA) program, which increased returns to human capital investment. They show that high school graduation and college attendance rates increased markedly for eligible individuals. Finally, studies have demonstrated that human capital investments increase when students are simply informed about returns (e.g., Jensen, 2010).

MacLeod et al.’s (2017) study of college exit exams also provides interesting evidence. As more information is provided to employers, and wages begin to more closely track individual ability, average wages rise by seven percent given the same formal educational investments. This rise in wages is consistent with a response of human capital investment to the higher return to individual ability, although it could also be explained by improved matching between workers and tasks.

1.5.4 Calibration to the United States Economy

I now calibrate the model to match both the evidence above and the empirical United States wage and income distributions. Table 1.2 summarizes the assumptions needed and my choices for them. Results with alternative calibrations are available in Appendix A.8.

The wage schedule that I target is the Pareto log-normal approximation provided by Mankiw et al. (2009) using March CPS data. However, a wage schedule cannot be

\textsuperscript{31}This may be conservative since Blomquist and Selin cannot capture long-run human capital responses.

\textsuperscript{32}Kibbutzim are small collective communities in Israel.
Table 1.2: Calibrated and Implied Objects

<table>
<thead>
<tr>
<th>Assumed object</th>
<th>Assumption</th>
<th>Implied object</th>
<th>Implied value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social welfare function</td>
<td>$\log(E(U))$</td>
<td>Income elasticity</td>
<td>$\varepsilon^R_i = 1.1$</td>
</tr>
<tr>
<td>Noise distribution</td>
<td>LN, $\text{var}(\theta</td>
<td>q) = 7q$</td>
<td>Wage elasticity</td>
</tr>
<tr>
<td>Labor supply elasticity</td>
<td>$\varepsilon_l = 0.25$</td>
<td>Cost distribution</td>
<td>Champernowne: $\lambda = 1.3$, $\alpha = 2.46$, $y_0 = 2.7$</td>
</tr>
<tr>
<td>Production concavity</td>
<td>$\beta = 0.3$</td>
<td>External fraction</td>
<td>of return 0.15 (average)</td>
</tr>
<tr>
<td>Wage distribution</td>
<td>Pareto LN: $a = 2$, $\sigma_q^2 = 0.56$, $\mu_q = 2.76$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table notes. This table summarizes the key assumptions underlying the simulation described in Section 1.5. Objects in the left column are calibrated directly, while the target objects in the right column are implied. See text and Appendix A.8 for further details and simulations with alternative calibrations.

assumed directly. Rather, equilibrium wages are jointly implied by productivity and signal distributions. The approach I take is to posit a conditional signal distribution, $f(\theta|q)$, and then find a distribution for productivity that yields a wage distribution as close as possible to the target.\textsuperscript{33} As panel (a) of Figure 1.7 shows, this exercise is successful. Notably, the Pareto right tail is closely replicated in addition to the overall shape.

Next, I choose a signal distribution so that, on average, a worker who increases her productivity by one dollar receives an 85 cent higher expected wage. This is at the upper end of the estimates provided by Kahn and Lange (2014). In line with my empirical results, I also ensure that there is a flatter slope at the low end of the income distribution. I achieve these aims by assuming a conditionally log-normal signal distribution with $E(\theta|q) = q$, and $\text{var}(\theta|q)$ linearly increasing in $q$. Panel (b) of Figure 1.7 displays the results.

I choose the remaining parameters to target income, wage and labor supply elasticities. First, I set $\varepsilon_l = 0.25$, which is approximately in line with estimates of the intensive-margin labor supply elasticity (e.g., Chetty, 2012). Second, I calibrate the long run elasticities of each variable to the retention rate.\textsuperscript{34} Although these cannot be directly assumed, they are closely connected to the concavity of the production function, $\beta$. I choose a value for $\beta$

\textsuperscript{33}Specifically, I parameterize a Champernowne (1952) distribution – a family of bell-shaped distributions designed to fit empirical income distributions – to minimize the Kullback-Leibler divergence between the equilibrium and target distributions under the 20 percent linear tax from which the simulation starts.

\textsuperscript{34}These long-run equilibrium responses take into account the multiplier effect from human capital investment: more productive workers work more, but working more further raises the return to investment.
Figure 1.7: Implications of the calibration procedure for the simulation described in Section 1.5. Panel (a) compares the empirical (target) and approximate (simulated) wage distributions. Panel (b) shows the relationship between expected wages and productivity in the baseline economy.
that produces long-run elasticities of wages and labor supply of 0.7 and 0.4 respectively.\textsuperscript{35} This implies that around 60 percent of the long run response of taxable income comes from changes in labor productivity, which is at the low end of the estimates above. The same estimates imply an overall elasticity of taxable income of 1.1.\textsuperscript{36}

1.5.5 Solving for Optimal Taxes

To simulate the model, I start with an initial tax schedule $T_0$ and a known equilibrium. I then consider adopting an alternative tax schedule, $T_1$, under which the marginal tax rate is raised or lowered by $\Delta T'$ over a range of incomes from $z$ to $\bar{z}$.

$$ T'_1(z) = \begin{cases} 
T'_0(z) + \Delta T' & \text{if } z \in (z, \bar{z}) \\
T'_0(z) & \text{otherwise}
\end{cases} $$

Given $T_1$, I re-calculate the expected utility of workers with each level of productivity, and let workers adjust their human capital investments. Next, I re-solve for employer beliefs, and wages, given the new productivity distribution. From here, I repeatedly re-optimize human capital decisions and re-calculate beliefs until a fixed point is obtained. At this fixed point, employers’ beliefs and workers’ investment decisions are mutually consistent. Finally, I calculate expected utility for each individual, weight using the social welfare function and adopt the new tax schedule if the welfare gain is positive.

This is the procedure that underlies Figures 1.2, 1.5 and 1.6. It can be continued repeatedly, starting with large perturbations and ending with smaller ones until the gain to the marginal perturbation is approximately zero. At this point, condition 1.17 of Proposition 3 approximately holds. I refer to this final tax schedule as \textit{optimal}. Further details of this process are available in Appendix A.8.

\textsuperscript{35}Long-run elasticities vary with the tax system and over the income distribution. The statistics quoted here are based on responses of aggregate wages and labor supply to a small change to a flat 20 percent tax.

\textsuperscript{36}This is in line with Mertens and Montiel-Olea (2018), whose estimates could be viewed as conservative since they do not take into account long-run human capital responses. I provide simulations with different parameter values in Appendix A.8. Holding $\ell_t^{LR}/\ell_t^{LR}$ constant, the level of $\ell_t^{LR}$ is not important to my results.
1.5.6 A Naïve Benchmark for Comparison

As a benchmark against which to compare the optimal tax schedule, I imagine a naïve planner who neglects to take into account the fact that part of the response of wages to taxation arises due to an externality, and is thus not internalized by investors. This means that she neglects the novel effect of a perturbation of the tax schedule, \( BE(z) \). Instead, she simply equates the fiscal externality and the mechanical effect, as would be the correct approach in a model with perfect employer information.

Comparison of the naïve and optimal tax schedules therefore facilitates an assessment of the quantitative importance of the belief externality. This is similar in spirit to Rothschild and Scheuer’s (2013) concept of a self-confirming policy equilibrium (SCPE). However, the planner is more sophisticated here in that she is aware when measuring the fiscal externality that she must take into account both wage and labor supply responses. Similarly, she knows that the mapping from productivity to wages is stochastic. However, she is unaware that part of the change in equilibrium wages arises due to a spillover that workers ignore when re-optimizing.

1.5.7 Simulation Results

Figure 1.8 shows the tax schedules produced by this procedure. The red line shows a tax schedule that satisfies equation 1.17, so that there is no first order gain from a small perturbation in any tax bracket. The blue line would satisfy a naïve social planner, because the mechanical effect and the fiscal externality sum to zero.\(^{37}\) Marginal tax rates are generally substantially lower under the optimal than the naïve schedule, reflecting the planner’s additional incentive to encourage investment by lowering taxes.

Both tax schedules have the familiar “U” shape, which comes from the trade-off between the mechanical effect and the fiscal externality when the income distribution has a Pareto right tail (Diamond, 1998). This shape is accentuated under optimal taxation because the

\(^{37}\)The algorithm to find the naïve schedule is conceptually identical to that used for the optimal schedule.
Figure 1.8: The optimal non-linear tax schedule in the simulation described in Section 1.5. The solid red line in the top panel shows the optimal tax schedule, while the dashed blue line shows a tax schedule that would be accepted by a naïve social planner who sets the sum of the mechanical effect and fiscal externality equal to zero. The lower panel shows a decomposition of small marginal tax cuts in each tax bracket. The tax function in this simulation is discretized into $20,000 brackets. Details of the procedure are available in Appendix A.8.
belief externality is more important at intermediate incomes (see Figure 1.8). In part, this is because a given wage impact from the belief externality is less important at high incomes where social welfare weight is low, and at low incomes where little labor is supplied; and in part the shape is due to variation in the wage impact of the externality.\textsuperscript{38}

At very high incomes, the optimal tax schedule is above the naïve tax schedule. This is for two reasons. First, as income rises, the belief externality becomes arbitrarily small so that the planner simply trades off the mechanical effect and the fiscal externality. Second, changes in marginal tax rates at high incomes shift investment incentives throughout the productivity distribution; and most of those who respond now face lower tax rates most of the time – implying a smaller fiscal externality from their re-optimization.\textsuperscript{39}

At very low incomes, optimal tax rates are also higher. In this case, the main reason is that the downward adjustment to taxes throughout the distribution raises expected utility for most workers, but not those at the bottom (see Figure 1.9). As a result, welfare weights rise at the lowest incomes. This increases the mechanical gain from raising marginal taxes at low incomes and collecting infra-marginal income all but the lowest-income workers.

The welfare gain from transitioning to the optimal tax schedule is equivalent to raising the consumption of all workers by one percent, holding labor supply and investment decisions fixed. However, this is not a Pareto improvement. As Figure 1.9 shows, individuals of moderate-to-high productivity experience large gains in utility. But workers with very low productivity are worse off because the government collects less revenue. This means that the transfer to the lowest-income household is five percent smaller. Workers with the highest productivity levels are also hurt, due to higher tax rates at top incomes.

\textsuperscript{38}The shape is further amplified because the belief externality scales with with the retention rate, $1 - T'(z)$.

\textsuperscript{39}Comparisons of the mechanical effect, fiscal externality and belief externality for marginal perturbations under the naïve and optimal tax schedules are available in Appendix A.8.
1.5.8 Approximately Optimal Taxation

Although the optimal tax formula cannot be written in terms of sufficient statistics, 60 percent of the gain from optimal taxation can be obtained using a simple approximation based on two principles. First, assume that a change in $T'(z)$ causes workers with income close to $z$ to respond. Second, assume that the incidence of the externality falls on workers with welfare weight, labor supply and tax rate similar to those with income $z$. In Appendix A.5, I show that these principles yield:

$$\text{FE}(z) + \text{ME}(z) - (1 - s(z)) \psi(z) l(z) \left[ 1 - T'(z) \right] \frac{d\bar{w}}{d[1 - T'(z)]} = 0 \quad (1.21)$$

where $l(z)$ is the labor supply of workers with income $z$, $d\bar{w}/[1 - T'(z)]$ is the response of average wages, and $s(z)$ is the share of the wage change that is not internalized.

An advantage of equation 1.21 is that it facilitates assumptions about how the belief externality varies with income, without having to find distributional assumptions which
produce that profile. As in Section 1.4, the correction term is larger if investment is more responsive, investing workers capture little of the return to investment; or if workers supply a large amount of labor, face a low tax rate and receive significant welfare weight.

1.6 Unproductive Signaling

I have so far assumed that investment is productive in the sense that its sole effect is to increase productivity. It is also possible for investment to play a ‘pure’ signaling role in the sense of Spence (1973). Investment then reflects both human capital accumulation and a worker’s immutable ability. In this case, the externality from investment may, in general, be more positive or more negative than in my baseline model.

To allow for this type of signaling, I replace the production function with $q = Q(x, k)$ so that productivity is a direct function of the worker’s type. Secondly, I assume that employers observe a signal of investment rather than productivity. Specifically, $\theta \in \Theta \subseteq \mathbb{R}_+$ has conditional density $f(\theta|x)$ twice differentiable in $x$, and full support for all $x$. As before, it satisfies the monotone likelihood ratio property: $\frac{\partial}{\partial \theta} \left( \frac{f(\theta|x_H)}{f(\theta|x_L)} \right) > 0$ for all $x_H > x_L$. Otherwise, I adopt all the assumptions from Section 1.2.

1.6.1 Unproductive Signaling: Example with Linear Taxation

I begin by adapting the example in Section 1.3. Specifically, let productivity $q = n^a h^{1-a}$ be a function of both human capital $h = x^\theta$, and inherent ability $n$, with ability negatively related to the worker’s investment cost: $n = 1/k$. As before, assume that the distribution of ability and the conditional signal distribution are log-normally distributed.

$$n \sim LN\left(\ln \mu_n - \frac{\sigma_n^2}{2}, \sigma_n^2\right) \quad \ln \theta = \ln x + \ln \zeta \quad \ln \zeta \sim N\left(0, \sigma_\zeta^2\right)$$

---

40 It is hard to assess the contribution of unproductive signaling to the return to education. Evidence from large-scale school reforms demonstrate large productive effects of education (Meghir and Palme, 2005; Aakvik et al., 2010; Oreopoulos, 2006) but there is also some evidence to suggest a role for unproductive signaling (Lang and Kropp, 1986; Bedard, 2001). See Lange and Topel (2006) for a discussion.
With these assumptions, there again exists an equilibrium in which income and productivity are log-normally distributed. Similar to the original example, the elasticities of productivity and income depend only on the labor supply elasticity $\varepsilon_l$, the concavity of the production function $\beta$ and, in this case, the relative importance of ability, $a$.

**Proposition 4.** For any tax rate $\tau$, there is an equilibrium in which productivity and income are log-normally distributed. Assuming this equilibrium is played, the elasticities of productivity and investment with respect $1 - \tau$ are as follows.

$$
\varepsilon_q = \frac{\beta (1 - a) (1 + \varepsilon_l)}{1 - \beta (1 - a) (1 + \varepsilon_l)} \quad \varepsilon_z = \frac{\varepsilon_l + \beta (1 - a) (1 + \varepsilon_l)}{1 - \beta (1 - a) (1 + \varepsilon_l)}
$$

This nests the original example in Section 1.3 in which investment is purely productive: when $a = 0$, the two elasticities $\varepsilon_q$ and $\varepsilon_z$ collapse to that case, and equation 1.22 collapses to equation 1.10. Alternatively, when $a = 1$, productivity does not respond to taxation, and the income elasticity collapses to the elasticity of labor supply.

The first-order condition for the optimal tax is given by Proposition 5. It features a second externality correction, $1 + sa (1 + \varepsilon_l)$, which pushes toward higher rather than lower taxes. Intuitively, there is no social benefit from the component of the private return to investment that comes from signaling innate ability, which in turn implies that this return comes at the expense of others. The logic is similar to the redistributive effect in Section 1.4: a worker who invests more negatively affects other workers, because she has higher productivity than the group she leaves, but lower productivity than the group she joins.

**Proposition 5.** Assume that the log-normal equilibrium described in Proposition 4 is played. Then the first-order condition for the optimal linear tax $\tau^*$ is:

$$
\frac{\tau^*}{1 - \tau^*} = \frac{1 - \gamma \left[ \frac{1}{1 + sa (1 + \varepsilon_l)} \right]}{\varepsilon_z}
$$

where $s = \frac{\sigma_z^2}{\zeta_z^2 + \sigma_l^2}$ and $\gamma = E_n \left[ \frac{\psi \frac{\zeta}{\gamma} \frac{z}{\tau}}{\zeta_z^2 + \sigma_l^2} \right]$. 

48
Since imperfect employer information now generates two opposite-signed externalities, there exist combinations of \( a \) and \( s \) that cause them to perfectly offset each other.

\[
\beta (1 - a) = s \left[ \frac{a + \beta (1 - a)}{1 + s \alpha (1 + \epsilon_1)} \right] \quad \iff \quad s \alpha (1 + \epsilon_1) = (1 - s) \epsilon_q
\]

The first condition states that the social and private benefits of investment are aligned. The second states that the unproductive component of the private return is equal in magnitude to the part of the productive component that is not captured by the individual. If these conditions hold, then condition 1.22 collapses to the standard optimal tax formula. However, any other parameter values imply an efficiency role for intervention.

An implication of the equations above is that less accurate employer information (lower \( s \)) implies a smaller private benefit of investment with an unchanged social benefit. Stated equivalently, lower \( s \) means that the signaling externality is smaller and the learning externality is larger. In this sense, evidence of residual employer uncertainty (Lange, 2007; Kahn and Lange, 2014) suggests a more positive externality, and lower taxes.

### 1.6.2 Non-Linear Taxation: Signaling with Observable Investment

To further build intuition, I now relax the parametric assumptions of the example and consider non-linear taxation, but in a special case of the model in which employers perfectly observe investment. There is then a deterministic equilibrium mapping from investment to wages, \( w(x) \). Taking this as given, the worker’s investment problem is:

\[
\max_{x \in \mathbb{R}_+} v (w(x) | T) - kx
\]

where:

\[
v (w(x) | T) = \max_{l \in \mathbb{R}_+} u(w(x) l - T(w(x) l), l).
\]

The solutions to problem 1.23 for each cost type jointly define a second mapping, \( x(k) \), from costs to investment levels.

To simplify the analysis, I assume \( w(x) \) is one-to-one. Then, given this assumption, I
provide conditions in Appendix A.6 to guarantee that \( x(k) \) and \( w(x) \) are differentiable, which ensures that the investment choice for a worker with cost \( k \) is characterized by:

\[
u_c(z(k) - T(z(k)), l(k)) \left[ 1 - T'(z(k)) \right] l(k) w'(x(k)) = k\]  \hfill (1.25)

where \( l(k) \) is the level of labor supply that solves problem 1.24, and \( z(k) = w(x(k)) l(k) \) is the equilibrium income of a worker with cost \( k \).

As is well known, the equilibrium relationship between innate ability and investment drives a wedge between the private and social returns, which I refer to as the unproductive component.

\[
\frac{Q_k(x(k), k)}{x'(k)} = w'(x(k)) - \frac{Q_x(x(k), k)}{x'(k)}
\]

If \( Q_k(x(k), k) < 0 \) so that costs are positively related to ability, there is a positive externality from investment: an individual who invests more makes others look better because she has higher productivity than those who invest at that level in equilibrium. Conversely, if \( Q_k(x(k), k) < 0 \), there is a negative externality from investment.

These results provide a foundation for policy analysis that mirrors Section 1.4. Specifically, consider again a perturbation that raises the marginal tax rate by \( d\tau \) on income between \( z \) and \( z + dz \), while raising the intercept of the tax schedule to ensure that the resource constraint still holds. A different but related form of belief externality arises.

\[
\text{BE}(z) = -d\tau dz \int_k \psi(k) \left[ 1 - T'(z(k)) \right] l(k) \frac{dx(k)}{d[1 - T'(z)]} \left[ w'(x(k)) - Q_x(x(k), k) \right] dG(k)
\]

This equation for \( \text{BE}(z) \) can again be written in terms of the observable income distribution, and combined with the fiscal externality and mechanical effect to obtain a necessary condition for optimality of the tax system:

\[
\text{FE}(z) + \text{ME}(z) + \int_Z \tilde{z} \psi(\tilde{z}) \left( \frac{1 - T'(\tilde{z})}{1 - T'(z)} \right) \epsilon_{\tilde{x}(\tilde{z}), 1 - T'(\tilde{z})} \left[ \frac{\epsilon_{\text{Private}}{\tilde{w}(\tilde{z}), \tilde{x}(\tilde{z})} - \epsilon_{\text{Social}}{\tilde{w}(\tilde{z}), \tilde{x}(\tilde{z})}}{1 - T'(\tilde{z})} \right] dH(\tilde{z}) = 0
\]

(1.27)

where \( \tilde{w}(\tilde{z}) \) and \( \tilde{x}(\tilde{z}) \) are the wages and investment levels of a worker with income \( \tilde{z} \), and
the elasticities are defined as follows.

\[
\begin{align*}
\varepsilon_{\text{Private}}^{\theta(z), x(z)} &= \frac{x'(k)}{x(k)} \frac{x(k)}{w(k)} \\
\varepsilon_{\text{Social}}^{\theta(z), x(z)} &= Q_x(x(k), k) \frac{x(k)}{w(k)}
\end{align*}
\]

Note the similarity between conditions 1.17 and 1.27. This is not coincidental: just as before, employer inference causes misalignment between the private and social returns to investment, and the resulting externality enters social welfare in the same way.

### 1.6.3 Non-Linear Taxation: Imperfectly Observable Investment

My final step is to return to the general model with both unproductive signaling and imperfectly observable investment. For any given set of wage externalities, the equation for the belief externality, \( \text{BE}(z) \), remains very similar to Section 1.4, and there remain distinct productivity and redistributive effects.

\[
\begin{align*}
\frac{dw(\theta|\pi)}{d[1 - T'(z)]} f(\theta) &= \int_{\mathcal{X}} \left( \frac{dx(k|\pi, T)}{d[1 - T'(z)]} \right) \left[ Q_x(x(k|\pi, T), k) f(\theta|x(k|\pi, T)) \right] \left[ Q(x(k|\pi, T), k) - E(q|\theta, \pi) \right] dG(k)
\end{align*}
\]

However, there are important differences in the interpretation of these two effects. First, the productivity effect may be small or even entirely absent if investment costs are negatively correlated with ability. For example, an extreme possibility is that \( q = Q(k) \) so that productivity is unaffected by investment. In this case, the productivity effect is zero and investment returns must come entirely from unproductive signaling of one’s ability. The private gain from investment is thus fully offset by negative impacts on the wages of other workers. In this extreme case, the planner would set higher rather than lower optimal taxes, given the same mechanical effect and fiscal externality.

A second possibility is that investment costs are positively rather than negatively related to ability, which is possible providing that investment also raises productivity. The
redistributive effect then becomes less negative, and may even be positive, since a worker who considers increasing her investment has higher productivity than those who invest at that new level in equilibrium. In this case, the “unproductive” component of the return reinforces rather than offsets the positive learning externality, and provides still further motivation to lower marginal tax rates and encourage investment.

1.7 Conclusion

A substantial body of evidence suggests that employers have imperfect information about the productivity of their workers. This paper provides a framework to study optimal income taxation in this environment. In the model I develop, employers observe an imperfect signal of workers’ human capital investments. I show how moral hazard caused by Bayesian inference introduces an externality: workers who invest more raise their own wage but also affect employers’ perceptions – and thus the wages – of other workers.

My quantitative results suggest that this new externality is of first-order importance. Taking it into account leads to marginal tax rates that are substantially lower on average. This downward adjustment to tax rates is concentrated at intermediate incomes, leading to an amplification of the classic “U” shape of the optimal tax schedule. There is a notable welfare gain from moving to optimal taxation.

My model provides a framework that could be extended to analyze the implications of many other features of the labor market. This could include asymmetric employer learning (Acemoglu and Pischke, 1998), screening by employers (e.g., Stantcheva, 2014), an extensive margin of labor supply (Saez, 2002), and richer labor market structures including tournaments or other dynamic contracts (Lazear and Rosen, 1981; Prendergast, 1993). These extensions would preserve the conclusion that wages or utility are compressed, lowering the private return to investment relative to the social return, but they will also lead to other insights. Some (e.g., asymmetric learning) will also feature multiple equilibria, which I provide a way of dealing with in an optimal taxation framework.

More broadly, the core insight of this paper is general: inference based on imperfect
information disconnects the private and social returns to engaging in positive behavior. For example, police officers interpret the actions of suspects based on their experiences with previous individuals; thus, compliance by one individual may reduce the likelihood that an officer uses force against a similar suspect in the future. Likewise, buyers form beliefs about the imperfectly observable qualities of goods and services based on past purchases; investment in quality by one seller may therefore raise a consumer’s willingness to pay for other similar products. In the future, the approach of this paper may be expanded to provide new insights into these contexts and many others.
Chapter 2

Complementary Bias: A Model of Two-Sided Statistical Discrimination

2.1 Introduction

Given the history of widespread gender, racial, and ethnic discrimination around the globe, it is natural to think of worker and firm actions as strategic complements. Examples abound. Statistical discrimination against women could be generated by employers’ asymmetric beliefs about the competence of men and women. But alternatively, given a history of bias in hiring and inflexible workplaces, women may believe that they will be treated unfairly, will encounter a hostile culture and will ultimately fail to be promoted. Women may therefore invest in such a way that causes employers to adjust their beliefs downward, even if they were initially homogeneous – confirming women’s suspicions. This mechanism could contribute to disparities in particular sectors, drive occupational segregation, or operate

---

1Co-authored with Roland Fryer

2Strategic complements, a term coined by Bulow et al. (1985), refer to decisions of at least two players that mutually reinforce one another. As Bulow et al. (1985) write, conventional substitutes and complements can be distinguished by whether a more “aggressive” strategy by firm A (e.g., a lower price in price competition or greater quantity in quantity competition) lowers or raises firm B’s total profits. Strategic substitutes or complements are analogously defined by whether a more “aggressive” strategy by A lowers or raises B’s marginal profits.
throughout the labor market more generally. As we demonstrate, such ingrained pessimism on both sides makes a given disparity in treatment much more difficult to identify and address.

A similar dynamic may undergird racial disparities. Racial inequality in the 20th century was generated by explicit racism and discrimination in almost every aspect of economic and social life. Jim Crow was all encompassing – parts of the South were plastered with signs that read “Negroes need not apply” (United States Congress Commitee on Education and Labor, 1963) and public lynchings of black bodies occured all too often. In our model, if employers openly discriminate against blacks then, as a best response, black workers decide it does not make economic sense to invest. Now imagine that with the signing of the Civil Rights Act of 1964, employers stopped discriminating and hired with homogeneous beliefs, but workers were not convinced firms were amenable to minority workers. In traditional models of discrimination, minority outcomes improve immediately. In our model, however, even if employers had homogeneous beliefs immediately following the civil rights legislation, the equilibrium remains unchanged. Minorities, after years of subjugation, continue to believe that firms are hostile to minority workers and consequently do not invest. If they do not invest, employers adjust their beliefs downward. This illuminates the basic economics of our approach.

In this paper, we expand on the intuition from the gender and race examples above by building a model in which worker and firm actions are strategic complements vis-à-vis a two-sided statistical discrimination model.\(^3\) Nature distributes costs of investment to workers and firms. We think of worker investment as classical Becker (1964) human capital. Firm investment is a fixed cost of creating a work environment conducive to workers (e.g. flexible work hours for women or affinity groups for minority workers).\(^4\) Workers (resp. 

\(^3\)It is important to note at the outset that our game is not supermodular. There is not a strategic complementarity between every worker and every firm, but there is strategic complementary between sides (workers and firms).

\(^4\)In the main model, we assume this cost is fixed. In Section B.4.2, we provide some intuition for how results change if costs are proportional to the number of hires.
firms) observe their costs and decide whether to invest. Conditional on this investment, nature distributes a signal to firms (regarding worker investment) and another to workers (regarding firm investment). Then workers, given their beliefs and observed signal, choose whether to apply to firms; and, conditional upon receiving an application, firms decide whether to hire.

Our model nests the classic one-sided treatment of statistical discrimination (e.g., Coate and Loury, 1993). Equilibria analogous to those in a one-sided model always exist in ours if the rate of firm investment is approximately fixed, which shuts down the strategic complementarity that drives our results. However, statistical discrimination can also be generated and sustained by worker pessimism. The resulting complementarity in beliefs between workers and firms makes the analysis of disparities between groups more complex but also considerably richer.

We begin our analysis of policy by considering affirmative action in the sense of a requirement that firms make job offers to members of both groups with equal probability. In models of statistical discrimination without strategic complementarity (classic one-sided models), such a requirement leads to homogeneous employer beliefs when lower hiring standards do not undermine worker investment, but negative stereotypes about minorities may persist if low standards are too demotivating (Coate and Loury, 1993). Affirmative action can have the same issues in our model. But worse, affirmative action can undermine firm investment incentives and trigger zero investment by minority workers. Such severe inequality can be sustained indefinitely in our model despite affirmative action. Moreover, it may be impossible for affirmative action to eliminate discrimination because firms have less incentive to invest in a numerical minority if investment costs are fixed. Our model also allows us to analyze a more ambitious form of affirmative action – employment quotas – which requires employers to hire members of each group in proportion to their representation in the population. Employment quotas can cause firms to be overly aggressive...

---

5 Altonji and Blank (1999) show that these “patronizing equilibria” are eliminated if investment is a continuous variable.
in their attempts to hire minority workers, which can severely undermine minority worker investment, as well as harming the majority.

Generally, we demonstrate a kind of “impossibility result”. Not only does every policy analyzed by Coate and Loury (1993) have the potential to be harmful, but any policy that fails to address the expectations of both sides (employers and workers) simultaneously will be ineffective. This result stems naturally from the two-sided nature of our model. Workers fail to invest both because of harsh hiring standards and because they are pessimistic about how they would fare in the workplace if they were hired. At the same time, firms are hesitant to make investments to support minority workers both because they think the workers they would attract would be unqualified and because minorities are not applying.

Consistent with this result, we provide suggestive evidence that two-sided policies are more effective at increasing the wages of disadvantaged groups. We focus on recent examples of job training programs (e.g., *Year Up* and *Per Scholas*) that have been unusually successful at increasing wages for disadvantaged youth. These programs combine worker investment with additional signals to firms, while at the same time demonstrating to workers that firms are investing. We argue that these programs are consistent with the policies suggested by our model. Firms can be confident that workers have the relevant cognitive and non-cognitive skills to be successful, and workers also know that their investments will pay off because they are matched with an employer with demand for their type.

Randomized evaluations of *Year Up* and *Per Scholas* demonstrate that treatment youth earn, on average, 30% more than control youth (Roder and Elliott, 2014; Hendra *et al*., 2016). In an analysis of 207 other (one-sided) job training programs in Card (2015), the treatment effect on monthly earnings is 9.3 percent. We view this as suggesting that two-sided policies are more effective. Although this evidence is incomplete, these results – coupled with the model – may help design future programs.

Based on our model, we propose a new policy – which we label “investment insurance” – as a simple solution to statistical discrimination. Imagine that the government can observe a noisy version of the signals employers and workers receive, and offers them contracts. If
the government believes an individual invested, it will subsidize them. The same assurance is provided to employers regarding their investment. This provides assurance to both workers and firms that their investments will pay off. Investment insurance can never be harmful to minority workers. Quite to the contrary, there is always a policy of this type that leads immediately to full equality. The underlying economics is similar to the concept of ‘insulating tariffs’ as discussed by Weyl (2010): the government effectively insulates workers and firms from uncertainty about investment by the other side.6

The paper concludes by deriving a model-based empirical test for statistical discrimination by employers. The test builds conceptually on the work of Altonji and Pierret (2001), Lang and Lehmann (2012), and Fryer et al. (2013), but it is designed to be robust to the confounds of worker belief formation and complementarity between firm and worker investment. Our analysis focuses directly on the mechanism through which rational stereotyping affects incentives: pessimistic employers shrink their estimates of worker productivity toward the group mean, causing a flattening of the relationship between productivity and wages. Based on this insight, we propose examining workers who switch firms. Under the assumption that firms gain some private information about a worker’s ability with tenure, we demonstrate that wage profiles should flatten more for minority than majority workers when they move.

The paper proceeds as follows. The next section provides a brief review of the literature. Section 2.3 introduces our model and derives the basic implications of two-sided statistical discrimination and how it differs from traditional one-sided models. Section 2.4 discusses policies such as affirmative action, employment quotas, and wage subsidies. Section 2.5 considers two-sided policies. Section 2.6 describes empirical implications of the two-sided approach. Section 2.7 concludes. Appendix C contains technical proofs, derivations omitted from the main analysis, and extensions of the basic model.

6We are grateful to Jesse Shapiro for pointing out this connection.
2.2 A Brief Review of the Literature

Our paper lies at the intersection of two important literatures: models of discrimination and models with strategic complementarities.\(^7\) We briefly discuss each in turn.

2.2.1 Models of Discrimination

The two main theories of discrimination are a theory based on tastes pioneered by Becker (1957) and a statistical theory posited by Phelps (1972) and Arrow (1973).\(^8\) Statistical models rely on imperfect observability of a worker’s productivity to account for employers’ use of a worker’s group identity in their decision-making.

Phelps (1972) assumes available measures of productivity to be noisier for minority workers. One prediction of this model – developed by Aigner and Cain (1977) – is that there will be a wage gap at the top of the income distribution favoring whites and another gap at the bottom of the income distribution favoring blacks. Arrow (1973) demonstrates that statistical discrimination can occur even when there is no such unexplained group heterogeneity. The key insight in Arrow (1973) is that when employee productivity is endogenous, employer prejudice can be self-fulfilling.

An important contribution to this literature is Coate and Loury (1993) who formalize the insights in Arrow (1973) using a job assignment model in the spirit of Milgrom and Oster (1987). Coate and Loury (1993) provide sufficient conditions for multiple equilibria to exist and then demonstrate that an affirmative action policy may fail in the presence of statistical discrimination by perpetuating stereotypes.

There have been several important extensions of the Coate and Loury (1993) model. Moro and Norman (2004) embed Coate and Loury (1993) in a general equilibrium framework

\(^7\)Our work is also related to the literature on two-sided markets: e.g., Caillaud and Jullien (2003), Rochet and Tirole (2003), Anderson and Coate (2005), Armstrong (2006), and Weyl (2010). The labor market in our game could be viewed as a ‘platform’, with each side (workers vs. firms) benefiting from participation by the other. Strategic interactions are complicated in our model by imperfect information but this complementarity between sides is fundamental to the model and our policy recommendations.

\(^8\)See Fang and Moro (2011) for a nice review of models of discrimination.
(with endogenous wages) and demonstrate that discrimination can occur even when the corresponding model with a single group has a unique equilibrium. Fang (2001) allows individuals to choose their group identity (i.e. social culture) and shows that allowing firms to give preferential treatment based on some seemingly irrelevant (chosen) group identity allows society to overcome an informational free-riding problem. Fryer (2007) develops a multi-stage model of statistical discrimination and explores what happens to individuals who nonetheless overcome the initial discrimination. If an employer discriminates against a group in the first stage, she may actually favor members of that group when she makes promotion decisions within the firm.

Our paper builds on this literature – being close in spirit to Coate and Loury’s work. The simple idea is that the original Arrow (1973) insight applies to both sides of the market: employers’ prejudicial beliefs can be self-fulfilling but so too can workers’ prejudicial beliefs about employers. This insight has far-reaching implications for policy and empirical analysis. As we show, policies that have been proposed based on one-sided models are ineffective and potentially harmful. However, our analysis suggests under-explored and potentially promising two-sided alternatives. Our paper also builds qualitatively on important work by Lang et al. (2005), who provide a model in which disparate outcomes can be sustained despite discriminatory employer preferences being arbitrarily weak.

Finally, our proposed method to identify statistical discrimination by employers is related to a small but burgeoning literature on empirical tests for discrimination. Altonji and Pierret (2001) provide a classic test for discrimination in wage-setting based on the dynamics of employer learning and implied trajectories of black and white workers. Lang and Lehmann

---

9There are also parallels between our model and others in which two sides invest before being matched. For example, Nöldeke and Samuelson (2015) develop a model with simultaneous investment followed by ex post matching. Although information is complete in their context, the models share a complementarity in decisions that underlies inefficient equilibria.

10Another related paper is Filippin (2009). In that model, unequal opportunities between groups are self-fulfilling, but driven by incorrect minority worker beliefs. Minority workers, perhaps due to a history of poor treatment, believe incorrectly that employers are biased. This causes them to supply low levels of effort, which precludes them from being promoted. Since minority workers never provide high levels of effort, they never observe whether they would have been promoted.
suggest an alternative test, which is based on the same data but is more robust to changing black-white relative productivity. Fryer et al. (2013) use a labor market search model to derive a conservative test for discrimination based on changes in the average wages of black and white workers who switch to new firms. Our contribution to this literature is to propose a test that would ensure that any finding of discrimination is robust to confounding variation in firm investment and worker belief formation.\textsuperscript{11}

\subsection{Strategic Complementarities}

As mentioned in the Introduction, the terms \textit{strategic complements} and \textit{strategic substitutes} were first used by Bulow et al. (1985) who used the two concepts to shed light on results in oligopoly theory.\textsuperscript{12} When two players’ actions are strategic complements, each player’s set of best responses weakly increases with the actions of the other. In our model, there is strategic complementarity between firm and worker investment: a higher level of firm (worker) investment raises the incentive for any worker (firm) to invest. This is the fundamental logic that underlies our ultimate policy prescription: two-sided investment insurance.

Vives (1990) builds on these concepts and analyzes \textit{supermodular} games, where all players’ payoffs satisfy monotonicity properties that are closely related to strategic complementarity.\textsuperscript{13} He shows that such games have appealing properties. However, despite the strategic complementarity between worker and firm investment, the game we describe is not supermodular due to the lack of strategic complementarity \textit{within} each side of the market. The results discussed by Vives therefore do not apply.

\textsuperscript{11}In addition to the papers cited here, alternative tests have been suggested in other contexts. For example, Knowles et al. (2001) use search and success rates to test for racial discrimination in searches for contraband by police officers.

\textsuperscript{12}In another early paper, Cooper and John (1988) use the concept of strategic complementarity to analyze macroeconomic coordination failures.

\textsuperscript{13}Supermodular games were originally developed by Topkis (1979) but were first applied to economics by Vives (1990) and further analyzed by Milgrom and Roberts (1990).
2.3 The Basic Model

2.3.1 Building Blocks

Imagine a large number of employers and an equal number of workers. Each worker is matched to one firm, drawn randomly from the population. Workers belong to one of two identifiable groups, \( j \in \{A, B\} \). Denote by \( \lambda_A \) the fraction of As in the population and \( \lambda_B = 1 - \lambda_A \) the fraction of Bs. One can imagine groups being race, gender, or any other protected class.

Nature moves first and assigns a type to each worker and a type to each employer. The worker’s type, denoted by \( c \in (0, \overline{c}], \overline{c} < \infty \), depicts her cost of investment in human capital. Let the fraction of workers with costs no greater than \( c \) be represented by \( G^W(c) \) – a smooth cumulative distribution function – with \( g^W(c) \) the associated density. Similarly, each employer has the opportunity to invest at cost \( k_j \in (0, \overline{k}], \overline{k} < \infty \) to make their workplace desirable and productive for workers of type \( j \). This is a fixed cost, drawn independently for each group. The fraction of employers with investment cost no greater than \( k_j \) is \( G^E(k_j) \), with density \( g^E(k_j) \). Superscripts “\( W \)” and “\( E \)” refer to workers and employers, respectively.

Consistent with Lang (1986), we assume that firm investment costs are fixed. This is done for analytical simplicity and symmetry. Appendix B.4.2 describes how the model changes if we assume that investment costs are proportional to the number of workers hired. This has no impact on our main results.

After observing their costs, workers (resp. employers) make a dichotomous investment decision, choosing to become “qualified” or “unqualified”, with no in-between. For workers, “qualified” implies that they are productive. For firms, “qualified” implies they are...
establishments that are desirable for a given type of worker. Nature then distributes a signal to employers regarding each worker’s investment decision and, simultaneously, a signal to workers regarding the employer’s investment decision. Specifically, let \( \theta \in \Theta = [0,1] \) denote a noisy, but informative, signal to employers about whether or not a particular worker chose to invest. There is an associated cumulative distribution function \( F_{i}^{W}(\theta) \), and smooth density, \( f_{i}^{W}(\theta) \), where \( i \in \{q, u\} \). We assume that \( \phi(\theta) = \frac{f_{i}^{W}(\theta)}{f_{j}^{W}(\theta)} \) is non-increasing in \( \theta \) (i.e., \( f_{i}^{W}(\theta) \) satisfies the monotone likelihood ratio property).

The signal structure for employer investment is similar: nature distributes a noisy but informative signal \( \psi \in \Psi = [0,1] \) to workers about whether or not the employer chose to invest. There is an associated cumulative distribution function, \( F_{i}^{E}(\psi) \), and smooth density, \( f_{i}^{E}(\psi) \), where \( i \in \{q, u\} \). We assume that \( \tau(\psi) = \frac{f_{i}^{E}(\psi)}{f_{j}^{E}(\psi)} \) is non-increasing in \( \psi \).

Next, workers observe the signal they receive from the employer and decide whether to “apply.” If they receive information that suggests that a given workplace will be a poor fit for their type, they may refrain from applying. Firms observe \( \theta \) for all who apply and make a deterministic hiring decision: hire or reject. Production occurs and payoffs are received.

2.3.2 Payoffs

If a worker is hired and the employer to whom she has been matched has made a group \( j \) investment, she receives a fixed payoff of \( \omega_{q} - c \) if she chose to invest, and \( \omega_{q} \) if not. If the worker is hired and the employer has not made a group \( j \) specific investment, she receives \(-\omega_{u} - c\) if she invested and \(-\omega_{u}\) if she did not. If she does not work for the employer, she receives \(-c\) if she invested or zero otherwise. We assume that both \( \omega_{q} \) and \( \omega_{u} \) are positive and exogenously determined. This, again, is purely for analytical convenience and ease of

\[ 14 \text{ We view this model as a static approach to what is likely a dynamic process. In a dynamic version, a worker’s choice to refrain from applying would reflect the option value of waiting for a better offer.} \]

\[ 15 \text{ A mathematically equivalent assumption (see Appendix B.3.1) is that the worker receives an unemployment payment \( \bar{U} \) – or other outside option – if she does not apply. In this case, the key assumption is that workers prefer unemployment to being matched to an employer who has not invested. A third alternative is to imagine that application to a firm is costly.} \]
exposition in our baseline model and does not change our main results.\textsuperscript{16}

The employer receives $\chi_q - k_j, \chi_q > 0$, if it hires a qualified worker and makes group $j$-specific investments, and $\chi_q$ if it hires a qualified worker and chooses not to invest in group $j$ amenities. Similarly, the employer’s payoffs are $-\chi_u - k_j$ where $-\chi_u < 0$ if it hires an unqualified worker and makes group $j$-specific investment, and $-\chi_u$ if it hires an unqualified worker and chooses not to invest in group $j$-specific amenities. If no worker is hired, the employer receives $-k_j$ if it invested and zero otherwise.

2.3.3 Strategies

The worker’s strategy consists of two functions: an investment decision that depends on her type; and an application decision that may be contingent on her type, the realization of $\psi$, and her own investment decision. We can write this as $I^W : \{A, B\} \times [0, \bar{c}] \rightarrow [0, 1]$ and $A^W : \{A, B\} \times \{q, u\} \times \Psi \times [0, \bar{c}] \rightarrow [0, 1]$. The employer’s strategy consists of two investment decisions and two assignment rules – one for each group $j \in \{A, B\}$ – which we can write as $I^E : [0, \bar{k}]^2 \rightarrow [0, 1]^2$ and $A^E : \{q, u\}^2 \times \Theta \times [0, \bar{k}]^2 \rightarrow [0, 1]^2$.

2.3.4 Expected Payoffs

Employer Offer Threshold

Let $\pi_j \in [0, 1]$ be the employer’s prior belief that a worker of group $j$ is qualified. The expected payoff for the employer is a function of its beliefs, investment decisions, the signal $\theta$, and net payoffs. An employer does not intrinsically care about which type of worker it hires – save investment costs, which are sunk at the time that it makes an offer – but it may have different priors about the likelihoods that workers of different types are qualified.

Given the prior $\pi_j$ and observed signal $\theta$, the employer formulates the posterior probability, using Bayes’ rule, that a worker of group $j$ is qualified: $\kappa(\pi_j, \theta) = \frac{\pi_j f^W_q(\theta)}{\pi_j f^W_q(\theta) + (1 - \pi_j) f^W_u(\theta)}$.

\textsuperscript{16}We endogenize wages in two places: (1) to demonstrate robustness of our approach, we discuss policy when workers are paid either by ex-post bargaining or by their expected marginal product in Appendix B.4.1; and (2) we allow for continuous wage-setting when discussing our proposed empirical test for statistical discrimination in Section 2.6.
The expected payoff to hiring a worker of group $j$ is: $\kappa (\pi_j, \theta) \chi_q - (1 - \kappa (\pi_j, \theta)) \chi_u$. Recall that the payoff for not hiring a worker is 0. The condition that this expected payoff be positive defines a standard, $s^*_j (\pi_j)$, which is a critical threshold in the signal $\theta$ such that the employer will choose to hire only if a worker’s signal exceeds this threshold.

$$s^*_j (\pi_j) \equiv \min \left\{ \theta \in [0, 1] | \frac{\chi_q}{\chi_u} > \left( \frac{1 - \pi_j}{\pi_j} \right) \phi (\theta) \right\} \quad (2.1)$$

**Worker Application Threshold**

Let $\delta_j \in [0, 1]$ denote the prior belief that a worker of type $j \in \{A, B\}$ has that an employer made the investment relevant to her group. The worker’s expected payoff is a function of her beliefs, investment decision, the signal she receives, and net payoffs. Given $\delta_j$ and observed signal $\psi$, workers calculate the posterior probability that a particular employer has invested, again using Bayes’ rule: $\xi (\delta_j, \psi) = \frac{\delta_j f_d^E (\psi)}{\delta_j f_d^E (\psi) + (1 - \delta_j) f_u^E (\psi)}$. The worker’s expected payoff of applying can be written as: $\xi (\delta_j, \psi) \omega_q - (1 - \xi (\delta_j, \psi)) \omega_u$. Thus, similar to before, the worker will only apply if the employer’s signal exceeds a threshold $t^*_j (\delta_j)$.

$$t^*_j (\delta_j) \equiv \min \left\{ \psi \in [0, 1] | \frac{\omega_q}{\omega_u} > \left( \frac{1 - \delta_j}{\delta_j} \right) \tau (\psi) \right\} \quad (2.2)$$

**Investment Decisions**

We begin with the worker. With probability $\delta_j \left( 1 - F_q^E \left( t^* (\delta_j) \right) \right)$ a group $j$ worker will be matched to an employer who made the group $j$ investment and will apply to that employer because the signal she receives exceeds her application threshold. However, with probability $(1 - \delta_j) \left( 1 - F_u^E \left( t^* (\delta_j) \right) \right)$ she will apply to an employer who did not invest. In total, the worker’s expected payoff from successfully obtaining a job is $\overline{\omega} (\delta_j) = \delta_j \left( 1 - F_q^E \left( t^* (\delta_j) \right) \right) \omega_q - (1 - \delta_j) \left( 1 - F_u^E \left( t^* (\delta_j) \right) \right) \omega_u$.

Investing in human capital increases the likelihood that a worker is accepted by an employer. If a worker of type $j$ invests, she gets expected gross payoff: $(1 - F_q^W (s_j)) \overline{\omega} (\delta_j)$. Conversely, if she does not invest, she gets $(1 - F_u^W (s_j)) \overline{\omega} (\delta_j)$. Thus, the net return on
investment for workers can be written as:

\[
\beta_W (s_j, \delta_j) \equiv \left[ F_u^W (s_j) - F_q^W (s_j) \right] \bar{w} (\delta_j). \tag{2.3}
\]

Now, consider the employer’s investment decision. Similar to the worker, the employer’s expected net payoff from hiring a worker is

\[
\pi_j = \pi_j \left( 1 - F_u^W \left( s_j^* (\pi_j) \right) \right) \chi_q - (1 - \pi_j) \left( 1 - F_u^W \left( s_j^* (\pi_j) \right) \right) \chi_u.
\]

If the employer makes the group \( j \) investment, it gets gross payoff \( \lambda_j [1 - F_q^E (t_j)] \bar{X} (\pi_j) \). If the employer does not invest, it gets \( \lambda_j [1 - F_u^E (t_j)] \bar{X} (\pi_j) \).

Recall that \( \lambda_j \) is the fraction of workers who are members of group \( j \). Thus, the net return to investment for firms is:

\[
\beta_E (t_j, \pi_j | \lambda_j) \equiv \lambda_j \left[ F_u^E (t_j) - F_q^E (t_j) \right] \bar{X} (\pi_j). \tag{2.4}
\]

2.3.5 Bayesian Nash Equilibrium

For each of the two groups, \( A \) and \( B \), a pair of beliefs – one for employers and one for workers – will be self-confirming if, by choosing standards optimal for those beliefs, the actions of each group of agents induce the other to become qualified at exactly the rate posited by the initial beliefs. This intuition leads to the following definition of equilibrium:

**Definition 1.** An equilibrium is a set of beliefs \((\pi_A, \pi_B), (\delta_A, \delta_B)\), worker standards \((t_A, t_B)\) and employer standards \((s_A, s_B)\) satisfying the following conditions for each \( j \in \{A, B\}\):

\[
\pi_j = G^W \left( \beta_W \left( s_j^* (\pi_j), \delta_j \right) \right) \tag{2.5}
\]

\[
\delta_j = G^E \left( \beta_E \left( t_j^* (\delta_j), \pi_j | \lambda_j \right) \right) \tag{2.6}
\]

where \( s_j^* (\pi_j) \) and \( t_j^* (\delta_j) \) are as defined by equations 2.1 and 2.2 respectively.

The equilibrium conditions for groups \( A \) and \( B \) are separable, since both investments and the threshold rules \( s_j^* (\pi_j) \) and \( t_j^* (\delta_j) \) are assumed to be group-specific, and each firm’s investment costs are drawn independently from each other. We thus omit the subscript \( j \) throughout this section, except when explicitly comparing equilibria between groups.
For each group \( j \in \{A, B\} \), equilibrium requires that both employer and worker beliefs are confirmed in equilibrium vis-à-vis a self-confirming feedback loop. Fix \( \delta \), and suppose that an employer believes that a fraction \( \pi \) of workers are qualified. Expecting this, each worker calculates her net benefit of investment and invests if and only if her cost is less than the benefit. In equilibrium, the fraction of workers who invest must be equal to the employer’s beliefs \( \pi \). Workers’ beliefs about firms must also be self-confirming.

For any fixed \( \delta \), our model nests the seminal Coate and Loury (1993) model where the wage is \( \overline{w}(\delta) \). Similarly, fixing \( \pi \) at some level induces a version of the same model in which roles are reversed: workers’ pessimistic beliefs drive disparate outcomes. In summary, discrimination can be generated by either side in the two-sided model and is generically sustained by both (e.g., in Silicon Valley, employers may discount the skills of women, while at the same time potential female software engineers discount employers’ proclamations about workplace flexibility, sexual harassment, and discrimination).

Beliefs in this model exhibit extensive complementarity. First, a belief that one side is more likely to invest increases the expected return to investment of the other side, since it strictly increases the likelihood (\( \pi \) or \( \delta \)) of getting a positive payoff from a given match. The signal threshold, \( s^*(\pi) \) or \( t^*(\delta) \), used by the opposing side is also lowered. For example, a rise in the fraction of firms investing causes \( t^*(\delta) \) to fall as workers become more optimistic and more willing to apply for jobs. Conversely, a rise in \( \pi \) causes \( s^*(\pi) \) to fall as firms become more optimistic about workers and more willing to hire.

To understand the mechanics of equilibrium, it is instructive to consider fixing the level of firm investment at some \( \hat{\delta} \). This fixes the worker application threshold at \( t^*(\hat{\delta}) \) and induces a model that is isomorphic to Coate and Loury (1993) with worker wage \( \overline{w}(\hat{\delta}) \). This is shown in Figure 2.2a. Holding fixed the level of firm investment, equilibrium is characterized by two graphs in \( \{\pi, s\} \) space: an EE curve, which embodies the employer’s hiring threshold; and a WW curve, which describes optimal worker investment as a function of that hiring threshold. The EE curve is downward sloping, since more optimistic firms set more generous (lower) thresholds. The WW curve is hump-shaped, reflecting the fact that
there is little incentive for workers to invest if employers set very high or low standards.

A high enough level of \( \delta \) ensures the existence of at least two non-zero solutions to equation 2.5. At each of these solutions, employers’ hiring standards are optimally set and every worker is making her investment decision optimally. The solutions for one arbitrary chosen \( \delta \) are shown as white dots in Figures 2.2a and 2.2b. As \( \delta \) rises, the WW curve shifts upward, and the value of \( \pi \) that solves equation 2.5 rises if the EE curve crosses from above and falls if it crosses from below. Varying \( \delta \) in this way traces out the solutions to equation 2.5 in \( \{\pi, \delta\} \) space as shown by \( \pi^*(\delta) \) in Figure 2.2b.

An analogous thought exercise can be conducted for any fixed \( \pi \), which induces a similar one-sided model in which employers invest and workers decide whether to apply. Varying \( \pi \) traces out the solutions to equation 2.6 as shown by \( \delta^*(\pi) \) in Figure 2.2b.

A “zero investment equilibrium” always exists in the two-sided model, since \( \pi = \delta = 0 \) and \( s = t = 1 \) always satisfy the equilibrium equations. Yet, other equilibria may also exist. This should not be surprising given the strategic complementarity between workers and firms that we have described. For example, there are exactly three equilibria for the parameterization shown in Figure 2.2b: \( Z, U \) and \( S \). The model generally has multiple solutions if workers and employers are responsive enough to each others’ investments.

**Proposition 6.** Assume that \( \phi(\theta) \) and \( \tau(\psi) \) are continuous, strictly decreasing and strictly positive on \([0, 1]\), and that \( G^W(c) \) and \( G^E(k) \) have full support on \([0, \tau]\) and \([0, K]\) with \( G^W(0) = G^E(0) = 0 \). Further assume that for some \( \hat{\delta}, \) there exists an \( s \) for which \( G^W(\beta_W(s, \hat{\delta})) > \phi(s) / [\chi_q / \chi_u + \phi(s)] \). Also assume that for some \( \hat{\pi} \), there exists a \( t \) for which \( G^E(\beta_E(t, \hat{\pi}|\lambda)) > \tau(t) / [\omega_q / \omega_u + \tau(t)] \). Then non-zero elements of \( \pi^*(\delta) \) and \( \delta^*(\pi) \) exist for any \( \delta \geq \hat{\delta} \) and \( \pi \geq \hat{\pi} \) respectively. If there is a set of beliefs \( \{\pi, \delta\} \) such that \( \delta \in \delta^*(\pi) \) and \( \pi < \max \{\pi^*(\delta)\} \) then there exist multiple solutions to the two-sided model.

All technical proofs are presented in Appendix B.1. To better understand the logic behind Proposition 6, see Figure 2.2b. First, the assumptions about \( \phi(\theta) \) guarantee that the solutions to equation 2.5 are bounded below one because there exists a threshold \( \pi < 1 \).
above which $s^*(\pi) = 0$, which in turn implies zero worker investment. An equivalent argument implies that the non-zero solutions to equation 2.5 are bounded strictly below one. Second, we assume that there exists some pair of beliefs $\{\pi, \delta\}$ such that $\delta < \delta^*(\pi)$ and $\pi < \max \{\pi^*(\delta)\}$, a condition that must hold for large enough $\chi_q$ and $\omega_q$. Combined with our regularity assumptions, this is enough to ensure that $\pi^*(\delta)$ and $\delta^*(\pi)$ intersect at multiple points with $\delta > 0$ and $\pi > 0$, which define equilibria with non-zero investment. This is the case in Figure 2.2b, and in the simple example that we consider below.

In our baseline model with fixed investment costs, there is an inherent disadvantage to being a numerical minority. If $\lambda_A = \lambda_B$ then the set of equilibria is fully symmetric. However, if $\lambda_A > \lambda_B$ then, for any given beliefs, firms have a strictly lower incentive to invest in amenities for $B$s. At the extreme, as $\lambda_B \to 0$, no firms will ever be motivated to accommodate $B$s. This intuition is straightforward to see from equation 2.2. For any fixed $\pi$ and $t$, $\lambda$ scales the return, while investment costs are independent of population size.\textsuperscript{17}

\textsuperscript{17}In Appendix B.4.2, we present an extension in which costs are incurred only for workers of group $j$ who apply and are hired. The analysis of this case is necessarily more complex, but the conclusions are largely
A discriminatory equilibrium is one in which employers do not have homogeneous beliefs (e.g., \( \pi_A > \pi_B \)), and therefore set different hiring standards. Discrimination implies that two workers who send the same signal to employers would be treated differently depending on their group membership. Even if the set of equilibria is symmetric, this can occur whenever the system described by Definition 1 has multiple solutions, for then both workers and employers understand that workers of group \( A \) are more qualified than workers of group \( B \), and employers are less likely to make the workplace suitable for \( B \) than they are for \( A \).\(^{18}\) One can imagine the familiar refrain: employers would be delighted to hire \( B \)s but they are just not qualified. And \( B \)s retort that they would invest if only they could trust that their efforts would be rewarded by employers. Discrimination, in this symmetric case, is a classic coordination problem.

### 2.3.6 Dynamics

To analyze dynamics, we define a learning process that describes how employers and workers adjust their beliefs and actions in response to a shock. This rule is backward-looking, with each generation of workers and firms choosing their actions based on the decisions of the preceding generation. We posit that this adjustment process plausibly and simply captures how agents would react to a policy change.\(^{19}\)

\[
\pi_{t+1} = G^W \left( [F^W_t (s^* (\pi_t)) - F^W_q (s^* (\pi_t))] \right) \\
\times \left[ \delta_t \left( 1 - F^E_q (t^* (\delta_t)) \right) \omega_q - (1 - \delta_t) \left( 1 - F^E_q (t^* (\delta_t)) \right) \omega_u \right] \\
\delta_{t+1} = G^E \left( \lambda \left[ F^E_t (t^* (\delta_t)) - F^E_q (t^* (\delta_t)) \right] \right) \\
\times \left[ \pi_t \left( 1 - F^W_q (s^* (\pi_t)) \right) \chi_q - (1 - \pi_t) \left( 1 - F^W_q (s^* (\pi_t)) \right) \chi_u \right]
\]

unchanged.

\(^{18}\)Although we focus on discrimination in hiring, we note that differences in firm investment rates between groups could also be viewed as directly discriminatory.

\(^{19}\)Our results are qualitatively robust to generalizations along the lines of Kim and Loury (2018), which features overlapping generations who are forward-looking. Introducing a more complex adjustment process is beyond the scope of this paper.
The same process can also be used to refine the set of equilibria to those that are stable in the sense that they are robust to small errors of perception. For example, equilibrium $U$ in Figure 2.2 is unstable, but the equilibrium with zero investment ($Z$) is stable. We discuss this type of refinement further in Appendix B.2.

### 2.3.7 An Example with Uniform Cost and Signal Distributions

To further fix ideas, we now introduce a simple example to provide intuition for the model. Let costs for workers and firms be distributed uniformly on $[0, 1]$ so that $G^W(c) = c$ and $G^E(k) = k$.

Worker signals are also uniformly distributed, but with the support depending on the investment decision. A qualified worker’s signal is distributed uniformly on $[\theta_q, 1]$, while an unqualified worker’s signal is distributed uniformly on $[0, \theta_u]$ with $\theta_q < \theta_u$. Thus, a worker is surely qualified if $\theta > \theta_u$, surely unqualified if $\theta < \theta_q$, and there is a constant likelihood ratio $\hat{\phi} = \frac{1 - \theta_q}{\theta_u}$ for $\theta \in [\theta_q, \theta_u]$.

We make analogous assumptions for firms. The signal sent by a firm that invested is uniformly distributed on $[\psi_q, 1]$, while that sent by a firm that did not invest is distributed uniformly on $[0, \psi_u]$ with $\psi_q < \psi_u$. Thus, a firm certainly invested if $\psi > \psi_u$, certainly did not invest if $\psi < \psi_q$, and there is a constant likelihood ratio $\hat{\tau} = \frac{1 - \psi_q}{\psi_u}$ for $\psi \in [\psi_q, \psi_u]$.

The employer will always reject workers with clear fail signals and always accept those with clear pass signals. However, employers will make an offer to unclear workers if and only if they are “optimistic” enough in the sense that their prior $\pi_j$ is greater than a fixed threshold $\hat{\pi}_j$. In symbols:

$$\pi_j \geq \left( \frac{\phi}{\chi_q/\chi_u + \phi} \right) = \hat{\pi}_j$$

Thus, the employer will set the hiring threshold at either $s^* = \theta_q$ or $s^* = \theta_u$. Similarly, workers will always apply to firms with clear pass signals, never to firms with clear fail signals, and will apply to firms with unclear signals if and only if they are optimistic enough about firms: $\delta_j \geq \hat{\delta}_j$. The worker will therefore set the threshold at either $t^* = \psi_q$ or $t^* = \psi_u$.

To make this example especially simple, we adopt parameter values that make firms
Figure 2.3: Equilibria in the clear / unclear example

and workers symmetric. Specifically, let $\theta_q = \psi_q = \frac{1}{3}$, $\theta_u = \psi_u = \frac{2}{3}$, $\omega_q = 3$, $\omega_u = 1$, $\chi_q = 6$, $\chi_u = 2$ and $\lambda = 0.5$. With these functional form assumptions and parameter values, the returns to investment for workers ($\beta_W$) and employers ($\beta_E$) are piecewise-linear functions.\(^{20}\)

$$\beta_W = \begin{cases} \frac{7}{4}\delta - \frac{1}{4} & \text{if } \delta \geq \hat{\delta} \\ \frac{3}{4}\delta & \text{if } \delta < \hat{\delta} \end{cases}$$

$$\beta_E = \begin{cases} \frac{7}{4}\pi - \frac{1}{4} & \text{if } \pi \geq \hat{\pi} \\ \frac{3}{4}\pi & \text{if } \pi < \hat{\pi} \end{cases}$$

The fraction of workers who invest is $\pi = \min\{\max\{G_W, 0\}, 1\}$ and the fraction of firms who invest is $\delta = \min\{\max\{G_E, 0\}, 1\}$.

The equilibria in this example are shown in Figure 2.3, which is the equivalent of Figure 2.2b for this example. For any fixed belief $\pi$ about the fraction of workers who are investing, the solid line, $\delta^* (\pi)$, shows the fraction of firms who optimally invest. Similarly, the dotted line, $\pi^* (\delta)$, shows the fraction of workers who invest for a given belief $\delta$ about the fraction of firms who are investing. There are three equilibria, shown as $S$, $U$ and $Z$. Only at these points are the actions and beliefs of firms and workers mutually consistent.

---

\(^{20}\)The unparameterized returns are derived in Appendix B.3.2.
Also evident in Figure 2.3 are the effects of changing signal thresholds. As firms become more optimistic (higher $\pi$), there are two effects. First, favorable beliefs about workers directly raise the return to firm investment, since the expected payoff from a match is higher. Secondly, firms eventually become so optimistic that they accept workers with ‘unclear’ test scores. At this point, firm returns become more sensitive to worker investment, since a match is more likely when workers are given the benefit of the doubt. This is reflected in a change in the slope of $\delta^* (\pi)$. The explanation of the shape of $\pi^* (\delta)$ is analogous.

Both Proposition 1 and this simple example demonstrate that multiple equilibria can occur in our model. In the next two sections, we explore policies that can help minimize the set of “bad” equilibria (i.e. those with low investment and pessimistic beliefs on both sides) and that furthermore achieve equality between the two groups.

### 2.4 Extending the Basic Model: One-Sided Policies

We now consider how the government or another third party intermediary might intervene with some policy to break equilibria that disadvantage some groups relative to others. We are particularly interested in a policy that: (i) eliminates equilibria without homogeneous beliefs; (ii) never harms its recipients; and (iii) achieves equality as quickly as possible. Given its historical and current prominence across the world – and the controversy that typically ensues – we begin with affirmative action.

Affirmative Action – Executive Order # 11246, signed by Lyndon B. Johnson – has been promulgated around the world from Malaysia to South Africa to Lebanon. Affirmative action policies generally entail the preferential treatment of persons who possess certain social traits based on a presumption that, on average, individuals of those traits are less effective in the competition for scarce resources because of some social or historical handicap.

The simplest affirmative action policy insists that employers make color-blind assignment – requiring that $A$s and $B$s with identical “signals” be treated equally. Unfortunately, this policy can only be enforced if – in every instance – a regulator can observe and verify all information upon which employers rely when making hiring decisions. We assume this type
of extreme informational requirement – essentially requiring government regulators to sit in all interviews – is impractical. Instead, we explore two potential definitions of affirmative action – equality in offers and equality in employment. We begin with the former.

2.4.1 Equality in Offers

In statistical discrimination models without strategic complementarity between worker and firm investment, affirmative action can be quite successful. In such models, affirmative action rules out the existence of any equilibrium with zero investment by workers of one group but positive investment by members of the other. Furthermore, firms in such models can satisfy the affirmative action requirement by setting $s_A = s_B$, which achieves full equality in investment rates and correspondingly homogeneous beliefs: i.e., $\pi_A = \pi_B$. Nonetheless, as Coate and Loury (1993) make clear, affirmative action may not lead to homogeneous beliefs if more generous hiring standards undermine investment and become demotivating. As we will show, affirmative action is even more problematic in a model with strategic complementarity.

The behavior of workers is not directly affected by affirmative action. They continue to make their decisions as before. However, affirmative action changes an employer’s problem, because standards and investments can no longer be chosen independently for the two groups.

Consider a group of workers about which an employer believes a fraction $\pi$ are qualified and uses assignment standard $s$. For each group, let $\rho(s_j, \pi_j) = \pi_j \left[1 - F^W_{q_j}(s_j)\right] + (1 - \pi_j) \left[1 - F^W_{u_j}(s_j)\right]$ be the probability that the employer assigns to making an offer to a random worker, and let $P(s_j, \pi_j, i_j)$ denote the expected payoff from making an offer to such a worker, where $i_j \in \{q, u\}$ captures the firm investment decision: $P(s_j, \pi_j, i_j) = \pi_j \left[1 - F^W_{q_j}(s_j)\right] \left[1 - F^E_{i_j}(t_j)\right] \chi_q - (1 - \pi_j) \left[1 - F^W_{u_j}(s_j)\right] \left[1 - F^E_{i_j}(t_j)\right] \chi_u$. Note the indirect dependence of this expected payoff the worker’s application threshold $t_j$, which plays an important role in the analysis below.

In the modified game, each employer must ensure that, whatever standards it uses,
anticipated hiring rates for each group are equal: i.e., $\rho(s_A, \pi_A) = \rho(s_B, \pi_B)$. Given beliefs $(\pi_A, \pi_B)$ and worker application standards $(t_A, t_B)$, the firm will thus choose hiring standards $(s_A, s_B)$ and an investment decision $i_j \in \{q, u\}$ for each group to solve the following optimization problem:

$$\max_{s_A, s_B, i_A, i_B} \left[ \lambda_B P(s_B, \pi_B, i_B) + \lambda_A P(s_A, \pi_A, i_A) \right] - \sum_{j \in \{A, B\}} k_j \times 1(i_j = q) \quad (2.7)$$

subject to $\rho(s_B, \pi_B) = \rho(s_A, \pi_A)$

Employers may set different standards depending on which investments they made, because investments affect the ability of an employer to attract workers. This in turn affects the relative attractiveness of adjusting each of the two signal thresholds, $s_A$ and $s_B$.

We therefore use $s_{j,i}^f$ to denote the hiring threshold set for group $j$ by a firm that made investment decisions $i_A$ and $i_B$. Using this notation, each firm will choose one of the following four payoffs.

$$\lambda_B P(s_{j,B}^q, \pi_B, q) + \lambda_A P(s_{j,A}^q, \pi_A, q) - k_A - k_B \quad (2.8)$$

$$\lambda_B P(s_{j,B}^u, \pi_B, u) + \lambda_A P(s_{j,A}^u, \pi_A, q) - k_A \quad (2.9)$$

$$\lambda_B P(s_{j,B}^q, \pi_B, q) + \lambda_A P(s_{j,A}^u, \pi_A, u) - k_B \quad (2.10)$$

$$\lambda_B P(s_{j,B}^u, \pi_B, u) + \lambda_A P(s_{j,A}^u, \pi_A, u) \quad (2.11)$$

For each group, this choice defines a critical cost threshold below which a firm will invest, but one that depends on the firm’s investment cost for the other group, $k_{-j}$. Defining this threshold as $k_j^f(k_{-j})$, the rate of investment for group $j$ can be written as

$$\delta_j = \int_0^1 G^E(k_j^f(k_{-j})) dG^E(k_{-j}).$$

The return expected by a worker, which we denote by $\beta^W$, will also be a more complicated function of firm costs and all four hiring thresholds for her group. Specifically, $\beta^W = \sum_{m \in \{q, u\}} \sum_{n \in \{q, u\}} \Pr(i_A = m, i_B = n) \beta_W \left(s_j^m, \delta_j\right)$. In turn, the probability $\Pr(i_A = m, i_B = n)$ is pinned down by multiple inequalities relating expressions 2.8 to 2.11. For example,

---

21 See Lemma 4 in Appendix B.1 for a characterization of the solution to problem 2.7.
Pr(i_A = q, i_B = q) is equal to the probability that expression 2.8 is less than expressions 2.9, 2.10 and 2.11, which defines a frontier for the two investment costs, k_A and k_B, inside which the firm undertakes both investments. Bringing everything together, this suggests the following definition of equilibrium under an employment quota.

**Definition 2.** An equilibrium under affirmative action is a set of beliefs \((\pi_A, \pi_B), (\delta_A, \delta_B)\), worker standards \((t_A, t_B)\) and employer standards \((s_A, s_B)\) satisfying the following conditions:

(a) Firms’ investment decisions \((i_A, i_B)\) and thresholds \((s_A, s_B)\) solve (2.7), given \((\pi_A, \pi_B, t_A, t_B)\).

(b) \(t_j = t_j^* (\delta), j \in \{A, B\}\)

(c) \(\pi_j = G^W (\pi_j)\), \(j \in \{A, B\}\)

(d) \(\delta_j = \int_0^1 G^E (k_j^* (k_j^*)) dG (k_j^*)\)

The only requirement that affirmative action adds is the constraint that \(\rho (s_B, \pi_B) = \rho (s_A, \pi_A)\). Without this, we obtain the unconstrained version of problem 2.7, which is solved by \(s_A = s^* (\pi_A), s_B = s^* (\pi_B)\) and firm investment rate \(\delta_j = G^E (\beta_t (t_j, \pi_j | \lambda_j))\). It is also clear that if an equilibrium with homogeneous beliefs exists without affirmative action, then an equilibrium with the same beliefs exists with the constraint. This follows directly from the fact that the affirmative action constraint is non-binding in any equilibrium in which employers have homogeneous beliefs \((\pi_A = \pi_B)\).

However, unlike Coate and Loury (1993), it is generally impossible to guarantee that homogeneous beliefs will prevail because an equilibrium with zero B investment but positive A investment always satisfies the affirmative action constraint. We formalize this result in Proposition 7, but the intuition is simple: if Bs do not apply, then changes in the firm offer threshold s have no effect.
Proposition 7. Assume that, without affirmative action, there exists an equilibrium with positive investment. Given the same parameters, there exists an equilibrium under affirmative action without homogeneous beliefs.

A second problem is that there is generally no equilibrium with positive investment and homogeneous employer beliefs unless unless $\lambda_A = \lambda_B$. The reason for this is that firm investment returns are lower for smaller groups. More formally, suppose that $\pi_A = \pi_B = \pi$. The affirmative action constraint is then only satisfied with equal firm hiring standards $s_A = s_B = s^*(\pi)$. Worker beliefs cannot be homogeneous in this case since firm investment incentives are strictly lower for the minority for any given firm beliefs $\pi$. However, strictly lower firm investment rates for Bs ($\delta_B < \delta_A$) combined with equal hiring thresholds must lead to lower investment returns for B workers ($\pi_B < \pi_A$). This is a contradiction, leading to the following result.

Proposition 8. Assume that $\phi(\theta)$ and $\tau(\psi)$ are continuous, strictly decreasing and strictly positive on $[0,1]$. Further assume that $\lambda_A \neq \lambda_B$ and that $G^E(k)$ and $G^W(c)$ are strictly increasing. Then no equilibrium with positive investment and homogeneous employer beliefs exists (with or without affirmative action).

A final drawback of this type of affirmative action in our model is that it can make outcomes for B workers strictly worse than under the status quo. Following the intuitive learning process we described in Section 2.3, suppose that employer and worker beliefs are fixed in the short run. At these fixed beliefs, the imposition of an affirmative action constraint can cause firm investment returns for Bs to become negative, ensuring that no firm invests and thus no B workers apply. This triggers reversion to zero investment by firms, and ultimately also workers ($\delta_B = \pi_B = 0$). We formalize this result in Proposition 9.

Proposition 9. Assume that $\phi(\theta)$ and $\tau(\psi)$ are continuous, strictly decreasing and strictly positive on $[0,1]$. Further suppose that the A and B markets start with $\pi_A > \pi_B > 0$ and $\delta_A > \delta_B > 0$. For
fixed beliefs \( \{\pi_A, \pi_B, \delta_A, \delta_B\} \) and low enough \( \delta_B \) and \( \pi_B \), imposing affirmative action causes zero firms to invest in B amenities and zero B workers to invest.

The intuition here is simple. Without affirmative action, firms were already hiring the few minority workers who they expected to be qualified. Affirmative action forces them to hire a potentially large number of additional minority workers if they apply, and these additional workers are expected to generate a loss for the firm on average. As a result, the few employers who were making investments to garner additional applications from minority workers now have less incentive to do so. If this effect is strong enough, they may even have an incentive to actively deter such applications. In summary, the affirmative action requirement may hurt minority workers’ interests by undermining employers’ efforts to attract them.

The mechanisms behind the failure of affirmative action here are substantively different from one-sided models and apply far more generally. For example, in Coate and Loury (1993), affirmative action eliminates equilibria with zero investment by one group but not the other, and equilibria with homogeneous beliefs always exist. The problem that arises is that homogeneous beliefs may not obtain if the parameters allow more generous employer hiring standards to be sufficiently demotivating. Instead, there may be a solution to the one-sided equivalent of problem 2.7 that features lower standards but also lower investment by the minority group. Since our model nests Coate and Loury (1993), this is also a concern here. However, affirmative action in the two-sided model is additionally complicated by worker belief formation and the inherent disadvantage that minorities face because employers have less incentive to adapt their workplaces to accommodate smaller groups.\(^{22}\) Affirmative action may even backfire in our model by causing employers to scale back their efforts to attract minorities.

\(^{22}\)Our result that members of smaller minorities are worse off aligns with the predictions of search models (e.g., Black, 1995), although Becker’s (1957) analysis of taste-based discrimination with perfect sorting of workers across firms predicts the opposite.
2.4.2 Equality in Employment

We now consider employment quotas, which require that members of groups \( A \) and \( B \) are hired in proportion to their population sizes.\(^{23}\) This articulation of affirmative action may be closer to the original spirit of early affirmative action (Department of Labor, 1969). Under this type of constraint, employers cannot use the excuse that they would like to hire minorities but are not receiving applications. However, employment quotas can trigger a severe version of *patronization* in which aggressive hiring by employers undermines minority workers’ incentive to invest.

Let \( \rho_H (s_j, \pi_j, i_j) = \pi_j \left[ 1 - F^W_i (s_j) \right] \left[ 1 - F^E_i (t_j) \right] + (1 - \pi_j) \left[ 1 - F^W_i (s_j) \right] \left[ 1 - F^E_i (t_j) \right] \) be the probability the employer assigns to hiring a randomly drawn group \( j \) worker, where \( i_j \in \{ q, u \} \). The employment quota requires that \( \rho_H (s_A, \pi_A, i_A) = \rho_H (s_B, \pi_B, i_B) \). Thus, given beliefs \((\pi_A, \pi_B)\) and worker application standards \((t_A, t_B)\), an employer chooses hiring standards \((s_A, s_B)\) and investment decisions \((i_A, i_B)\) to solve the following problem.

\[
\max_{s_A, s_B, i_A, i_B} [\lambda_B P (s_B, \pi_B, i_B) + \lambda_A P (s_A, \pi_A, i_A)] - \sum_{j \in \{ A, B \}} k_j \times 1 (i_j = q) \quad (2.12)
\]

s.t. \( \rho_H (s_B, \pi_B, i_B) = \rho_H (s_A, \pi_A, i_A) \)

This is identical to problem 2.7 but \( \rho (s_j, \pi_j) \) has been replaced with \( \rho_H (s_j, \pi_j, i_j) \).

Employers may again set different standards depending on which investments they made. Using the same notation as above, equilibrium under an employment quota can be defined as follows.

**Definition 3.** An equilibrium under an employment quota is a set of beliefs \((\pi_A, \pi_B)\), \((\delta_A, \delta_B)\), worker standards \((t_A, t_B)\) and employer standards \((s_A^q, s_A^u, s_B^q, s_B^u)\), \( j \in \{ A, B \} \) satisfying the following conditions:

(a) Firms’ investment decisions \((i_A, i_B)\) and thresholds \((s_A, s_B)\) solve (2.12), given \((\pi_A, \pi_B, t_A, t_B)\)

(b) \( t_j = t_j^* (\delta_j), \ j \in \{ A, B \} \)

\(^{23}\)We are grateful to Lawrence Katz for suggesting this exercise.
An advantage of an employment quota over a regulation that simply requires equality in offers is that an employment quota obviously eliminates the possibility of an equilibrium with zero investment by B workers but positive investment by A workers. It may even eliminate all discriminatory equilibria. For example, any equilibrium under an employment quota must entail homogeneous beliefs if two conditions hold: (i) the worker signal of firm investment is very informative; and (ii) the employer signal of worker investment is very uninformative. We formalize this in Proposition 10.

**Proposition 10.** Assume that $G^E$ has full support on $[0, \overline{c}]$ with $\overline{c} > \omega_q$, let $\phi(\theta)$ be strictly decreasing, and define $\tilde{s}$ as the firm signal threshold such that $\phi(\tilde{s}) = 1$. If firm investment is close enough to perfectly observable, any equilibrium under an employment quota must entail homogeneous beliefs if:

$$\eta(\overline{p}(s)) < \frac{\phi(s_j)}{\phi(s_j) - 1}$$

for all $s \in [0, \tilde{s})$ where $\eta(c) = \frac{d[c \cdot G(c)]}{dc}$ and $\overline{p}(s) = \left[ F^W_u(s) - F^W_q(s) \right] \omega_q$.

Fixing employer investment decisions, the inequality in Proposition 10 guarantees that no two levels of worker investment are consistent with the same probability of being hired. It is always satisfied if $\phi(0) = f^W_u(0) / f^W_q(0)$ is small enough, which implies that the employer signal of worker productivity is relatively uninformative. Next, near-perfect observability of employer investment ensures that firms will not be able to satisfy the employment quota unless they make both or neither of the investments. Thus, $\delta_A \approx \delta_B$. Combined, these two assumptions ensure that an employment quota eliminates any possibility of discrimination in equilibrium.

The result above is subject to an important caveat: even if homogeneous beliefs are achieved, this need not improve the outcomes of any individual. The policy may harm
the majority rather than helping the minority, and can worsen outcomes for both groups. This is easiest to see from an extreme example in which \( \pi_B = \delta_B = 0 \) initially, implying that no minority workers apply. Holding beliefs fixed, the only way for a firm to satisfy an employment quota is to hire zero workers of type A. This is in stark contrast to Coate and Loury (1993), where firms can satisfy the quota by lowering the minority hiring standard.

The same logic applies more generally. Intuitively, employers are constrained in their ability to attract minority workers who are themselves pessimistic about firms. They are therefore forced by the employment quota to aggressively lower their standards, which can severely undermine worker investment incentives. We formalize these intuitions in Proposition 11.

**Proposition 11.** Assume that \( \phi(\theta) \) and \( \tau(\psi) \) are continuous, strictly decreasing and strictly positive on \( [0, 1] \). Further suppose that the A and B markets start with \( \pi_A > \pi_B > 0 \) and \( \delta_A > \delta_B > 0 \). For low enough \( \delta_B \) and \( \pi_B \), imposing an employment quota lowers employment of A workers. Furthermore, there exists an open set of parameters such that the policy leads to zero investment by B workers.

2.4.3 Wage and Employment Subsidies

Another policy proposal put forward in the literature is to subsidize worker wages or employment. Indeed, wage subsidies are highlighted as particularly effective by Coate and Loury (1993). However, not only do these policies fail to eliminate zero investment as an equilibrium in our model, but both can actually be harmful. To see why, suppose a wage subsidy \( s \) is introduced, raising a worker’s positive and negative payoffs to \( s + \omega_q \) and \( s - \omega_u \) respectively. A worker will now apply if and only if:

\[
\xi(\delta_j, \psi) (s + \omega_q) + (1 - \xi(\delta_j, \psi)) (s - \omega_u) > 0.
\]

This subsidy lowers the worker’s application threshold, which can undermine firms’ incentive to invest. While this intuition is general, it can again be seen most clearly from the
extreme: if \( s \geq \omega_u \), the worker will always apply and zero firms will invest. Depending on the subsidy chosen, this policy therefore has the potential to harm its intended beneficiaries. The basic intuition is that the effect of a wage subsidy on worker application behavior can reduce the impact of firm investment on the number of workers that it attracts, lowering firms’ return on investment. An analogous problem occurs if an employment subsidy is provided to firms, which raises both \( \chi_d \) and \( \chi_u \).

2.4.4 An “Impossibility” Result

Given the failure of the specific policies we have considered thus far, we now search more systematically for a simple and reliable way to rule out equilibria with zero investment, and to quickly eliminate discrimination without potential for unintended harm. Since there are two decisions for workers and two for employers, we ask – abstractly at first – which of these margins one should target with policy. The answer is surprisingly definitive. Policy must target at least one of the first-stage investment decisions, and should simultaneously target at least one of the variables in each of the equilibrium equations, 2.5 and 2.6: i.e., it should address both sides of the coordination problem between workers and firms.

**Proposition 12.** Suppose that a government has access to policies that operate by affecting some combination \( C \) of \( s \), \( t \), \( \pi \) and \( \delta \). Regardless of the specific policy, the planner cannot guarantee that a positive fraction of both workers and firms invest unless \( \{ \delta, \pi \} \in C, \{ t, \pi \} \in C \) or \( \{ s, \delta \} \in C \).

An immediate implication of Proposition 12 is that policies that only affect one decision margin will fail to achieve the goals we have set forth. This includes not just affirmative action but also many policies that we have not considered explicitly. The result also provides some guidance for where to look for policy solution, which is a problem we take up in Section 2.5.

---

24 An employment subsidy suffers from this problem in one-sided models but wage subsidies do not (see Coate and Loury, 1993).
Suggestive Evidence of the Efficacy of Two-Sided Interventions

Empirical evidence suggests that two-sided policies are indeed more effective than their one-sided equivalents. We focus on job training programs, of which Job Corps is a canonical example. It is a residential program funded by the Department of Labor but operated mostly by private contractors, and typically lasts around eight months. Participants receive vocational and academic training, counseling, social skills training, health education and job search assistance. There is some limited input from business to incorporate specific proficiencies but little involvement of employers. A large-scale randomized evaluation suggested that Job Corps increased earnings by around four percent, one year after the program, but with little long term impact and no effect on hourly wages (Schochet *et al.*, 2008).

Conversely, *WorkAdvance* programs are narrowly targeted and employers are deeply involved in designing the training. Participants are strongly encouraged to participate in work-based learning with an employer who offers a job, good benefits and the possibility of career advancement. Randomized evaluation suggests that *WorkAdvance* programs increase earnings by 14 percent on average and the effect is mostly driven by higher wages (Hendra *et al.*, 2016).

The success of these programs stands in stark contrast to the average job training program, a result that aligns with the predictions of our model. A trainee who participates *WorkAdvance* has an incentive to invest and gain the skills being offered: not only does this investment lead to being hired, but the worker knows she will be rewarded with a position with an employer who is offering a real opportunity. At the same time, employers can trust that they will receive workers who have both the cognitive and non-cognitive skills that they need to be productive. While we cannot prove that these are the reasons why these programs succeed, there does seem to be a pattern in which programs are most successful (Hossain and Bloom, 2015).
2.5 Extending the Basic Model: Two-Sided Policies

Building on the impossibility result we presented in Proposition 12, the next section considers specific interventions that target both sides of the coordination problem faced by workers and firms in our model.

2.5.1 Two-Sided Investment Insurance

We begin with a new policy: two-sided investment insurance. Specifically, we suppose that the government has access to informative (but possibly imperfect) signals of worker and firm investment, and that it offers incentive payments conditional on these signals. This solution is very effective in our model. We also believe some approximation is likely to be implementable in reality, given policymakers have access to increasingly rich administrative data that could be used to measure both worker qualification and firm investment.

**Proposition 13.** Let \( \{\pi_A, \pi_B, \delta_A, \delta_B\} \) be equilibrium beliefs as defined by equations 2.5 and 2.6, with \( \pi_A > \pi_B, \delta_A > \delta_B, \) and signal thresholds \( \{s_A, s_B, t_A, t_B\} \) defined by equations 2.1 and 2.2. Assume that the government observes informative signals, \( \theta^g \) and \( \psi^g \), of worker and firm investment respectively. There exist group-B-specific incentive payments \( \omega^g \) and \( \chi^g \), which are paid if \( \theta^g \) and \( \psi^g \) exceed fixed thresholds \( s^g \) and \( t^g \), and which ensure equality in investment rates and signal thresholds. If and only if \( \lambda_A \neq \lambda_B \), a non-zero permanent investment subsidy is required to preserve equality.

An especially attractive version of the worker subsidy is feasible if the government observes the same signal as firms: i.e., \( \theta^g = \theta \). In this case, it can set \( s^g = s_A \) and condition the worker payment on rejection by a firm, thereby insuring workers against the possibility that employers are discriminatory. The advantage of this policy is that no worker payments are made by the government, rendering the worker intervention costless. Intuitively, for a worker to receive a government payment, she would have to be rejected by an employer and then be “hired” by the government. But this is impossible if the employer and government signals and signal thresholds are identical. Essentially, a non-discriminatory government’s use of its “market power” in standard setting achieves equality.
If $\theta^g$ is a noisy approximation of $\theta$, the intuition is similar. In this case, we can characterize investment incentives using the conditional distribution of the government signal $\theta^g$. The probability of rejection by the government is $\tilde{F}_i^W(s^g|\theta < s_B)$. If the government sets $s^g = s_A$, the fraction of B workers who invest is as follows.

$$\pi_{B,t} = G^W \left( \beta_W(s^*(\pi_{B,t-1}), \delta_{B,t-1}) + \left[ F_u^W(s_A|\theta < s^*(\pi_{B,t-1})) - F_q^W(s_A|\theta < s^*(\pi_{B,t-1})) \right] \omega^g \right)$$

Just like the policy described in Proposition 13, there is an incentive payment that ensures that $\pi_{B,t} = \pi_A$ for any $\pi_A$. Since this is achieved immediately, the actual cost of the worker payments is as follows.

$$\delta \left[ 1 - F_q(s_A|\theta < s_A) \right] \omega^g + (1 - \delta) \left[ 1 - F_u(s_A|\theta < s_A) \right] \omega^g$$

This expected cost clearly shrinks to zero if the government signal is approximately identical to that of the firm. A small amount of additional noise adds to the cost.

In summary, two-sided investment insurance achieves all of our stated goals. It eliminates equilibria with discriminatory beliefs weakly faster than any other policy, with no potential for negative side effects. Some elements of investment insurance may already be provided via state subsidization of merit-based scholarships, systematic efforts to eliminate discrimination and harassment, and – for women – policies that promote workplace flexibility. But, no policy to date has simply guaranteed a market wage for those who appear to have invested.

*Year Up* – an organization that offers disadvantaged youth a combination of skills training and a six month internship with corporate partners such as JPMorgan, State Street or Google – is an example of an ambitious two-sided investment insurance program. Unlike a traditional job training program, *Year Up* training is targeted to an industry or even a specific employer. Employers are then actively involved in its operation; for instance, some design case studies and conduct mock interviews or customer interactions. A key difference

---

25 A practical implementation of investment insurance would need to determine details such as which signals of investment should be subsidized, but this is beyond the scope of this paper.
between Year Up and other training programs is that they also train these employers on how to best deal with minority youth (i.e., they require firm investment). Upon completing the program satisfactorily, participants are then rewarded with a well-paid internship with genuine opportunities for career advancement and ongoing support from Year Up. They even guarantee an internship for every student: if a student successfully completes the program but does not obtain an internship at the end of their six month training period, Year Up itself will hire them.

This model has been remarkably successful, with a randomized evaluation indicating that treated individuals had 30 percent higher earnings over two years, and that this was mostly driven by higher wages (Roder and Elliott, 2014).

2.5.2 Affirmative Action

Proposition 12 suggests that an alternative policy is to simultaneously target both the investment and hiring decisions of firms. One way to implement this is to combine affirmative action on the ‘extensive’ and ‘intensive’ margins. Essentially, firms would be encouraged to invest in amenities for the minority group (i.e., one-sided investment insurance) and change their hiring practices (i.e., affirmative action). If the gap between A and B workers is small, one-sided investment insurance alone can be effective, but it would have to be accompanied by affirmative action if B workers begin with zero investment. Intuitively, affirmative action ensures that at least some minorities are hired, and investment incentives ensure that workplaces are attractive for minority workers so that they will apply.

Note 2. A one-sided investment subsidy can eliminate discrimination after a one period delay if: (i) $s_B < 1$; (ii) $s_B \approx s_A$; and (iii) $\delta_A$ is small. If $s_B = 1$ but the remaining conditions hold, a one-sided subsidy can eliminate discrimination after one period if combined with affirmative action (equality in offers).

A larger gap between groups limits the effectiveness of a one-sided investment subsidy. Even if combined with affirmative action, it may be impossible to achieve equality in finite
time. There are two reasons for this. First, to be successful, the subsidy needs to raise firm investment in the B amenity above the rate of investment in the A amenity. There is little scope to do this if nearly all firms are already making the A investment. Second, a harsh hiring standard for B workers limits the impact on worker returns of higher firm investment. These concerns imply that “two-sided affirmative action” is unlikely to be a policy that could eliminate severe inequalities. Two-sided investment insurance would be more robust.

**Proposition 14.** Assume that \( \phi(\theta) \) and \( \tau(\psi) \) are continuous, strictly decreasing and strictly positive on [0, 1]. For any \( \pi_B \in (0, 1) \) and \( \pi_A \in (0, 1) \) with \( \pi_B < \pi_A \), there exist cost distributions \( G^W \) and \( G^E \), a signal distribution \( F^W_{i}(\theta) \) and parameters such that: (i) \( \pi_B \) and \( \pi_A \) are part of an equilibrium; and (ii) no one-sided investment subsidy can raise \( \pi_B \) to \( \pi_A \) in any finite number of periods \( T \), even if combined with affirmative action.

### 2.6 Interpreting Group Differences in the Presence of Two-Sided Statistical Discrimination

Two-sided statistical discrimination complicates empirical analysis, since differences between groups are generically a combination of both employer and worker decision-making. For example, consider a setting with employer learning as in Altonji and Pierret (2001). Under conditions they outline, the conditional expectation for log-wages can be written as a time-varying function of the form:

\[
E(w_t|s_i, z_i, t) = b_{s,t}s_i + b_{z,t}z_i + H(t)
\]

where: \( s_i \) is observable to the employer and econometrician; \( z_i \) is observable to econometrician but not initially observed (or at least not used) by the employer; \( t \) denotes experience; and \( H(t) \) is an experience profile of productivity that does not depend on \( s_i \) and \( z_i \).

Altonji and Pierret (2001) show that, if their assumptions hold, \( b_{s,t} \) falls with experience, while \( b_{z,t} \) rises. Their empirical results show that the racial wage gap rises with experience,
which suggests that employers do not fully incorporate racial differences in productivity into their initial wage offer. The conclusion of their analysis is that statistical discrimination cannot explain racial differences in wage profiles.

As Altonji and Pierret (2001) acknowledge, the assumption that the experience profile of productivity is independent of race is restrictive; this motivates them to test for racial differences in training opportunities. Such a difference could arise from employer discrimination but could also be driven by worker expectations. For example, suppose that workers make costly investments in human capital. The return on investment depends on whether higher productivity will be rewarded by employers. Early in a worker’s career, the forces in our model would predict that pessimistic black workers would invest less. This would cause the racial wage gap to widen with experience in the early years, an effect that would be sharpened by the fact that black workers often have shorter tenure. However, information about a specific employer should eventually overwhelm any racial difference in priors. If investment returns are diminishing, we would expect to see convergence between the wages of black and white workers at “good” firms after many years of tenure. This aligns with the results found by Fryer et al. (2013), who demonstrate that racial wage gaps widen with experience but narrow after many years of tenure within the same firm.

An implication of a model in which ongoing investments depend on beliefs or otherwise depend on race is that empirical analysis designed to detect statistical discrimination may be misleading. Even if race is a variable – i.e., employers statistically discriminate – underinvestment by black workers due to their own pessimism would lead to the false conclusion that employers do not discriminate. The same facts could be explained by a model in which training opportunities depend on employer beliefs (see Altonji and Pierret, 2001). The two reasons for underinvestment are thus inseparable with this approach.

Lang and Lehmann (2012) discuss a test that is robust to differing experience profiles of black and white workers. Let \( B_i \) indicate whether a worker is black. As before, \( z_i \) is correlated with productivity and initially unobserved by the employer. In a simplified
model, Lang and Lehmann propose comparing two regressions.

\[ E^* (w_t | B_i, z_i, t) = a_1 + a_2 B_i + a_3 t + a_4 B_i t + a_5 z_i \]

\[ E^* (w_t | B_i, z_i, t) = \beta_1 + \beta_2 B_i + \beta_3 t + \beta_4 B_i t + \beta_5 z_i + \beta_6 z_i t \]

Since employers gradually learn about \( z_i \), low \( z_i \) workers would initially be overpaid but their wages would converge to their productivity over time. If black workers have lower \( z_i \) on average and employers statistically discriminate, we would expect \( \gamma_4 < 0 \) and \( \gamma_2 > 0 \) in the following auxiliary regression.

\[ E^* (z_t | B_i, z_i, t) = \gamma_1 + \gamma_2 B_i + \gamma_3 t + \gamma_4 B_i t + \gamma_5 z_i \]

Assuming that the weight on \( z_i \) increases over time as predicted by employer learning (i.e., \( \beta_6 > 0 \)), this implies that \( a_2 < \beta_2 \) and \( a_4 > \beta_4 \), which is precisely what Altonji and Pierret (2001) find using the Armed Forces Qualification Test (AFQT) as \( z_i \). As Lang and Lehmann (2012) argue, these results therefore suggest a model in which black-white productivity differences widen over time and employers statistically discriminate.

### 2.6.1 Detecting Employer Discrimination

Even in the presence of complementarity, one can make progress identifying employer discrimination. The approach we suggest is to focus directly on the mechanism through which statistical discrimination affects incentives: pessimistic employer beliefs lower the return to investment for blacks relative to whites. Specifically, we propose a test of whether there is a racial difference in the degree to which imperfect information lowers the return to improving one’s own productivity.

Consider the following highly stylized thought experiment. Statistical discrimination should imply that a group \( j \) worker who is 10 percent more productive would be paid \( \beta_j \leq 10 \) percent more because employers shrink their estimates of productivity toward the mean of the group. If statistical discrimination causes \( \beta_B \) to be lower than \( \beta_W \) then investment is undermined for blacks relative to whites. The statistical discrimination
literature suggests two reasons why we might expect this to be true.

1. Productivity may be harder to assess for minority workers.

2. Lower investment returns should compress the productivity distribution for blacks.

This implies that priors – if correct – are tighter and that lower returns are self-fulfilling.

Although the latter effect is what we focus on in our model, we do not attempt to provide a way of empirically distinguishing the two competing reasons for statistical discrimination.

The test that we derive involves measuring the relationship between a worker’s past wage and her current wage, and comparing the coefficient for black and white workers of similar tenure at their previous firms. The intuition is that if past wages better reflect productivity, and productivity is imperfectly observed at the time of hiring by a new firm, then black workers’ past wages should be less predictive of current wages. Models of taste-based discrimination do not share this prediction.

To facilitate empirical analysis, we adapt our model to allow for continuous investment and adopt the assumption that workers are paid their expected marginal product. To begin, assume that output is produced at constant returns to scale using a mass of quality-adjusted labor $Q_j$ and group-specific ‘capital’ $K_j$ provided by the firm. Each worker provides a unit of physical labor but individuals are heterogenous in their ability $a_i$. Effective labor is $Q_j = L_j \cdot \bar{a}_j$ where $L_j$ is the aggregate amount of physical labor supplied by group $j$ workers, and $\bar{a}_j$ is the average productivity of those workers. For convenience we adopt a Cobb-Douglas specification for production: $Y_j = K_j^{1-\gamma} Q_j^\gamma$. Thus, letting $k_j = K_j/\bar{a}_j L_j$ be the amount of group-specific capital provided by the firm per unit of effective labor, the marginal product of a worker with ability $a_i$ at firm $j$ is $MP_i = a_i \gamma k_j^{1-\gamma}$. A given firm may provide different levels of $k_j$ for members of each group.

For tractability, we assume that worker ability is distributed log-normally: $\ln a_i \sim N\left(\mu_{a_i,j}, \sigma_{a_i,j}^2\right)$. Firms receive a noisy but unbiased signal $\theta_i$ about each worker’s productivity. Specifically, $\ln \theta_i = \ln a_i + \ln \epsilon_i$ where $\ln \epsilon_i \sim N\left(0, \sigma_{\epsilon_i,j}^2\right)$. If workers are paid their expected

---

26 The logic underlying this analysis here is similar to Kahn’s (2013) test for asymmetric employer information.
marginal product, the wage paid by a firm to a worker with ability \( a_i \) at a firm with \( k_{j,F} \) can be shown to be as follows.

\[
\ln w_i = \left( \frac{\sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \ln a_i + \left( \frac{\sigma_{e,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \mu_{a,j} + \frac{1}{2} \left( \frac{\sigma_{e,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \sigma_{a,j}^2 + \left( \frac{\sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \ln \epsilon_i \\
+ \ln \gamma + (1 - \gamma) \ln k_{j,F}
\]

Purely for pedagogical purposes, we now temporarily adopt a strong assumption about employer learning, which we will subsequently relax.

**Assumption 4.** For a worker of long enough tenure at her previous employer, her past wage exactly reflects her ability at that firm and is not observable to a new firm. Ability at the old and new firms are equivalent.

This is restrictive for two reasons. First, we are assuming that learning is complete with long enough tenure. Secondly, we are assuming that a worker’s ability at a new firm is equivalent to her ability at her old firm. These assumptions may both be reasonable in some contexts, but it is easy to imagine violations.

The assumption above allows us to write the wage \((w_i)\) offered to an experienced worker who moves to a new firm as a particularly simple function of her wage at her previous firm \((w_i^{OLD})\), group-specific fixed effects for the source and destination firms \((\alpha_{j,fNEW} \text{ and } \alpha_{j,fOLD})\), and an error term \(v_i\):

\[
\ln (w_i) = \beta_j \ln \left( w_i^{OLD} \right) + \alpha_{j,fNEW} + \alpha_{j,fOLD} + v_i
\]

(2.13)

where \(\beta_j = \frac{\sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2}\), \(v_i = \left( \frac{\sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \ln \epsilon_i\) and the fixed effects are functions of the model’s parameters.

\[
\alpha_{j,fNEW} = \left( \frac{\sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \mu_{a,j} + (1 - \gamma) k_{j,FNEW} + \frac{1}{2} \left( \frac{\sigma_{a,j}^2 \sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \\
\alpha_{j,fOLD} = -(1 - \gamma) \ln k_{j,FOLD}
\]
The coefficient on a worker’s previous wage, $\beta_j$, is the elasticity of the wage with respect to ability. This is a measure of the return to productivity-enhancing investment, with $\beta_j = 1$ corresponding to a worker always receiving her marginal product. The two fixed effect terms allow for potentially different levels of group-specific capital at the new and old firms.

To assess the impact of statistical discrimination, we propose a test of whether the return to ability is lower for blacks. Since $\beta_j$ is exactly the degree to which statistical discrimination lowers the return to productivity, this amounts to the following statistical test.

$$H_0: \Gamma = \beta_W - \beta_B \leq 0$$
$$H_1: \Gamma = \beta_W - \beta_B > 0$$

With data on past and present wages, and adequate movement between firms, it is straightforward to estimate equation (2.13) for each group and calculate an estimate of the racial difference in returns $\hat{\Gamma} = \hat{\beta}_W - \hat{\beta}_B$.

A potential complication for the attainment of a consistent estimate of $\Gamma$ in a regression is that movement between firms may be non-random. Yet, selective movement does not necessarily affect the estimated relationship between current and past wages ($\hat{\beta}_j$), conditional on including firm fixed effects. For example, there may be correlation between the investments made by a worker’s current and previous firm but this is accounted for by including fixed effects for both firms. Alternatively, idiosyncratic match effects or a connection between firm-wide shocks and mobility would bias estimates of the firm fixed effects themselves (see Card et al., 2016) but do not necessarily affect $\hat{\Gamma}$.

We next relax the assumption that past wages fully reflect productivity. First, we allow for the possibility that the previous employer also has imperfect information about a worker’s ability. However, we continue to assume that this information is better than the new firm in the sense that $\sigma_{e_{ij},\text{OLD}}^2 < \sigma_{e_{ij}',\text{NEW}}^2$ since some private learning has occurred over the worker’s tenure. Second, we allow for the possibility that ability at the new and old firms are correlated but not equivalent.

Under these alternative assumptions, we argue that our proposed test for statistical dis-
criterion is conservative in the sense that differences in returns to ability are understated. Specifically, we argue that \( \hat{G} \) is a downward-biased measure of the difference in returns to ability induced by statistical discrimination.

**Proposition 15.** Assume that ability at the new and old firms are correlated: \( \ln a_i = c_j + \rho \ln a_{i,OLD} + \ln \eta_i \), where \( 0 < \rho \leq 1 \). Then the difference in coefficients from equation (2.13) is \( \hat{G} \), where:

\[
\hat{G} = \rho \left[ \frac{\sigma_{a,W}^2}{\sigma_{a,W}^2 + \sigma_{a,W}^2} - \frac{\sigma_{a,B}^2}{\sigma_{a,B}^2 + \sigma_{a,B}^2} + \frac{\sigma_{a,W,OLD}^2}{\sigma_{a,W,OLD}^2 + \sigma_{a,B}^2} - \frac{\sigma_{a,B,OLD}^2}{\sigma_{a,B,OLD}^2 + \sigma_{a,B}^2} \right].
\]

It is evident from Proposition 15 that imperfect correlation between ability at the old and new firms \( (0 < \rho < 1) \) biases \( \hat{G} \) toward zero. Second, any statistical discrimination against black relative to white workers reduces the return to investment, compressing the productivity distribution and leading to a tighter employer prior \( (\sigma_{a,B}^2 \leq \sigma_{a,W}^2) \). This further pushes toward a downward-biased estimate of \( \Gamma \). Finally, as long as workers of similar tenure are compared across races, we would expect a similar amount of learning to have occurred so that \( \frac{\sigma_{a,B}^2}{\sigma_{a,B,OLD}^2} \approx \frac{\sigma_{a,W}^2}{\sigma_{a,W,OLD}^2} \). This is enough to conclude that our proposed test remains a conservative measure of the disparate impact of statistical discrimination.

### 2.7 Conclusion

Statistical discrimination is a foundational concept in the economic analysis of discrimination. Intuitively, the information problem inherent in such models seems two-sided. Yet, current models do not take this into account. Two-sided belief formation makes the interpretation of any empirical data on group differences more complicated – though not impossible – because for any disparity, differences can be driven by either side of the market: workers or firms; universities or applicants; police or civilians.

Furthermore, policies designed to break equilibria with negative beliefs about certain groups are complicated by complementarity between the beliefs and actions of workers...
and firms. Affirmative action, employment quotas, wage subsidies, and unemployment insurance all perform poorly relative to traditional statistical discrimination models. Indeed, we demonstrate that any one-sided policy fails to reliably ensure homogeneous beliefs.

We posit a new policy – two-sided investment insurance – as a solution to statistical discrimination. Investment insurance is a method for the government or another entity to guarantee returns for workers it deems as investors, while rewarding firms for making their workplaces productive for all types. We demonstrate that this policy can immediately improve outcomes for a disadvantaged group without any risk of unintended harm. This stands in stark contrast to the traditional “one-sided” policies that we consider. Year Up is a strikingly successful example of this type of opportunity for urban youth. Similar policies might be envisioned for broader classes of workers.
Chapter 3

The Impact of Student Suspension on Human Capital Accumulation

3.1 Introduction

School discipline policy is highly controversial among education policy-makers, in part because there is limited evidence on the effectiveness of alternative approaches. Punitive policies – typically relying heavily on suspensions – could help deter bad behavior and maintain a peaceful learning environment for students. Indeed, “no excuses” charter schools, which have been shown to raise their students test scores (Angrist et al., 2013), attribute their success in part to harsh disciplinary regimes. However, punishing vulnerable students by removing them from the classroom may reduce their opportunity to learn or disconnect them from the school system and their peers. Some authors have therefore gone as far as arguing that disproportionately harsh discipline for black students explains a large fraction of the racial gap in academic achievement (Morris and Perry, 2016).

We contribute to this debate by studying the causal impact of suspensions on students’ test scores using administrative data from public middle school students in New York City. First, we document the negative unconditional relationship between suspensions and test

---

1Co-authored with David Martin

95
scores. Second, we present a fixed effects strategy, which shows that much of this negative association can be attributed to variation across schools and fixed differences between students. Nonetheless, there remains a negative conditional relationship after controlling for school-grade and individual fixed effects. Finally, we harness quasi-experimental variation due to the elimination of suspensions for low-level behavior in 2012. Our results using this natural experiment suggest that the abolition of suspensions was beneficial, on average, for students in the school-grades that were affected by them. This suggests that the recent focus by policymakers on alternative approaches to school discipline may indeed be justified.

There is a powerful negative unconditional association in our data between suspension rates and test scores. School-grades with a one percentage point higher suspension rate (nine percent of one standard deviation) have test scores that are 0.64 of a standard deviation lower. This association explains around 2.5 percent of the population variance in test scores in our sample, and around 4.5 percent of the gap in test scores between black and white students. However, the implications of these facts depend critically on the extent to which this empirical relationship reflects causal impacts of different discipline policies.

A significant fraction of this negative relationship can be explained by individual and school-grade fixed effects, which absorb variation across school-grades and differences on average between students. Nonetheless, the conditional relationship between suspensions and test scores remains negative, although it is weaker than the unconditional relationship. This implies that students who are suspended have lower test scores at the time of their suspension than would be expected given their own history and the school grade they attend. Our results here mirror those of Lacoe and Steinberg (2019) who use a similar approach with two years of data from Philadelphia. Arguably, these fixed effects estimates come closer to measuring the causal effect of a student being suspended. However, it remains possible that students who violate the discipline code have been hit by a negative shock, which causes them both to misbehave and to perform poorly on exams.

Our final step is to analyze the impact of a change in discipline policy in 2012 by the New

---

2 The suspension rate includes in-school and out-of-school suspensions. See Section 3.3 for further details.
York City Department of Education, which eliminated suspensions for low-level infractions. Our results suggest significant gains from this reform. Specifically, we show that test scores in school-grades that were affected – i.e., those in which teachers and administrators had previously used the class of suspensions that was eliminated – rose markedly in subsequent years relative to those in which such suspensions had not previously been used. Although we caution against interpreting this result as necessarily arising solely from the elimination of suspensions for students who would have been suspended, it does suggest that policymakers’ focus on alternative approaches to discipline may be warranted.

In addition to providing input into to an active policy debate surrounding school discipline, our results contribute toward a better understanding of the human capital production function more generally. While it is well-established that test scores are causally improved by some schools (Angrist et al., 2013) and by better teachers (Chetty et al., 2011), it is much less clear what specific policies and methods are important in driving these gains. Our results suggest that the implementation of harsher discipline alone does not help students in traditional public schools, at least in New York City.

### 3.2 Suspensions in the Literature and in Practice

Most school districts in the United States use suspension as a way of punishing students for misbehavior. A student who receives an out-of-school suspension is physically separated from the school for a predetermined period of time. In contrast, in-school suspensions remove students from the classroom but not from the school. Unless indicated otherwise, we use the term “suspension” to refer to both types throughout the paper. The U.S. Department of Education estimates that over 2.6 million students nationwide received at least one out-of-school suspension in the 2013-2014 school year, with 570,000 students receiving an in-school suspension (United States Department of Education, 2018). Black students were suspended at a far higher rate than white students.

It has been widely documented that there is a powerful negative correlation between suspensions and student outcomes (Morris and Perry, 2016), but there remains controversy
about whether suspensions cause lower test scores. This is a challenging question to resolve non-experimentally, since many of the same factors that put students at risk of suspension – e.g., difficult home environments or behavior regulation issues – are likely to contribute to poor performance on exams. Even the sign of the causal effect of suspension on average student outcomes is theoretically ambiguous. Suspensions may deter bad behavior and improve learning outcomes, including those of classmates who would never have been suspended. However, suspensions may negatively affect students who are punished by reducing their opportunities to learn, or by disconnecting them from the school system.

3.2.1 Effects on Individual Outcomes

The most salient direct cost of suspensions is missed instructional time for suspended students. To the extent that school attendance builds human capital, a physical presence in the classroom is likely necessary to reap these benefits. In fact, the broader literature on school absences finds that failing to attend school has a large enough effect on test scores to explain around one fourth of the academic achievement gap by income (Goodman, 2014).³

Spending significant time separated from his school and classmates may also weaken a student’s commitment to education. Specifically, suspension (or the expectation of future suspension) may reduce the marginal benefit to costly attendance or effort, or raise the marginal benefit to investment in undesirable substitute activities such as membership in a gang. This could explain why frequently-suspended students drop out at greater rates than their peers (see Skiba, 1997), although the direction of causality is again unclear. Even conditional on staying in school, subsequent effort may be lower.

On the other hand, one of the purposes of a strict discipline policy is to deter misbehavior à la Becker (1968). If students dislike suspensions, either because they internalize the human capital costs or because they face external pressure from parents or stigma from peers, they will be less likely to misbehave in the first place. In fact, if the threat of punishment

³Many school districts, including New York City, provide some sort of alternative instruction for suspended students, but in practice this is likely to be an imperfect substitute for their regular education.
is sufficiently severe, students’ attention and effort may increase enough to raise their performance. This is the result found by Kinsler (2013) in North Carolina using a structural model that takes into account spillovers onto other students from bad behavior.

It is difficult to estimate the causal effect of suspensions on individual outcomes non-experimentally, and existing attempts reach conflicting conclusions. For example, Lacoe and Steinberg (2019) find a negative net effect on test scores but no effect on subsequent absences among students, using two years of data covering students in Philadelphia Public Schools. Other papers find no evidence of a causal effect (Cobb-Clark et al., 2015) or even a positive effect (Kinsler, 2013). Our estimates indicate that banning suspensions for low-level infractions led to significant gains for the affected students in New York City.

3.2.2 Spillovers to Other Students

A key purpose of strict suspension policy is to remove disruptive influences from the classroom, and thereby facilitate the learning of other students. Such spillovers are potentially large: in a series of papers using domestic violence in the home as an instrument for disruptive behavior, Carrell and Hoekstra find that exposure to disruptive peers reduces both short-term academic achievement (Carrell and Hoekstra, 2010) and long-term earnings (Carrell et al., 2018). In fact, a single disruptive peer in a class of 25 students reduces the present discounted value of future earnings by $80,000 (Carrell et al., 2018).

Spillovers from suspensions are difficult to identify separately from the direct effect of suspensions of student outcomes. Nonetheless, Lacoe and Steinberg (2019) provide some evidence that higher suspension rates lead to lower test scores for non-suspended students. Moreover, Kinsler’s results suggest positive spillovers that so large that they more than offset the negative direct effect of suspension on students who are suspended.

3.2.3 Racial Discipline Gap

Suspension is far more common for minority students. Across the United States, black students comprised 15.5 percent of total enrollment but 40.3 percent of students who received
an out-of-school suspension in 2013-2014 (United States Department of Education, 2018). In our sample, the unconditional odds of receiving any type of suspension are three times higher for black than white students in middle or high school. Moreover, there is evidence elsewhere to suggest that black students receive more severe punishments conditional on the same behavior (McFadden and Marsh, 1992; Rocque, 2010; Kinsler, 2011).

The observed disparities in suspension rates raise the question of whether the ‘racial discipline gap’ can explain part of the ‘racial achievement gap’. Indeed, Morris and Perry (2016) suggest that differences in suspension rates can account for up to one fifth of the gap in test scores between black and white students. On the other hand, Kinsler (2013) argues that the racial discipline gap actually narrows the racial achievement gap. Concerns about the racial incidence of suspensions was an important part of the stated motivation for the 2012 reform that we use for our quasi-experimental analysis that we describe in Section 3.4. In our data from New York City, however, conditioning on indicators of whether a student is suspended reduces the black-white gap in test scores by less than 5 percent.

3.2.4 Zero Tolerance Policies

Variation in the philosophies of policy makers has led to substantial changes over time in the harshness of school discipline. From the 1980s to the 2000s, discipline policies became stricter, as districts enacted “zero-tolerance policies” (Hanson, 2005). Under these policies, even minor infractions were punished with a suspension. For example, in New York City, the number of suspensions per student in grades 6-8 rose by 160 percent from 2001 to 2008. The proponents of zero tolerance policies argued that harsh discipline would change the cultures of under-performing schools and improve student performance.

These policies have since been partially reversed, as school districts and policymakers

4This may understate the racial discipline gap if it is also more common for black than white students to receive multiple suspensions.

5As Kinsler (2011) shows, this need not constitute evidence of discrimination.

6This statistic excludes charter and special-ed-only schools. See Section 3.3.3 for other sample restrictions.
more generally have moved away from zero tolerance approaches. In part, this was due to concerns about the racial incidence of student suspensions, and suspicion that higher suspension rates were counterproductive. Indeed, these concerns were supported by the work of Skiba (1997), Morris and Perry (2016) and others. Evidence on the effect of the movement toward more relaxed student discipline is limited, but existing evidence suggests that student behavior and performance may have worsened (Lacoe and Steinberg, 2018). Our evidence suggests the opposite for New York City.

3.3 Institutional Context and Data

Throughout our analysis, we use data from the New York City Department of Education (NYCDOE) on all students in the New York City public school system. This is the largest public school district in the United States, with over one million students and 1,800 schools.

3.3.1 Discipline in NYC Schools

The disciplinary code in the New York City public school system is laid out by the Citywide Standards of Intervention and Discipline Measures, which is published each year by the NYCDOE. The code enumerates five levels of disciplinary infraction based on their severity. Level 1 infractions are the least severe, covering various types of non-compliance such as being late for class, making excessive noise, or engaging in verbally rude or disrespectful behavior. At the opposite extreme, Level 5 infractions include the most severe behavior, which is generally dangerous or violent. Although these levels are associated with infractions, we occasionally refer to a suspension for Level x behavior as a “Level x suspension”.

The discipline code prescribes progressively more punitive interventions for each infraction level. As a first step, schools are encouraged to use guidance interventions such as outreach to parents or intervention by school counseling staff. If guidance interventions are insufficient, disciplinary measures may be taken; this may include exclusion from extracurricular activities, suspension (in-school or out-of-school), or expulsion. During an in-school suspension, the offending student is sent to a different location within the school.
to receive continued educational services and classwork. For an out-of-school suspension, the offending student is not allowed on school property for a predetermined number of days, but completes work at another school or an alternative learning center.

**Figure 3.1:** This figure provides a simplified outline of the set of decision points that lead to a suspension.

In practice, teachers and school administrators exercise considerable discretion. Figure 3.1 provides a simplified outline of the decision points that lead to a suspension. When misbehavior occurs, teachers first decide whether to handle the incident internally or to escalate it to the Principal. School administrators may then assign a “Principal’s Suspension”, which lasts for up to five days. For more serious infractions (Level 3 and above), school administrators can escalate the incident to the Office of Safety and Youth Development. The student may then receive a “Superintendent’s Suspension” of 6 or more days.\(^7\)

### 3.3.2 Evolution of Disciplinary Policy

Table 3.1 shows the recent history of school discipline reform in New York City. Prior to 1998, school safety fell under the purview of the New York City Board of Education.

---

\(^7\)This is a simplified statement of the suspension process. For example, a parent conference occurs before a Principal’s suspension, and a hearing is held before a Superintendent’s suspension. Also note that school administrators can choose from a range of guidance or disciplinary interventions for any particular incident.
(NYCBOE), which preceded the current Department of Education. Echoing Mayor Rudy Giuliani’s more general “tough on crime” message, the BOE voted to transfer responsibility for school safety to the New York Police Department (NYPD) in September 1998. This marked the beginning of a period of increasingly aggressive discipline policy, which was continued by Giuliani’s successor, Mayor Michael Bloomberg. In 2004, Bloomberg enacted the Impact Schools Initiative, which doubled police presence at targeted high-crime schools and called for a zero tolerance policy towards disciplinary infractions. Then, in 2006, he announced a roving metal detector program at all middle and high schools. By 2008, the number of suspensions used per year had peaked at nearly 74,000 (see Figure 3.2).

Figure 3.2: This figure shows the long-term trend in the total number of in-school and out-of-school suspensions used per year in the New York City public school system, including all suspensions in all grades. Data for this figure come from reports by the New York Civil Liberties Union (NYCLU).

By the 2010s, the NYCDOE began taking steps to ease its reliance on punitive discipline in response to public pressure. Social justice groups such as the American Civil Liberties Union
Table 3.1: Timeline of School Discipline Reform in New York City

1998  Responsibility for discipline transferred to the New York Police Department
2004  Zero tolerance policy introduced, and police presence in schools doubled
2006  Roving metal detector program instituted
2012  No suspensions for Level 2 offenses, plus shorter Level 3 suspensions for K-5 students
2015  Additional case-based review required for some types of Level 3 suspensions, and suspensions removed for some Level 3 offenses

Table notes. This table shows a timeline of key events in the recent history of school discipline reform in the New York City public school system. Our quasi-experimental analysis focuses on the reform in 2012 (see Section 3.4).

(ACLU) lobbied vigorously for discipline reform. Working under the implicit assumption that high suspension rates are causally harmful, these groups argued that the burden of the existing policies fell disproportionately on minority students. Our data confirm that the rate of suspension is higher for minority students: between 2005 and 2016, black students made up less than 30 percent of enrollment but over 45 percent of suspended students (see Figure 3.3). In middle school and high school, the unconditional odds of being suspended were nearly three times higher for black students than for white students.

In 2012, the number of suspensions per year dropped sharply because the NYCDOE prohibited suspensions for Level 2 infractions. The suspensions banned as part of this reform were in-school, and included those for cutting class, the use of obscene language or gestures, and persistent low-level (Level 1) behavior. In the 2011-2012 school year, 8.6 percent of suspensions in grades 6-12 were for Level 2 infractions. But in 2012-2013, the Level 2 suspension rate dropped to zero (see Figure 3.4). This sharp change in suspension rates provides quasi-experimental variation that allows us to study the causal impact of higher suspension rates. We discuss our quasi-experimental analysis in Section 3.4.

---

8Suspensions had already been prohibited for Level 1 infractions.
9The revised disciplinary code also reduced the maximum suspension length for Level 3 infractions from 10 days to 5 days for K-5 students.
10There was also a temporary fall in suspensions for Level 3 behavior in 2012 (see Appendix C.3).
Figure 3.3: This figure shows the proportion of students who are black, among students who are suspended for an infraction at each level, and across all students.

3.3.3 Data

Our data are comprised of NYCDOE administrative records from 2006-2015, which include demographic information, attendance, disciplinary events, test scores in mathematics and reading for grades 3 through 8, and subject-specific test scores for higher grades.

Suspensions

Each suspension in the DOE reporting system provides a student identifier, the start and end date of the suspension, the duration of the suspension, an alphanumeric code corresponding to a particular infraction in the discipline code, and a school identifier. Many disciplinary events involve multiple infractions, so we can see other infraction codes attributed to a given suspension. Table 3.2 lists the most common infractions that led to a suspension.

Test scores

Each year, students in grades 3-8 take exams in “English / Language Arts” (ELA) and Math. We standardize test scores within each grade-subject for each year to facilitate
Figure 3.4: This figure shows the long-term trend in the number of suspensions for Level 2 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system. Similar figures for other suspension levels are available in Appendix C.3.

comparisons across tests of varying difficulty. Throughout the paper, we consider two methods of standardization: percentiles within a grade-subject-year cell, and z-scores which are normalized to have a zero mean and unit standard deviation.

Student characteristics

We observe demographic characteristics including race and sex. We also see an indicator for whether the student has learned English as a second language (ELL), or has an Individual Education Plan due to special needs. As is standard in the education literature, we use an indicator for eligibility for free or reduced price lunch as a proxy for family income.

Sample

Over our sample period, enrollment in charter schools increased from just over 15,000 students to almost 100,000 students. However, since charter schools have substantially more autonomy than traditional public schools in setting both their discipline policies
Table 3.2: Most Common Infractions Leading to Suspension

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>91277</td>
<td>Dangerous physical behavior</td>
</tr>
<tr>
<td>3</td>
<td>53156</td>
<td>Insubordination</td>
</tr>
<tr>
<td>3</td>
<td>25954</td>
<td>Shoving, pushing or other similar behavior</td>
</tr>
<tr>
<td>2</td>
<td>17933</td>
<td>Persistent low-level (Level 1) bad behavior</td>
</tr>
<tr>
<td>4</td>
<td>17820</td>
<td>Intimidating or bullying behavior</td>
</tr>
<tr>
<td>3</td>
<td>14828</td>
<td>Leave class or school without permission</td>
</tr>
<tr>
<td>4</td>
<td>14061</td>
<td>Serious injury through reckless behavior</td>
</tr>
</tbody>
</table>

Table notes. This table shows the most common infractions that lead to any type of suspension for all students in grades 6 to 12 between 2006 and 2011.

and curriculum, the NYCDOE does not collect disciplinary records for these schools. We therefore omit charter school students from the analysis. In addition, we omit all home-schooled students and students in special-education-only schools.

We focus on middle school students in grades 6 to 8. Although standardized test data are available for all students from grades 3 to 8, students in earlier grades are subject to a different disciplinary regime. Moreover, these younger students were affected by a contemporaneous policy change in 2012 that reduced the length of suspensions for Level 3 behavior from 10 days to 5. Students in grades 6 to 8 were unaffected by this other reform.

Throughout the paper, we use all students who satisfy these restrictions. However, in Appendix C.1, we replicate our analysis without students who are accelerated in mathematics, and who therefore go on to complete the New York City regents exam in Math one year early (in grade 8). We do this to rule out any possibility that our results are driven by a waiver that was received by the NYCDOE from the U.S. Department of Education in 2013 to avoid ‘double testing’. Our results are qualitatively unchanged with this restriction.

Descriptive statistics

Table 3.3 shows summary statistics for our sample of middle-school students (grades 6-8), broken down by sex and race. As a share of all enrollment, 15 percent of students are white, while about 30 percent are black and 40 percent hispanic. Although 80 percent of all middle school students are eligible for free or reduced-price lunch, poverty rates are more than
30 percent higher among black or hispanic students than among white students. We also see evidence of academic achievement gaps by race and by sex. The unconditional gaps between the test scores of white and black students are $0.69\sigma$ in ELA and $0.79\sigma$ in math; and girls score $0.27\sigma$ better than boys on ELA exams and $0.086\sigma$ better in math.

### Table 3.3: Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>All Students</th>
<th>Boys</th>
<th>Girls</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Share of Enrollment</td>
<td>1.00</td>
<td>0.51</td>
<td>0.49</td>
<td>0.14</td>
<td>0.29</td>
<td>0.40</td>
</tr>
<tr>
<td>Free/Reduced Price Lunch</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.51</td>
<td>0.84</td>
<td>0.89</td>
</tr>
<tr>
<td>Test Scores (ELA)</td>
<td>0.01</td>
<td>−0.12</td>
<td>0.15</td>
<td>0.51</td>
<td>−0.17</td>
<td>−0.22</td>
</tr>
<tr>
<td>Test Scores (Math)</td>
<td>0.00</td>
<td>−0.04</td>
<td>0.05</td>
<td>0.48</td>
<td>−0.31</td>
<td>−0.22</td>
</tr>
<tr>
<td>Suspension Rate</td>
<td>0.10</td>
<td>0.14</td>
<td>0.06</td>
<td>0.06</td>
<td>0.17</td>
<td>0.10</td>
</tr>
<tr>
<td>Suspension Rate - Level 2</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Suspension Rate - Level 3</td>
<td>0.03</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>Suspension Rate - Level 4</td>
<td>0.06</td>
<td>0.08</td>
<td>0.04</td>
<td>0.03</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>Suspension Rate - Level 5</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*Table notes.* This table shows summary statistics for our sample, which include students in grades 6-8. Test scores standardized within subject-grade-year cells. Level 2 suspension rates are those corresponded to years prior to the 2012 policy change. Data are from the New York City Department of Education.

Overall suspension rates, defined as the number of suspensions divided by enrollment, are about 10 percent. The largest share of suspensions are for Level 4 infractions, for aggressive or injurious behavior. Suspensions for Level 2 infractions, which provide us with our quasi-experimental variation, make up five percent of the total prior to 2012. Like test scores, suspension rates vary dramatically by sex and race. Boys are suspended twice as often as girls (14 percent, compared to 6.5 percent); and black students are suspended three times as often as white students (17.4 percent, compared to to 5.1 percent).

### 3.4 Empirical Strategy

Figure 3.5 shows that there is a powerful negative relationship between suspension rates and test scores in New York City. For both Math and ELA, suspension rates fall as exam performance rises: students with a one percentile higher Math score are 0.26 percent less
Figure 3.5: This figure shows the relationship between suspension rates and test scores in our sample of middle school students (grades 6-8). Students are grouped into bins based on percentiles of their math and ELA exam scores. The figure shows suspension rates in each bin plotted against the average in each bin. Graphs of suspension rates versus standardized scores are available in Appendix C.3.
likely to be suspended, with a marginally weaker relationship between suspension and ELA scores. This relationship is consistent with suspensions having a negative causal effect on test scores. However, it is also consistent with students who are disadvantaged or less able performing worse on standardized tests, and also behaving badly, without any causal effect of suspension rates. We attempt to distinguish between these alternative hypotheses.

### 3.4.1 Fixed Effects Strategy: Estimating the Direct Effects of Suspensions

We begin by using a fixed effects strategy to eliminate variation in suspension rates that is driven by fixed student characteristics or differences on average between the discipline policies used at different schools. Specifically, we estimate the following model:

\[
y_{ijt} = \eta_i + \theta_j + \gamma_t + \beta (s_{ijt} > 0) + \epsilon_{ijt}
\]

where: \( y_{ijt} \) is the test score of student \( i \) in school-grade \( j \) in year \( t \); \( 1 (s_{ijt} > 0) \) indicates that student \( i \) was suspended during year \( t \); and \( \eta_i, \theta_j, \) and \( \gamma_t \) are fixed effects.

The coefficient \( \beta \) measures the association between suspension rates and test scores for a given student over time. As such, this approach eliminates the possibility that \( \beta \) simply captures variation across students (e.g., due to differences in family background) that leads to both lower test scores and higher suspension rates. Instead, it is an estimate of the extent to which being suspended predicts lower than usual test scores for the affected student — holding fixed her grade and the school she attends.

Since longer suspensions for more severe behavior may have larger effects, we also estimate a second model that include a separate indicator for severity level:

\[
y_{ijt} = \beta_2 1 (s_{ijt}^{L2} > 0) + \beta_3 1 (s_{ijt}^{L3} > 0) + \beta_4 1 (s_{ijt}^{L4} > 0) + \beta_5 1 (s_{ijt}^{L5} > 0) + \eta_i + \theta_j + \gamma_t + \epsilon_{ijt} \tag{3.2}
\]

where \( 1 (s_{ijt}^{Lx} > 0) \) is an indicator for having at least one suspension for Level \( x \) behavior during that school year, and so on. The coefficient \( \beta_x \) is the effect of having a suspension for Level \( x \) behavior conditional on one’s record of suspensions of other levels.
Fixed effects eliminate many potential sources of selection bias, and still allow us to fully exploit our rich panel data. Specifically, we can calculate how much of the unconditional correlation between suspension rates and test scores is due to variation across versus within schools, and we can trace out how effects appear to differ across suspensions levels.

Nonetheless, there are two obstacles to interpreting $\beta$ as the causal effect of suspensions. First, the effects we observe may be from common shocks that affect both behavior and performance. For example, trouble at home may cause a student to act out, while also making it hard to concentrate on school work.\textsuperscript{11} Second, we don’t directly observe behavior – a limitation shared by other existing analyses of suspensions. This approach is therefore unable to separate the effect of suspension from the effect of the behavior. To the extent that certain kinds of infractions – or the circumstances that led to them – have lingering direct effects on achievement, our fixed effect model will attribute those effects to the suspension.

3.4.2 Quasi-experiment: Harnessing a Change in Suspension Policy

In 2012, the NYCDOE revised its discipline code to eliminate suspensions for Level 2 infractions, which – as shown in Figure 3.4 – are for non-violent misbehavior such as smoking, lying or gambling. The key to identifying the effect of this reform is that the impact of the policy varied across schools, depending on the extent to which that school used suspensions as a disciplinary device. A school that never used suspensions for Level 2 behavior had no effective change in its discipline policy, whereas a school with a 5 percent pre-period suspension rate for Level 2 infractions experienced a large shift in policy.

Although the 2012 policy change took place in a climate of increasing lobbying for discipline reform, it appears to have been largely unanticipated by teachers and administrators. In fact, our conversations with school employees suggest that some did not realize that the policy had changed until they tried to suspend a student for a Level 2 infraction. To the extent that this reflects the experience of the broader population of school staff, the changes

\textsuperscript{11}Adolescence is also a time of rapid developmental changes (see e.g., Steinberg and Morris, 2001; Meeus, 2016; Foulkes and Blakemore, 2018). Changing behavioral traits as a result of adolescence or as a result of changes in home or school environments will not be absorbed by student fixed effects.
Table 3.4: Most Common Level 2 Infractions Leading to Suspension

<table>
<thead>
<tr>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>17933</td>
<td>Persistent low-level (Level 1) bad behavior</td>
</tr>
<tr>
<td>9875</td>
<td>Profane or abusive language or gestures</td>
</tr>
<tr>
<td>1287</td>
<td>Lying to school personnel</td>
</tr>
<tr>
<td>662</td>
<td>Smoking</td>
</tr>
<tr>
<td>575</td>
<td>Misusing property belonging to others</td>
</tr>
<tr>
<td>339</td>
<td>Gambling</td>
</tr>
<tr>
<td>217</td>
<td>Disruptive behavior on school bus</td>
</tr>
</tbody>
</table>

Table notes. This table shows the most common Level 2 infractions that lead to a suspension for all students in grades 6 to 12 between 2006 and 2011.

to Level 2 suspension levels induced by the policy are plausibly exogenous.

We use this variation to estimate the causal effect of a reduction of suspensions for low-level infractions on academic performance in a school-grade. First, we define \( dy_{ijt} = y_{ijt} - y_{ij,t-1} \) as the growth in student \( i \)'s test score from one year to the next. We then decompose average test score growth in school-grade \( j \) into the causal effect of suspensions, \( \rho \tilde{s}_{jt} \), and a term that accounts for all other variation between school-grades over time, \( \theta_{ijt} \):

\[
dy_{ijt} = \rho \tilde{s}_{jt} + \theta_{ijt} \tag{3.3}
\]

Next, we divide Level 2 suspensions into the part we can predict with past data, \( \hat{s}_{jt} \), and the rest. The actual Level 2 suspension rate is restricted to be zero after the policy change.

\[
\tilde{s}_{jt} = \overset{1}{1} \times (t < 2012) \times [\hat{s}_{jt} + \eta_{jt}] \tag{3.4}
\]

We allow the other factors that affect test scores, \( \theta_{ijt} \), to be correlated with \( \hat{s}_{jt} \), even conditional on time and school-grade fixed effects, and individual-level covariates \( X_{it} \).

\[
\theta_{ijt} = \tilde{a}_j + \tilde{\gamma}_t + \lambda \hat{s}_{jt} + \tilde{\beta}X_{it} + \tilde{\eta}_{ijt} \tag{3.5}
\]

Finally, we can rewrite test scores as a function of predicted Level 2 suspensions:

\[
dy_{ijt} = a_j + \gamma_t + \underbrace{[\rho \times \overset{1}{1} (t < 2012) + \lambda \hat{s}_{jt} + \tilde{\beta}X_{it} + \tilde{\eta}_{ijt}]}_{\hat{\theta}(t)} + \epsilon_{ijt} \tag{3.6}
\]
where \( \alpha_j \) and \( \gamma_t \) are school-grade fixed effects.

In this framework, which amounts to a difference-in-difference (DiD) design, the pre-period coefficient on predicted suspensions is a combination of a causal effect \( \rho \) and a selection term \( \lambda \). After 2012, the causal channel is shut off. If high suspension rates have a negative causal effect on student outcomes on average, we would therefore expect students in school-grades with high predicted suspension rates for Level 2 behavior to converge toward those with low predicted suspension rates. Conversely, if high suspension rates have a positive effect on student outcomes, we would expect the opposite.

The most important threat to our identification is that schools implemented other contemporaneous policy changes – or were hit by contemporaneous shocks – that differentially affected schools with higher versus lower predicted suspensions. In this case, the selection effect, \( \lambda \), would shift sharply in 2012, preventing us from identifying \( \rho \) from the pre-post shift in \( \delta^{(t)} \). As far as we are aware, there was no such simultaneous shift in policy that affected students in grades 6 to 8, and which can explain the results that we find below.

### 3.4.3 Predicting Suspension Rates

To obtain estimates of \( \hat{s}_{jt} \), we simply average Level 2 suspension rates for each school-grade from the 2006-2007 school year to the 2011-2012 school year. Time-varying school-grade level covariates make little difference.\(^{12}\) Since \( \hat{s}_{jt} = \bar{s}_j \) is constant over time with this approach, \( \lambda \hat{s}_j \) is absorbed by the school-grade fixed effects, \( \gamma_j \), in equation 3.6.

### 3.5 Results

#### 3.5.1 Fixed Effects Strategy

The results from our fixed effects strategy are shown in Tables 3.5 and 3.6 for Math and ELA respectively. First, columns (1) and (4) of these tables show unconditional correlations between suspensions and test scores for middle school students. Students with at least one

---

\(^{12}\)See Appendix C for a comparison of in-sample performance with and without controls.
### Table 3.5: Regressions of Standardized Math Scores on Suspensions

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Suspension</td>
<td>-0.641***</td>
<td>-0.483***</td>
<td>-0.060***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.007)</td>
<td>(0.003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 2</td>
<td>-0.281***</td>
<td>-0.259***</td>
<td>-0.017*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.015)</td>
<td>(0.010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 3</td>
<td>-0.428***</td>
<td>-0.348***</td>
<td>-0.051***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.008)</td>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 4</td>
<td>-0.517***</td>
<td>-0.392***</td>
<td>-0.051***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.007)</td>
<td>(0.003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 5</td>
<td>-0.571***</td>
<td>-0.364***</td>
<td>-0.090***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.008)</td>
<td>(0.006)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

School-Grade FE: No, Yes
Student FE: No, No, Yes
Year FE: No, Yes, Yes
Observations: 2030213, 2030213, 1816969
Clusters: 1944, 1944, 1915
Adjusted $R^2$: 0.025, 0.263, 0.801

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table notes. Dependent variable is the student’s test score, standardized within a grade-year cell. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Suspension in a given year perform 0.64σ worse on average in Math and 0.55σ worse in ELA. This correlation becomes more negative at higher levels. Holding other suspensions constant, students with a suspension for Level 2 behavior perform 0.28σ worse in Math and 0.17σ worse in ELA; but these effects are more than twice as large if the suspension is for Level 5 behavior – 0.57σ in Math and 0.47σ in ELA.

Columns (2) and (5) show the impact of adding school-grade and year fixed effects. This reduces the magnitudes of the effects somewhat, but never by more than 40 percent. Most of the relationship between suspensions and test scores cannot therefore be explained by differences across schools in discipline policies. Rather, even within school-grades, students who get suspended fare much worse than those who do not get suspended.

Even when student fixed effects are added in columns (3) and (6), there remains a negative and statistically significant relationship between suspensions and test scores. The
### Table 3.6: Regressions of Standardized ELA Scores on Suspensions

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Suspension</td>
<td>-0.548***</td>
<td>-0.403***</td>
<td>-0.038***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.007)</td>
<td>(0.002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 2</td>
<td>-0.169***</td>
<td>-0.149***</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.015)</td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 3</td>
<td>-0.355***</td>
<td>-0.277***</td>
<td>-0.030***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.007)</td>
<td>(0.003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 4</td>
<td>-0.447***</td>
<td>-0.333***</td>
<td>-0.032***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.007)</td>
<td>(0.003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Level 5</td>
<td>-0.471***</td>
<td>-0.295***</td>
<td>-0.058***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.008)</td>
<td>(0.005)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School-Grade FE</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student FE</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Year FE</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>2024429</td>
<td>2024429</td>
<td>1817756</td>
<td>2024429</td>
<td>2024429</td>
<td>1817756</td>
</tr>
<tr>
<td>Clusters</td>
<td>1943</td>
<td>1943</td>
<td>1915</td>
<td>1943</td>
<td>1943</td>
<td>1915</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.018</td>
<td>0.218</td>
<td>0.768</td>
<td>0.018</td>
<td>0.218</td>
<td>0.769</td>
</tr>
</tbody>
</table>

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

**Table notes.** Dependent variable is the student’s test score, standardized within a grade-year cell. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Magnitudes of the coefficients are substantially reduced, but effect sizes are meaningful. Students who are suspended at least once score 0.06σ lower in Math and 0.04σ lower in ELA. These negative relationships are stronger for higher-level suspensions.

In summary, students within schools who get suspended are different from those who do not, yet a given student’s performance is indeed lower in years in which they are suspended. Most of the covariance between suspensions and test scores is explained by differences between individuals rather than across schools.

#### 3.5.2 Quasi-experiment

We next assess the causal impact of the 2012 reform, which eliminated suspensions for Level 2 behavior across the entire public school district. To do this, we analyze the change in the

---

13Similar estimates are produced by regressions with individual fixed effects but no school-grade effects.
relationship between test score growth and predicted Level 2 suspension rates over time. We proceed in several steps, beginning with a graphical analysis. Specifically, we estimate the following equation, but omit the indicator for the last pre-treatment year, 2011.\(^{14}\)

\[
 dy_{ijt} = \alpha_j + \gamma_t + \sum_{s=2006}^{2015} \rho_s \mathbb{1}(t = s) \times \hat{s}_{jt} + \beta X_{it} + \epsilon_{ijt}
\]  

This approach allows the coefficient on predicted suspensions to vary year-to-year rather than only differing between the pre-period and the post-period as in equation 3.6.

The coefficients from this regression, \(\rho_s\), are shown in the top and bottom panels of Figure 3.6 for Math and ELA respectively, along with 95 percent confidence intervals. Prior to the 2012 policy change, the relationship between suspensions and relative Math test score growth is stable. In the years following the reform, however, there is a substantial improvement in school-grades that would have had high suspension rates, compared to those that would have had low suspension rates. The results for ELA follow the same qualitative pattern, but the gains are less clear.

The results in Figure 3.6 are exactly what would be expected if it is, on average, beneficial for students to have suspensions removed as a punishment option for Level 2 infractions. However, the fact that there is a one year delay in the effect of the reform suggests that it may have taken time for school administrators to adjust to the reform, possibly replacing suspensions with a more effective alternative. Indeed, we advise caution in interpreting these results as the effects of removing suspensions without changing any other aspect of school policy. We discuss this further in Section 3.5.3.

As an alternative way of visualizing the impact of the 2012 policy change, we also discretize \(\hat{s}_{jt}\) and estimate coefficients for groups of cohorts that are pooled based on their predicted suspension rates. Specifically, we estimate the following equation:

\[
 dy_{ijt} = \alpha_j + \gamma_t + \sum_{k=1}^{K} \sum_{s=2006}^{2015} \rho_{sk} \mathbb{1}(t = s) \times \mathbb{1}(\hat{s}_{jt} \in R_k) + \beta X_{it} + \epsilon_{ijt}
\]  

\(^{14}\)The estimated coefficients therefore measure the change in the relationship between test score growth and suspension rates compared to 2011.
Figure 3.6: This figure plots the coefficients, $\rho_s$, from regression 3.7. The results for Math are shown in the top panel, and ELA below. Each point measures the change in the relationship between test score growth and suspension rates across school-grades, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. Analogous graphs using test score percentiles are available in Appendix C.3.
where $R_k$ is the range of suspension rates of school-grades included in group $k$. Discretization in this manner has the advantage that we do not need to assume that the effect of treatment is directly proportional to a school-grade’s pre-period suspension rate. As above, we omit the indicator for 2011. We also omit a control category of school-grades with zero pre-period suspensions, so that the estimated coefficients are relative to this group.

The results with discretized predictions are shown in Figure 3.7, for two groups. The first group (‘below median’) contains students in school-grades with pre-period Level 2 suspension rates below the median among those with at least one suspension for Level 2 behavior. Pooling across these school-grades, there are on average 0.2 suspensions per student per year in this group. The second group (‘above median’) contains school-grades with more suspensions; these had an average suspension rate of 1.3 suspensions per student.

The coefficients plotted in Figure 3.7 largely mirror those with a continuous treatment measure (Figure 3.6), but show evidence of some non-linearity. Prior to the policy change, test score growth in school grades with both below-median and above-median suspension rate cohorts were stable relative to the control group. After the reform, both groups experience relative improvements in their Math scores – but especially the school-grades with the highest suspension rates. The relative gains in Math are significant, both economically and statistically. However, there is again less evidence of gains in ELA.

Our final step is to pool across years to gain precision, estimating equation 3.6. The coefficient on predicted suspensions in then an estimate of the change in the conditional relationship between predicted suspensions and average test scores between the pre-period and the post-period, after removing fixed differences between school-grades.

The results are shown in columns (1) and (3) of Table 3.7, for Math and ELA respectively. Columns (2) and (4) then break down the treatment effects by year of the post period, estimating a separate effect for each year from 2012 to 2015. The estimates mirror the graphical analysis above. Students who would have been most affected by the suspensions that were removed experience faster test score growth in the post period, once these suspensions are eliminated – especially two to three years after the reform.
Figure 3.7: This figure plots the coefficients, $\rho_{sk}$, from regression 3.8. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. Analogous graphs using test score percentiles are available in Appendix C.3.
Table 3.7: Change in Relationship Between Predicted Suspensions and Test Score Gains

<table>
<thead>
<tr>
<th></th>
<th>(1) Math</th>
<th>(2) Math</th>
<th>(3) ELA</th>
<th>(4) ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.00944**</td>
<td>0.000672</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00432)</td>
<td>(0.00213)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_{(t+1)}$</td>
<td>0.00223</td>
<td>-0.00593*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00433)</td>
<td>(0.00339)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_{(t+2)}$</td>
<td>0.0144***</td>
<td>0.00185</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00540)</td>
<td>(0.00338)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_{(t+3)}$</td>
<td>0.0134**</td>
<td>0.00439</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00555)</td>
<td>(0.00319)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_{(t+4)}$</td>
<td>0.00833</td>
<td>0.00289</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00583)</td>
<td>(0.00359)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1668620</td>
<td>1668620</td>
<td>1671041</td>
<td>1671041</td>
</tr>
<tr>
<td>Clusters</td>
<td>1735</td>
<td>1735</td>
<td>1735</td>
<td>1735</td>
</tr>
</tbody>
</table>

Table notes. Columns (1) and (3) of this table show estimates from equation 3.6. These pooled estimates are then broken down in columns (2) and (4). The dependent variable is the change in a student’s test score, with scores standardized within a grade-year cell. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

In Table 3.8, we repeat the same analysis on sub-populations who are at higher risk of suspension. Interestingly, the point estimates for $\rho$ are smaller rather than larger for black students and boys, although the differences are not statistically significant.

3.5.3 Discussion

Our fixed effect models suggest that suspensions of all severity levels are associated with lower contemporaneous academic achievement for suspended individuals. However, they do not provide definitive evidence that suspensions have a negative causal impact on suspended students, even though we control for both individual and school fixed effects.

In contrast, our quasi-experimental analysis shows that eliminating suspensions for Level 2 infractions had a significant positive effect on the average test scores of students in the school-grades most affected by the reform. This suggests that researchers and policymakers
Table 3.8: Change in Relationship Between Predicted Suspensions and Math Score Gains – Sub-samples

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.00705</td>
<td>0.0120***</td>
<td>0.00386</td>
<td>0.0120**</td>
<td>0.0168</td>
<td>0.00862**</td>
</tr>
<tr>
<td></td>
<td>(0.00431)</td>
<td>(0.00451)</td>
<td>(0.00471)</td>
<td>(0.00548)</td>
<td>(0.0126)</td>
<td>(0.00422)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Male</th>
<th>Female</th>
<th>Black</th>
<th>Hispanic</th>
<th>White</th>
<th>Free Lunch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>849648</td>
<td>818972</td>
<td>469182</td>
<td>681097</td>
<td>242535</td>
<td>1343693</td>
</tr>
<tr>
<td>Clusters</td>
<td>1705</td>
<td>1713</td>
<td>1701</td>
<td>1698</td>
<td>1588</td>
<td>1732</td>
</tr>
</tbody>
</table>

*p < 0.10, ** p < 0.05, *** p < 0.01

Table notes. This table shows estimates from equation 3.6 for specific sub-populations. The dependent variable is the change in a student’s Math test score, with scores standardized within a grade-year cell. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

should continue to pay close attention to alternative approaches to discipline, since the results from this reform are promising. Nonetheless, we would not necessarily suggest that the estimated effect of this reform be interpreted solely as the effect of removing suspensions for those who would have received them.

There are two reasons for the caution we advise in interpreting our results. First, there is a delay in the effect of the 2012 reform, which suggests that school administrators and teachers took time to learn how to respond to student infractions without suspensions at their disposal. This could have included more effective alternatives to suspensions that we cannot measure, such as ‘restorative justice’. Second, the effects we measure are very large on a per-suspension basis. In the absence of spillovers, column (1) of Table 3.7 suggests that one less suspension corresponds to an individual test score gain of 90 percent of one standard deviation if all of the effect comes from suspensions alone. The effect per suspension is even larger for the high suspension rate group in Figure 3.7.

An additional reason for caution is that our estimated effects measure the impact of banning suspensions for low-level behavior, and may not generalize to suspensions for serious infractions. Our fixed effect models revealed increasingly negative direct effects on achievement of suspensions for more severe infractions, which are usually longer. However, we cannot rule out the possibility that banning suspensions for higher-level behavior could have a weaker positive causal effect, or even a negative effect, on student outcomes.
3.6 Conclusion

Across all of our empirical approaches, our estimates suggest that more relaxed student discipline can be beneficial. Although much of the relationship between suspensions and achievement can be attributed to fixed differences across schools and across students, a significant negative relationship persists even after controlling for these differences. Moreover, our quasi-experimental analysis suggests that relaxing the disciplinary code was causally beneficial for students in New York City.

Further research is required, however, to determine how teachers and administrators responded to the removal of suspensions for Level 2 behavior. This would help understand the delay in the effect of the policy change, and why our estimated effect sizes are so large on a per-suspension basis. In addition, analyzing other similar reforms will help establish whether our findings generalize to other locations, and other types of school discipline including out-of-school suspensions or corporal punishment. Finally, it would be helpful to decompose the average effect that we measure into direct effects on the students who would have been suspended, and spillovers that were avoided.
References


Appendix A

Appendix to Chapter 1

A.1 Generalized Contracts

The model outlined in Section 1.2 assumes that employers offer a wage to workers, as opposed to offering a general contract that specifies both a wage and labor supply. In this appendix, I show that this is not restrictive. To do so, I adopt all the assumptions of the baseline model except that I allow each employer to offer a contract \( C_j = \{ z_j, l_j \} \in \mathbb{R}_+ \times \mathbb{R}_+ = C \) to the worker. Each contract specifies a salary \( z_j \in \mathbb{R}_+ \) and a quantity of labor \( l_j \in \mathbb{R}_+ \), which jointly imply a price per unit (wage) \( w_j = z_j / l_j \). As before, the worker accepts her preferred offer, supplies labor and consumes \( c = z - T(z) \).

The worker’s strategy is now a set of two functions – an investment decision and an acceptance rule – which can be written as: \( x : K \times T \to \mathbb{R}_+ \); and \( A : K \times T \times \Theta \times C^{[I]} \to J \). Each employer’s strategy is a a function that maps signals and tax systems to contract offers \( O_j : \Theta \times T \to C \). Despite the increased complexity, it remains true that every firm earns zero expected profit. Moreover, contracts can always be equivalently characterized as an offer of a wage \( w_j = w(\theta|\pi) \) equal to the worker’s expected marginal product given the signal \( \theta \), with the worker freely choosing how much labor to supply. In this sense, nothing substantive is changed from the baseline model.

*Lemma 3.* Fix a realized value of \( \theta \) and assume that \( E[q|\theta, \pi] \) is strictly positive and finite.
given equilibrium beliefs \( \pi (q) \). In any pure-strategy equilibrium: all firms \( j \in J \) earn zero expected profit; the wage \( w_j = z_j/l_j \) implied by every contract offered to the worker is equal to her expected marginal product \( E [q_j | \theta, \pi] \); and the worker’s labor supply \( l_j \) satisfies \( l_j \in L_j^* = \text{argmax}_{l_j \in \mathbb{R}_+} u (w_j I_j - T (w_j I_j), I_j) \).

A.2 Continuity and Stability

A.2.1 Continuity of Investment Responses

In this appendix, I discuss conditions under which equilibrium indeterminacy is avoided, and a given equilibrium shifts continuously in response to the perturbations that I consider. I assume throughout that there is a finite number of cost types. Let \( i = 1, \ldots, |K| \) index these types, let \( x \) be the vector of investment decisions, and define \( q_i = Q(x_i) \).

For each \( i \), Assumption 3 ensures that the following binding first-order condition characterizes the optimal investment decision.

\[
 f_i (x, T) = Q'(x_i) \int_{\Theta} v (\theta | \pi, T) \left. \frac{f (\theta | q)}{\partial q} \right|_{q=q_i} d\theta - k_i = 0
\]  

(A.1)

Differentiating \( f_i (x, T) \) with respect to \( x_j \), we obtain the effect of higher investment by type \( j \) on the investment returns of type \( i \). There are two cases:

\[
 \frac{\partial f_i}{\partial x_j} (x) = \begin{cases} 
 f_{ii}^d + f_{ij}^w & \text{if } i = j \\
 f_{ij}^w & \text{if } i \neq j 
\end{cases}
\]  

(A.2)

where \( f_{ii}^d \) is type \( k \)’s second-order condition, and \( f_{ij}^w \) is the effect via employer beliefs.

\[
 f_{ii}^d = Q''(x_i) \int_{\Theta} v (\theta | \pi, T) \left. \frac{\partial f (\theta | q)}{\partial q} \right|_{q=q_i} d\theta + Q'(x_i)^2 \int_{\Theta} v (\theta | \pi, T) \left. \frac{\partial^2 f (\theta | q)}{\partial q^2} \right|_{q=q_i} d\theta
\]  

(A.3)

\[
 f_{ij}^w = Q'(x_j) \int_{\Theta} u_c (\theta) \left[ 1 - T' (z (\theta | \pi, T)) \right] l (\theta | \pi, T) \left. \frac{\partial w (\theta | \pi)}{\partial q_j} f (\theta | q_i) \right|_{\theta=q_i} d\theta
\]  

(A.4)
Letting $p(k_j)$ be the probability of drawing type $k_j$, the equation for $\frac{\partial w(q_j p)}{\partial q_j}$ is as follows.

$$\frac{\partial w(q_j p)}{\partial q_j} = f(q_j q_j) + q_j w(q_j p) \left( \frac{\partial f(\theta q_j)}{\partial q} \right)_{q=q_j} p(k_j)$$  \hspace{1em} (A.5)

The partial derivatives (equation A.2) can be arranged to form the Jacobian $J_{f,x}$.

$$J_{f,x} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_{|K|}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{|K|}(x)}{\partial x_1} & \cdots & \frac{\partial f_{|K|}(x)}{\partial x_{|K|}} \end{bmatrix} \hspace{1em} (A.6)$$

Next, let $dc(\theta|\pi, T) = -dT(z(\theta|\pi, T))$ be the Fréchet derivative with respect to $T$ of consumption by a worker with signal $\theta$. The Fréchet derivative of $v(\theta|\pi, T)$ is then:

$$dv(\theta|\pi, T) = u'(z(\theta|\pi, T) - T(z(\theta|\pi, T))) \times dc(\theta|\pi, T)$$

And in turn, the Fréchet derivative of $f_i(x, T)$ is given by $df_i(x, T)$.

$$df_i(x, T) = Q'(x_i) \int_{\Theta} dv(\theta|\pi, T) \left( \frac{f(\theta|q)}{\partial q} \right)_{q=q_i} \, d\theta$$  \hspace{1em} (A.7)

These derivatives can be stacked into a $|K| \times 1$ vector $df(x, T)$.

Providing that $J_{f,x}$ invertible, the Implicit Function Theorem implies that there is a neighborhood around $x$ and $T$ in which there is a unique Fréchet differentiable function mapping $T$ to $x$, and the response of investments is given by $-J_{f,x}^{-1} \times df(x, T)$. As I argue below, invertibility of $J_{f,x}$ is the generic case.

**A.2.2 Invertibility of $J_{f,x}$**

I next show that, if $J_{f,x}$ is not invertible, it can be rendered invertible by an arbitrarily small perturbation to the investment technology $Q(x)$, which preserves both the key properties of that technology and the existing equilibrium. Moreover, starting with any equilibrium in which $J_{f,x}$ is invertible, this clearly remains the case after a similarly small perturbation. In these two senses, invertibility of $J_{f,x}$ is generic.

First, I construct a parameterized family of functions, $\bar{Q}(x|c)$, where $c$ is a vector
of strictly negative real numbers $c_1, \ldots, c_K$. Each function in this family retains the key properties of $Q(x)$, but the second derivative of $\tilde{Q}(x|c)$ evaluated at $x_j$ is $c_j$.

1. Take each $x_j$ and define a narrow domain $x_j \pm r$ where $r > 0$ is arbitrarily small. On this domain, define a function $B_j(x|c_j) = Q(x_j) + Q'(x_j)(x - x_j) + \frac{1}{2} c_j (x - x_j)^2$. $B_j(x|c_j)$ has the same level and derivative as $Q(x)$ at $x_j$, but $B_j''(x_j|c_j) = c_j$.

2. Link the functions $B_j(x|c_j)$ to form any twice-differentiable function $\tilde{Q}(x|c)$ with $\tilde{Q}(0|c) = 0$, $\tilde{Q}'(x|c) > 0$, $\tilde{Q}''(x|c) > 0$ and $\lim_{x \to 0} \tilde{Q}'(x|c) = \infty$. This is always possible, since $r$ is small and $Q$ strictly concave.

3. Let $\hat{Q}(x|c, a) = a \tilde{Q}(x|c) + (1 - a) Q(x)$ with $a \in (0, 1)$.

Next, I replace $Q(x)$ with $\hat{Q}(x|c, a)$ in the economy described in Section 2. For any $c$, there remains an equilibrium with the same investment decisions. However, the diagonal elements of the Jacobian $J_{f,x}$ are replaced by:

$$f_{ii} = c_i \int_{\Theta} v(\theta|\pi, T) \left. \frac{\partial f(\theta|q)}{\partial q} \right|_{q = Q(x_i)} d\theta + Q'(x_i)^2 \int_{\Theta} v(\theta|\pi, T) \left. \frac{\partial^2 f(\theta|q)}{\partial q^2} \right|_{q = Q(x_i)} d\theta.$$  

Moreover, $f_{ii}$ scales with $c_i$ since $\int_{\Theta} v(\theta|\pi, T) \left. \frac{\partial f(\theta|q)}{\partial q} \right|_{q = q_i} d\theta > 0$. Non-diagonal elements of $J_{f,x}$ are unchanged.

Finally, let $c_j = Q''(x_j) + \epsilon_j < 0$ where $\epsilon_j$ are distinct real numbers with $\epsilon_j < -Q''(x_j)$. For small enough $a$, $\hat{Q}(x|c, a)$ is an arbitrarily close approximation of $Q(x)$. However, the Jacobian $J_{f,x}$ of the new economy is invertible. Specifically, any two rows that were collinear are no longer collinear; and, since $a$ is small, no two rows are newly collinear.

### A.2.3 Stability of Equilibria

Restricting the set of equilibria to those that are stable is one way to ensure that the economy does not switch equilibria in response to a perturbation such as that described in Section 1.5. To define such a notion of stability, suppose that the economy evolves according to the
following backward-looking dynamic adjustment process:

\[ x_{k,t+1} \in X_{k,t+1}^* = \arg\max_{\hat{x} \in \mathbb{R}_+} \int_{\mathcal{Q}} v(\theta|\pi_t, T) f(\theta|Q(\hat{x})) \, d\theta - k\hat{x} \quad (A.8) \]

where:

\[ v(\theta|\pi_t, T) = w(\theta|\pi_t, T) l(\theta|\pi_t, T) \]

\[ l(\theta|\pi_t, T) \in L^*(\theta|\pi_t, T) = \arg\max_{\tilde{l} \in \mathbb{R}_+} u(w(\theta|\pi_t) \tilde{l} - T(w(\theta|\pi_t) \tilde{l}), \tilde{l}) \]

\[ w(\theta|\pi_t) = \frac{\int_K Q(x_{k,t}) f(\theta|Q(x_{k,t})) \, dG(k)}{\int_K f(\theta|Q(x_{k,t})) \, dG(k)} \]

In general, this does not necessarily define a unique path for the economy. However, Assumptions 1 to 3 ensure that this is true locally because both \( x_{k,t+1} \) and \( l(\theta|\pi_t, T) \) are both uniquely pinned down and vary continuously with other agents’ investment decisions.

Thus, letting \( \bar{x}(T) \) be a set of equilibrium investment decisions, the dynamic adjustment process above can be approximated locally around \( \bar{x}(T) \) by a first-order linear system \( x_{t+1}(T) = Ax_t(T) \). If all the eigenvalues of the matrix \( A \) have moduli strictly less than one, then the equilibrium is locally asymptotically stable. Providing that \( J_{f,x} \) is invertible (see part A above) so that there is a locally unique Fréchet differentiable function mapping \( T \) to \( x \), local asymptotic stability in turn ensures that the economy does not switch equilibria in response to a small change in the tax schedule.

### A.3 Beyond the First Order Approach

Proposition 3 provides the derivative of social welfare with respect to a perturbation in the tax schedule, providing that there is a locally continuous selection around the initial point, \( (E(T), T) \). I adopted assumptions that ensure this is true for an arbitrary tax system. The proposition also states a condition that holds at an optimum, providing that the planner does not systematically locate at a point where the regularity conditions break down.

In this appendix, I discuss complications that arise when the planner does in fact have a
reason to locate at a discontinuity, in which case the derivatives in Proposition 3 are not defined. I also discuss reasons why the planner’s first-order condition is not sufficient for optimality. For expositional clarity, I focus on a particularly simple case of the general model, in which the planner is restricted to a linear tax, labor supply is perfectly inelastic, and investment decisions are binary.\(^1\) This greatly simplifies the analysis of this subset of issues, while providing insights that are conceptually general.

\section*{A.3.1 Special Case of the Model with Binary Investment}

In this special case of the model, investment is dichotomous. A worker decides to become qualified \(q\) at cost \(k\), or remain unqualified \(u\) at no cost. A qualified worker who is hired produces a fixed payoff \(\omega > 0\) for the firm who hires her, while an unqualified worker produces zero. As before, the cost distribution \(G(k)\) is the probability that a worker has investment cost no greater than \(k\); here, I additionally assume that \(G(0) = 0\) and that \(G(k)\) is continuously differentiable, with density \(g(k)\).

With binary investment, an employer’s prior belief is summarized by the fraction of workers it believes have invested. In addition, employers see a common non-contractible signal \(q \in [0, 1]\), which in this case has CDF \(F_i(\theta)\) and PDF \(f_i(\theta)\) where \(i \in \{q, u\}\) and \(f_u(\theta) / f_q(\theta)\) is strictly decreasing in \(\theta\). In equilibrium, firms’ prior beliefs coincide with the true equilibrium probability \(\pi\) that a worker invests; and each firm offers to pay the worker a wage \(w(\theta|\pi)\) equal to her expected marginal product.

\[
\begin{align*}
  w(\theta|\pi) &= \omega \times \frac{\pi f_q(\theta)}{\pi f_q(\theta) + (1 - \pi) f_u(\theta)}
\end{align*}
\]

The worker accepts her best offer, supplies a unit of labor and receives that wage. If she invested, she obtains utility \(v(\theta|\pi, \tau) - k = u((1 - \tau) w(\theta|\pi) + \tau \bar{w}) - k\), where \(\tau\) is a linear income tax, and \(\bar{w} = \pi \omega\) is the average wage. If she did not invest, she receives \(v(\theta|\pi, \tau) = u((1 - \tau) w(\theta|\pi) + \tau \bar{w})\). I assume that \(u(c)\) is strictly increasing, strictly concave and satisfies Inada conditions: \(\lim_{c \to 0} u'(c) = \infty\) and \(\lim_{c \to \infty} u'(c) = 0\).

\(^1\)The model with binary investment is similar to Moro and Norman (2004).
Integrating over \( q \), the expected utilities for an investor (\( v_q \)) and non-investor (\( v_u \)) are given by equations A.9 and A.10.

\[
\begin{align*}
    v_q (\pi | \tau) &= \int_0^1 v (\theta | \pi, \tau) dF_q (\theta) - k \quad \text{(A.9)} \\
    v_u (\pi | \tau) &= \int_0^1 v (\theta | \pi, \tau) dF_u (\theta) \quad \text{(A.10)}
\end{align*}
\]

Since workers invest if their expected return is greater than their cost, this implies an investment rate of \( G (\beta (\pi | \tau)) \) where \( \beta (\pi | \tau) = v_q (\pi | \tau) - v_u (\pi | \tau) \).

The final requirement of equilibrium is that workers invest at a rate that coincides with employers’ beliefs. This is embodied in equation A.11, which states that the fraction of investors must be equal to the fraction of workers that employers believe are qualified.

\[ \pi = G (\beta (\pi | \tau)) \quad \text{(A.11)} \]

For a given tax rate \( \tau \), equation A.11 defines a fixed point as shown in Figure A.1. An employer belief \( \pi \), combined with the tax \( \tau \), pins down the investment return and an investment rate, \( G (\beta (\pi | \tau)) \).

Any point on the 45 degree line constitutes an equilibrium, since employers’ beliefs are confirmed. At the extremes, either \( \pi = 0 \) or \( \pi = 1 \) ensure that there is no return to investment, since employers who are certain of a worker’s decision place no weight on the signal. There is thus always an equilibrium in which no workers invest, and all workers receive a zero wage. Proposition 16 provides sufficient conditions for there to be others. For example, the economy in Figure A.1 has four equilibria: 0, \( E_1 \), \( E_2 \) and \( E_3 \).

**Proposition 16.** Assume that \( f (\theta) = f_u (\theta) / f_q (\theta) \) is continuous and strictly positive on \([0, 1]\). If there exists \( \pi \) such that \( G (\beta (\pi | \tau)) > \pi \) then there are multiple solutions to condition A.11.

Intuitively, these conditions are satisfied if the returns to investment are high enough, as ensured by a large value of \( \omega \) or a low enough tax rate. In turn, this means there is some employer belief \( \pi \) such that the fraction of investors given that belief, \( G (\beta (\pi | \tau)) \), is higher
Figure A.1: This figure shows an example economy with binary investment. In the left panel, the aggregate rate of investment implied by worker and firm optimization, $G(\beta(\pi))$, is plotted against the employer prior, $\pi$. Any intersection between this line and the 45 degree line is an equilibrium. The arrows show the direction in which each equilibrium moves as $\tau$ rises. The right panel shows the set of equilibria over a range of values of $\tau$. Pareto dominant equilibria are shown by the black line segments.

than $\pi$. Since $G(\beta(1|\tau)) = 0$, and the regularity assumptions ensure that $G(\beta(\pi|\tau))$ is continuous $\pi$, this guarantees that there is a belief $\pi^* > 0$ such that $\pi^* = G(\beta(\pi^*|\tau))$.

### A.3.2 Optimal Taxation with Binary Investment

Tax policy can be analyzed in the same way as in the general model. Raising the linear tax $\tau$ causes $G(\beta(\pi|\tau))$ to shift down for every employer belief $\pi$. As a result, the location of an equilibrium falls if $G(\beta(\pi|\tau))$ crosses the 45 degree line from above, and rises if it crosses from below, as shown in panel (b) of Figure A.1.

For simplicity, I assume that agents play the planner’s preferred equilibrium, which ensures that investment and welfare always increase as $\tau$ is lowered.\(^2\) The arguments that follow do not depend on this assumption. However, it provides a concrete equilibrium

\(^2\)The set of equilibria can alternatively be refined by requiring stability under a simple adjustment process: $\pi_{t+1} = G(\beta(\pi_t|\tau))$. This amounts to a requirement that the absolute value of the slope of $G(\beta(\pi|\tau))$ is less than one (see Appendix A.4), which in turn implies that investment must fall when $\tau$ rises. In Figure A.1, both the zero investment equilibrium and $E_2$ are ruled out by the stability condition.
selection criterion that is especially compelling here because equilibria for a given tax rate are Pareto-ranked, with higher investment corresponding to higher welfare. In Figure A.1, the black line traces out the Pareto-dominant equilibria.

**Proposition 17.** Assume that multiple values of $\pi$ satisfy equation A.11 for a given tax rate $\tau$. Let $\pi_i$ and $\pi_j$ be two solutions. Welfare is higher for every worker under $\pi_i$ than $\pi_j$ iff $\pi_i > \pi_j$. Moreover, investment in the planner’s preferred equilibrium increases as $\tau$ is lowered.

Next, to characterize optimal taxation, define $\varepsilon_z$ as the elasticity of average income with respect to the retention rate. Second, let $u'_\theta$ be the marginal utility of consumption of an individual who sends signal $\theta$ and therefore receives wage $w(\theta|\pi)$. Finally, let $\tilde{u}'_\theta$ be the same individual’s marginal utility relative to the average: i.e., $\tilde{u}'_\theta = u'_\theta / \bar{u}'_\theta$. For simplicity, I assume here that the planner’s social welfare function is linear, but additional concavity from the social welfare function does not change the analysis.

Proposition 18 provides a necessary condition for the optimality of $\tau$, in the same form as Propositions 2 and 3. As before, there is a trade-off between redistribution from high-wage to low-wage workers, a fiscal externality and a belief externality. Ignoring the belief externality, an optimal $\tau$ at which this condition holds would always be strictly positive. However, the belief externality $\bar{w}_z$ provides an efficiency motive for intervention and pushes toward lower and possibly negative tax rates.

**Proposition 18.** Fix a value of $\tau$ and an investment rate $\pi^*(\tau) > 0$, which satisfies equation A.11. If $g(\beta(\pi^*(\tau)|\tau)) \beta'(\pi^*(\tau)|\tau) \neq 1$ and $\tau$ is optimal, then the following condition holds:

$$\frac{\tau}{1 - \tau} = \frac{\pi_z - \varepsilon_z \bar{w}_z}{\varepsilon_z}$$

(A.12)

where $\pi_z = (1 - \pi) \int_0^1 \tilde{u}'_\theta \left[ f_u(\theta) - f_q(\theta) \right] d\theta$, $\varepsilon_z$ is the elasticity of income to the retention rate $1 - \tau$, and $\bar{w}_z = \frac{1}{\omega} \int_0^1 \tilde{u}'_\theta \frac{\partial \omega(\theta|\pi)}{\partial \pi} \left[ \pi f_q(\theta) + (1 - \pi) f_u(\theta) \right] d\theta$ is the belief externality.

Proposition 18 parallels the results from the linear tax example (Proposition 2) and non-linear taxation (Proposition 3). The requirement that $g(\beta(\pi^*(\tau)|\tau)) \beta'(\pi^*(\tau)|\tau) \neq 1$
simply suffices to ensure the investment rate varies continuously with \( \tau \) at the optimum, which is equivalent to invertibility of the Jacobian, \( J_{f,x} \), discussed in Appendix A.2.

A.3.3 Limitations of the First Order Approach

The model with binary investment provides a transparent and flexible platform to discuss complications that could lead to discontinuity at the optimum or prevent my necessary conditions from being sufficient for optimality. The first caveat is that condition 18 may hold at other points. For example, the planner’s optimal tax rate may be \( A_1 \) in panel (b) of Figure A.1, but the first order condition may also hold at \( C \). This a natural limitation of the first-order approach and is easily resolved by examining a second-order condition.

The second caveat is more interesting: in some economies, there may be an incentive for the planner to choose a tax rate that places the economy at a discontinuity. For example, consider again panel (b) of Figure A.1. By Proposition 18, we know that \( B_1 \) dominates \( B_2 \). The complication is that it is possible for social welfare to be increasing in \( \tau \) as we approach \( \tau_B \) from below and also as we approach \( \tau_B \) from above, so that \( \tau_B \) is the optimal tax rate. However, equation A.12 does not hold at the discontinuity. This is not a violation of Proposition 18, since \( g(\beta(\pi|\tau))\beta'(\pi|\tau) = 1 \) at \( B_1 \). However, it highlights a conceptually important limitation of the first-order approach in this context.

A.4 Multiple Groups and Self-fulfilling Disparities

A possibility with multiple equilibria is that employers have different beliefs about members of distinct groups (e.g., black and white workers). Although this is ruled out if agents always play the planner’s preferred equilibrium and the groups are identical, asymmetric equilibria could well arise in reality. This is the classic case of self-fulfilling statistical discrimination, as analyzed by Arrow (1973), Coate and Loury (1993), and others. In this appendix, I discuss the implications of this for optimal taxation.

My first step is to adapt the model in Appendix A.3 by dividing workers into an advantaged (A) group and a disadvantaged (D) group. Specifically, I assume that a worker
Thus, a worker’s expected utility is $v_j(\theta|\pi_j, T)$. The two groups are identical in fundamentals. As in Appendix A.3, the planner is restricted to linear taxation. However, she can set a different tax rate $\tau_j$ for each group $j \in \{A, D\}$, and a lump sum transfer $T_{A\to D}$ from As to Ds. These three variables constitute a tax system $T$.

**Definition.** A tax system $T$ is a triple $(\tau_A, \tau_D, T_{A\to D})$, comprised of marginal tax rate $\tau_j$ for each group combined with an intergroup transfer $T_{A\to D}$.

Equilibrium in the model with two distinct groups can be characterized as follows. First, net of investment costs, a worker of type $j$ with signal $\theta$ receives utility $v_j(\theta|\pi_j, T)$.

$$v_A(\theta|\pi_A, T) = u \left( (1 - \tau_A) \omega \frac{\pi_A f_q(\theta)}{\pi_A f_q(\theta) + (1 - \pi_A) f_u(\theta)} + \tau_A \pi_A \omega - \frac{T_{A\to B}}{\lambda_A} \right)$$

$$v_D(\theta|\pi_D, T) = u \left( (1 - \tau_D) \omega \frac{\pi_D f_q(\theta)}{\pi_D f_q(\theta) + (1 - \pi_D) f_u(\theta)} + \tau_D \pi_D \omega + \frac{T_{A\to D}}{\lambda_D} \right)$$

Thus, a worker’s expected utility is $\bar{v}_q^j(\pi_j|T)$ if she invests, and $\bar{v}_u^j(\pi_j|T)$ if she does not.

$$\bar{v}_q^j(\pi_j|T) = \int_0^1 v_A(\theta|\pi_j, T) dF_q(\theta) - k \quad \bar{v}_u^j(\pi_j|T) = \int_0^1 v_B(\theta|\pi_j, T) dF_u(\theta)$$

The model remains otherwise unchanged from Appendix A.3. Workers invest if the return, $\beta_j(\pi_j|T) = \bar{v}_q^j(\pi_j|T) - \bar{v}_u^j(\pi_j|T)$, is greater than their cost, implying an investment rate of $G(\beta_j(\pi_j|T))$. Equilibrium requires that $\pi_j = G(\beta_j(\pi_j|T))$, $j \in \{A, D\}$.

Unlike Appendix A.3, I do not assume that agents coordinate on the planner’s preferred equilibrium. Instead, I follow the approach of Section 1.4, which applies given any continuous selection of equilibria. Specifically, for any given tax schedule $T$, let $\pi(T)$ be the set of pairs $(\pi_A, \pi_D)$ such that $\pi_j(T) = G(\beta_j(\pi(T)|T))$ for $j \in \{A, D\}$. The correspondence $\pi(T)$ suffices to characterize the set of equilibria for each tax schedule. I define a selection by choosing one equilibrium pair $\pi^+(T)$ for each tax schedule from this set.

Optimal taxation is then similar to the case with one group. The planner values both groups equally, so welfare is the weighted average $W = \lambda_A W_A + \lambda_D W_D$, where:

$$W_j = \pi_j \bar{v}_q^j(\pi_j|T) + (1 - \pi_j) \bar{v}_q^j(\pi_j|T) - \int_0^{\pi_j} \bar{v}_q^j(\pi_j|T) - \bar{v}_u^j(\pi_j|T) kdG_j(k).$$
Within each group, the same perturbation arguments apply and the condition required for \( \tau_j \) to be optimal is unchanged. The only additional complication is the inter-group transfer, which is set so that the average marginal utility is the same for As and Ds.

**Proposition 19.** If \( \pi^\dagger(T) \) is locally continuous and \( T \) is optimal, the following conditions hold.

\[
\frac{\tau_j}{1 - \tau_j} = \frac{\tau_{j,\tau} - \varepsilon_z^j w_z^j}{\varepsilon_z^j}
\]

\[
\int_\theta u'_{A,\theta} dF (\theta) = \int_\theta u'_{B,\theta} dF (\theta)
\]

where \( \tau_{j,\tau} = (1 - \tau_j) \int_0^1 \tilde{u}'_{j,\theta} \left[ f_u (\theta) - f_q (\theta) \right] d\theta \), \( \varepsilon_z^j \) is the income elasticity of group \( j \), and \( w_z^j = \frac{1}{\alpha} \int_0^1 \tilde{u}'_{j,\theta} \frac{\partial \alpha (\theta; \pi_j)}{\partial \pi_j} \left[ \pi_j f_q (\theta) + (1 - \tau_j) f_u (\theta) \right] d (\theta) \) is the belief externality.

To build intuition, consider the case in which \( T_{A \rightarrow D} \) is constrained to be zero and \( \pi^\dagger(T) \) selects equilibria that are symmetric in the sense that \( \pi_A = \pi_B \). This is always possible, because the groups are identical. The planner’s choice of \( \tau_j \) is then isomorphic to the model with a single group, so \( \tau_A = \tau_B \) and \( \pi_A = \pi_B \). Moreover, if condition A.13 holds, equation A.14 must as well. Starting from equal treatment (\( \tau_A = \tau_B \) and \( T_{A \rightarrow D} = 0 \)), there is therefore no first order gain from slightly changing the tax system. This implies that the planner would not want to set \( T_{A \rightarrow D} \neq 0 \), even if she could. Intuitively, if the two groups are identical and equilibria are symmetric, there is no motive for the planner to choose a tax system that favors one group over the other.

In general, however, it is possible that \( \pi^\dagger(T) \) includes non-symmetric equilibria, which raises the possibility of “self-fulfilling” differences between groups. In this case, even through groups \( A \) and \( D \) are ex ante identical, it is not generally true that \( \pi_A = \pi_B \) even at the planner’s optimal choice of \( T \). The optimal \( T \) may then involve different marginal tax rates for \( A \) and \( B \) workers, and an inter-group transfer.

Although Proposition 19 still holds in this non-symmetric case, the potential for self-fulfilling asymmetries raises the question of whether there are policies that can eliminate this problem. One possibility is for the planner to set a tax that conditions on the aggregate level of investment, which would always allow the planner to ensure Pareto efficiency.
Alternatively, one could imagine a dynamic policy that transitions the economy from one equilibrium to another. For example, one could temporarily implement a very low tax rate and then ratchet it back up, ensuring convergence to a Pareto efficient equilibrium.

### A.5 Approximately Optimal Taxation

In this appendix, I derive a way of approximating the optimal tax schedule given only a few measurable statistics. Two principles underlie the approximation. First, I assume that a change in $T'(z)$ primarily causes individuals with income close to $z$ to respond. Second, I assume that the incidence of the externality falls on workers with similar welfare weight, labor supply and tax rate to those with income $z$.

Letting $l(z)$ be the labor supply at income $z$, I first define:

$$\tilde{W}_z, \tilde{q} = y_z(z(\tilde{q}_j, T))^{-1} T_0(z(\tilde{q}_j, T)) l(\tilde{q}_j) y_z(z)$$

The belief externality can then be re-written as an approximation, plus a covariance bias.

$$BE(z) = \int_\Theta \frac{dw(\hat{\theta} | \pi)}{d[1 - T'(z)]} f(\hat{\theta}) d\hat{\theta}$$

Next, without loss of generality, I write the externality as a share of the average wage rise.

$$\int_\Theta \frac{dw(\hat{\theta} | \pi)}{d[1 - T'(z)]} f(\hat{\theta}) d\hat{\theta} = (1 - s(z)) \frac{d\bar{w}}{d[1 - T'(z)]}$$

Bringing everything together, condition 1.17 can then be approximated by:

$$FE(z) + ME(z) - (1 - s(z)) \psi_z(z) l(z) [1 - T'(z)] \frac{d\bar{w}}{d[1 - T'(z)]} = 0.$$  

Figure A.2 shows the results when equation A.17 is implemented in the simulated economy, using the values of $s(z)$ implied by the simulation. Starting from the naïve
Figure A.2: This figure shows the results of the simulation described in Appendix A.8. The solid red line shows the optimal tax schedule, the dashed blue line shows the naïve schedule, and the dotted black line shows a schedule what would be accepted by a planner who implemented equation A.17.

A benchmark, 60 percent of the gains from optimal taxation are achieved.

A.6 Unproductive Signaling with Observable Investment

In this appendix, I provide conditions under which $w(x)$ and $x(k)$ are differentiable. As in Section 1.4, I assume that problem 1.24 is strictly concave given a wage $w = w(x)$ so that the labor supply choice can be characterized by a first-order condition (equation A.18):
\[
wr_c(wl^*(w) - T(wl^*(w)), l^*(w)) [1 - T'(wl^*(w))] + u_l(wl^*(w) - T(wl^*(w)), l^*(w)) = 0
\]
(A.18)

where $l^*(w) = \text{argmax}_{l \in \mathbb{R}_+} u(wl - T(wl), l)$.

Next, I define $\delta(x) = \nu(w(x)|T)$, and let $x_{FB}(k) = \text{argmax}_x \nu(Q(x,k)|T) - kx$ be the investment level chosen by an agent with cost $k$ in the equivalent problem with perfect employer information. Using these definitions, I adopt three assumptions regarding problem
1.23, which can be viewed as restrictions on the investment technology, $Q(x,k)$.

**Assumption 5.** The solution to the first best contracting problem, $x_{FB}(k)$, is unique for all $k$.

**Assumption 6.** For all $k \in K$, $\vartheta(x)$ is strictly concave around $x_{FB}(k)$.

**Assumption 7.** $\exists \kappa > 0$ such that $\vartheta''(x) \geq 0 \Rightarrow \vartheta'(x) > \kappa$ for all $(k,x) \in K \times \mathbb{R}_+$.

A sufficient condition for assumption 5 to hold is that the first best contracting problem is strictly concave, which is always true given sufficient concavity of the investment technology. Assumption 6 simply states that problem 1.23 is *locally* strictly concave around the first-best investment choice, while assumption 7 is a *global* equivalent that is weaker than strict concavity but stronger than strict quasi-concavity.

Assumptions 5, 6 and 7 jointly ensure that $x(k)$ is differentiable for all $k \in K$ (see Mailath and von Thadden, 2013), which in turn implies that $w(x)$ is differentiable and that the following condition holds for all $k$:

$$u_c(z(k) - T(z(k)), l(k)) \left[1 - T'(z(k))\right] l(k) w'(x(k)) = k \quad (A.19)$$

where $l(k) = l^*(w(x(k)))$ and $z(k) = w(x(k)) l(k)$.

### A.7 Proofs and Derivations

**Proof of Lemma 1.** Firm beliefs about the distribution of productivity in the population must be confirmed in equilibrium and identical across firms. Let $\pi$ denote the equilibrium set of beliefs. Firm $j$'s expectation of the worker’s productivity is $E[q|\theta, \pi, A_j = 1] \geq 0$. Next, let $\tilde{u}(w_j) = u(w_j,l^*(w_j) - T(w,l^*(w_j)), l^*(w_j))$ represent the utility that the worker receives from accepting wage $w_j$ and supplying labor optimally.

Suppose that some firm $j$ makes strictly positive expected profits given its wage offer $w_j$. It must then be the case that $\tilde{u}(w_j) \geq \tilde{u}(w_k)$ for all wages $w_k$ offered by other firms. There are several cases to consider, each of which lead to a contradiction.
Case 1: $\tilde{u}(w_j) > \tilde{u}(w_k)$ for some $w_k$.

In this case, firm $k$ initially earns zero expected profit, since no workers accept its offer. However, it can offer a wage slightly higher than $w_j$. It then attracts the worker with probability one and earns strictly positive profits. This is a profitable deviation.

Case 2: $\tilde{u}(w_j) = \tilde{u}(w_k)$ for all $w_k$, and $P_{k,\theta} \leq 0$ for some $k$.

If any firm makes weakly negative profits, then the same deviation as Case 1 applies.

Case 3: $\tilde{u}(w_j) = \tilde{u}(w_k)$ and $P_{k,\theta} > 0$ for all $k$.

Since the worker always accepts an offer, $E[q|\theta, \pi, A_j = 1]$ is bounded weakly below $E[q|\theta, \pi]$ for at least one firm. This firm’s expected profit is bounded below $P_{\text{MAX}}$.

$$P_{\text{MAX}} = \max_w [E[q|\theta, \pi] - w] I^*(w) \text{ s.t. } u(wl^*(w) - T(\omega l^*(w)), l^*(w)) \geq u(T(0), 0)$$

The assumptions on the worker’s utility function ensure that this yields finite labor supply for any finite $E[q|\theta, \pi]$. Since $w_j$ is greater than zero and $E[q|\theta, \pi]$ is finite, $P_{\text{MAX}}$ is also bounded. Finally, this firm can strictly increase its profit by raising $w_j$ slightly and attracting the worker with probability one.

Since every case in which a firm makes a strictly positive expected profit implies a profitable deviation, and all firms can obtain zero expected profit by offering a zero wage, it must be true that every firm makes zero expected profit. Finally, the wage, $w$, must be the same at every firm who hires the worker with positive probability. We have therefore established that $[E[q|\theta, \pi] - w] I^*(w) = 0$, which is only satisfied if $w = E[q|\theta, \pi]$. \hfill \Box

Proof of Proposition 1. Assume – subject to verification – that investment is distributed log-normally as hypothesized.

$$\ln q_i \sim N \left( \ln \mu_q - \frac{\sigma_q^2}{2}, \sigma_q^2 \right)$$

147
Given this, employers face a log-normal signal extraction problem. The expectation of log-productivity is as follows.

\[
E [\ln q | \theta] = \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \ln \theta + \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \left( \ln \mu_q - \frac{\sigma_q^2}{2} \right) 
\]

\[
= \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \ln q + \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \left( \ln \mu_q - \frac{\sigma_q^2}{2} \right) + \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \ln \xi
\]

Since employers offer workers their expected marginal product, the after-tax wage is:

\[
\ln [(1 - \tau) w] = \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \ln q + \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \ln \mu_q + \left( \frac{\sigma_q^2}{\sigma_q^2 + \sigma_q^2} \right) \ln \xi + \ln (1 - \tau).
\]

Exponentiating, we obtain the level of wages: \( w = q^s \mu_q^{1-s} \xi \), where \( s = \sigma_q^2 / (\sigma_q^2 + \sigma_q^2) \). Given this wage, labor supply is \( l = (1 - \tau)^{\xi} w^{\xi} \), which implies an after-tax income of:

\[
(1 - \tau) z = (1 - \tau) w l = (1 - \tau)^{1+\xi} w^{1+\xi} = (1 - \tau)^s \mu_q^{1-s} \xi^{1+\xi}.
\]

Next, since \( q = Q(x) = x^\beta \) and costs are linear, expected utility is as follows.

\[
\left( (1 - \tau)^{1+\xi} \mu_q^{(1-s)(1+\xi)} \right) E \left[ \xi^{s(1+\xi)} \right] \frac{x^{\beta s (1+\xi)}}{1+\xi} - kx + \tau z
\]

Since I assume that \( \beta s (1 + \xi) < 1 \), we can differentiate to find the agent’s choice of \( q \).

\[
q = \left[ \frac{\beta s (1 - \tau)^{1+\xi} \mu_q^{(1-s)(1+\xi)}}{k} E \left[ \xi^{s(1+\xi)} \right] \right] \left( 1 - \tau \right)
\]

Then, since \( \ln q \) is the sum of two normally distributed variables and a constant term, \( q \) is itself log-normally distributed. Specifically, it has the following distribution.

\[
\ln q \sim N \left( \frac{\beta}{1 - \beta s (1 + \xi)} \ln \beta + \frac{\beta}{1 - \beta s (1 + \xi)} \ln \mu_q + \frac{\beta (1 + \xi)}{1 - \beta s (1 + \xi)} \ln (1 - \tau) + (1 - s) \frac{\beta (1 + \xi)}{1 - \beta s (1 + \xi)} \ln \mu_q + \frac{\beta}{1 - \beta s (1 + \xi)} \ln E \left[ \xi^{s(1+\xi)} \right] \right)
\]

\[
- \frac{\beta}{1 - \beta s (1 + \xi)} \left( \ln \mu_q - \frac{\sigma_q^2}{2} \right), \left( \frac{\beta}{1 - \beta s (1 + \xi)} \right)^2 \sigma_q^2
\]
Finally, we can obtain expressions for \( m_q \) and \( s_{2q} \) by matching coefficients.

\[
\begin{align*}
\sigma_{q}^2 &= \left( \frac{\beta}{1 - \beta s (1 + \epsilon_l)} \right)^2 \sigma_k^2 \\
\mu_q &= \left\{ \frac{\beta s (1 - \tau)^{1+\epsilon_l} E[\xi^s(1+\epsilon_l)]}{\mu_k} \right\} \exp \left[ \left( 1 + \frac{\beta}{1 - \beta s (1 + \epsilon_l)} \right) \frac{\sigma_k^2}{2} \right]^{-\frac{\beta}{1 - \beta s (1 + \epsilon_l)}} \tag{A.20}
\end{align*}
\]

Equation A.20 implicitly pins down \( s_{2q} \) in terms of \( s_{2k}, \beta, \epsilon_l \) and \( \sigma_k^2 \). It is independent of \( m_k \).

In turn, equation A.21 characterizes \( m_q \) as a function of the same set of parameters plus \( m_k \).

The elasticity of \( m_q \) with respect to \( m_k \) is \( -\beta / [1 - \beta (1 + \epsilon_l)] \).

**Proof of Lemma 2.** There are two effects on \( q \) of increasing the retention rate \( 1 - \tau \): a direct effect, and an effect via average productivity. Combining these yields the total elasticity.

\[
\sigma_q = \frac{d q}{d (1 - \tau)} \times \frac{1 - \tau}{q} = \left( \frac{d q}{\partial (1 - \tau)} + \frac{d q}{\partial \mu_q} \frac{d \mu_q}{d (1 - \tau)} \right) \frac{1 - \tau}{q}
= \left( \frac{\beta (1 + \epsilon_l)}{1 - \beta s (1 + \epsilon_l)} + (1 - s) \frac{\beta (1 + \epsilon_l)}{1 - \beta s (1 + \epsilon_l)} \frac{\beta (1 + \epsilon_l)}{1 - \beta (1 + \epsilon_l)} \right)
= \frac{\beta (1 + \epsilon_l)}{1 - \beta (1 + \epsilon_l)}
\]

Similarly, we can derive the elasticity of income \( z \) to the retention rate.

\[
\sigma_z = \frac{d z}{d (1 - \tau)} \times \frac{1 - \tau}{z} = \left( \frac{d z}{\partial (1 - \tau)} + \frac{d z}{\partial q} \frac{d q}{\partial (1 - \tau)} + \frac{d z}{\partial \mu_q} \frac{d \mu_q}{d (1 - \tau)} \right) \frac{1 - \tau}{z}
= \epsilon_l + (1 + \epsilon_l) \frac{\beta (1 + \epsilon_l)}{1 - \beta (1 + \epsilon_l)}
= \frac{\epsilon_l + \beta (1 + \epsilon_l)}{1 - \beta (1 + \epsilon_l)}
\]

**Proof of Proposition 2.** The utility of a worker with noise realization \( \xi \) and cost \( k \) is:

\[
v = \left[ \frac{(1 - \tau) q^s \mu_q^{1-s} \xi^{1+\epsilon_l}}{1 + \epsilon_l} \right] - k x + \tau z
\]

149
where $x$ is chosen optimally according to the following first order condition.

$$k = \beta s \left( (1 - \tau)^{1+\epsilon_l} \mu_q \right) \frac{1 + s}{1 + \beta (1+\epsilon_l)} E \left[ \xi^{s(1+\epsilon_l)} \right] \chi^{s(1+\epsilon_l) - 1}$$

Taking the expectation over $\xi$, the expected utility for an individual with cost $k$ is:

$$E \left[ \xi^{s(1+\epsilon_l)} \right] (1 - \tau)^{1+\epsilon_l} \mu_q (1-s) \left( 1 + \beta (1+\epsilon_l) \right) q^{s(1+\epsilon_l)} + \tau \bar{z}$$

Then, substituting in the optimal choice of $q$, and weighting by the worker’s welfare weight $\psi_k$, we get expected welfare in terms of $\mu_q$ and $\xi$.

$$E \xi [\psi_k \psi_k] = \psi_k \frac{1 - \beta s (1 + \epsilon_l)}{1 + \epsilon_l} \left( (1 - \tau)^{1+\epsilon_l} \mu_q \right) \left( 1 - \tau \right)^{1+\epsilon_l} \mu_q \left( 1 - s \right) \left( 1 + \beta (1+\epsilon_l) \right) q^{s(1+\epsilon_l)} + \psi_k \tau \bar{z}$$

Finally, we can integrate over cost realizations to obtain average welfare.

$$E [\psi_k \psi_k] = (1 - \tau) E \left[ \psi_k \xi \right] \frac{1 - \beta s (1 + \epsilon_l)}{1 + \epsilon_l} + \tau \bar{z}$$

Building on this result, there are three effects from raising the retention rate. First, there is a fiscal externality from the change in average income, $\bar{z}$.

$$FE = \tau \bar{z} \psi \xi \bar{z} \frac{\bar{z}}{1 - \tau}$$

Second, welfare rises due to the externality via employer beliefs. Specifically, differentiating with respect to $\mu_q$ and aggregating over $k$, the gain in social welfare is as follows.

$$BE = (1 - s) E_k (\psi_k \bar{z}_k) \epsilon_q$$

Finally, there is a mechanical welfare loss due to the transfer from the average worker to high-income workers:

$$ME = E_k \left( \psi_k \bar{z}_k \right) - \bar{z}$$
Summing the three effects we obtain an expression for the total welfare gain.

\[ \text{FE} + \text{ME} + \text{BE} = \frac{\tau}{1 - \tau} \varepsilon q \bar{z} + E_k (\psi_k \bar{z}_k) \left[ 1 + (1 - s) \varepsilon_q \right] \bar{z} - \bar{\psi} \bar{z} \]

Then setting this to zero yields the first order condition shown in the proposition. \( \square \)

**Proof of Proposition 3.** The objective of the social planner is to maximize welfare \( W(T) \) subject to the four constraints of Problem 1.5. This problem is restated here for convenience.

\[
\max_T W(T) = W(\overline{V}(k,T)) \, dG(k)
\]

where:

\[
\overline{V}(k,T) = \int_\Theta (v(\theta|k,T) - kx(k,\pi,T)) f(\theta,q(k,\pi,T)) \, d\theta
\]

subject to:

\[
x(k|\pi,T) \in \arg\max_{x \in \mathbb{R}_+} \int_\Theta v(\theta|\pi,T) f(\theta|Q(\xi)) \, d\theta - k\bar{x}
\]

\[
l(\theta|\pi,T) \in \arg\max_{l \in \mathbb{R}_+} u( l (w(\theta|k)) \bar{l} - T(w(\theta|k)) \bar{l} )
\]

\[
w(\theta|\pi) = \frac{\int_k q(k|\pi,T) f(\theta|q(k,\pi,T)) \, dG(k)}{\int_k f(\theta|q(k,\pi,T)) \, dG(k)}
\]

\[
R = \int_\Theta T(z(\theta|\pi,T)) f(\theta) \, d\theta
\]

For ease of discussion, it will also be helpful to recall that \( v(\theta|\pi,T) \) can be expanded and written as a function of a worker’s wage, labor supply and tax liability.

\[
v(\theta|\pi,T) = u\left( w(\theta|k) l(\theta|\pi,T) - T(w(\theta|k)) l(\theta|\pi,T) \right), l(\theta|\pi,T)
\]

(A.22)

A perturbation to \( T \) as described has three effects that I will consider in turn. First, there is a welfare loss (WL) from taking money from individuals with income higher than \( z \).

\[
\text{WL} = -d \tau dz \left\{ \int_{\theta(z|\pi,T)} u_c(\theta) \int_k \psi(k) dG(k|\theta) f(\theta) \, d\theta \right\}
\]

(A.23)

Since the revenue raised is returned to all individuals equally via an increase in the intercept
of the tax schedule, it is worth \( \lambda \) per dollar in terms of social welfare, where:

\[
\lambda = \int_{\Theta} u_{c}(\theta) \int_{K} \psi(k) dG(k|\theta) f(\theta) \ d\theta \tag{A.24}
\]

Multiplying by the amount of revenue raised, the welfare gain (WG) from this transfer is:

\[
WG = d\tau dz \left\{ \frac{\tau}{1 - T'(z)} f(\theta) \ d\theta \right\} \lambda. \tag{A.25}
\]

Summing WL and WG, then dividing by \( l \) yields the mechanical gain in welfare, ME (z).

The second effect to consider is the fiscal externality, FE (z), which arises when individuals re-optimize. The value of the fiscal externality can be obtained by differentiating the resource constraint, yielding the impact on government revenue from re-optimization.

Since the focal selection \((E(T), T)\) is assumed to be locally continuously differentiable with respect to \( T \), \( l(\theta|\pi, T) \) and \( x(k|\pi, T) \) respond continuously to the perturbation. Next, since \( x(k|\pi, T) \) responds continuously and \( Q \) is differentiable, so does \( q(k|\pi, T) = Q(x(k|\pi, T)) \). Finally, since \( f(\theta) = \int_{K} f(\theta|q(k|\pi, T)) dG(k) \) is continuous in \( q(k|\pi, T) \), \( f(\theta) \) responds continuously. In turn, this implies that \( w(\theta|\pi) \) responds continuously. The change in income given a signal realization \( \theta \) can therefore be written as follows.

\[
- \frac{dz(\theta|\pi, T)}{d[1 - T'(z)]} = -w(\theta|\pi, T) \frac{dl(\theta|\pi, T)}{d[1 - T'(z)]} - l(\theta|\pi, T) \frac{dw(\theta|\pi, T)}{d[1 - T'(z)]}
\]

These results allow the fiscal externality to be written as a combination of the effects of changes in \( z(\theta|\pi, T) \) and \( f(\theta) \), capturing the effect on government revenue from both investment and labor supply decisions. After dividing through by \( \lambda \), the total fiscal externality is as follows.

\[
FE(z) = -d\tau dz \left\{ T'(z(\hat{\theta}|\pi)) \left( \frac{dz(\hat{\theta}|\pi, T)}{d[1 - T'(z)]} \right) f(\hat{\theta}) - T(z(\hat{\theta}|\pi, T)) \frac{df(\hat{\theta})}{d[1 - T'(z)]} \right\} d\hat{\theta}
\]

The final effect of taxation is the effect on individual utility of changing wages in response to shifts in the distribution of productivity (BE). Since individuals take the wage paid given any signal realization as fixed, they ignore this effect. Differentiating the belief
consistency constraint, the effect of a rise in individual $k$’s productivity on the wage of a worker with signal realization $\theta$ is as follows.

$$
\frac{dw(\theta|\pi)}{dq (k|\pi, T)} = f(\theta, q(k|\pi, T)) + \left( \frac{\partial f(\theta,q)}{\partial q} \bigg|_{q=q(k|\pi, T)} \right) [q (k|\pi, T) - E(q|\theta, \pi)]
$$

Applying the envelope theorem and again dividing by $\lambda$, the effect of this wage change on social welfare is simply scaled by the affected worker’s labor supply, retention rate and the average welfare weight of an individual with signal realization $\theta$.

$$
\frac{dw(\theta|\pi)}{dq (k|\pi, T)} \psi_z(z(\theta|\pi, T)) \left[ 1 - T'(z(\theta|\pi, T)) \right] l(\theta|\pi)
$$

To obtain the total belief externality shown in the main text, we then integrate over the distributions of $\theta$ and $k$.

These three effects jointly capture the total change in welfare from a perturbation, since the effects of individuals’ re-optimization on their own welfare are second-order. Thus, given any continuous selection, if $FE + BE + ME \neq 0$, welfare increases in response either to an arbitrarily small positive perturbation or an equivalent negative perturbation. Except at a discontinuity at which $ME$, $FE$ and $BE$ are not defined, a necessary condition for an optimum is therefore that the sum of the three effects is zero.

$$
\text{Proof of Proposition 4.} \text{ Assume – subject to verification – that productivity and investment are log-normally distributed.}

q \sim LN \left( \ln \mu_q - \frac{\sigma^2_q}{2}, \frac{\sigma^2_q}{2} \right)
$$

Next, suppose the relationship between productivity and investment can be written as:

$$
\ln q = \ln A + B \ln x
$$

where $A$ and $B$ are scalars that will be found by matching coefficients. This allows the signal
to be written as a linear combination of productivity $q$ and noise $\xi$.

$$\ln \theta = \left( \frac{1}{B} \right) \ln q - \left( \frac{1}{B} \right) \ln A + \ln \xi$$

For convenience, define $\ln \tilde{\xi} = B \ln \xi$ and let $\ln \tilde{\theta}$ be the following linear transformation of the original signal.

$$\ln \tilde{\theta} = B \ln \theta + \ln A = \ln q + B \ln \xi = \ln q + \ln \tilde{\xi}$$

The expected log-marginal product of an individual follows from the fact that the employer faces a standard normal signal extraction problem:

$$E \left[ \ln q | \tilde{\theta} \right] = s \ln \tilde{\theta} + (1 - s) \left( \ln \mu_q - \frac{\sigma^2_q}{2} \right)$$

where $s = \sigma^2_q / (\sigma^2_q + \sigma^2) = \sigma^2_x / (\sigma^2_x + \sigma^2)$. A worker’s expected level of productivity is therefore a geometric weighted average of $A, x, \xi$ and $\mu_q$.

$$w = \tilde{\theta}^s \mu_q^{1-s} = A^s x^{sB} \xi^{sB} \mu_q^{1-s}$$

Optimal labor supply is $l = \left( 1 - \tau \right)^{\epsilon_i} w^{\epsilon_i}$, which means that after tax income is:

$$(1 - \tau) z = (1 - \tau)^{1+\epsilon_i} w^{1+\epsilon_i}$$

$$= (1 - \tau)^{1+\epsilon_i} \left[ A^s x^{sB} \xi^{sB} \mu_q^{1-s} \right]^{1+\epsilon_i}.$$  

In turn, this implies a value of expected utility for any investment level.

$$v = \left[ A^s \left( 1 - \tau \right) \mu_q^{1-s} \right]^{1+\epsilon_i} E \left[ \xi^{sB(1+\epsilon_i)} \right] \frac{x^{sB(1+\epsilon_i)}}{1+\epsilon_i} - kx + \tau z$$

Assuming again that $\beta s(1+\epsilon_i) < 1$, it will also turn out to be true that $sB(1+\epsilon_i) < 1$. This in turn ensures that the worker’s optimal choice of $\ln x$ is as follows.

$$\ln x = \frac{1}{1 - sB(1+\epsilon_i)} \left[ \ln n + \ln (sB) + (1+\epsilon_i) \ln \left( 1 - \tau \right) + (1 - s) (1 + \epsilon_i) \ln \mu_q + \ln E \left[ s^{s(1+\epsilon_i)} \right] + s (1 + \epsilon_i) \ln A \right]$$

154
Next, using the fact that \( \ln q = \alpha \ln n + \beta (1 - \alpha) \ln x \), and matching coefficients, \( B \) is:

\[
B = \frac{\alpha + \beta (1 - \alpha)}{1 + s \alpha (1 + \varepsilon_l)}.
\]

This can in turn be used to solve for \( \ln A \) in terms of \( x \).

\[
\ln A = \alpha \ln n - \frac{\alpha - \beta (1 - \alpha) s \alpha (1 + \varepsilon_l)}{1 + s \alpha (1 + \varepsilon_l)} \ln x
\]

\( A \) can then be eliminated to yield a new expression for \( \ln x \).

\[
\ln x = \frac{1 + s \alpha (1 + \varepsilon_l)}{1 - s \beta (1 - \alpha) (1 + \varepsilon_l)} \ln(n) + \frac{1}{1 - s \beta (1 - \alpha) (1 + \varepsilon_l)} \left[ \ln s + \ln \left( \frac{\alpha + \beta (1 - \alpha)}{1 + s \alpha (1 + \varepsilon_l)} \right) \right.
\]

\[
+ (1 + \varepsilon_l) \ln (1 - \tau) + (1 - s) (1 + \varepsilon_l) \ln \mu_q + \ln E \left( \tilde{x}^{s(1 + \varepsilon_l)} \right) \right]
\]

Finally, since \( x \) inherits the log-normality of \( n \), and \( \ln q = \alpha \ln n + (1 - \alpha) \beta \ln x \), \( q \) is also log-normal. This means that the values of \( \mu_q \) and \( \sigma^2_q \) can be found by matching coefficients.

\[
\sigma^2_q = \left[ \frac{\alpha + \beta (1 - \alpha)}{1 - \beta s (1 - \alpha) (1 + \varepsilon_l)} \right]^2 \sigma^2_n
\]

\[
\ln \mu_q = \left[ \frac{\alpha + \beta (1 - \alpha)}{1 - (1 - \alpha) \beta (1 + \varepsilon_l)} \right] \ln n + \left[ \frac{\beta (1 - \alpha)}{1 - (1 - \alpha) \beta (1 + \varepsilon_l)} \right]
\]

\[
\times \left\{ \ln s + \ln \left[ \frac{\alpha + \beta (1 - \alpha)}{1 + s \alpha (1 + \varepsilon_l)} \right] + (1 + \varepsilon_l) \ln (1 - \tau) + \ln E \left( \tilde{x}^{s(1 + \varepsilon_l)} \right) \right\}
\]

\[
+ \frac{\sigma^2_q}{2} \left[ \frac{\alpha + \beta (1 - \alpha)}{1 - s (1 - \alpha) \beta (1 + \varepsilon_l)} \right] \left[ \frac{\alpha + \beta (1 - \alpha)}{1 - (1 - \alpha) \beta (1 + \varepsilon_l)} \right]
\]

The elasticity of productivity follows directly.

\[
\frac{d \ln \mu_q}{d \ln (1 - \tau)} = \left( \frac{\beta (1 - \alpha) (1 + \varepsilon_l)}{1 - \beta (1 - \alpha) (1 + \varepsilon_l)} \right)
\]

Finally, the elasticity of income can be found as follows.

\[
\frac{d \ln z}{d \ln (1 - \tau)} = \frac{\partial z}{\partial (1 - \tau)} + \frac{\partial \ln z}{\partial \ln q} \frac{\partial \ln q}{\partial \ln (1 - \tau)} + \frac{\partial \ln z}{\partial \ln \mu_q} \frac{\partial \ln \mu_q}{\partial \ln (1 - \tau)}
\]

\[
= \varepsilon_l + (1 + \varepsilon_l) \left[ \frac{\beta (1 - \alpha) (1 + \varepsilon_l)}{1 - \beta (1 - \alpha) (1 + \varepsilon_l)} s + \frac{\beta (1 - \alpha) (1 + \varepsilon_l)}{1 - \beta (1 - \alpha) (1 + \varepsilon_l)} (1 - s) \right]
\]

\[
= \varepsilon_l + (1 + \varepsilon_l) \frac{\beta (1 - \alpha)}{1 - \beta (1 - \alpha) (1 + \varepsilon_l)}
\]

\( \square \)
Proof of Proposition 5. Using the results from Proposition 4, a worker’s expected utility, \( \mathcal{V}_n \), can be derived in the same way as in Proposition 2.

\[
\mathcal{V}_n = n \left[ 1 + \beta (1 - \alpha) \right]^{(1 + \epsilon_l)} \mu_q^{(1 - s)} (1 + \epsilon_l) E \left[ \epsilon^s (1 + \epsilon_l) \right] \frac{1}{1 - \beta (1 - \alpha) (1 + \epsilon_l)} \left[ sB \right]^{\beta \left(1 + \epsilon_l\right) (1 - \alpha) \left(1 + \epsilon_l\right)}
\]

where \( B = \frac{\alpha + \beta (1 - \alpha)}{1 + s \alpha \left(1 + \epsilon_l\right)} \). The expected after-tax income for an individual with investment cost \( n \) can be derived similarly.

\[
(1 - \tau) \mathcal{Z}_n = n \left[ 1 + \beta (1 - \alpha) \right]^{(1 + \epsilon_l)} \mu_q^{(1 - s)} (1 + \epsilon_l) E \left[ \epsilon^s (1 + \epsilon_l) \right] \frac{1}{1 - \beta (1 - \alpha) (1 + \epsilon_l)} \left[ sB \right]^{\beta \left(1 + \epsilon_l\right) (1 - \alpha) \left(1 + \epsilon_l\right)}
\]

The welfare of workers with ability \( n \) can then be re-written in terms of income, and weighted by \( \psi_n \).

\[
\psi_n \mathcal{V}_n = (1 - \tau) \psi_n \mathcal{Z}_n \left[ 1 - \left(1 + \epsilon_l\right) sB \right] \frac{1 + \epsilon_l}{1 + \epsilon_l} + \tau \psi_n \mathcal{Z}
\]

Differentiating \( \psi_n \mathcal{V}_n \) with respect to \( 1 - \tau \), we obtain the effects on welfare of both the mechanical transfer and the distortion from the unproductive component of investment, which is built into \( \mathcal{V}_n \). Then taking the expectation over ability types, \( n \), we obtain:

\[
\text{MEU} = E \left[ \mathcal{Z}_n \psi_n \right] \left[ \frac{1}{1 + s \alpha \left(1 + \epsilon_l\right)} \right] - \bar{\psi} \mathcal{Z}
\]

Next, we can calculate the belief externality. This is again captured by the effect via \( \mu_q \).

Using the elasticities from Proposition 4 and the expression for \( \mathcal{V}_n \), the effect on the welfare of a worker with ability \( n \) is:

\[
\frac{(1 + \epsilon_l) (1 - s)}{1 - \beta \alpha (1 - \alpha) (1 + \epsilon_l)} \mathcal{V}_n - \tau \mathcal{Z} \frac{\beta (1 - \alpha) (1 + \epsilon_l)}{1 - \beta (1 - \alpha) (1 + \epsilon_l)} \mu_q \frac{1}{1 - \tau}.
\]

Weighting by \( \psi_n \), using the expression for \( \mathcal{V}_n \) and taking the expectation over ability types, this gives us the total belief externality.

\[
\text{BE} = (1 - s) E \left[ \mathcal{Z}_n \psi_n \right] \left[ \frac{1}{1 + s \alpha \left(1 + \epsilon_l\right)} \right] \frac{\beta (1 - \alpha) (1 + \epsilon_l)}{1 - \beta (1 - \alpha) (1 + \epsilon_l)}
\]
Finally, the fiscal externality follows from the elasticity of income.

\[ \text{FE} = \tau \left[ \varepsilon_l + (1 + \varepsilon_l) \beta (1 - \alpha) \right] \frac{z}{1 - \beta (1 - \alpha) (1 + \varepsilon_l)} \]

By the same argument as Proposition 2, the sum of BE, MEU and FE must be zero for \( t \) to be optimal, which yields the result.

\[ \frac{\tau}{1 - \tau} = 1 - E_n \left( \frac{z_j}{z} \frac{p_j}{q} \right) \left[ \frac{1}{1 + s \alpha (1 + \varepsilon_l)} \right] \left[ 1 + (1 - s) \left( \frac{1 + \varepsilon_l) \beta (1 - \alpha)}{1 - \beta (1 - \alpha) (1 + \varepsilon_l)} \right) \right] \]

**Proof of Lemma 3.** Firm beliefs about the distribution of productivity in the population must be confirmed in equilibrium and identical across firms. Let \( \pi \) denote the equilibrium set of beliefs. Firm \( j \)'s expectation of the worker's productivity is \( E [q | \theta, \pi, A_j = 1] \geq 0 \). Finally, let \( \tilde{u} (C_j) = u (z_j - T (z_j), l_j) \) represent the utility that the worker receives from accepting offer \( C_j \).

Suppose that some firm \( j \) makes strictly positive expected profits given its contract offer \( C_j \). It must then be the case that \( \tilde{u} (C_j) \geq \tilde{u} (C_k) \) for all contracts \( C_k \) offered by other firms. There are several cases to consider, each of which will lead to a contradiction.

**Case 1:** \( \tilde{u} (C_j) > \tilde{u} (C_k) \) for some \( C_k \).

Firm \( k \) initially earns zero expected profit, since not workers accept its offer. However, it can replicate \( C_j \) but slightly reduce \( l_j \). By doing so, it attracts the worker with probability one and earns strictly positive profits. This is a profitable deviation.

**Case 2:** \( \tilde{u} (C_j) = \tilde{u} (C_k) \) for all \( C_k \), and \( \overline{p}_{k, \theta} \leq 0 \) for some \( k \).

If any firm makes weakly negative profits, then the same deviation as Case 1 applies.

**Case 3:** \( \tilde{u} (C_j) = \tilde{u} (C_k) \) and \( \overline{p}_{k, \theta} > 0 \) for all \( k \).

Since the worker always accepts an offer, \( E [q | \theta, \pi, C_j] \) is bounded weakly below
for at least one firm. This firm’s expected profit is bounded below $\bar{P}_{\text{MAX}}$.

$$\bar{P}_{\text{MAX}} = \max_{l,z} E[q|\theta, \pi] l - z \quad \text{s.t.} \quad u(z - T(z), l) \geq u(T(0), 0)$$

The assumptions on the worker’s utility function ensure that this yields finite labor supply for any finite $E[q|\theta, \pi]$. Since $z_j$ is restricted to be greater than zero and $E[q|\theta, \pi]$ is finite, $\bar{P}_{\text{MAX}}$ is also bounded. Finally, this firm can strictly increase its profit by reducing $l_j$ slightly and attracting the worker with probability one.

Since every case in which a firm makes a strictly positive expected profit implies a profitable deviation, and all firms can obtain at least zero expected profit by offering a contract with $z_j = 0$, it must be true that every firm makes zero expected profit.

Next consider two cases for the worker’s effective wage and labor supply.

Case A: One firm hires the worker with probability one.

If one firm $j$ always hires the worker in equilibrium, zero profit implies directly that the worker’s wage is her expected marginal product.

$$w_j = \frac{z_j}{l_j} = E[q|\theta, \pi]$$

Next, suppose that $C_j$ specifies a labor supply $l_j \notin L^*$ where:

$$L^* = \arg\max_{l_j} u(E[q|\theta, \pi] l_j - T(E[q|\theta, \pi] l_j), \tilde{l}_j).$$

Some other firm $k$ could offer a contract with the same implied wage as $C_j$ but with $l_k \in L^*$. Since $w_j = E[q|\theta, \pi]$, this produces zero profits but the worker’s utility is strictly higher. Firm $k$ can now increase $l_k$ slightly, thereby attracting the worker with probability one and earning strictly positive profit. Thus, it must be that $l_j \in L^*$.

Case B: Multiple firms hire the worker with positive probability.

Since each firm earns zero profit, a similar wage condition must hold for firms who
hire a worker with positive probability.

\[ w_j = \frac{z_j}{l_j} = E [q|\theta, \pi, A_j = 1] \forall j \]

Moreover, similar logic to above implies that \( l_j \in L_j^* \) where:

\[ L_j^* = \arg\max_{l_j} u (E [q|\theta, \pi, A_j = 1] \tilde{l}_j - T (E [q|\theta, \pi, A_j = 1] \tilde{l}_j), \tilde{l}_j). \]

Otherwise, firm \( j \) could offer a contract with the same implied wage but with \( l_j \in L_j^* \), so that \( \tilde{u} (C_j) \) is higher than before. It could then slightly increase \( l_j \). The worker would always accept the firm’s offer and it earns strictly positive expected profit.

Next, suppose \( E [q|\theta, \pi, A_j = 1] > E [q|\theta, \pi, A_k = 1] \) for some firms \( j \) and \( k \). For at least one pair, it must be that \( E [q|\theta, \pi, A_j = 1] > E [q|\theta, \pi] > E [q|\theta, \pi, A_k = 1] \). Let \( l_j^* \in L_j^* \) be the labor supply offered by firm \( j \). By the definition of \( L_j^* \) we know that:

\[ u (w_j l_j^* - T (w_j l_j^*), l_j^*) \geq u (w_k l_k^* - T (w_k l_k^*), l_k^*) . \]

Suppose now that \( u (w_j l_j^* - T (w_j l_j^*), l_j^*) \leq u (w_k l_k^* - T (w_k l_k^*), l_k^*) \). Then firm \( j \) can alter its offer to \( z_j = w_k l_k^* < w_j l_j^* \) and set \( l_j \) below but arbitrarily close to \( l_k \). Firm \( j \) then attracts the worker with probability one. Since \( E [q|\theta, \pi] > E [q|\theta, \pi, A_k = 1] \), firm \( j \) can make strictly positive profit with this strategy.

Alternatively, suppose that \( u (w_j l_j^* - T (w_j l_j^*), l_j^*) > u (w_k l_k^* - T (w_k l_k^*), l_k^*) \), which implies that \( u (w_j l_j^* - T (w_j l_j^*), l_j^*) > u (w_k l_k^* - T (w_k l_k^*), l_k^*) \). This is a contradiction since we assumed that both firms attract the worker with positive probability, which requires that \( u (w_j l_j^* - T (w_j l_j^*), l_j^*) = u (w_k l_k^* - T (w_k l_k^*), l_k^*) \).

In conclusion, firms must earn zero expected profit, and \( E [q|\theta, \pi, A_j = 1] = E [q|\theta, \pi] \). \( \square \)

**Proof of Proposition 16.** I begin by establishing that there is an equilibrium with zero investment. The stated assumptions ensure that \( w (\theta|\pi) \) is strictly increasing in \( \pi \), that \( w (\theta|0) = 0 \) for all \( \theta \) and that \( w (\theta|1) = \omega \) for all \( \theta \). This guarantees that \( v_q (0|\tau) = v_u (0|\tau) \) and
\( v_q(1|\tau) = v_u(1|\tau) \), which in turn implies that \( G(\beta(0|\tau)) = 0 \) and \( G(\beta(1|\tau)) = 0 \). Thus, there is a solution with no investment and no solution in which all agents invest.

Finally, if \( G(\beta(\pi|\tau)) > \pi \) for some \( \pi^* \) then the continuity of \( \phi(\theta) \) and \( G \) combined with the fact that \( G(\beta(1|\tau)) = 0 \) implies that \( G(\beta(\hat{\pi}|\tau)) = \hat{\pi} \) for some \( \hat{\pi} > \pi^* \). There are therefore at least two solutions to equilibrium condition A.11.

**Proof of Proposition 17.** Social welfare is given by:

\[
\pi \bar{v}_q(\pi) + (1 - \pi) \bar{v}_u(\pi) - \int_0^{v_q(\pi) - v_u(\pi)} kdG(k).
\]

where:

\[
\bar{v}_q(\pi|\tau) = \int_0^1 v(\theta|\pi) dF_q(\theta) - k
\]

\[
\bar{v}_u(\pi|\tau) = \int_0^1 v(\theta|\pi) dF_u(\theta).
\]

By differentiating the equation for the worker’s wage, it can be shown that the wage is increasing in \( \pi \).

\[
\frac{\partial w(\theta|\pi)}{\partial \pi} = \omega \times \frac{f_u(\theta) f_q(\theta)}{[\pi f_q(\theta) + (1 - \pi) f_u(\theta)]^2} > 0
\]

In turn, this means that \( v(\theta|\pi, \tau) = u((1 - \tau) w(\theta|\pi) + \tau \tau w) \) is increasing in \( \pi \). Thus, holding investment decisions and \( \tau \) constant, welfare increases with \( \pi \). The accompanying change in individual investment decisions can only make those marginal individuals better off. Thus, welfare is higher for all workers.

Next, let \( \pi^*(\tau) \) be the investment rate in the planner’s preferred equilibrium for each tax rate. The proof that \( \pi^*(\tau) \) rises as with \( \tau \) falls is simple. First, if \( \pi^*(\tau) = 0 \), it cannot fall. Alternatively, suppose that \( \pi^*(\tau_0) > 0 \). Since lowering \( \tau \) from \( \tau_0 \) to \( \tau_1 \) raises \( G(\beta(\pi|\tau)) \) for any \( \pi \), it must be true that \( G(\beta(\pi^*(\tau_0)|\tau_1)) > \pi^*(\tau_0) \). Since \( G(\beta(\pi|\tau)) \) is continuous and \( G(\beta(1|\tau)) = 0 \), there must be some higher investment rate \( \hat{\pi} \) such that \( G(\beta(\hat{\pi}|\tau_1)) = \hat{\pi} > \pi^*(\tau_0) \). Since the equilibrium with the highest investment rate Pareto dominates all others, the planner’s preferred equilibrium now features a higher investment rate.
Proof of Proposition 18. Just as in Sections 1.3 and 1.4, there are three effects from a fall in $\tau$.

First, there is a mechanical effect. For a worker with signal $\theta$, this is as follows.

$$\frac{\partial v(\theta)}{\partial (1-\tau)} = u' \left[ (1-\tau) \omega \frac{\pi f_q(\theta)}{\pi f_q(\theta) + (1-\pi) f_u(\theta)} + \tau \pi \omega \right] \left[ \frac{\pi f_q(\theta)}{\pi f_q(\theta) + (1-\pi) f_u(\theta)} - \pi \right] \omega$$

$$= u' \omega (1-\tau) \left[ \frac{f_q(\theta) - f_u(\theta)}{\pi f_q(\theta) + (1-\pi) f_u(\theta)} \right]$$

Aggregating up, we obtain the total mechanical effect on social welfare.

$$ME = \omega \pi (1-\pi) \int_0^1 u'_\theta [f_q(\theta) - f_u(\theta)] d\theta = -\omega \pi \bar{v}_{\tau}$$

Next, there is a fiscal externality. Assuming $\pi^+(\tau)$ is locally continuous, this is given by:

$$FE = \tau \frac{d\pi}{d(1-\tau)} \omega \int_0^1 u'_\theta [\pi f_q(\theta) + (1-\pi) f_u(\theta)] d\theta = \frac{\tau}{1-\tau} \pi \varepsilon \omega \bar{v}'_{\theta}$$

Finally, there is the externality via employer beliefs, which raises wages for all workers but is not taken into account when workers optimize. Using the continuity of $\pi^+(\tau)$ again:

$$BE = (1-\tau) \frac{d\pi}{d(1-\tau)} \int_0^1 u'_\theta \left[ \frac{\partial w(\theta)}{\partial \pi} \right] \left[ \pi f_q(\theta) + (1-\pi) f_u(\theta) \right] d\theta$$

$$= \varepsilon \pi \omega \int_0^1 u'_\theta \left[ \frac{f_u(\theta) f_q(\theta)}{\pi f_q(\theta) + (1-\pi) f_u(\theta)} \right] d\theta = \varepsilon \pi \bar{w}$$

Adding the three effects and re-arranging yields the following first-order condition.

$$\frac{\tau}{1-\tau} = \frac{(1-\pi) \int_0^1 u'_\theta [f_u(\theta) - f_q(\theta)] d\theta - \varepsilon \int_0^1 u'_\theta \left[ \frac{f_u(\theta) f_q(\theta)}{\pi f_q(\theta) + (1-\pi) f_u(\theta)} \right] d\theta}{\varepsilon \int_0^1 u'_\theta \left[ \pi f_q(\theta) + (1-\pi) f_u(\theta) \right] d\theta} = \frac{\bar{v}_{\tau} - \varepsilon \bar{w}}{\varepsilon}$$

Proof of Proposition 19. Fixing a value of $T_{A \rightarrow D}$, the proof that condition A.13 must hold at the optimum is analogous to the proof of Proposition 18. A similar perturbation argument can be used to establish that condition A.14 must hold. An increase in $T_{A \rightarrow D}$ leads to the
following gain in welfare for type $A$ and $D$ individuals:

$$
-\Delta_A = \frac{1}{\lambda_A} \int_0^1 u'_{A,\theta} dF(\theta) \\
\Delta_D = \frac{1}{\lambda_D} \int_0^1 u'_{D,\theta} dF(\theta)
$$

The welfare gain, $\lambda_D \Delta_D - \lambda_A \Delta_A$, must be zero at interior optima if $\pi^*(T)$ is locally continuous, implying condition A.14.

A.8 Simulation of the Model

This appendix provides detailed information on the methods I use to simulate the full model. The first step is to discretize the signal space into $n_\theta$ possible values, and categorize individuals into $n_q$ groups, each with a different productivity decision. I then use the noise and productivity distributions to define an $n_q \times n_\theta$ matrix $B_0$, which maps productivity decisions to distributions of realized signals.

A.8.1 Evaluation of a Single Perturbation

Evaluation of a perturbation proceeds as follows. First, define a perturbation that raises the tax rate on income between $\underline{z}$ and $\overline{z}$ by $\Delta T'$. This yields a new tax schedule, $T_1$.

$$
T_1'(z) = \begin{cases} 
T_0'(z) + \Delta T' & \text{if } z \in [\underline{z}, \overline{z}) \\
T_0'(z) & \text{otherwise}
\end{cases}
$$

Take the existing wage given each $\theta$ but apply $T_1$ instead of $T_0$. Re-optimize labor supply decisions and calculate $v (w (\theta | \pi_0) | T_1)$ for each $\theta$, yielding a candidate vector of utilities $v_1^{(0)}$. Using $v_1^{(0)}$, calculate $E_\theta (v (\theta | \pi_0, T_1) | q)$ and adjust workers’ investment decisions toward their preferred choice. This yields a new distribution of productivity, $\delta_1^{(0)} (q | \pi_0, T_1)$.

In the discretized space, $\delta_1^{(0)} (q | \pi_0, T_1)$ implies a new candidate vector of productivity choices $q_1^{(1)}$. Use these choices to reconstruct a new candidate $B_1^{(1)}$ matrix. Then solve for employers’ rational productivity inferences at each value of $\theta$, yielding a candidate set of
employer beliefs \( \pi_1^{(1)}(q) \) and a hypothesized vector of wages \( w_1^{(1)} \).

\[
w_1^{(1)} = \left[ \text{diag} \left( B_1^{(1)r} \times \delta_1^{(1)} \left( q|\pi_1^{(1)}, T_1 \right) \right) \right]^{-1} \times \left[ B_1^{(1)r} \times \text{diag} \left( q_1^{(1)} \right) \times \delta_1^{(1)} \left( q|\pi_1^{(1)}, T_1 \right) \right]
\]

Recalculate utilities to obtain \( v_1^{(1)} \) and adjust workers’ investment decisions again, yielding \( q^{(2)} \). Iterate this process until individuals do not want to adjust their investment decisions given the hypothesized employer beliefs: i.e., when \( \pi_1^{(k)}(q) \approx \delta_1^{(k)} \left( q|\pi_1^{(k)}, T_1 \right) \). At this point, the process has converged.

Once this inner fixed point has been obtained, compare the new value of expected utility for each level of costs, weight using the assumed social welfare function, and adopt the perturbation if and only if it produced an increase in average social welfare.

### A.8.2 Decomposition of a Perturbation

The effect of a perturbation on equilibrium social welfare can be decomposed into its three components: the mechanical effect (ME), the fiscal externality (FE) and the belief externality (BE). To calculate the mechanical effect, simply hold all decisions (wages, labor supply and investment) constant and evaluate the mechanical change in utility. The belief externality can be calculated by comparing the true gain in expected utility to the gain holding fixed the wage paid at each level of \( \theta \). Finally, the fiscal externality can be evaluated by subtracting the behavioral effect on tax revenue from all individuals’ incomes.

### A.8.3 Solving for the Optimal Tax Schedule

To solve for the optimal tax schedule, simply consider a series of perturbations as defined above. Define a size for each perturbation, \( \Delta T \). Then divide the income distribution into \( n_b \) tax brackets. Loop through the tax brackets and calculate the gain in welfare from a perturbation in each direction. Adopt the perturbation that increases welfare, then move to the next bracket. Repeat until there are no perturbations that increase welfare. Optionally, reduce the size of each perturbation and repeat.
A.8.4 Recovery of Fundamentals

To back out fundamentals for the simulation described in Section 1.5, I begin with the Pareto log-normal approximation of the United States wage distribution provided by Mankiw et al. (2009). Next, I use this wage distribution, and the posited log-normal conditional signal distribution, to infer a productivity distribution that produces this wage schedule.

The specific procedure that I follow is to parameterize a Champernowne distribution for log wages with density proportional to:

\[
\frac{1}{\frac{1}{2} \exp (\alpha (z - z_0)) + \lambda + \frac{1}{2} \exp (-\alpha (z - z_0))}
\]

To choose the parameters, I use MATLAB’s \texttt{fminunc} function to solve for the set of parameters that jointly minimize the Kullback-Leibler divergence between the target wage distribution \(f_w\) and the simulated distribution \(f_{\text{sim}}^w\).

\[
D_{KL} \left( f_w || f_{\text{sim}}^w \right) = \sum_w f_w (w) \ln \left( \frac{f_w (w)}{f_{\text{sim}}^w (w)} \right)
\]

As Figure 1.7 shows, this process is effective.

For each wage, I can then calculate utility \(v (w (\theta | \pi) | T_0)\), given an initial tax system \(T_0\), by solving workers’ labor supply problems for each value of \(\theta\). Expected utility for each level of productivity is then given by:

\[
E_{\theta} (v (\theta | \pi_0, T_0) | q) = \frac{B_0}{n_q \times n_p} \times \frac{v_0}{n_p \times 1}
\]

where \(v_0\) is a vector that stacks the utility realized at each value of \(\theta\) and \(\pi_0\) denotes employers’ current and correct beliefs about the distribution of productivity. Combined with individuals’ productivity choices and a value of \(\beta\), this vector of expected utilities can then be used to back out an implied cost distribution.

A.8.5 Additional Figures

Table A.1 provides summary statistics for data used to test for employer learning in Section 1.5. Figures A.3 to A.5 compare the mechanical effect, fiscal externality and belief externality
between the naïve and optimal tax schedules, in each tax bracket for the simulation in Section 1.5. Additionally, Figure A.6 plots the change in marginal social welfare weights starting from naïve taxation and transitioning to optimal taxation.

### Table A.1: Summary Statistics for High and Low AFQT Workers

<table>
<thead>
<tr>
<th></th>
<th>Low AFQT</th>
<th></th>
<th>High AFQT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>AFQT</td>
<td>-0.67</td>
<td>(0.67)</td>
<td>1.00</td>
<td>(0.40)</td>
</tr>
<tr>
<td>Log(wage)</td>
<td>6.68</td>
<td>(0.46)</td>
<td>7.03</td>
<td>(0.54)</td>
</tr>
<tr>
<td>Experience</td>
<td>7.21</td>
<td>(5.99)</td>
<td>8.13</td>
<td>(6.05)</td>
</tr>
<tr>
<td>Years since left school</td>
<td>10.55</td>
<td>(6.49)</td>
<td>9.74</td>
<td>(6.29)</td>
</tr>
<tr>
<td>Urban (%)</td>
<td>74.4</td>
<td></td>
<td>78.6</td>
<td></td>
</tr>
<tr>
<td>Education (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– 12 years</td>
<td>59.7</td>
<td></td>
<td>35.7</td>
<td></td>
</tr>
<tr>
<td>– 16 years</td>
<td>3.8</td>
<td></td>
<td>25.4</td>
<td></td>
</tr>
<tr>
<td>– Other</td>
<td>36.5</td>
<td></td>
<td>38.9</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>18921</td>
<td></td>
<td>18903</td>
<td></td>
</tr>
</tbody>
</table>

* * p < 0.10, ** p < 0.05, *** p < 0.01

**Table notes.** Data are from the National Longitudinal Survey of Youth (NLSY79). The sample is restricted to working black and white men who have wages between one and one hundred dollars and at least eight years of schooling. AFQT is a worker’s score on the armed forces qualification test, standardized by age to have zero mean and unit standard deviation. Experience is measured in years.

### A.8.6 Alternative Parameter Values

This subsection discusses simulations with alternative sets of parameters. As predicted by the theoretical results in Sections 1.3 to 1.4 and Appendix A.5, the most important factors in determining the impact of taking into account the belief externality are the responsiveness of productivity relative to income, and the accuracy of employer information. Other factors, such as the level of the elasticity of taxable income, are less important.

Figure A.7 shows the results of these alternative calibrations. First, panel (a) shows a simulation identical to the baseline exercise in Section 1.5 but with changes to \( \epsilon_l \) and \( \beta \) to ensure that the elasticity of taxable income is lower, at \( \epsilon_{z LR} = 0.8 \), but that the ratio \( \epsilon_{w LR} / \epsilon_{z LR} \) is approximately unchanged at 0.6. The level of the optimal tax schedule is higher, but the shift between the naïve and optimal schedules is qualitatively unchanged. This is despite there being a very large difference in \( \epsilon_{z LR} \) between the two scenarios.

Panel (b) of figure A.7 shows the adjustment to marginal tax rates between the naïve
Figure A.3: This figure compares the fiscal externality in each tax bracket under naïve and optimal taxation, in the simulation described in Section 1.5.

Figure A.4: This figure compares the belief externality in each tax bracket under naïve and optimal taxation, in the simulation described in Section 1.5.
Figure A.5: This figure compares the mechanical effect in each tax bracket under naïve and optimal taxation, in the simulation described in Section 1.5.

Figure A.6: This figure plots the change in marginal social welfare weights starting from naïve taxation and transitioning to optimal taxation, for the simulation described in Section 1.5.
and optimal tax schedules in four scenarios. The gray dotted line shows the baseline results from Section 1.5. Next, the black line shows the adjustment in the scenario from panel (a), with a lower income elasticity. The red line shows a third case with $\varepsilon_w^{LR}/\varepsilon_z^{LR} = 0.5$, but $\varepsilon_z^{LR}$ unchanged. Finally, the blue line shows a fourth scenario with $dE(w)/dq = 0.9$, which implies a smaller belief externality. As expected, these latter two scenarios lead to a smaller adjustment to the tax schedule.
Figure A.7: This figure shows the results of exercises similar to that described in Section 1.5 but with alternative parameter values. The top panel shows optimal and naïve tax schedules when $\varepsilon_1 = 0.2$ and $\beta = 0.25$, which still yields $\varepsilon_w^{LR} / \varepsilon_z^{LR} = 0.6$ but with $\varepsilon_z^{LR} = 0.8$. The lower panel shows the adjustment to marginal tax rates in: (i) the baseline exercise; (ii) the case with $\varepsilon_z^{LR} = 0.8$; (iii) a third case with $\varepsilon_w^{LR} / \varepsilon_z^{LR} = 0.5$ but $\varepsilon_z^{LR}$ unchanged; and (iv) a fourth with $dE(w)/dq = 0.9$, which implies a smaller belief externality.
Appendix B

Appendix to Chapter 2

B.1 Proofs

Proof of Proposition 6. Given the assumptions, the worker and employer EE curves lie above their WW curves for $s$ and $t$ near zero and one respectively. For the worker EE and WW curves, we know this because: (i) our assumptions on $\phi(\theta)$ guarantee that $s^*(\pi) = 1$ for any $\pi$ below some strictly positive threshold value; (ii) there exists a threshold $\pi < 1$ above which $s^*(\pi) = 0$; and either $s = 0$ or $s = 1$ imply zero worker investment. Equivalent arguments apply for the employer WW and EE curves. In conclusion, we know that the non-zero solutions in $\pi^*(\delta)$ and $\delta^*(\pi)$ are bounded strictly between zero and one.

The conditions that $G^W(\beta^W(s,\delta)) > \phi(s)/[\chi_q/\chi_u + \phi(s)]$ and $G^E(\beta^E(t,\pi)|\lambda) > \tau(t)/[\omega_q/\omega_u + \tau(t)]$ guarantee that the EE curves and WW curves cross at least once, implying at least two non-zero solutions to each of equations 2.5 and 2.6. Since $\overline{\omega}(\delta)$ and $\overline{\chi}(\pi)$ are increasing, the same is true for any $\delta > \delta_0$ and $\pi > \overline{\pi}$. Assume, then, that $\delta_0$ and $\overline{\pi}$ are the lowest values for which these conditions hold. Below $\delta_0$ and $\overline{\pi}$, there is no non-zero solution to equations 2.5 and 2.6 respectively.

For any value of $\delta$, let $\overline{\pi}^*(\delta) = \max \{\pi^*(\delta)\}$. Similarly define $\overline{\delta}^*(\pi) = \max \{\delta^*(\pi)\}$. Both $\overline{\pi}^*(\delta)$ and $\overline{\delta}^*(\pi)$ are obviously defined on $[0,1]$. Both are also increasing in their arguments. To see why, start at $\overline{\pi}(\delta_1)$, at which $s = s_1$. Consider increasing $\delta$ to $\delta_2 > \delta_1$. For
any given $s$, $G \left( \left[ F_u (s) - F_q (s) \right] \varpi (\delta) \right)$ increases since $\varpi (\delta)$ is increasing in $\delta$. This means that $G \left( \left[ F_u (s_1) - F_q (s_1) \right] \varpi (\delta_2) \right) > \overline{\pi} (\delta_1)$. In other words, the WW curve is above the EE curve at $s_1$. Thus, since the EE curve is strictly decreasing and $G \left( \left[ F_u (0) - F_q (0) \right] \varpi (\delta_2) \right) = 0$, there must be at least one solution to the left of $s_1$, which implies a value of $\overline{\pi} (\delta_2)$ greater than $\overline{\pi} (\delta_1)$. An analogous argument can be used to show that $\delta^* (\pi)$ is increasing in $\pi$.

We directly assume that is some $\{\pi, \delta\}$ such that $\delta \in \delta^* (\pi)$ and $\pi < \overline{\pi}^* (\delta)$. Combined with the monotonicity of $\overline{\pi}^* (\delta)$ and $\delta^* (\pi)$, this implies that there is some $\pi$ such that $\overline{\pi}^* \left( \delta^* (\pi) \right) > \pi$. We also know that $\overline{\pi}^* \left( \delta^* (\pi) \right) < 1$ since $\overline{\pi}^* (\delta)$ is bounded below 1. There must therefore be a $\pi$ such that $\overline{\pi}^* \left( \delta^* (\pi) \right) = \pi$. To see why, suppose that there is not. Then there must be a downward discontinuity in $\overline{\pi}^* \left( \delta^* (\pi) \right) - \pi \leq 0$. This is impossible since $\pi$ is continuous and $\overline{\pi}^* \left( \delta^* (\pi) \right)$ is positive monotic. Since $\pi$ is a non-zero solution and $\delta = \pi = 0$ always satisfies both equations 2.5 and 2.6, there are multiple solutions to the two-sided problem.

Proof of Proposition 7. Assume that there exists an equilibrium without affirmative action in which there is positive investment in the A market: $\pi_A = \pi_A^* > 0$ and $\delta_A = \delta_A^* > 0$. In the B market, there is always an equilibrium with zero investment: $\pi_B = \delta_B = 0$. Now suppose that, under affirmative action, $\pi_A = \pi_A^*$, $\delta_A = \delta_A^*$, $\pi_B = \delta_B = 0$ and $t_B = t_B^* (0) = 1$. Then affirmative action is non-binding since, with zero workers applying, an employer’s profits are independent of $s_B$. It therefore optimally sets $s_B = s_B^{AA}$ such that the affirmative action constraint holds.

$$\rho \left( s^* (\pi_A^*), \pi_A^* \right) = \pi_A^* \left[ 1 - F_{q A}^W (s^* (\pi_A^*)) \right] + \left( 1 - \pi_A^* \right) \left[ 1 - F_u^W (s^* (\pi_A^*)) \right]$$

$$= 1 - F_u^W \left( s_B^{AA} \right)$$

$$= \rho \left( s_B^{AA}, 0 \right)$$

Regardless of $s_B^{AA}$, $\pi_B = \delta_B = 0$ is still an equilibrium, since both firms and workers have zero investment returns.
**Proof of Proposition 8.** First consider an equilibrium without affirmative action. Suppose that \( \pi_A = \pi_B = \pi > 0 \). Then employers’ unique optimal signal threshold is \( s_A = s_B = s^*(\pi) \). Worker beliefs \( \delta \) cannot be homogeneous since \( G^E(e^{t^*(\delta)}, \pi) | \lambda \) must be strictly lower for the minority for a given any threshold \( t \). Combined with \( s_A = s_B \) this is incompatible with \( \pi_j = G^W(e^{s_j \delta_j}) \) being the same for both groups. This is a contradiction. Finally, consider imposing an affirmative action target. Since the constraint does not bind if employers have homogeneous beliefs, they still set \( s_A = s_B = s^*(\pi) \). By the same logic as above, worker beliefs cannot be homogenous, which is incompatible with \( \pi_A = \pi_B \).

**Proof of Proposition 9.** If \( \delta_B \) is low enough, \( t_B \) approaches one, and Lemma 4 below implies that affirmative action lowers every firm’s threshold for Bs to some \( s < s_B \) if beliefs are held constant. Now consider the firm’s payoff conditional on a worker application.

\[
\pi_B \left( 1 - F^W_q(s) \right) \chi_q - \left( 1 - \pi_B \right) \left( 1 - F^W_u(s) \right) \chi_u \tag{B.1}
\]

Under our assumptions, low enough \( \pi_B \) ensures that \( s^*(\pi_B) = 1 \). Suppose that \( \pi_B \) is above but close to this critical value such that \( s^*(\pi_B) \approx 1 \). The firm’s expected payoff (equation B.1) is then arbitrarily close to zero. The imposition of \( s < s_B \) adds a strictly positive probability of being required to hire marginal workers who have negative expected payoffs for the firm, ensuring that the total firm payoff from hiring type B workers is negative. As a result, zero firms subsequently invest in the B amenity. In turn, this ensures that no B workers have an incentive to invest.

**Lemma 4.** Assume that \( \phi(\theta) \) and \( \tau(\psi) \) are continuous, strictly decreasing and strictly positive on \([0,1]\). For fixed current beliefs with \( \pi_A > \pi_B, \delta_A > \delta_B \) and \( t_B \) close enough to one, affirmative action lowers \( s_B \) for all firms with \( s_A \) approximately unchanged.

**Proof of Lemma 4.** The Lagrangean for an affirmative action target is as follows.

\[
L(s_A, s_B, i_A, i_B, \gamma | \pi_A, \pi_B) = \lambda_A P(s_A, \pi_A, i_A) + \lambda_B P(s_B, \pi_B, i_B) + \gamma [\rho(s_B, \pi_B) - \rho(s_A, \pi_A)]
\]
where \( \rho(s_j, \pi_j) \) is the probability that the employer assigns to hiring a randomly drawn worker from group \( j \in \{A, B\} \) and \( P(s_j, \pi_j, i_j) \) is the expected payoff from making an offer to said worker (which depends on whether the firm has invested – \( i_j \in \{q, u\} \)).

\[
\rho(s_j, \pi_j) = \pi_j \left[ 1 - F^W_q(s_j) \right] + (1 - \pi_j) \left[ 1 - F^W_u(s_j) \right]
\]

\[
P(s_j, \pi_j) = \pi_j \left[ 1 - F^E_q(t_j) \right] \left[ 1 - F^F_{i_j}(t_j) \right] \chi_q - (1 - \pi_j) \left[ 1 - F^W_u(s_j) \right] \left[ 1 - F^F_{i_j}(t_j) \right] \chi_u
\]

This is enough for us to write down the expressions for the key FOCs.

\[
\gamma \left[ \pi_A f^W_A(s_A) + (1 - \pi_A) f^W_u(s_A) \right] = \lambda_A \left[ 1 - F^E_A(t_A) \right] \left[ \pi_A f^W_q(s_A) \chi_q - (1 - \pi_A) f^W_u(s_A) \chi_u \right]
\]

\[-\gamma \left[ \pi_B f^E_q(s_B) + (1 - \pi_B) f^E_u(s_B) \right] = \lambda_B \left[ 1 - F^E_B(t_B) \right] \left[ \pi_B f^E_q(s_B) \chi_q - (1 - \pi_B) f^E_u(s_B) \chi_u \right]
\]

These can be re-arranged as follows.

\[
\left( \frac{1 - \pi_A}{\pi_A} \right) f^W_q(s_A) = \frac{\chi_q - \left( 1 - F^E_A(t_A) \right) \gamma}{\lambda_A \left[ 1 - F^E_A(t_A) \right]} = r_A(\gamma)
\]

\[
\left( \frac{1 - \pi_B}{\pi_B} \right) f^W_q(s_B) = \frac{\chi_q + \left( 1 - F^E_B(t_B) \right) \gamma}{\lambda_B \left[ 1 - F^E_B(t_B) \right]} = r_B(\gamma)
\]

These FOCs characterize the firm’s signal thresholds for any given investment decision. The threshold \( t_B \) being close to one means that \( 1 - F^E_B(t_B) \) is close to zero and all the adjustment occurs on the B side: the multiplier approaches zero in this case. Intuitively, if very few Bs apply then it is nearly costless to adjust on their margin relative to adjustment on the A side.

More formally, we know that the two signals must change in the following proportion to satisfy the affirmative action constraint.

\[
\frac{\partial s_B}{\partial s_A} = \frac{\pi_A f^W_q(s_A) + (1 - \pi_A) \pi_A f^W_u(s_A)}{\pi_B f^W_q(s_B) + (1 - \pi_B) \pi_B f^W_u(s_B)}
\]

This implies that the change in profits from an increase in \( s_A \) is proportional to:

\[
\lambda_B \left[ \frac{\pi_B f^W_q(s_B) \chi_q - (1 - \pi_B) f^W_u(s_B) \chi_u}{\pi_B f^W_q(s_B) + (1 - \pi_B) \pi_B f^W_u(s_B)} \right] \left[ 1 - F^E_B(t_B) \right] - \lambda_A \left[ \frac{\pi_A f^W_q(s_A) \chi_q - (1 - \pi_A) f^W_u(s_A) \chi_u}{\pi_A f^W_q(s_A) + (1 - \pi_A) \pi_A f^W_u(s_A)} \right] \left[ 1 - F^E_A(t_A) \right]
\]
Our assumptions on \( \phi(\theta) \) and \( \tau(\psi) \) imply that as \( t_B \to 1 \), the firm’s optimal \( s_A \) approaches \( s^* (\pi_A) \). Since the affirmative action constraint implies that \( s_B \) is strictly less than \( s_A \), \( s_B > s_A \) without affirmative action and \( s_A \) is approximately unchanged, \( s_B \) is lower for all firms with the additional constraint.

**Proof of Proposition 10.** First note that the firm investment rate is bounded strictly below \( \delta = G^E (\omega_q) < 1 \). With close-to-perfect observability of firm investment, workers can and will optimally set \( t = t^* (\delta) \) such that \( F^E_u (t^* (\delta)) \to 1 \) and \( F^E_q (t^* (\delta)) \to 0 \). The probability that a worker is successfully hired by any firm that only makes one of the two investments must therefore be approximately zero (for both types), which implies that the expected return for a firm that makes a single investment is approximately zero. Nearly all firms therefore make both investments or neither, which means that \( \delta_A \approx \delta_B \). Finally, firms that do not invest at all hire approximately zero workers.

Combined, this means that as firm investment becomes near-perfectly observable, there is only one type of firm for which the employment quota could lead to different signal thresholds that impact worker investment: firms that make both investments. Specifically, the fraction of workers who invest approaches:

\[
\pi_j \to G^W \left( \delta_j \left[ F^W_u (s^q_A) - F^W_u (s^q_B) \right] \omega_q \right).
\]

Next, for firms that made both investments, the affirmative action constraint amounts to:

\[
\pi_A \left[ 1 - F^W_q (s^q_A) \right] + (1 - \pi_A) \left[ 1 - F^W_q (s^q_A) \right] = \pi_B \left[ 1 - F^W_q (s^q_B) \right] + (1 - \pi_B) \left[ 1 - F^W_q (s^q_B) \right]
\]

Using \( \pi = G^W (\beta_W (s, \delta)) \), the probability of employment for any given \( s \) is \( \hat{\rho} (s | \delta) \).

\[
\hat{\rho} (s | \delta) = G^W \left( \hat{\delta} \left[ F^W_u (s) - F^W_q (s) \right] \omega_q \right) \left[ 1 - F^W_q (s) \right] \\
+ \left( 1 - G^W \left( \hat{\delta} \left[ F^W_u (s) - F^W_q (s) \right] \omega_q \right) \right) \left[ 1 - F^W_q (s) \right]
\]

The slope of \( \hat{\rho} (s | \delta) \) with respect to \( s \) is as follows.

\[
\hat{\rho}' (s | \delta) = G^W \cdot \delta \left[ f^W_u (s) - f^W_q (s) \right] \omega_q \left[ F^W_u (s) - F^W_q (s) \right] - \pi f^W_q (s) - (1 - \pi) f^W_u (s)
\]

This is always strictly negative if \( \phi (s) \leq 1 \), so a sufficient condition for strict monotonicity
is that \( \hat{\rho}'(s|\delta) < 0 \) for all \( \delta \in [0,1] \) and \( s : \phi(s) > 1 \). Re-arranging the expression for \( \hat{\rho}'(s|\delta) \), this requirement amounts to the following condition.

\[ \eta(\bar{\rho}(s)) < \frac{\phi(s)}{\phi(s) - 1} \]

for all \( s : \phi(s) > 1 \) where \( \eta(c) = \frac{d[c \cdot G(c)]}{dc} \) and \( \bar{\rho}(s) = \left[F^W_u(s) - F^W_q(s)\right] \omega_q \).

Finally, as \( \delta_A \to \delta_B \), non-homogeneous beliefs requires that \( \hat{\rho}(s_A|\delta_A) = \hat{\rho}(s_A|\delta_B) \) for some \( s_A \neq s_B \). But with \( \delta_A \approx \delta_B \), a necessary condition for this is that \( \hat{\rho}(s|\delta) \) is non-monotonic. If the above condition is satisfied, this is not possible. \( \square \)

**Proof of Proposition 11.** Consistent with our proposed dynamic adjustment process, fix beliefs at their original values. Given these beliefs, there are two actions that a firm can take to boost employment of minorities: (a) it can make the B investment if it was not doing so already; and (b) it can lower its hiring standard for B workers. However, if few enough minority workers are applying (i.e., low enough \( \delta_B \)), a standard of \( s_B = 0 \) still does not allow the firm to satisfy the employment quota, regardless of its investment decision. Thus, the equality constraint must entail an immediate reduction in employment of type A workers.

Next, consider the firm’s choices of \( s_A \) and \( s_B \). The firm’s problem is as follows.

\[
\begin{align*}
\max_{s_A, s_B, i_A, i_B} & \quad \lambda_A P(s_A, \pi_A, i_A) + \lambda_B P(s_B, \pi_B, i_B) \\
\text{s.t.} & \quad [\rho(s_B, \pi_B, i_B) = \rho(s_A, \pi_A, i_A)] \\
& \quad s_B \geq 0
\end{align*}
\]

In principle, there are also constraints that \( s_B \leq 1 \), \( s_A \geq 0 \) and \( s_A \leq 1 \) but these never bind.

We now prove that an internal solution does not exist for some parameter values. First, assume the inequality constraint does not bind and let \( \gamma \) be the multiplier on the equality constraint. Differentiating with respect to \( s_j \) yields the effect of raising each threshold.

\[
\begin{align*}
-\lambda_A \left[ \pi_A f^W_q(s_A) \chi_q - (1 - \pi_A) f^W_u(s_A) \chi_u \right] + \gamma \left[ \pi_A f^W_q(s_A) + (1 - \pi_A) f^W_u(s_A) \right] & \quad \text{(B.2)} \\
-\lambda_B \left[ \pi_B f^E_q(s_B) \chi_q - (1 - \pi_B) f^E_u(s_B) \chi_u \right] - \gamma \left[ \pi_B f^E_q(s_B) + (1 - \pi_B) f^E_u(s_B) \right] & \quad \text{(B.3)}
\end{align*}
\]
For low enough $\delta_B$, $t_B \to 1$ and the maximum hiring probability that a firm can achieve for B worker is $\rho_B (0, \pi_j, q) = \left[ 1 - F_t^E (t_j) \right] \to 0$. We therefore also require that $\rho (s_A, \pi_A, i_A) \to 0$, which in turn implies that $s_A \to 1$. Setting expression B.2 to zero allows us to obtain the limit for the multiplier $\gamma$ in this scenario.

$$
\gamma \to \gamma^* = \lambda_A \left[ \frac{\pi_A \chi_q - (1 - \pi_A) \phi (1) \chi_u}{\pi_A + (1 - \pi_A) \phi (1)} \right]
$$

A sufficient condition for expression B.3 to be strictly less than zero for any value of $s_B$ is that $\lambda_B \chi_u > \gamma^*$. Fixing $\pi_A$, this condition must hold for an open set of parameters and ensures that we obtain a boundary solution with $s_B = 0$. This contradicts the assumption that there is an internal solution to the firm’s problem, proving that the firm sets $s_B = 0$. This in turn ensures that no workers have an incentive to invest and that $\pi_B = 0$ subsequently.

Proof of Proposition 12. Consider the effect of setting $\{s, t\}, \{t, \delta\}$ or $\{s, \pi\}$ if $\pi_0 = \delta_0 = 0$. Even if $s \in [0, 1]$ and $t \in [0, 1]$, the investment returns of workers ($\beta^W$) and firms ($\beta^E$) are weakly negative as long as $\delta = 0$ and $\pi = 0$ respectively. Similarly, $s = 1$ and $\pi = 0$ ensures that $\beta^W = 0$ and $\beta^E = 0$, regardless of the values of $t$ and $\delta$. The same logic applies whenever both $t = 1$ and $\delta = 0$. The failure of these pairs also implies that no intervention on a single margin can succeed.

Proof of Proposition 13. First, set government signal thresholds $s^g \in (0, 1)$ and $t^g \in (0, 1)$. Next, set worker and firm incentive payments $\omega^g$ and $\chi^g$. Let $\hat{F}^W_i (\theta^g)$ and $\hat{F}^E_i (\theta^g)$, $i \in \{q, u\}$, be the distributions of $\theta^g$ and $\psi^g$ respectively, the increase in investment returns for workers and firms are:

$$
\left[ \hat{F}^W_u (s^g) - \hat{F}^W_q (s^g) \right] \cdot \omega^g
$$

$$
\left[ \hat{F}^E_u (t^g) - \hat{F}^E_q (t^g) \right] \cdot \chi^g
$$

Providing that the government signals are strictly informative, these payments can be set such that the expected returns to investment for B workers and firms investing in the B
amenity are equal to those that would prevail at \{s_A, t_A, \pi_A, \delta_A\}. In response, the fraction of B workers who invest is \(\pi_B = \pi_A\) and the fraction of firms who invest in the B amenity is \(\delta_B = \delta_A\). Then B workers set \(t = t_A\) and firms \(s = s_A\). Once \(\pi_B = \pi_A\) and \(\delta_B = \delta_A\) have been achieved, they can be retained with only the following permanent investment subsidy.

\[
\left[ F^E_u (t^g) - F^E_q (t^s) \right] \cdot \chi^s = (\lambda_A - \lambda_B) \cdot \left[ F^E_u (t) - F^E_q (t) \right] \chi (\pi_A)
\]

Clearly \(\lambda_A = \lambda_B\) ensures that the required permanent investment subsidy is zero.

**Proof of Proposition 14.** Assume \(G^W\) and \(G^E\) are strictly increasing with \(G^E (0) = G^W (0) = 0\), and take any \(\pi_B \in [0, 1)\) and \(\pi_A \in (0, 1)\) with \(\pi_B < \pi_A\). These worker investment levels, combined with signal distributions and threshold rules \(t^s\) and \(s^s\) pin down firm investment returns for any \(\delta_j\). The fractions of firms that invest in each amenity are:

\[
\delta_j = G^E \left( \left[ F^E_u (t^s (\delta_j)) - F^E_q (t^s (\delta_j)) \right] \chi (\pi_j) \right)
\]

If \(\pi_B = 0\) then \(\delta_B = 0\) for any \(G^E\) satisfying our assumptions, which ensures equilibrium in the B market. For any \(\pi_j > 0\), there always exists a set of worker payoffs \((\omega_q, \omega_u)\) and a distribution function \(G^E\) such that this equation is solved by any \(\delta_A \in (0, 1)\) and \(\delta_B \in (0, 1)\) with \(\delta_B < \delta_A\), given \(\pi_A\) and \(\pi_B\). Combined with the worker and firm threshold rules, \(\delta_A\) and \(\delta_B\) pin down worker investment. The fractions of workers who invest are:

\[
\pi_j = G^W \left( \left[ F^W_u (s^s (\pi_j)) - F^W_q (s^s (\pi_j)) \right] \chi (\pi_j) \right)
\]

Since \(\delta_A > \delta_B > 0\), there always exists a function \(G^W\) that satisfies our assumptions and for which \(\pi_A\) and \(\pi_B\) satisfy this equation given firm investment rates \(\delta_A\) and \(\delta_B\) respectively.

Next, the maximum level of worker investment that can be achieved with firm investment incentives alone is as follows.

\[
\pi_{B,t} = G^W \left( \left[ F^W_u (s^s (\pi_{B,t-1})) - F^W_q (s^s (\pi_{B,t-1})) \right] \omega_q \right)
\]

Clearly the maximum difference between worker investment rates occurs as \(\pi_{B,t-1} \to 0\)
and \( \pi_{A,t-1} \to 1 \). Moreover, equilibrium worker investment \( \pi_{B,t-1} \) close to zero implies, for strictly positive \( \omega_q \), that \( F^W_t(s^* (\pi_{B,t-1})) - F^W_q(s^* (\pi_{B,t-1})) \approx 0 \). This ensures that, for any finite \( \omega_q \), \( \pi_{B,t} \) is also arbitrarily close to zero. The same logic implies that \( \pi_{B,t+1} \approx 0 \), given that \( \pi_{B,t} \approx 0 \). Fixing a finite time horizon \( T \), a low enough \( \pi_{B,t} \) therefore ensures that no one-sided investment incentive can achieve equality by time \( T \).

Similar but more complex logic applies when affirmative action is allowed. As \( \pi_B \to 0 \), firm investment returns approach zero and thus \( \delta_B \to 0 \). In turn, this implies that \( t^* (\delta_B) \to 1 \). Firms therefore respond to AA by changing \( s_B \) to \( \hat{s}_B \), with \( s_A \) unchanged (see Lemma in Appendix B.1). Now assume a signal distribution \( F_i (\theta) \) such that \( \phi (s_A) > 1 \). This implies that \( F^W_t(\hat{s}_B) - F^W_q(\hat{s}_B) < F^W_t(s_A) - F^W_q(s_A) \) for any \( \hat{s}_B < s_A \). If we then assume that \( G^W (x) \approx 0 \) for any \( x < F^W_t(s_A) - F^W_q(s_A) \), this ensures that \( \pi_{B,t} \approx 0 \). The same logic implies that \( \pi_{B,t+1} \approx 0 \), given that \( \pi_{B,t} \approx 0 \). Fixing a finite time horizon \( T \), a low enough \( \pi_{B,t} \) therefore ensures that no one-sided investment incentive can achieve equality by time \( T \), even if combined with affirmative action. \( \square \)

**Proof of Proposition 15.** The log wage of a worker at her new firm is given by the following equation.

\[
\ln w_i = \left( \frac{\sigma^2_{a,j}}{\sigma^2_{e,j} + \sigma^2_{a,j}} \right) \ln a_i + \left( \frac{\sigma^2_{e,j}}{\sigma^2_{e,j} + \sigma^2_{a,j}} \right) \mu_{a,j} + \ln \gamma + (1 - \gamma) \ln k_{j,NEW} \\
+ \frac{1}{2} \left( \frac{\sigma^2_{e,j} \sigma^2_{a,j}}{\sigma^2_{e,j} + \sigma^2_{a,j}} \right) \ln \varepsilon_i
\]

Similarly, her wage at her past firm is as follows.

\[
\ln w^i_{t} = \left( \frac{\sigma^2_{a,j}}{\sigma^2_{e,j,OLD} + \sigma^2_{a,j}} \right) \ln a_i + \left( \frac{\sigma^2_{e,j,OLD}}{\sigma^2_{e,j,OLD} + \sigma^2_{a,j}} \right) \mu_{a,j} + \ln \gamma + (1 - \gamma) \ln k_{j,OLD} \\
+ \frac{1}{2} \left( \frac{\sigma^2_{e,j,OLD} \sigma^2_{a,j}}{\sigma^2_{e,j,OLD} + \sigma^2_{a,j}} \right) \ln \varepsilon_i
\]
This can be re-arranged to isolate ability.

\[
\ln a_i^{OLD} = \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{a,j}^2} \right) \ln w_i^{OLD} - \left( \frac{\sigma_{e,j,OLD}^2}{\sigma_{a,j}^2} \right) \mu_{a,j} - \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{a,j}^2} \right) \ln \gamma \\
- (1 - \gamma) \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{a,j}^2} \right) \ln k_{j,OLD} - \frac{1}{2} \sigma_{e,j,OLD}^2 - \ln \epsilon_i^{OLD}
\]

Since \( \ln a_i = c_j + \rho \ln a_i^{OLD} + \ln \eta_i \), this allows us to write current ability as a function of the past wage.

\[
\ln a_i = c_j + \rho \left[ \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \ln w_i^{OLD} - \left( \frac{\sigma_{e,j,OLD}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \mu_{a,j} - \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{a,j}^2} \right) \ln \gamma \\
- (1 - \gamma) \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{a,j}^2} \right) \ln k_{j,OLD} - \frac{1}{2} \sigma_{e,j,OLD}^2 - \ln \epsilon_i^{OLD} \right] + \ln \eta_i
\]

Finally, we can substitute this measure of ability into the equation for the wage at the current firm to obtain the following regression equation:

\[
\ln w_i = \rho \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \ln w_i^{OLD} + \alpha_{j,OLD} + \alpha_{j,NEW} + \nu_i
\]

where:

\[
\alpha_{j,NEW} = \rho \left[ \left( \frac{\sigma_{e,j}^2 - \sigma_{e,j,OLD}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \mu_{a,j} + \left[ 1 - \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \right] \ln \gamma + (1 - \gamma) k_{j,NEW} \right] + \left( \frac{\sigma_{e,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) c_j
\]

\[
\alpha_{j,OLD} = -\rho \left[ (1 - \gamma) \left( \frac{\sigma_{e,j,OLD}^2 + \sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \ln k_{j,OLD} \right]
\]

\[
\nu_i = \rho \left( \frac{\sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \ln \epsilon_i - \ln \epsilon_i^{OLD} \right) - \left( \frac{\sigma_{a,j}^2}{\sigma_{e,j}^2 + \sigma_{a,j}^2} \right) \ln \eta_i
\]

It follows directly that the difference in coefficients will be as shown in the proposition. □
B.2 Stability

We can use the learning process described in Section 1.2 to analyze the robustness of equilibria to an arbitrary but small perturbation to both firm and worker investments. Under the hypothesized adjustment process and the assumptions of our existence proposition, the zero investment equilibrium is always locally stable. To see this, observe that below some strictly positive \( d \), no worker applies because workers are too pessimistic about firm investment. Similarly, low enough \( \pi \) guarantees that no firms make offers. As long as \( d \) and \( \pi \) remain below these critical values, there is no incentive for either party to invest.

To analyze stability more generally, we can linearize this two-dimensional system around the equilibrium. For ease of exposition, define the following derivatives.

\[
\begin{align*}
W W_0^1 &= G^W \left[ f_u^W (s^*(\pi)) - f_q^W (s^*(\pi)) \right] \\
E E_0^1 &= 1/s''(\pi) \\
R R_0^1 &= \varphi' (\delta) \left[ F_u^W (s^*(\pi)) - F_q^W (s^*(\pi)) \right] G^W \\
W W_0^2 &= G^E \left[ f_u^E (t^*(\delta)) - f_q^E (t^*(\delta)) \right] \\
E E_0^2 &= 1/t''(\delta) \\
R R_0^2 &= \varphi' (t^*(\delta)) \lambda \left[ F_u^E (t^*(\delta)) - F_q^E (t^*(\delta)) \right] G^E
\end{align*}
\]

Intuitively, \( W W_0^1 \) is the slope of the \( WW \) curve in Figure 2.2 and captures the impact of a less favorable (higher) firm signal threshold on worker incentives. Similarly, \( E E_0^1 \) is the slope of the \( EE \) curve and captures the effect of lower worker investment on the signal threshold that firms optimally set. The direct impact of higher firm investment on the worker’s payoff from being hired is \( R R_0^1 \) and could be shown by an upward shift in the \( WW \) curve. The firm equivalents – \( W W_0^2, E E_0^2 \) and \( R R_0^2 \) – are analogous.

These definitions allow us to write the Jacobian of the linearized system compactly.

\[
\begin{bmatrix}
W W_0^1 \frac{1}{E E_0^1} & R R_0^1 \\
R R_0^2 & W W_0^2 \frac{1}{E E_0^2}
\end{bmatrix}
\]

The system is stable if both eigenvalues of this matrix have absolute values strictly less than one, and the following condition is necessary and sufficient for this (Neusser, 2016).

\[
\left| W W_0^1 \frac{1}{E E_0^1} + W W_0^2 \frac{1}{E E_0^2} \right| < 1 + \left( W W_0^1 \frac{1}{E E_0^1} \cdot W W_0^2 \frac{1}{E E_0^2} \right) - (R R_0^1 \cdot R R_0^2) < 2.
\]
Since both eigenvalues strictly less than one guarantees that the equilibrium is hyperbolic, this is also sufficient for the non-linear system to be locally asymptotically stable.

To understand the stability condition, consider two special cases. First, suppose that worker and firm signal thresholds are locally unresponsive to changes in investment: i.e., the two thresholds $s^*$ and $t^*$ are approximately fixed. In the limit, this implies that $1/EE_1' \to 0$ and $1/EE_2' \to 0$, which causes the condition for stability to collapse to: $-1 < RR_1' : RR_2' < 1$. This condition is intuitive. If worker and firm payoffs $\overline{w}$ and $\overline{c}$ change too sharply with each others’ investments, then a small perturbation causes a reinforcing dynamic through beliefs, which moves the system further away from the original equilibrium.

It is also instructive to consider another extreme in which worker investment responds very strongly to the firm signal threshold $s^*$, which in turn is highly responsive to worker investment. This implies that $WW'/EE' > 1$, violating the first inequality of the stability condition. Instability arises in this case because a small perturbation to worker investment is compounded through changes in firms’ hiring thresholds.

This simple example with uniform distributions (Section 1.3) permits a particularly transparent discussion of dynamics, since it in fact corresponds to the special case in which signal thresholds are locally unresponsive to beliefs. The phase arrows in Figure 2.3 show the direction of adjustment in any given region. With the assumptions of this example, an equilibrium is stable if and only if the solid line is flatter than the dotted line: i.e., only the two extreme equilibria, $S$ and $Z$, are stable. To understand the instability of the equilibrium at $U$, imagine a small upward perturbation to both $\pi$ and $\delta$ such that the economy is at a point above the dotted line and below the solid line. From here, the fraction of firms and workers investing both increase further, moving us further away from $U$. 

181
B.3 Derivations

B.3.1 Normalization of Worker Payoffs

A worker will apply for a job if the expected benefit is better than her outside option.

\[
\left(1 - F^W_q(s)\right) \xi(\delta, \psi) w_q + \left(1 - F^W_q(s)\right) (1 - \xi(\delta, \psi)) w_u + F^W_q(s) \bar{U} > \bar{U}
\]

\[
\xi(\delta, \psi) (w_q - \bar{U}) - (1 - \xi(\delta, \psi)) (\bar{U} - w_u) > 0
\]

Providing that \(w_q - \bar{U} > 0\) and \(\bar{U} - w_u > 0\), this amounts to the following condition.

\[
\frac{w_q - \bar{U}}{\bar{U} - w_u} > \left(\frac{1 - \delta}{\delta}\right) \tau(\psi)
\]

The utility that the individual expects to get from investment is as follows.

\[
\left(1 - F^W_q(s)\right) \left[ \delta \left(1 - F^E_q\right) w_q + (1 - \delta) \left(1 - F^E_u\right) w_u \right]
\]

\[
+ \left[F^W_q(s) + \left(1 - F^W_q(s)\right) \delta F^E_q + (1 - \delta) F^E_u\right] \bar{U} - c
\]

The utility from not investing is:

\[
\left(1 - F^W_u(s)\right) \left[ \delta \left(1 - F^E_q\right) w_q + (1 - \delta) \left(1 - F^E_u\right) w_u \right]
\]

\[
+ \left[F^W_u(s) + \left(1 - F^W_u(s)\right) \delta F^E_q + (1 - \delta) F^E_u\right] \bar{U}.
\]

The worker will invest if and only if the following condition holds.

\[
\left(F^W_u - F^W_q\right) \left[ \delta \left(1 - F^E_q\right) (w_q - \bar{U}) + (1 - \delta) \left(1 - F^E_u\right) (w_u - \bar{U}) \right] > c
\]

If we normalize the payoffs in this example by defining \(\omega_q = w_q - \bar{U}\) and \(\omega_u = \bar{U} - w_u\), these conditions exactly match those discussed in Section 1.2.

B.3.2 Returns in Our Example

The probabilities that a worker sends an unclear signal if he did or did not invest are respectively \(p_q = \frac{\theta_q - \theta_q^*}{1 - \theta_q^*}\). Similarly, the probabilities that an employer sends an unclear signal if he did or did not invest are respectively \(q_q = \frac{\psi_q - \psi_q^*}{1 - \psi_q^*}\) and \(q_u = \frac{\psi_u - \psi_u^*}{\psi_u^*}\). These can be used
to derive return to investment for workers and employers. With the parameter values we provide, these then collapse to the returns we discuss in Section 2.1.

\[ \beta_W = \begin{cases} \left( \frac{\theta_q}{\psi_u} \right) \cdot \left[ \delta_j \omega_q - \left( \frac{\psi_u - \psi_q}{\psi_u} \right) (1 - \delta_j) \omega_u \right] & \text{if } \delta_j \geq \delta_j \text{ and } \pi_j \geq \hat{\pi}_j \\ \left( \frac{1 - \theta_q}{\psi_u} \right) \cdot \left[ \delta_j \omega_q - \left( \frac{\psi_u - \psi_q}{\psi_u} \right) (1 - \delta_j) \omega_u \right] & \text{if } \delta_j \geq \delta_j \text{ and } \pi_j < \hat{\pi}_j \\ \left( \frac{1 - \theta_q}{\psi_u} \right) \cdot \left[ \frac{1 - \psi_u}{\psi_u} \delta_j \omega_q \right] & \text{if } \delta_j < \delta_j \text{ and } \pi_j \geq \hat{\pi}_j \\ \left( \frac{1 - \theta_q}{\psi_u} \right) \cdot \left[ \left( \frac{1 - \psi_u}{\psi_u} \right) \delta_j \omega_q \right] & \text{if } \delta_j < \delta_j \text{ and } \pi_j < \hat{\pi}_j \end{cases} \]

\[ \beta_E = \begin{cases} \lambda_j \left( \frac{\psi_u}{\psi_q} \right) \cdot \left[ \tau_j \chi_q - \left( \frac{\psi_u - \psi_q}{\psi_u} \right) (1 - \tau_j) \chi_u \right] & \text{if } \delta_j \geq \delta_j \text{ and } \pi_j \geq \hat{\pi}_j \\ \lambda_j \left( \frac{\psi_u}{\psi_q} \right) \cdot \left[ \left( \frac{1 - \theta_q}{\psi_u} \right) \tau_j \chi_q \right] & \text{if } \delta_j \geq \delta_j \text{ and } \pi_j < \hat{\pi}_j \\ \lambda_j \left( \frac{1 - \psi_u}{\psi_q} \right) \cdot \left[ \tau_j \chi_q - \left( \frac{\psi_u - \psi_q}{\psi_u} \right) (1 - \tau_j) \chi_u \right] & \text{if } \delta_j < \delta_j \text{ and } \pi_j \geq \hat{\pi}_j \\ \lambda_j \left( \frac{1 - \psi_u}{\psi_q} \right) \cdot \left[ \left( \frac{1 - \psi_u}{\psi_u} \right) \tau_j \chi_q \right] & \text{if } \delta_j < \delta_j \text{ and } \pi_j < \hat{\pi}_j \end{cases} \]

### B.4 Further Extensions of the Model

#### B.4.1 Endogenous Wages

**Ex-Post Bargaining**

Consider the following modification to the model described in Section 2. Rather than payoffs from a match being fixed at \( \{ \omega_q, \omega_n, \chi_q, \chi_n \} \), we can add a third stage at which worker and firm investment decisions become common knowledge. To model the bargaining process, we assume for simplicity that total worker and firm payoffs are linear in monetary transfers that can be made between the two parties. The firm and worker investment decisions \( i^W, i^E \in \{ q, u \} \) determine the total surplus to be split, \( x^{i^W,i^E} \). Workers receive a fixed fraction \( \alpha \in (0, 1) \) of this surplus.

Workers can, at the time of application, exercise a more valuable outside option \( w^W_0 \) if they invested than if they did not, with an equivalent assumption regarding the outside option for firms \( x^E_0 \). However, at the time of bargaining, the outside options of both parties

---

1For example, a firm and worker that both invested split a surplus \( x^W \).
are zero. To exactly replicate the payoff structure of our baseline model, further assume that the benefit to workers (resp. firms) from being matched to a good firm (resp. worker) is independent of their own investment decision. This allows us to define $\omega_q, \omega_u, \chi_q, \chi_u$.

\[
\omega_q = ax_{qq} - w_q^0 = ax_{uq} - w_u^0 \geq 0 \\
\omega_u = w_q^0 - ax_{uu} = w_u^0 - ax_{uu} \geq 0 \\
\chi_q = (1 - \alpha) x_{qq} - x_q^0 = (1 - \alpha) x_{uu} - x_u^0 \geq 0 \\
\chi_u = x_u^0 - (1 - \alpha) x_{uu} = x_u^0 - (1 - \alpha) x_{uu} \geq 0
\]

Finally, if the lowest worker cost is $c \geq w_q^0 - w_u^0$ and the lowest firm cost is $k \geq x_q^0 - x_u^0$ then this structure exactly replicates our baseline model.

It is possible to relax some of these assumptions without any qualitative changes to the model. For example, if $(1 - \alpha) x_{qq} - x_q^0 \neq (1 - \alpha) x_{uu} - x_u^0$ or $x_q^0 - (1 - \alpha) x_{uu} \neq x_u^0 - (1 - \alpha) x_{uu}$ then the firm hiring threshold would depend on whether the firm invested. This does not introduce any substantive change to our results. The restriction that firm and worker costs are bounded above zero is more important, but also sensible: without it, a worker would have an incentive to invest even if doing so never increased the probability of being hired.

**Wage Equals Marginal Product**

We assumed throughout the analysis that net worker and firm payoffs are exogenous parameters. We show above that this payoff structure can be rationalized by ex-post bargaining. Here, we instead explore the possibility of variable wage offers at the hiring stage. Specifically, we consider a simple benchmark model in which workers are paid their marginal product, although investment remains binary. The resulting policy implications are qualitatively the same as those of our baseline model.

**Model**

Begin by assuming that output (with price normalized to one) is produced at constant
returns to scale using a mass of qualified workers $Q_j$, combined with group-specific capital provided by the firm $K_j$. For convenience, we adopt a Cobb-Douglas specification.

$$Y_j = K_j^{1-\gamma} Q_j^\gamma$$

We can derive the average product of a worker by dividing by the total labor force $L_j = Q_j + U_j$ where $Q_j$ is the mass of qualified workers, $U_j$ is the mass of unqualified workers and $S_j = Q_j / L_j$ is the share of qualified workers.

$$\frac{Y_j}{L_j} = \left( \frac{K_j}{L_j} \right)^{1-\gamma} S_j^\gamma$$

Mirroring the assumptions of our baseline model, the firm can choose to provide exactly one unit or zero units of capital per worker for each group $j \in \{ A, B \}$ so that $K_j / L_j \in \{ 0, 1 \}$ depending on which investments the firm chooses to make. If the firm doesn’t invest, no output is produced for that group.

Next, we can derive the marginal product of a worker. For a qualified group $j$ worker at a firm that made the group $j$ investment, the marginal product is:

$$MP_j = \gamma \left( \frac{Y_j}{Q_j} \right) = \gamma S_j^{\gamma-1}.$$ 

The marginal product for a worker who did not invest is zero, since such workers never add value to production. This implies, given the same signal structure as in our baseline model, that a random worker’s expected marginal product – which we assume is also the wage that a firm offers – is as follows.

$$\kappa (\pi_j, \theta) \gamma S_j^{\gamma-1} = w (\pi_j, \theta) \geq 0$$

Assuming that workers have no outside option, they are always willing to accept this offer, since it is always weakly positive.

Aggregating up, the share of qualified workers is just $\pi_j$. This means that the average wage payment for group $j$ is $\gamma \pi_j^\gamma$. Thus, the revenue that the firm earns per worker, net of
wage payments, is \( \lambda_j \left[ (1 - \gamma) \pi_j^\gamma \right] \). The fraction of firms who invest is therefore given by:

\[
\delta_j = G^E \left( \lambda_j \left[ (1 - \gamma) \pi_j^\gamma \right] \right).
\]

Since a fraction \( \delta_j \) of firms made the group \( j \) investment, the return to investment for group \( j \) workers is simply \( \delta_j \) multiplied by impact that worker investment has on the average wage offer. Thus, the fraction of workers of group \( j \) who invest is as follows.

\[
\pi_j = G^W \left( \delta_j \int_0^1 w(\pi_j, \theta) \left[ f_u^W(\theta) - f_q^W(\theta) \right] d\theta \right)
\]

Clearly if \( \delta_j = 0 \), then the return to investment is zero. Similarly, if \( \pi_j = 0 \), then there is no return to investment because the wage is zero for any signal.

**Discriminatory Equilibria**

Since this model can have multiple equilibria, there is potential for an equilibrium in which there is zero investment by Bs and positive investment by As. If \( \pi_B = 0 \), firms would never offer a positive wage to B workers here, since \( \kappa(\pi_j, \theta) = 0 \). In turn, this means that there is never an incentive for workers to invest. Since hiring a B worker never adds to output, the return to B investment for firms is zero as well. Thus \( \delta_B = 0 \). Turning to the A market, if \( \pi_A > 0 \), wages are positive and the fraction of A workers who invest is as follows.

\[
G^W \left( \delta_A \int_0^1 w(\pi_A, \theta) \left[ f_u^W(\theta) - f_q^W(\theta) \right] d\theta \right)
\]

The return to investment for firms is also positive, and a fraction \( \delta_A \) invest.

\[
\delta_A = G^E \left( \lambda_A \left[ (1 - \gamma) \pi_A^\beta \right] \right)
\]

We can prove that there can be such an equilibrium by positing a value of \( \pi_A \), and a function \( G^E \) such that \( \delta_A > 0 \). For any such \( \pi_A \) and \( \delta_A \), there is some function \( G^W \) that that satisfies our assumptions and which yields the required worker investment levels.

\[
\pi_A = G^W \left( \delta_A \int_0^1 w(\pi_A, \theta) \left[ f_u^W(\theta) - f_q^W(\theta) \right] d\theta \right)
\]
**Symmetric Investment**

We next examine the conditions under which a non-discriminatory equilibrium exists. Assume that $\pi_A = \pi_B = \pi$. This means that the fraction of workers who invest (for both types) is:

$$G^W \left( \delta \int_0^1 w(\pi, \theta) \left[ f^W_u(\theta) - f^W_q(\theta) \right] d\theta \right).$$

If $G^W$ is strictly increasing and $\delta_j > 0$, than this fraction can only be the same for both groups if $\delta_A = \delta_B = \delta$. However, the return to investment for firms is $\delta = G^E(\lambda_j [(1 - \gamma) \pi^\gamma])$. If $G^E$ is strictly increasing, then firm investment levels cannot be the same for both groups unless $\lambda_A = \lambda_B$. This precludes $\pi_A = \pi_B = \pi$ if $\lambda_A \neq \lambda_B$, implying that an equilibrium with positive investment but no discrimination is impossible.

**Affirmative Action**

One definition of affirmative action in this model is a requirement that the average wage paid to workers, conditional on their being hired, is equal across groups.

$$\int_0^1 w(\pi_A, \theta) = \int_0^1 w(\pi_B, \theta)$$

This has many of the same problems as affirmative action in our baseline model. First, it does not eliminate the possibility of zero investment by Bs but positive investment by As, with no B workers receiving any wage offer. Under affirmative action, there is an equilibrium with $\pi_B = \delta_B = 0$ combined with any set of beliefs $\{\pi_A, \delta_A\}$ that constituted an equilibrium in the A market without affirmative action.

The second question we can ask is whether it is possible for this type of AA to lead to homogeneous beliefs. First, note that $\pi_A = \pi_B = \pi$ implies that wages are identical across groups for every $\theta$ and that affirmative action does not bind.

$$w(\pi, \theta) = \kappa(\pi, \theta) \gamma \pi^{\gamma-1}$$

Assuming again that $G^W$ and $G^E$ are strictly increasing, a requirement for positive and equal rates of worker investment is again that $\delta_A = \delta_B = \delta$, which is only possible if
\( \lambda_A = \lambda_B \). Otherwise, affirmative action again has no prospect of eliminating discrimination in equilibrium.

**Investment Insurance**

Our main policy prescription, two-sided investment insurance, is similarly effective in the model with variable wages. Assume initially that Bs are a numerical minority \((\lambda_B \leq \lambda_A)\) and that they are in an inferior equilibrium compared to As: \(\pi_A > \pi_B\). This implies that wages are lower for this group.

\[
w(\pi_A, \theta) = \kappa(\pi_A, \theta) \gamma \pi_A^{\gamma - 1} < \kappa(\pi_B, \theta) \gamma \pi_B^{\gamma - 1} = w(\pi_B, \theta)
\]

Assuming that \(G^E\) is strictly increasing, it also implies that firm investment is lower.

\[
\delta_A = G^E (\lambda_A [(1 - \gamma) \pi_A^\gamma]) < G^E (\lambda_B [(1 - \gamma) \pi_B^\gamma]) = \delta_B
\]

As we did in our baseline model, imagine that the government has access to its signals of worker and firm investment: \(\theta^G\) and \(\psi^G\), which satisfy the same assumptions as \(\theta\) and \(\psi\). The government can use these signals to target potentially variable “wage” payments to workers, with similar incentive payments for firms. This must be effective here as well, because large enough wage payments can achieve any investment return for both workers and firms.

Consider the following policy, which will lead to immediate elimination of discrimination. First, set government wages \(w^G(\theta)\) such that the fraction of Bs who invest is \(\pi_A\).

\[
\pi_{B,t} = G^W \left( \delta \int_0^1 [w(\pi_{B,t-1}, \theta) + w^G(\theta)] \left[ f^W_u(\theta) - f^W_q(\theta) \right] d\theta \right) = \pi_A
\]

Similarly, set weakly positive payments \(p^G(\psi)\) firms such that the fraction who invest is \(\delta_A\).

\[
\delta_{B,t} = G^E \left( \lambda_B \left[ (1 - \gamma) \pi_{B,t-1}^\gamma \right] + \int_0^1 p^G(\psi) \left[ f^E_u(\psi) - f^E_q(\psi) \right] d\psi \right) = \delta_A
\]

Once this has been achieved, firms will set \(w(\pi_A, \theta) = w(\pi_B, \theta)\). This will ensure that \(\pi_{B,t+1} = \pi_A\) with no subsidy, and it can be removed. The aggregate firm investment subsidy
that is still required to maintain equal firm investment returns is as follows.

\[
\int_0^1 p^g(\psi) \left[ f_u^E(\psi) - f_q^E(\psi) \right] d\psi = (\lambda_A - \lambda_B) \left[ (1 - \gamma) \pi \right]
\]

Thus, if \( \lambda_A = \lambda_B \) then \( p^g(\psi) = 0 \) for all \( \psi \): i.e., no investment subsidy is needed. Otherwise, some level of firm investment subsidy must be maintained to preserve an equilibrium without homogenous beliefs.

### B.4.2 Marginal Firm Investment Costs

In our baseline model, we assumed that firms paid a fixed cost \( k_j \) for each investment that they chose to make. Here we provide intuition for an alternative case in which the cost of investing in group \( j \) is proportional to the number of workers from group \( j \) who end up being hired. With this change, there is no longer an inherent disadvantage to being a minority, but the model is otherwise qualitatively unchanged in most respects.

The investment cost is only paid for workers who apply and receive offers from the firm, which implies that the expected investment cost for a given group is as follows.

\[
\lambda \left[ 1 - F^E_q(t^*(\delta)) \right] \cdot \left[ \pi \left( 1 - F^W_q(s^*(\pi)) \right) + (1 - \pi) \left( 1 - F^W_u(s^*(\pi)) \right) \right] \cdot k
\]

The gross returns to investment are unaffected. For any equilibrium without zero investment, the fraction of firms who invest is therefore as follows.

\[
\delta = G^E \left( \frac{\lambda \left[ F_u^E(t^*(\delta)) - F_q^E(t^*(\delta)) \right] \left[ \pi \left( 1 - F^W_q(s^*(\pi)) \right) \chi_q - (1 - \pi) \left( 1 - F^W_u(s^*(\pi)) \right) \chi_u \right]}{\lambda \left[ 1 - F^E_q(t^*(\delta)) \right] \left[ \pi \left( 1 - F^W_q(s^*(\pi)) \right) + (1 - \pi) \left( 1 - F^W_u(s^*(\pi)) \right) \right]} \right)
\]

Compared to our baseline model, the net return is simply scaled up by the proportion of workers hired.

It is clear from equation B.4 that the returns to firm investment, and the fraction of firms who invest, are both independent of the population fraction \( \lambda \). This is enough to conclude that for otherwise identical groups, the set of equilibria no longer depends on population size. In this sense, if investment costs are marginal, there is no inherent disadvantage to
being a member of a minority group.

Aside from this point, the change in assumptions does not substantively alter the model, although this version is much less convenient to analyze. Compared to the case with fixed investment costs, firm investment returns are scaled up by a factor that varies with $\delta$ and $\pi$. As $\delta$ and $\pi$ both approach one, the denominator in equation (B.4) is simply $\lambda$. At lower levels, returns are further scaled up, since costs are only paid for individuals who are hired.

The zero investment equilibrium clearly still exists and is stable. With the same regularity assumptions as we adopt for our existence proposition, the firm will set its signal threshold to one if $\pi$ falls below some low but positive level. Similarly, if $\delta$ falls below some positive critical value, no workers apply. As long as $\pi$ and $\delta$ are low enough, there is therefore no incentive for any firm or worker to invest. Similarly, other equilibria may still exist, although stability and existence are much more complex to verify. This is already enough to conclude that modified versions of Propositions 6 and 7 continue to hold.

The logic behind the proof of Proposition 9 also remains intact, since the numerator of equation (B.4) is negative for a low enough value of value of $\pi$, while investment costs remain strictly positive. Perhaps most importantly, the logic of investment insurance (Proposition 13) is fundamentally unchanged.
Appendix C

Appendix to Chapter 3

C.1 Adjustment for 2013 Testing Waiver

In this appendix, we discuss complications due to a waiver that was received by the NYCDOE to avoid ‘double testing’. This waiver allowed accelerated students in grade 8 who were sitting the New York City regents exam in mathematics to avoid sitting the grade 8 common core exam in the same subject. In our data, this leads to a sharp reduction in the number of grade 8 students present in 2013, compared to previous years.

To avoid any concern that this sudden change in composition is biasing our results, we adjust for it here by omitting any student who is eventually observed to sit the regents exam for Math in grade 8. Restricting to this sample eliminates the jump in our sample size in 2013. As Figure C.1 shows, there is also no sharp change in the percentage of students sitting the regents exam in grade 8 around the reform. Our approach thus effectively mitigates the impact of the change in composition in 2013.

The results from this robustness exercise are shown in Figure C.2. The results are qualitatively unchanged, although the measured effect sizes are marginally smaller than in our baseline analysis. Just as before, there appear to be significant treatment effects for Math, but these appear primarily to take effect from 2013 onwards.
Figure C.1: This figure shows the share of grade 8 students who take the Regents exam in mathematics, with and without taking the regular grade 8 exam.

C.2 Alternative Prediction Period

As an additional robustness check, we also include results using an alternative base period (2006–2009) for the purpose of calculating pre-period suspension rates. The results are shown in Figure C.3, and are again qualitatively unchanged.

C.3 Supplementary Figures

Our final appendix contains additional figures to supplement the main text.
Figure C.2: This figure plots the coefficients, $r_{sk}$, from regression 3.8. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. In this version of the graph, we exclude all students who go on to sit the Regents exam for Math in grade 8.
Figure C.3: This figure plots the coefficients, $\rho_{sk}$, from regression 3.8. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period. In this version of the graph, we define pre-period suspension rates based on 2006-2009 data.
Figure C.4: This figure shows the relationship between suspension rates and test scores in our sample of middle school students (grades 6-8). Students are grouped into bins based on their standardized math and ELA exam scores. The figure shows suspension rates in each bin plotted against the average in each bin.
Figure C.5: This figure shows the accuracy of in-sample predictions of school-grade suspension rates, depending on which covariates are used. Available covariates include the race and gender composition of each school-grade, average age, the share of students who are English Language Learners (ELL), the share of students qualifying for free or reduced-price lunch, and the share of students who have previously repeated a grade.

Figure C.6: This figure shows the long-term trend in the number of suspensions for Level 3 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system.
Figure C.7: This figure shows the long-term trend in the number of suspensions for Level 4 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system.

Figure C.8: This figure shows the long-term trend in the number of suspensions for Level 5 infractions dispensed per year in the New York City students in grades 6-8 of the New York City public school system.
Figure C.9: This figure plots the coefficients, $\rho_s$, from regression 3.7 using test score percentiles as an outcome rather than z-scores. The results for Math are shown in the top panel, and ELA below. Each point measures the change in the relationship between test score growth and suspension rates across school-grades, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period.
Figure C.10: This figure plots the coefficients, $\rho_{ik}$, from regression 3.8 using test score percentiles as an outcome rather than z-scores. The results for Math are shown in the top panel, and ELA below. Each point measures the change in test score growth in school-grades with high and low pre-period suspension rates, relative to students in school-grades with no pre-period suspensions at all, compared to 2011, conditional on year and school-grade fixed effects. The vertical blue lines show 95 percent confidence intervals, and the red lines separate the pre-period from the post-period.